Hand Gesture Control for First-Person Tabletop Interaction

Master’s thesis in partial fulfilment of the requirements for the degree of

“Master of Science”

Author: Dipl.-Ing. Dominik Patrick Hofer, BSc

Submitted to the Joint Master Programme on Human-Computer Interaction

Salzburg University of Applied Sciences

Paris-Lodron-University Salzburg

Assessed by:

Univ.-Prof. Dipl.-Ing. Christopher Frauenberger Ph.D.

Department of Artificial Intelligence and Human Interfaces

Salzburg, January 2024

Zusammenfassung

In dieser Arbeit – die in Kooperation mit dem Projekte „Hybrid Learning Environments in Tourism Education „ (HyLTE) erarbeitet wurde - wird die Entwicklung und Implementierung einer Gestensteuerung für den Kontext von Tischprojektionen aus der Vogelperspektive vorgestellt, mit der versucht wird, die Ich-Perspektive zu imitieren, ohne dass tragbare Technologien benötigt werden. Als Funktionalitäten für das System wurden Videoaufzeichnung und -wiedergabe definiert. Wobei zum Themengebiet der bildbasierten Gestensteuerung bereits wissenschaftliche Publikationen existieren, sind die definierten Funktionalitäten, gegeben dem Kontext, nicht erforscht. Um diese wissenschaftliche Lücke zu erforschen, wurde die Forschungsfrage „Wie können Tischgesten im HyLTE-Kontext gestaltet werden?“ definiert. Für die Beantwortung wurden zwei Studien durchgeführt. Bei der ersten Studie handelt es sich um eine Erhebungsstudie, bei der in Zusammenarbeit mit den Teilnehmer:innen potenzielle Gesten gesammelt wurden, die sich intuitiv bedienen lassen. Die Ergebnisse dieser Studie führten zur zweiten Studie: einer Nutzerstudie, um Daten für die Implementierung der Steuerung zu sammeln und Einblicke in die Präferenzen der Teilnehmer:innen gegenüber den Gesten zu gewinnen, die von den Teilnehmer:innen der ersten Studie vorgeschlagen wurden. Die endgültigen Ergebnisse dieser Arbeit sind: (1) ein Gestenwörterbuch für den gegebenen Kontext, (2) eine Methode zur Bewertung von Gesten nach Benutzerpräferenzen, (3) eine open source Video- und KI-Modell Datenbank und (4) Gestaltungsempfehlungen und Best Practices für ein solches Vorhaben. Diese Arbeit stellt somit einen wissenschaftlichen Beitrag zur Erstellung, Implementierung, und Evaluierung von gestengesteuerten Systemen dar.

Abstract

In this work - which was developed in cooperation with the project "Hybrid Learning Environments in Tourism Education " (HyLTE) - the development and implementation of a gesture control system for the context of table projections from a bird's eye view is presented, which attempts to imitate the first-person perspective without the need for portable technologies. Video recording and playback were defined as functionalities of the system. Although scientific publications already exist on the topic of image-based gesture control, the defined functionalities have not been researched, given the context. To explore this scientific gap, the research question "How can table gestures be designed in the HyLTE context?" was defined. Two studies were conducted to answer this question. The first study was a survey study in which potential gestures that can be used intuitively were collected in cooperation with the participants. The results of this study led to the second study: a user study to collect data for the implementation of the control and to gain insight into the participants' preferences towards the gestures suggested by the participants of the first study. The final results of this work are: (1) a gesture dictionary for the given context, (2) a method for evaluating gestures according to user preferences, (3) an open-source video and AI model database, and (4) design recommendations and best practices for such an endeavour. This work thus represents a scientific contribution to the creation, implementation, and evaluation of gesture-driven systems.

List of Figures

List of Tables

List of Equations

Contents

1 Introduction

Besides the common interaction forms of computer mouse, keyboard, and touch screen, other, less often used ones, exist. A group of those interaction forms is called touchless technology[1] and includes gestures, object, face, and speech detection. They became more dominant during the COVID-19 pandemic.

Gesture Detection (GD) or Control (GC) represents the form of interaction that allows systems to interpret gestures – body, head, and/or hands - as a means to convey intent. Notably, GC has garnered attention with the Microsoft Kinect[2] and Leap Motion[3]; As well as, within the eXtended Reality (XR) community, with Meta's Oculus Quest[4] and Microsoft Hololens[5] series being recognized for their support of GC.

Although GC is considered a natural form of interaction and has been researched since the 1980s [1], its widespread adoption has been hindered due to different reasons such as cultural differences in gestures[6], hardware limitations[7], and sharing of GC implementations [2]. However, in the past decade, advancements in hardware have alleviated these constraints and facilitated the broader utilization of Artificial Intelligence (AI) applications, particularly those centred around Deep Neural Networks (DNN). With the rise of AI, new possibilities for investigating more complex applications, including GC, have emerged [2].

This thesis was written while being employed for the Hybrid Learning Environments in Tourism Education (HyLTE) project at the Paris Lodron University Salzburg (PLUS) which is in collaboration with the Salzburg University of Applied Sciences (SUAS) and the Tourism School Klessheim. The goal of the project is to define use cases where technology and tourism education overlap and do so in a user-centred manner. Two use cases were defined: first-person educational video

[1] https://www.intel.com/content/www/us/en/internet-of-things/iot-solutions/touchless-technology.html
[2] https://learn.microsoft.com/en-us/windows/apps/design/devices/kinect-for-windows
[3] https://leap-2.ultraleap.com
[4] https://www.meta.com/at/en/quest/quest-pro/
[5] https://www.microsoft.com/en-us/hololens
[6] https://www.linkedin.com/pulse/20140320012035-12181762-gestural-control-the-good-the-bad-and-the-ugly/
[7] https://intranet.birmingham.ac.uk/it/innovation/documents/public/Gesture-Control-Technology.pdf

recording and examination simulation; Though only the first use case will be further discussed in this thesis because it serves as the contextual basis for the GC.

This thesis aims to answer the research question “How can tabletop gestures be designed within the HyLTE context?”, by creating a gesture dictionary (GD)—a repository recommending gestures for specific actions. The context is a tabletop projection from a bird’s eye view, trying to imitate a first-person perspective without needing wearables. To develop the GD, two studies were conducted. The first was an elicitation study, which sought to gain insights into intuitive gestures for the given context. The second study was a user study, conducted to compare two preliminary gesture dictionaries derived from the elicitation study. The first preliminary gesture dictionary was created by following design guidelines proposed by Vatavu et al. [2], Saffer [3], and Wigdor et al. [4]. The second preliminary dictionary was created by using the mental models (MM) discussed with the participants during the elicitation study [5]. For the user study, custom software was developed that supported a full interactive prototype as well as video and audio recording.

To guide the creation of the GD, four sub-research questions were developed:

1. What gestures are intuitively associated with video recording and player functionalities in a tabletop projection setting?
2. Given the elicitation study outcome, what AI approach can serve as a proof of concept for implementing gesture recognition models?
3. How can participant preferences for gestures be evaluated using two AI models – one informed by the Agreement Rate and the other by Mental Models?
4. What insights and recommendations can be derived from developing and evaluating AI-driven gesture recognition models in tabletop projection settings for future projects?

The final output of this work is the following: (1) a gesture dictionary for the given context, (2) a method to rate gestures according to user preferences, (3) an open-source video and AI model database and (4) design recommendations and best practices for such a project. It is worth noting that this work distinguishes itself as one of the few scientific endeavours that openly shares its results by making them freely available on GitHub8, by the principles advocated by Vatavu et al. [2].

The structure of the thesis is the following: Firstly, as part of the introduction, a short overview of the HyLTE project will be given. This provides insights into the use case that was defined, as well as the prototype of the project, which will be used as a base for this thesis. Secondly, related work will be discussed. These include aspects of scientific literature about tabletop digitalization, classification of human gesture, elicitation studies, and artificial intelligence (AI) for gesture detection. Thirdly, the method of the thesis will be deliberated. This involves the subdivision of the primary research question into four sub-questions, and the conduction of two studies – elicitation and user. Fourthly, the results of the studies will be debated, which includes the final GD and design recommendations. Fifthly, the overall discussion of this thesis will be held. Finally, the potential for future work based on the thesis findings, along with the conclusion, will be addressed.

1.1 HyLTE Prototype

In this section the approach and use case will be discussed, followed by a short explanation of the setup and how hard- and software are involved. Overall, this section should be perceived as a summary, giving a brief understanding of the context, setup, and technical implementation in which, the results of this work will be applied.

1.1.1 Approach and Use Case

Based on insights gained via observations, interviews, and workshops - acquired during the project - the use-case of “first-person video recording” was identified, which is the basis for this thesis. The first-person video recording use-case was born of the idea that pupils should see concepts such as cutting, folding, or removing fish bones, from the perspective of the teacher. This would allow them to directly imitate the movements of the teachers and give the pupils time-invariant access to videos of which they would be the target group. It needs to be also mentioned that certain tasks such as fishbone removement, flambe crepe, or opening and resealing wine are tasks that pupils might (but not necessarily) train once a semester.

[8] Github.com/DPHofer-HAII/gesture-control-for-tabletop-interaction

Additionally, the idea of post-processing the video via annotations was created. This means that users can, after the recording of the video is concluded, directly write on the video – like holding a pen and adding notes or highlighting important content – to draw the attention of the pupils onto important content.

In summary, the concept of "first person video recording" that allows for video annotation was created by conducting observations, interviews, and workshops with teachers and pupils.

1.1.2 Set-Up

The prototype consists of a projector, three cameras – including the study setup -, aluminium profiles, a personal computer, and an interchangeable table. The complete setup of the prototype can be seen in Figure 1. Similar setups, with varying contexts, have already been created and are discussed in chapter 2.1.

A digital interface, with a 16:9 format, is projected onto the table by the projector. The projected content can be manipulated by the users via hand gestures. The recording will be done by the two cameras in parallel, creating a bird's eye view and "side view" video. The second video should provide a focused view, allowing the content creators to highlight important details. The third camera – the Webcam for Study Recordings – was only present during the two studies and is not a part of the prototype design.

The vision-based approach was chosen, because electro-capacitive surfaces and touch screens are not a possibility in a kitchen setting, due to regulations, cost, and the safety of pupils. Therefore, a combination of projector and camera-based analysis was perceived as the most fitting approach. Microsoft Kinect Azure[9] and other depth cameras could not be used because the metallic surface of the kitchen counters distorted the infrared rays that were needed for tracking, or the sensors couldn't reach the distance that was needed – Leap Motion for example.

Due to the usage of a DNN, which depending on the setup needs a strong CPU or preferably GPU, a personal computer (PC) needs to be used. A server architecture for the setup was also thought of but will be part of the future work discussion in chapter 6.

[9] https://azure.microsoft.com/en-us/products/kinect-dk

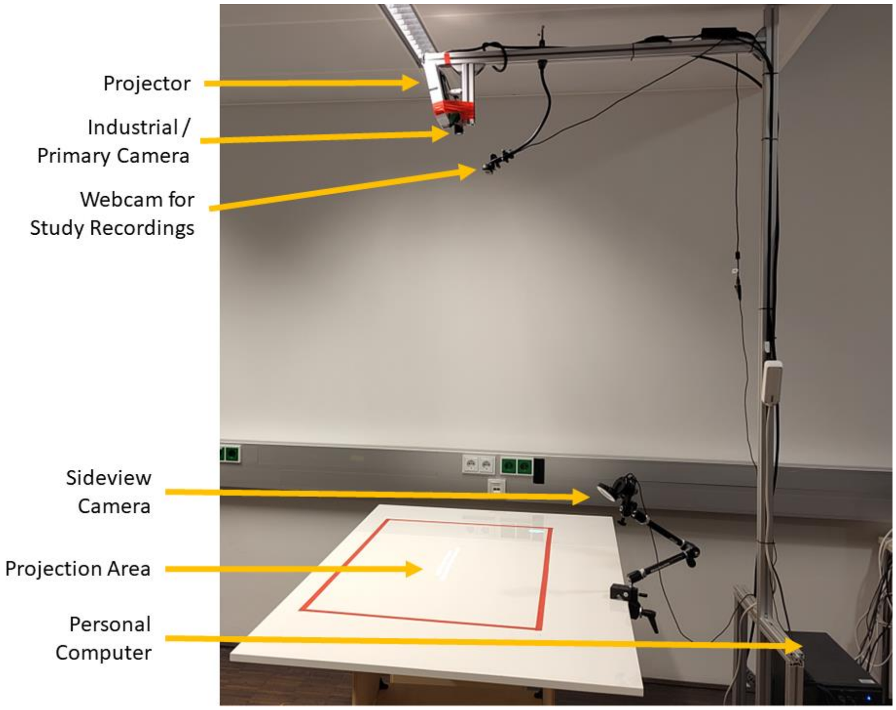


Figure 1: Prototype setup during the second study

1.1.3 Implementation

The implementation of the prototype was done in Unity[10] and Python[11]. This combination was chosen to focus on the strengths and capabilities of the respective tools.

Unity is primarily used for the User Interface (UI), allowing to easily create simple UIs. Additionally, it has a vast community that provides open-source assets that could potentially be used. There are advances in the use of AI for Unity, but when the choice needed to be made, they seemed experimental, with a steep learning curve. Additionally, it should be stated that the decision for Unity was made before the policy announcement in September 2023.

Python on the other hand was chosen, due to open-source software libraries such as OpenCV[12], MediaPipe[13] and TensorFlow[14]. OpenCV is used for camera controls, the saving of video and audio

[10] https://unity.com/
[11] https://www.python.org/
[12] https://opencv.org/
[13] https://developers.google.com/mediapipe
[14] https://www.tensorflow.org/

material, and image transformation. MediaPipe – or to be more precise the Python project by Takahashi Shigeki[15], which uses MediaPipe - provides the DNN model for gesture control. MediaPipe is a software framework for building AI applications. Takahashi Shigeki's project was used because it provides an infrastructure to easily add new gestures for the system to recognize. Finally, TensorFlow is the software basis the model was trained / iterated with.

The above-mentioned image transformation was needed for the mapping of vector spaces. This is needed to align the 3D vector spaces of the user with the 2D vector spaces of the projector and Unity. This is achieved by cropping the image to the projection on the table, calculating the pixel value position of the user's hand in the cropped image, and inverting it (1920 x 1080 resolution) for the position to be sent to Unity.

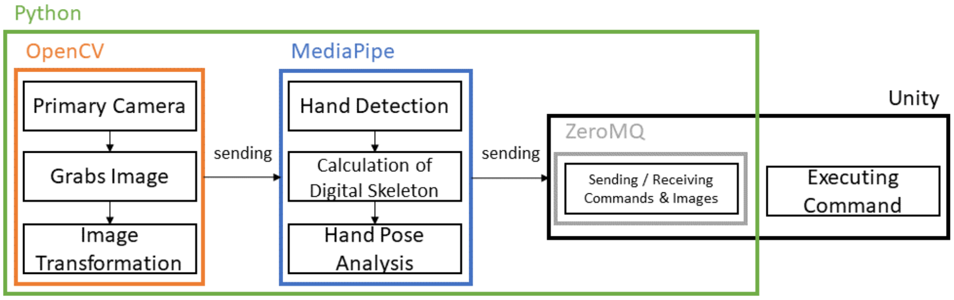


Figure 2: Relationship of the software tools

The bidirectional communication between Unity and Python was implemented in ZeroMQ[16] via a Publisher-Subscriber architecture. ZeroMQ allows to send data via a TCP/IP protocol, though the user only needs to set up the ports and define the gates that send/receive data. The capability of sending image data quickly was an important reason for picking ZeroMQ.

All those software libraries - except for ZeroMQ - were known beforehand, which might be a bias, but due to the brief timeframe of the project and not aiming for a production-ready prototype, known concepts were favoured over the acquisition of completely new technologies.

Regarding the processing flow of the gesture control, the following approach was chosen: The primary camera continuously grabs an image, which is transformed and cropped to fit the

[15] https://github.com/Kazuhito00/hand-gesture-recognition-using-mediapipe/tree/main
[16] https://zeromq.org

projected area. and sends it to MediaPipe. The hand detection of MediaPipe calculates if a hand – or hands, depending on the setup though only one hand could control the mouse cursor / was perceived as the dominant hand – could be detected in the transformed image. If hand(s) can be detected, the digital skeleton is created. A digital skeleton consists of normalized data points – x and y axis - indicating the relative joint positions of the hand. With the joint positions of the skeleton, the gesture can be calculated. As a final step in Python, the detected gesture(s) are sent to Unity, which will then execute the command associated with the gesture. For a graphical representation of the gesture control process, please have a look at the flow chart depicted in Figure 2.

In summary, the starting point of this work is based on one of the HyLTE use cases. The interaction via GC is a requirement coming from the need to find alternatives to electro-capacitive surfaces, touch screens, and infrared-enabled cameras. A prototype with a focus on augmenting tabletop surfaces for digital interaction has already been created in this project and serves as the technological basis for this work. The prototype was developed with Unity and Python as technical components, allowing data transfer between them via a TCP/IP connection.

2 Related Work

To answer the research questions, this work combines different aspects of scientific research. In this chapter, an overview of the most important components of these aspects, concerning this work, will be given. First, the concept of digitalizing surfaces will be discussed, showing the scientific basis on which the prototype was created. As a second aspect principles of gesture control will be discussed, introducing concepts which are relevant to the aspect of elicitation studies. These forms of studies build the basis on which the final GC is built upon. As a final sub-chapter, DNN concerning GC will be discussed, because one of the scientific contributions of this work is an open-source model. The Venn diagram, shown in Figure 3 can be seen as a representation of where to position the thesis in the scientific landscape.

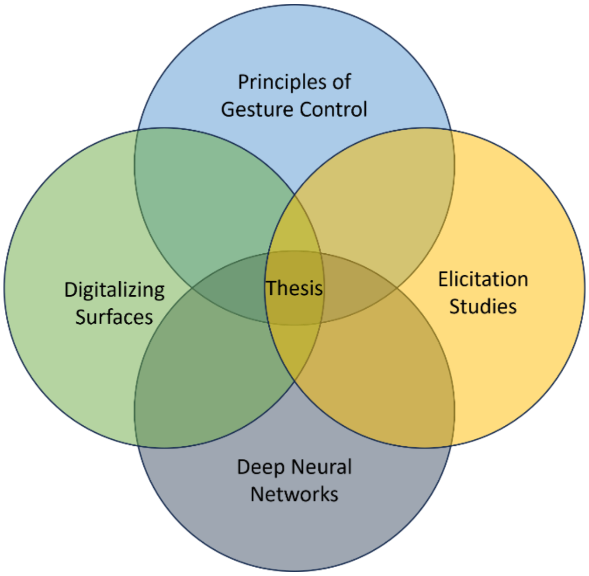


Figure 3: Venn diagram positioning this thesis as a combination of four scientific aspects

2.1 Digitalization of Surfaces

The approach of digitalizing desks or other surfaces is not a novelty. Various publications - beginning with the 1980s - have created prototypes with the capability of projector interaction based on different input modalities.

The first publications that were found are from Krueger [6] and Wellner [7]. Krueger used a table, camera, and screen. Users could see a silhouette of themselves and interact via gestures with objects on the screen [6]. Wellner created the DigitalDesk Calculator which uses a projector and camera to allow to calculate and project digital information onto paper for calculation. The gesture for interaction is clicking to insert the numbers into the calculator. He also states some issues that can arise when working with projectors. The shadow of the users, as well as the light in the room, can have a significant influence on usability and therefore the use of two projectors is proposed – one projecting from the top and the other from the bottom onto one semi-transparent surface.

The concept CounterActive by Ju [8] proposes a digital cookbook, which is projected onto the kitchen counters. Speakers for videos are the second output modality, and an electric field sensing array would be the input modality to manipulate the cookbook and videos.

DiamondTouch is a system created by Dietz and Leigh for multiple users to work in a digital collaborative work environment [9]. The UI is projected on a table which is augmented by tiny antennas. The signal of the antennas is sent via the user to receivers in the chairs, to map the interaction to a user. The concept limits the interaction of users to the fingertips and allows no complex gestures.

Bonanni and Lee [10] discussed how everyday objects in the kitchen can be augmented by digital information. They designed a moveable information table which gains a UI via a ceiling-mounted projector. The projector is supported by a camera which detects objects, tasks and simple gestures for the user to communicate what they would like to execute; Which, leads to the change of the digital information projected onto the table.

An example of how different surfaces can be used for digital interactions would be OmniTouch [11]. Harrison et al. combined a pico projector in combination with a depth camera to imitate smartphone interactions with surfaces such as a table, college block, and hands. Different use cases defined by Harrison et. al. include annotating / marking sheets, dialling a phone number, and a digital clock. Regarding the click interaction, they took a depth sensor approach, where they applied flood filling from the fingertip to the hand, calculating the fingertip position, and deducing a click motion from there.

The Extended Multitouch was a project by Murugappan et al. [12] creating a multi-user system that was able to detect different gestures. As in the above-mentioned examples, a projector was used for the UI on a table. The Microsoft Kinect was used as an input modality, allowing depth information to be used for the detection of table touching. One of the insights discussed was the possibility of expanding the interaction area with the system above the surface. Making space, which can be easily reached by humans, into an interaction area.

Chen et al. [13] use a depth camera – Microsoft Kinect and the corresponding software - that allows lecture recordings and digital distribution. Five functionalities are stated: Tilt-Up, Tilt-Down, Zoom-In, Zoom-Out, and Camera Movement. Though the gestures – which correlate with the number of functionalities – are referred to as intuitive, no usability study is stated.

In summary, scientific publications that focused on the augmentation of surfaces with projectors have been published since the 1980s with Krueger and Wellner being one of the first. Different contexts and set-ups have been explored with the CounterActive and DiamonTouch systems by Ju and Dietz et al. Respectively. Cameras, depth sensors as well as electro-capacitive surfaces are used as input modules. Problems and limitations that have been mentioned in the 1980s are still prevalent.

2.2 Classification of Human Gestures

This section focuses on the research for the classification of human gestures. Five works will be discussed, giving a brief overview of the topic.

Mitra et al. [14] define that there are three distinguishable types of gestures: body gestures; hand and arm gestures; and head and facial gestures. Body gestures refer to full-body movement and actions such as tracking of dance moves for musical adaption, and human gaits for rehabilitation. Hand and arm gestures focus on the recognition of hand poses such as sign language or virtual reality interaction (without controllers). Finally, examples of head and face gestures are the nodding and shaking of a head or looks of happiness, surprise, sadness, and fear.

In addition to the types of gestures, it is stated by Wigdor et al. [4] that gestures can be classified into two main types: static and dynamic. A static gesture refers to a particular hand position or configuration being captured in a single frame. Recognizing static gestures involves identifying static poses of the hand or whole body, making it a pattern recognition challenge. Dynamic gestures, on the other hand, involve movements of the hands, head, or entire body, which are captured as a sequence of frames. Recognizing dynamic gestures focuses on identifying frames within a specific time sequence. While static gestures tend to be more easily recognized, they are limited in their versatility. Dynamic gestures would allow for more diverse interaction between humans and machines but are accordingly more difficult to technically implement.

Efron [15] conducted one of the pioneering studies on communicative human gestures, leading to the establishment of five categories that served as the foundation for later classifications: physiographics, kinetographics, ideographics, deictics, and batons. The first two, in McNeill's system, are grouped as iconics [16]. McNeill also introduces metaphorics, deictics, and beats. However, because Efron's and McNeill's studies were centred on human conversation, their categories have limited relevance to interactive surface gestures.

Tang [17] tasked people to collaborate around a large drawing surface, analyzing their interactions. It was noted that gestures emerged as an important form of simulating operations, indicating areas of interest, and referring to other members of the group. Similar to the guessability methodology later described in this chapter, Tang defined *actions* and *functions* which refer to behaviour and the corresponding effect.

To summarize, gesture interaction analysis can be applied to the full body, or parts thereof such as the hands, arms, and facial expressions. There have been approaches to define different taxonomies for the classification of gestures, but the ones specific to surface interaction via gestures will be discussed in chapter 2.3.2.

2.3 Elicitation Studies

An elicitation study belongs to the methodology of guessability study, which belongs to the participatory design approach [18]–[20][17]. The approach exposes participants to referents or effects on an action and has them act out signs. This is contrary to an expert lead design approach, where the expert – e.g.: designer or developer – without the input of the target group, decides upon the interaction. Specifically in the gesture literature, elicitation studies have been used for different application domains and devices [5], [21]–[24]. Think-aloud protocols and

[17] It is argued that an elicitation study belongs to the participatory design approach due to the fact that referents[18] are independent of the experimenters, allowing participants to focus on the proposition of symbols[19], therefore the experiment generates data rooted in the subject’s phenomenology [20].

video recordings are usually used to foster a guessability study and gain insights into the users' cognitive models [5].

It needs to be pointed out that guessability and immediate usability are two different metrics [5]. Guessability focuses only on the quality of the input, rather than the learnability thereof or the entire system for which the input is used. Immediate usability, on the other hand, expects prior learning and evaluates the entire system. The benefits of a guessability study have been documented in different contexts:

One of the initial studies on guessability centred around text inputs in command lines [25], [26]. In this context, experts, namely designers, provided only a single command-line term for each referent[18]. However, relying on just one term, regardless of how "intuitive" it may be, led to guessability failures ranging from 80% to 90% [27]. A suggested remedy is the concept of "unlimited aliasing" [26], wherein the system attempts to make the best inference regarding the intended referent if an unfamiliar symbol is encountered. It has also been acknowledged that having multiple synonyms plays a crucial role in achieving a high level of guessability in command-line interfaces [25], [26].

Guessability has also been researched on text labels and graphical icons [28]. Wiedenbeck explored the importance of gussability for buttons, toolbars, and menus by letting participants design text labels and graphical UI elements.

A similar study to the work presented here was done by Epps et al. [29] and Wobbrock et al. [5]. The former showed static Windows desktop images on a table and asked participants to execute different tasks with their hands. It was concluded that the use of the index finger was the most common gesture. Latter used a table TV, which showed non-interactable referents. Participants acted out with one or both hands their symbols[19]. The study indicated that users are finger-

[18] A referent is an example of the effect an interaction should cause; For example, a button in a web interface that changes the color shortly, indicating that it was pressed, or as in the case of this study, a command-line term.

[19] Symbols in the context of elicitation studies means the movement / gesture participants would expect to use to fit the referent.

insensitive – meaning that for a gesture the amount of fingers matters little – and that one-hand gestures are preferred over two-hand gestures.

Mignot et al. [30] studied a multimodal approach of speech and gestures for a PC-based furniture layout. Insights indicate that gestures are preferably used for the executing of simple, direct, and physical commands; Whereas speech is for more abstract and high-level commands. Robbe [31] followed this experiment by comparing unconstrained and constrained commands, learning that constraints improve usability due to the reduction of complexity.

A point of critique that has appeared in the last few years regarding elicitation studies would be the reproducibility of such outcomes. Nebeling et al [32] recreated a study, which was followed by a user study trying to implement the gestures proposed by participants. Their outcome was that a guessability study alone was not enough to design a system, but that also system designers need to be involved in the process. Their role is primarily to focus on the feasibility of suggested interactions.

In summary, elicitation studies belong to the participatory design approach and can be used to collect interaction proposals by participants, as well as the mental model thereof. The literature provides different contexts for elicitation studies which include, but aren't limited to: user interfaces, gestures, voice commands, and multimodality. There are also critical voices in the community proposing to include system designers in the analysis of elicitation studies to define the feasibility of user proposals.

2.3.1 Measures: Agreement Rate

The degree of consensus among participants has been assessed and documented in the literature using the agreement rate formula established by Wobbrock et al. [33]. Agreement rates yield standardized values within the [0..1] range, indicating the extent of concurrence among users. Equation 1 illustrates the formula.

The concurrence among users - or agreement - has been defined by Villarreal-Narvaez et al. [20] as a situation where the proposed gestures of two or more participants are evaluated as identical or similar, given a preset of rules, criteria, and / or similarity functions; Therefore, being perceived as equivalent in regards to the target application. The defined criteria are linked to the taxonomy, which is why they are further explained there – section 2.3.2 Measures: Taxonomies.

$$AR(r) = \frac{\sum_{P_i \subseteq P} \frac{1}{2}|P_i|(|P_i|-1)}{\frac{1}{2}|P|(|P|-1)}$$

Equation 1: Agreement rate equation by Wobbrock et al. [33]

In Equation 1, "r" represents a referent from the set of all referents "R". |P| denotes the size of the suggested gestures for referent "r", while "Pi" represents a subset of gestures from |P| that are identical. To illustrate, an example calculation of a referent that has been part of the elicitation study conducted for this work: Overall there were 32 propositions. The distribution for gestures would be: 22, 6, 2, 1, and 1 respectively. The AR(r) would be calculated the following:

$$AR(r) = \frac{\frac{22*21}{2} + \frac{6*5}{2} + \frac{2*1}{2} + \frac{1*0}{2}}{\frac{32*31}{2}} = 0{,}52$$

For the analysis of an AR Vatavu and Wobbrock [33] propose the use of the AR(r) interval, which is based on the probability distribution function of AR. The function can be seen in Figure 1.

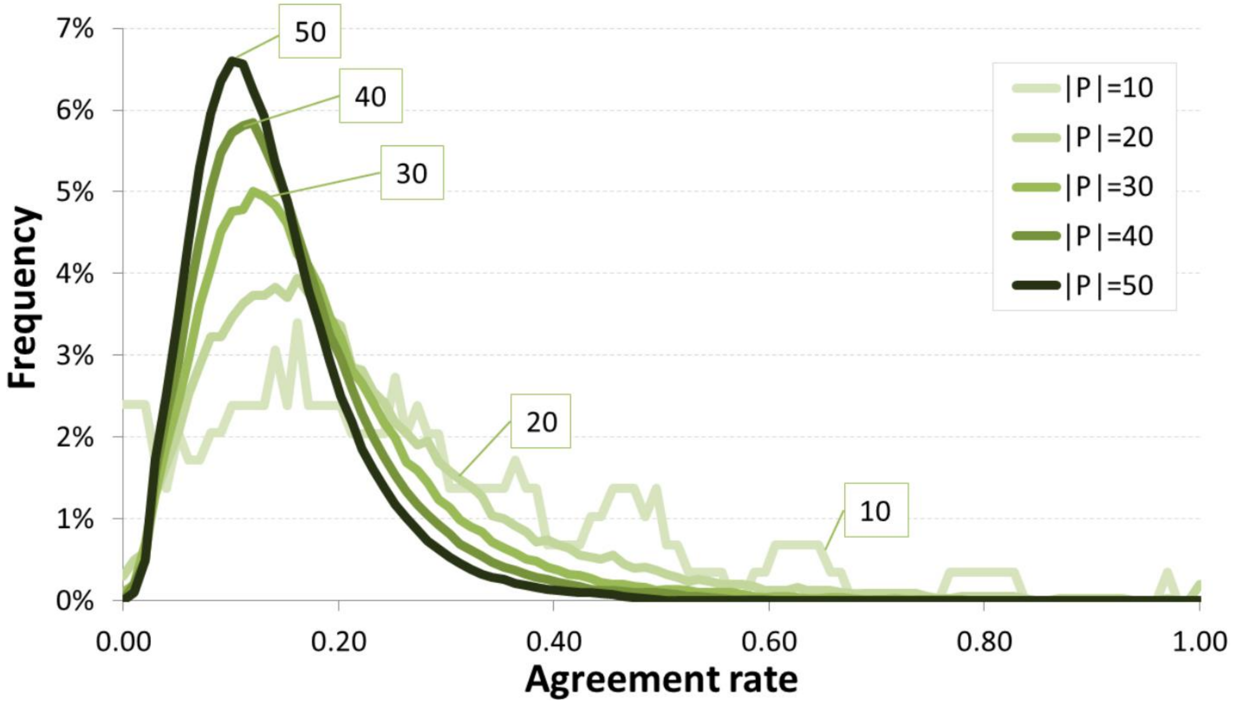


Figure 4: Probability distribution functions of AR computed by Vatavu and Wobbrock for [33] various numbers of participants |P| from 10 to 50.

Vatavu and Wobbrock [33] argue that a cumulative probability of 90% is reached for AR <= 0,374, while a cumulative 99% is reached for AR <= 0,636 given |P| = 20 participants. It is further stated that with the increase in the number of participants, the peak of the probability distribution shifts towards a lower value e.g., 90% cumulative probability for AR y= 0,222 and 99% for AR y= 0,424 for |P| = 50 participants. From an interpretational point of view, this would mean that e.g., an AR (r) > 0.500 would have a 3.9% chance of occurrence. Meaning that potential results for referents are more likely to be in disagreement than in agreement. To help with the interpretation of overall AR(r) the margins for interpretations are provided by Vatavu and Wobbrock [33] and will be used in this work. They can be found in Table 1.

AR(r) interval	Probability	Interpretation
<= 0,100	22,9%	Low agreement
0,100 – 0,300	59,1%	Medium agreement
0,300 – 0,500	14,1%	High agreement
> 0,500	3,9%	Very high agreement

Table 1: Margins for interpreting the magnitude of agreement as proposed by Vatavu and Wobbrock [33]

2.3.2 Measures: Taxonomies

There are different propositions of how to define the gestures in an elicitation study. Wobbrock et al. [5], inspired by Efron [15], propose four dimensions: form, nature binding, and flow. Each dimension has multiple categories, defining the proposal per referent further – see Table 2. Because of the complexity of this taxonomy, which makes it not feasible for one person to use it within the short time frame of a master's thesis, a different taxonomy was chosen. It is still mentioned in this work because it is a taxonomy for tabletop interaction close to the HyLTE context.

Taxonomy of Surface Gestures		
Form	Static pose	Hand pose is held in one location.
	Dynamic pose	Hand pose changes in one location.
	Static pose and path	Hand pose is held as hand moves
	Dynamic pose and path	Hand pose changes as hand moves
	One-point touch	Static pose with one finger.
	One-point path	Static pose & path with one finger
Nature	symbolic	Gesture visually depicts a symbol
	physical	Gesture acts physically on objects
	metaphorical	Gesture indicates a metaphor.

	abstract	Gesture-referent mapping is arbitrary
Binding	Object-centric	Location defined w.r.t object features
	World-dependent	Location defined w.r.t world features
	World-independent	Location can ignore world features
	Mixed dependencies	World-independent plus another
Flow	discrete	Response occurs after the user acts
	continuous	Response occurs while the user acts.

Table 2: Taxonomy of the surface gestures based on Wobbrock et al. [5]

Gheran et al. [18] also proposed a five-dimension taxonomy – nature, structure, complexity, symmetry, and locale - based on Wobbrock et al. [5], Ruiz et al. [22], and Piumsomboon et al. [21]. As Table 3 shows, the categories of the dimensions are kept more general, reducing the complexity of classifying gesture proposals. Due to the combination of three taxonomies and the definition of more general dimensions, the taxonomy by Gheran et al. [18] was chosen, though any categories solely related to rings were eliminated, due to them not being relevant for this work.

Taxonomy of Smart Ring Gestures		
Nature	Symbolic	Gesture visually depicts a symbol
	Metaphorical	Gesture indicates a metaphor.
	Abstract	Gesture-referent mapping is arbitrary
Structure	Buttons-only	Is a button pressed and for how long
	Hand poses only	The form of the hand is important not the movement
	Hand motion	The movement of the hand is important not the form
	Hand poses & motion	The form of the hand and the movement is important
	Mixed locals	Hand poses and motion have a varying degree of importance during the execution.
Complexity	Simple	Gesture consists of a single gesture
	Compound	Gesture can be decomposed into simple gestures

Symmetry	Dominant unimanual	Only the dominant hand is used
	Nondominant unimanual	Only the nondominant hand is used
	Symmetric bimanual	Both hands are being used in a symmetric fashion
	Asymmetric bimanual	Both hands are used in an asymmetric fashion
Locale	On the ring	Gesture that needs to touch the ring
	On other surface	Gesture that uses a surface for the execution
	In-the-air	Any gesture that uses the air / space in front of the participant / above the tabletop.
	Mixed locales	A combination of the locale categories mentioned above.

Table 3: Taxonomy of ring gesture controls proposed by Gheran et al. [17]

The following will briefly describe the dimensions proposed by Gheran et al. [18]:

1) *Nature*: This categorization characterizes the meaning behind gestures, classifying them into three groups: (a) symbolic, (b) metaphorical, and (c) abstract.
    a) **Symbolic gestures** employ well-known symbols to convey information. Examples include cultural gestures like the "call me" gesture, where the thumb and little finger are extended to symbolize a phone call. Another instance is swiping the index finger from left to right, a widely recognized action on touchscreens to go back to the previous item in a sequence.
    b) **Metaphorical gestures** give a tangible form to an abstract idea or concept. For instance, using the thumb to simulate “pressing a button” on an imaginary remote control to switch a TV on or off, or turning an invisible knob in the air.
    c) **Abstract gestures** lack any symbolic or metaphorical associations with their referents. The mapping is entirely arbitrary. An example would be three taps on a table to jump to the next video in a playlist – given the HyLTE prototype.
2) *Structure:* This classification system assesses the significance of hand poses and hand motion in the execution of gestures, distinguishing between five categories:

   a. **Buttons-only**: The primary focus of this category is on the act of pressing the button and the duration for which it is pressed. The specific gesture used to press the button is irrelevant.
   b. **Hand poses only**: This category encompasses gestures where the precise arrangement of the hand holds meaning, while the motion of the hand itself is not a critical factor. An example is the "thumbs up" gesture.
   c. **Hand motion**: In this category, the movement of the hand takes priority, while the specific poses of the hand are of lesser importance.
   d. **Hand poses & motion**: Here, both the hand's/hands' poses and its/their motion/s play significant roles in conveying the gesture's meaning.
   e. **Mixed locales**: This category involves a combination of various elements, where different aspects of the gesture, such as hand poses and motion, hold varying degrees of importance.

3) *Complexity:* This dimension defines if a gesture consists out of a simple gesture, or can be decomposed into simple gestures originating from a compound gesture.
   a. **Simple gesture**: This is a gesture that holds meaning on its own, such as sketching a “circle” in mid-air.
   b. **Compound gesture**: Such gestures are comprised of individually meaningful gestures that can be broken down, for example, first pressing a button and then drawing a “circle” around the pressed button.
4) *Symmetry:* This dimension focuses on how the two hands are employed to produce gestures. Four categories can be defined:
   a. **Dominant unimanual**: A gesture in this category uses only the dominant hand.
   b. **Nondominant unimanual**: A gesture only executed by the nondominant hand.
   c. **Symmetric bimanual**: Such gestures are comprised by the use of both (dominant and nondominant) hands, executing the same gesture in parallel.
   d. **Asymmetric bimanual**: Such gestures also include both hands but may execute gestures differently, or at least time-shifted.
5) *Locale:* Defines the location in space where the gesture is performed:

a. **On the ring***:* Gesture will be executed by touching the ring. This category is excluded from the analysis for this work!
b. **On the surface***:* Using a surface, such as a table, for the execution of the gesture.
c. **In-the-air**: Using the air / free space above the tabletop for the execution of the gesture. Any gesture not touching the table will be perceived as in-the-air, even if only hovering minimally above it.
d. **Mixed locales***:* This category mixes the above-mentioned categories belonging to locale. This can include a combination of two or more.

For gesture clustering – to define if gestures are similar enough to be seen as the same – and inadvertently to calculate the agreement rate, the following process was followed: First, a gesture gets classified according to **complexity**. If a gesture falls into the compound category, it is split into its simple gestures, which are then analyzed separately. Secondly, the simple gesture gets classified according to their **symmetry**. If not explicitly stated by the participants, it was always assumed that the dominant hand was used. Thirdly, the **structure** of the gesture was analyzed, which was followed by the **locale** of the gesture and / or hands, if bimanual. The final step was to define the **nature** of the gesture. The decision tree in Figure 5 is a graphical representation of the process.

In summary, two proposals for taxonomies that can be used for the analysis of gestures have been discussed. Wobbrock et al. [5] defined one taxonomy that is specifically designed for tabletop interaction. Unfortunately, this taxonomy is too complicated in its application for one person, given the time constraints and the additional work that needed to be executed for this master's thesis, which is why an alternative had to be used; Therefore, the taxonomy proposed by Gheran et al. [18] has been used for this work – leaving out the **on the ring** *Locale*, because it does not apply to the HyLTE context. The categories of the chosen taxonomy are: Nature, Structure, Complexity, Symmetry, and Locale.

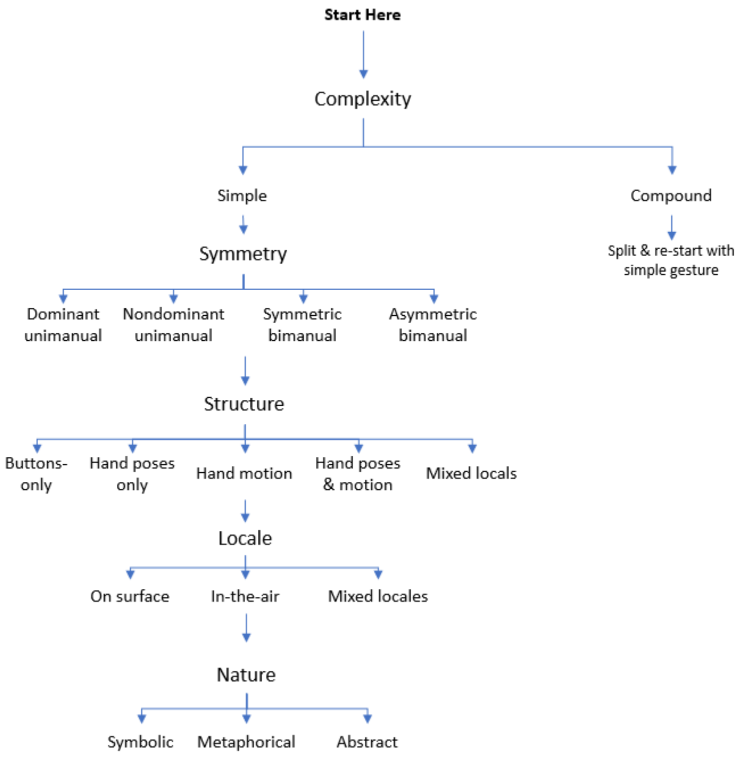


Figure 5: Decision tree based on the taxonomy for gesture clustering

2.3.3 Mental Models

Mental models can be referred to as cognitive frameworks formed by prior knowledge and experiences [34]. In the context of user interfaces (UIs), an intuitive UI aligns with users' expectations, which are guided by these mental models, influencing the user's interactions with the system. In essence, a UI or system becomes intuitive to the user when they possess prior experience with a similar UI featuring comparable interaction methods, or when it employs real-world analogies, such as utilizing tools like a brush, eraser, or magnifier in photo editing software [34].

Wobbrock et al. [5] discussed their observations of mental models in relation to their elicitation study. Six mental model observations were defined: (1) inversing gestures for the opposite effect, (2) using identical gestures with similar effects – zoom in and enlarge – only being distinguished by the context, (3) the amount of fingers for a gesture is usually not relevant, (4) Windows operating system influencing the proposed gestures, (5) that the space beyond projection and screens is also used, and (6) the space above the tabletop also being interpreted as an interaction area.

Peshkova et al. [35] argue that three major forms of mental models are relevant to gesture controls: (1) instrumented, (2) imitative, and (3) intelligent. Due to their primary research topic being on Unmanned Aerial Vehicles (UAV), it seems, at least for the moment without any further research, that the proposed categorization is not applicable for a bigger generalization. They discriminate between single mental model and mixed mental model gesture vocabularies though. If all gestures within a gesture set (vocabulary) are based on a single underlying mental model e.g., “space beyond the projection is also for interaction”, the set is called single mental model vocabulary, otherwise, it is a mixed mental model vocabulary. Hitz et al. [36] state that a single mental model vocabulary indicates to need less cognitive load for learning gestures than a mixed mental model vocabulary.

Peshkova et al. [34] propose a coherence score to calculate how easily deductible gestures from a vocabulary would be, fitting to one mental model chosen, if one gesture, belonging to the respective vocabulary, was revealed to the participants. The score is defined as a proportion of correctly guessed gestures ($N_{Guessed}$) to the total number of gestures (N_{Total}) minus the number of entries given as hints (N_{Hint}) for n participants. The formula for the score can be found in Equation 2. The max of the score is 1 and respectively the minimum score would be 0. Overall, the mean coherence score serves as an indicator of vocabulary coherence.

$$c = \frac{\sum_{i=1}^{i=N} \frac{N_{Guessed}^{i}}{N_{Total} - N_{Hint}}}{n}$$

Equation 2: Coherence score equation by Peshkova et al. [34]

The score is not applicable in this work because the user study is not designed for gesture guessing, but can be used as an inspiration to calculate the percentage of gestures participants can remember. The only variable excluded would be the number of hints (N_{Hint}) resulting in the following formula – 3:

$$c = \frac{\sum_{i=1}^{i=N} \frac{N_{Guessed}^{i}}{N_{Total}}}{n}$$

Equation 3: Adapted coherence score equation

In this work, the analogical representation of mental models will be applied. The analysis of the gesture proposals for the second GD will be solely based on this. This is because not only one mental model could be defined such as "Use your gestures like the UAV would be a puppet", but rather a combination of a "whiteboard" and "video player" concept.

In summary, Wobbrock et al. [5] as well as Peshkova et al. [35] stated that mental models may be a part of elicitation studies. Peshkova et al. argue that a GD can be classified into two categories regarding mental models: single mental model and mixed mental model GD. In the case of this work, two mental models could be identified – "whiteboard" and "video player" – as a mixed mental model GD.

2.4 Questionnaires

Two questionnaires were used in this work: the NASA Task Load Index (NASA-TLX) [37] and the After-Scenario Questionnaire (ASQ) [38]. They are used to define if one GD is more difficult to use than the other. The ASQ was used by Francese et al. [39] in combination with the System Usability Scale (SUS), which was not used in this work because the goal was not to gain insights into the usability of the system. Both of the questionnaires can be found in the section 8.1 Questionnaire.

2.4.1 NASA-TLX

NASA-TLX is the most widely used mental workload scale [40]. Mental workload is regarded as an important metric when assessing human work because levels that are too high or too low can adversely affect human performance: reducing efficiency, increasing the likelihood of human error, and creating undesirable conditions for human workers [40]. The NASA-TLX is a subjective self-assessment along the following six dimensions: (1) *mental demand*: How mentally demanding the human perceived the task; (2) *physical demand*: how physically demanding the human perceived the task; (3) *temporal demand*: How temporally demanding the human perceived the task; (4) *performance*: How successful the human felt they were at accomplishing the task goals; (5) *effort*: How hard the human felt they worked to accomplish their level of performance; and (6) *frustration*: How insecure, discouraged, irritated, stressed, and annoyed the person was during the task [37]. The dimensions – each of them having a rating from 1 to 20 - are then synthesized into a final, overall workload score.

Following the recent publication by Bolton et al. [40], only the analysis of the six dimensions will be done in this work. In the work, it is argued that there is no evidence that the different dimensions need to be combined/further calculated into one value and suggests analysing the different dimensions separately via descriptive analysis.

2.4.2 After-Scenario Questionnaire

The ASQ is a three-item after-scenario questionnaire. After participants finish a scenario – in this case the user study -, the questionnaire can be used. The three questions are: "Overall, I am satisfied with the ease of completing the tasks in this scenario.", "Overall, I am satisfied with the amount of time it took to complete the tasks in this scenario.", and "Overall, I am satisfied with the support information (on-line help, messages, documentation) when completing the tasks?".

2.5 Artificial Intelligence for Gesture Control

Artificial Intelligence (AI) methods for GC have been an ongoing research topic in different domains. Oudah et al. [41] and Guo et al. [42] each executed a meta-analysis on hand gesture recognition. The former stated that two primary classification methods exist: instrumented glove and computer vision. Similarly, Guo et al. [42] conducted a meta-analysis to gain an overview of how hand gesture recognition can be approached. Overall, six technological approaches are being explored: data gloves, vision, surface electromyography, ultrasound, other methods, and hybrid methods. Only the two overlapping approaches, gloves and computer vision will be further discussed because the other approaches are not relevant to this project due to the HyLTE prototype.

Oudah et al state that gloves are enhanced by sensors that precisely calculate the coordinates of the palm and finger locations, orientations, and configurations. They tend to be connected via wires to a computer which may limit the interaction area, though wireless solutions also exist.

They may additionally limit the fine motor skills of the wearer. Fiorini et al. [43] analyzed image and wearable-based feature analysis. They compared four benchmark algorithms – Support Vector Machine, Random Forest, K-Nearest Neighbour, and Long-Short Term Memory DNN – and concluded that a wearable – such as a SensHand [44] – improves the identification of gestures.

The computer vision method has seven sub-categories: (1) colour recognition – skin colour or marked glove, (2) appearance recognition, (3) motion recognition, (4) skeleton recognition, (5) depth recognition, (6) 3D model recognition, (7) and deep learning recognition [41]. Because of the diversity of approaches, only one example per approach will be discussed, to create an understanding of how such an approach can be implemented.

Colour recognition can be done by either calculating the skin colour of the user or detecting a marked glove. The method has already been used to research digital graphics, image process applications, TV transmissions, and the application of computer vision techniques [45], [46]. The major problem of skin colour detection is not only the diversity of skin colours but also the combination of colour channels and intensity information of an image.

For appearance-based recognition, different features are calculated from the input image and compared to features extracted from a base image – which can be seen as a ground truth. The difference to other ML-based approaches would be that no prior segmentation of the hand is executed; Reducing the steps of the gesture analysis pipeline [41], [47]. Fang et al. [48] employed an extended AdaBoost technique to detect hands, incorporating both optical flow and colour information for tracking. Additionally, they gathered hand colour data from the vicinity of feature positions' average using a single Gaussian model to represent hand colour in the HSV colour space. They conducted multi-feature extraction and recognized gestures through the segmentation of palms and fingers. To address the challenge posed by aspect ratio limitations commonly encountered in hand gesture learning, they integrated scale-space feature detection into their gesture recognition approach.

Motion-based recognition is the premise of dynamic gesture analysis and was briefly described in chapter 2.2. The primary focus is to analyze if a human moves and the prediction of the motion. Motion detection can only be done by analyzing multiple images that are time-sensitive and dependent on one another – meaning the order is important. In [49], a real-time dynamic hand gesture recognition system utilizing Time-of-Flight (TOF) technology was introduced. This system detected motion patterns by processing hand gestures captured as depth images. These motion patterns were then compared with hand motion classifications derived from actual dataset videos, eliminating the need for a segmentation algorithm. The system demonstrated satisfactory performance, although it was constrained by the depth range limitation of the TOF camera.

The skeleton-based recognition of gestures focuses on the calculation of a skeleton which overlays the detected hand – see Figure 6. The method aims to articulate geometric characteristics and constraints, facilitating the seamless translation of features and data correlations. This approach prioritizes both geometric and statistical features, enhancing the detection of difficult-to-track features such as fingers. Konstantinidis et al. [50] propose a novel single-lens reflect camera (SLR) system based on the processing of video sequences to extract precise body and hand skeletal data. The subsequent analysis and classification are executed using a DNN.

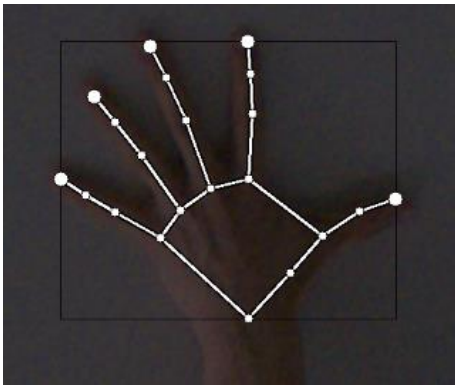


Figure 6: Example of a digital skeleton for a hand

The depth-based recognition uses 3D geometric information to predict the x, y, and z coordinates of a hand. All of the examples analyzed by Oudah et al. [41] use a Microsoft Kinect and a variation of classification algorithms such as k-Nearest Neighbour, state machine, and hidden Markov model. Qi et al. [51] propose the combination of an LSTM and Recurrent Neural Network (RNN) for the fusion of two Leap Motion controllers, which calculates digital skeletons, for the control of a surgical robot.

The 3D model-based approach uses, like depth-based recognition, 3D geometric data from e.g.: a Microsoft Kinect, and translates the data of the hand – volumetric or skeletal - into the 2D appearance of the 3D hand on e.g.: a TV. Limitations that have been stated are large datasets of images, especially when multi-view; alignment process is time-consuming; computational costs, and unclear views. [41]

The final approach focuses on deep learning, which uses DNNs for gesture detection. DNNs are a class of artificial neural networks characterized by their multiple layers of interconnected nodes, or "neurons." These networks are designed to process complex information in a manner inspired by the human brain's neural structure. Each layer of neurons processes input data, extracting progressively higher-level features as information flows through the network. This approach has been shown to excel at tasks like image and speech recognition, natural language processing, and even playing complex games. The term "deep" in DNNs refers to the depth of layers, distinguishing them from shallower neural networks. Though seen as powerful tools, Deep Learning needs extensive computing resources and massive datasets to train these networks. Due to that, deep learning may not always be possible, especially in niche applications where open-source datasets may be scarce. [52]

MediaPipe by Google also uses the approach of Deep learning [53]. MediaPipe is a software framework providing an architecture for DNN training as well as pretrained DNN models. Models provided by them cover use cases such as object detection, face identification, and gesture detection; Though more use cases are covered. They also support the concept of transfer learning where a pre-trained model can be used to be trained for a more specific task, such as the gesture control for the HyLTE prototype.

In various literature [41], [54], [55] it is pointed out that the computer vision approach contains several challenges that still occur – lighting variation, effect of occlusions, processing time tradeoff against resolution -, though some of them having already been mentioned by Wellner

[7] in the 80s. This stands in contrast to [42], who states that vision is a mature technology that is used to *"identify, extract, and classify the gesture features in the images"*.

In conclusion, this section provides a comprehensive overview of AI methods for GC research. The meta-analyses by Oudah et al. [41] and Guo et al. [42] shed light on two primary classification methods: instrumented gloves and computer vision. While gloves offer precise hand position calculations, they may have limitations in interaction areas and fine motor skills. On the other hand, computer vision encompasses various sub-categories like colour recognition, appearance recognition, motion recognition, skeleton recognition, depth recognition, 3D model recognition, and deep learning recognition, each with its unique approach and challenges. Notably, the depth-based recognition and 3D model-based approach provide a detailed understanding of how 3D geometric data can be utilized for gesture recognition. Meanwhile, the deep learning approach shows the potential of hierarchical information processing in AI applications. However, it's essential to acknowledge that challenges persist in computer vision, including lighting variations, occlusions, and processing time considerations. Despite differing perspectives on the maturity of vision technology, this chapter demonstrates the development of GC research in the last few years.

To conclude the chapter on literature research, the work presented here combines different research areas such as interactive tabletops, classification of human gestures, elicitation studies, and AI for gesture detection. A special focus was set on elicitation studies and AI because these topics were important for the design and / or execution of the studies, which will be further discussed in the next chapter Methods, and the analysis thereof.

3 Methods

With the insights gathered from the literature review, an approach was designed to answer the four sub-research questions. Question one – *What gestures are intuitively associated with video recording and player functionalities in a tabletop projection setting?* – will be answered via an elicitation study – section 3.1. It was chosen because it is the state-of-the-art approach for collecting interaction proposals. The data collected from the elicitation study resulted in the creation of two GDs – further discussed in chapter 4.1. For the second question – *Given the elicitation study outcome, what approach can serve as a proof of concept for implementing gesture recognition models?* – a combination of multiple concepts – general hand gesture interaction, machine vision, and DNNs – were combined to find a solution for the task, which led to the use of the MediaPipe software library – as shortly discussed in chapter 2.5. A description of the approach can be found in section 3.2.3 and the result in section 4.6. The third question – *How can participant preferences for gestures be evaluated using two GDs, one informed by the AR and the other by MMs?* – led to the conduction of a second study and will be discussed in section 3.2. The fourth and final question - *What insights and recommendations can be derived from developing and evaluating AI-driven gesture recognition models in tabletop projection settings for future projects?* – lead to the discussion of design recommendations of the overall process presented in this thesis and will be discussed in the chapter Results – section 4.5.

Based on the four sub-research questions this chapter is divided into two parts: First the elicitation study will be discussed. This is followed by the user study. For the results of the respective study and / or the DNN please have a look at chapter 4.

3.1 Elicitation Study

To gain proposals for potential gestures, an elicitation study was chosen. The study was set to take approximately 30 to 40 minutes – excluding organizational components such as the consent form. The technical setup is equivalent to the prototype described in chapter 2.2. The referents were designed in PowerPoint to be able to show representations that weren't implemented in the prototype at that time. It needs to be pointed out that the referents were non-interactable and changed their state over time. Gheran et al. [18] argued that they allowed contact-based and mid-air input in their elicitation study to give rise to more diverse gestures. This is also the case for this elicitation study.

3.1.1 Participants

Overall, twenty (20) participants took part in the study, which were all acquired by e-mail via the SUAS student e-mail server. All the active SUAS students were written to have as much a diverse participant group as possible. Half (10) of the participants identified as female, and all participants were right-handed. The age span is defined from 19 to 29 years (M = 22.75, SD = 2.9 years). None of the participants had any sort of physical movement limitations. Nearly all participants (80% - 16/20) have a familiarity with technology by either studying computer science or human-computer interaction. The rest have a background in design and media art. Additionally, 15% (3/20) of the participants had prior experience with gesture control technologies such as Leap Motion, Oculus Quest, and Microsoft HoloLens. The experiences were gathered during lectures of their respective studies.

3.1.2 Procedure

For training and to familiarize with the technology / setup, participants were given two training referents: Object selection and grouping. Both referents have been thoroughly researched and therefore little insight was lost by not including them in the study [20]. In each session, participants were presented with referents – e.g. turning on/off the video recording – for which they discussed and executed gestures (multiple suggestions were allowed). The participants were informed that there are no hardware or gesture limitations – besides obscene gestures commonly known in the western hemisphere – and that every gesture would be recognized by the system because it was believed that this would otherwise limit the creative output. The repetition of gestures for different referents was also allowed. The order of the referents was fixed, imitating the user journey of the HyLTE prototype of the video recording use case.

3.1.3 Design

The study was within-subjects and had referent as the only variable. 13 conditions, representing common interactions with a video player were analyzed: (1) video recording on / off, (2) audio recording on / off, (3) audio volume, (4) open / close main menu, (5) play / pause video, (6 & 7) jump forward / backward in video, (8) zoom video area, (9) annotate video, (10) change colour of annotation, (11) insert text to video, (12) delete ONE annotation, and (13) delete ALL annotations.

The choice of referents was based on well-known video platform interactions such as YouTube[20] and Dailymotion[21], and standard drawing software i.e. Microsoft Paint[22]. Some of the tasks are mutually exclusive (i.e. video recording on / off), which allows for the use of the same gesture to achieve, given the context, either functionality. This design decision was made based on prior studies that identified users' preferences for toggle gestures [23], [24].

3.2 User Study

The second study had two goals: gaining insights into the preferences of participants when it comes to gestures to define a final GD and gathering data to train the AI model for the final gesture dictionary. The study was conducted as an A/B-Test because certain gestures in the GD would overlap / negate each other, making a test implementation with both dictionaries impossible. Additionally, one run-through took approximately 45 minutes – without organizational components e.g., informed consent – potentially reducing the number of potential participants if the study had been approximately 2 hours. It was decided to conduct this second study, because few scientific publications in that regard are available, and it correlates with the criticism that proposed gesture dictionaries don't always have a technological implementation - or the possibility thereof - in mind [32]. The study was conducted over a week and took approximately one hour per participant with an equal distribution of eight (8) participants per GD. It was important that none of the participants of the first study took part in the second study to avoid bias.

The storyline for the study is that a child that stands in some kind of relation to oneself (own child, sibling, grandchild, etc.) would like to have a video explaining how the interlocking plastic bricks product Gyro Copter is ensembled. The concept of assembling interlocking plastic bricks was chosen because it was believed that such a concept is well-known, and the difficulty of the task can be reduced by pre-building certain parts of the construct. Contrary to belief, some

[20] https://www.youtube.com/

[21] https://www.dailymotion.com/at

[22] https://en.wikipedia.org/wiki/Microsoft_Paint

participants of the study never played with interlocking plastic bricks before, which made certain parts of the study last longer and potentially more complicated than intended. Fortunately, neither time nor the result of the Gyro Copter was a measurement of the study.

3.2.1 Participants

Eighteen (18) participants (5 female), aged between 19 and 37 years old (M = 28,24; SD = 4,36 years), volunteered for the study, which were all acquired by e-mail via the SUAS e-mail server. All the active SUAS students were written to have as much a diverse participant group as possible. Participants were all students studying subjects such as Computer Science, Business Studies, and Social Innovation. Of all participants, 50% (9/18) had a technical background, and all participants were right-handed. None of the participants had any sort of physical movement limitations. The majority of participants – 77% (14/18) – had little to no prior knowledge about gesture control as an interaction form with technology – excluding touch technology such as smartphones. The other third had already worked with Oculus, Microsoft Kinect, and Leap Motion.

3.2.2 Procedure

The study consisted out of four parts: *pre-study questionnaire*, *explanation & training phase*, *video recording*, and *post-study questionnaire & interview*.

A *pre-study questionnaire* was used in the beginning to collect sociographic data – age, gender, handedness – and collect background information such as prior experience with gesture interaction technology.

For the second part of the study – *explanation & training phase* - participants were introduced to the prototype, setup, and functionalities. The explanation of the functionalities were fostered by the respective GD – depending on which dictionary the participants received – and a demonstration by the experiment led to demonstrate that all functions were technically implemented. After the participants had time to think through and discuss the gestures, they were given time to train and familiarize themselves with the gestures for a maximum of approx. 20 minutes. This was done since participants were not familiar with this setup and / or in-air gestures. For this, a separate Unity program was developed, to prevent participants from accidentally starting the following task before feeling comfortable enough with the gesture control.

After the participants felt comfortable enough with the gesture control, they began the *video recording* step of the study. Similar to the HyLTE setup, this task had three scenes in a Unity application. The first scene allowed for the participants to start / stop the video and audio recording, followed by the building of the Gyro Copter. After finishing and stopping the recordings, a second scene for annotation could be entered via the menu. Here the participants had a chance to interact and annotate with the before-recorded video. The final scene showed the participants the finished video (with annotations). This was meant to give the participants and understanding of what they just created.

The final task of the study was a *post-study questionnaire & interview*, collecting both quantitative and qualitative feedback. For the post-study questionnaire, the NASA-TLX and ASQ questionnaire were used. As for the interview, the following five questions were asked: (1) What was the most positive aspect of your experience?, (2) What was the most negative aspect of your experience?, (3) Could you please recount all the gestures you just learned and what functionality they have?, (4) Were there gestures you liked and if yes, what did you like about them?, (5) Where there gestures you did not like and if yes, what did you not like about them?, and (6) Any additional remarks you would like to make?

3.2.3 Prototype

For the conduction of the study, a custom experiment software was developed to be able to interact with the prototype and simultaneously record video and audio data. The test prototype UI was implemented as a Unity application, the gesture tracking was developed expanding the Python implementation of the HyLTE prototype, and another Python script was implemented for the video and audio recording. To make the prototype recognize the gestures without having a model that has been already trained for that use case and / or a dataset to gain such a model, a combination of invisible UI elements and gesture detection was used.

Overall, the pre-defined gestures of the base model could be used. If specific gestures were not implemented a minor expansion of the model's capabilities was done. The expansion of the basic

AI model was conducted by expanding the model capabilities via transfer learning, using only a little data – gathered from the researcher - to train it. Gestures of the expansion were: hand side, index finger and thumb touch / opening, and stop gesture.

To summarize this chapter, from the initial research question, three sub-questions were derived. As a first step, literature research was conducted to gain an understanding of the state of the art, which was then followed by two studies: elicitation and user study. The first study – based on participatory design - was conducted to gain gesture proposals, whereas the second was to collect data for the final model and get an indicator which GD and/or gestures participants liked. The prototype for the second study was created by using components from the HyLTE prototype with Unity being the UI and Python taking over the AI part. Functionalities that were not yet available, were implemented.

4 Results

This chapter presents the findings from both conducted studies. It initiates with an examination of the insights gained from the elicitation study, with a detailed discussion of the analytical approaches outlined in Chapter 3.3. Subsequently, the chapter discusses the two GDs as intermediate outcomes. Following this, the chapter provides an evaluation of the user study results, followed by an overview of the two questionnaires and further fostering the insights through an analysis of the qualitative data. The culminating point is the presentation of the final GD outcome.

4.1 Elicitation Study

The results from the elicitation study – discussed in section 4.1.1 – show a low average AR according to Vatavu et al. [33]. Following the guidelines provided by Vatavu et al. [33] for interpreting agreement rates, the findings fall within, though close at the lower end of, the range of medium agreement (0.100 – 0.300). This might seem discouraging at first, but supports the ambition to create a second GD based on mental models. Furthermore, there might be different factors that influence the overall low rate of agreement. Examples of why this could be the case are also discussed by [18].

4.1.1 Results

Once all 20 participants had contributed gestures for the referents, the gestures were organized into groups for each referent, where each group shared identical gestures. The size of each group was then utilized to calculate the AR, as proposed by Wobbrock and Vatavu [2], for the first GD. Overall, a total of 319 gesture proposals were collected, which were clustered into categories of similar types inspired by the category criteria defined by [18]:

1) *Handedness.* Gestures where it was stated that the use of either the left or right hand – or the differentiation of the dominant and non-dominant hand - is important, are considered differently.
2) *Scale.* Gestures executed at various scales are classified separately. For instance, a "circle" with a wide amplitude carried out using the entire arm is distinct from a small-scale "circle" performed using just the finger. In this work, three scales (large,

medium, and small) were aligned with the ranges of motion for the arm, wrist, and fingers, respectively.

3) *Direction.* The same gestures performed in different directions are considered different, e.g., clockwise and counter-clockwise “circle” shapes drawn in mid-air.

In total 112 distinct gestures, for their respective referent, were identified. After the clustering of gestures, the AR was calculated. The outcome of the AR calculations can be seen in Figure 7.

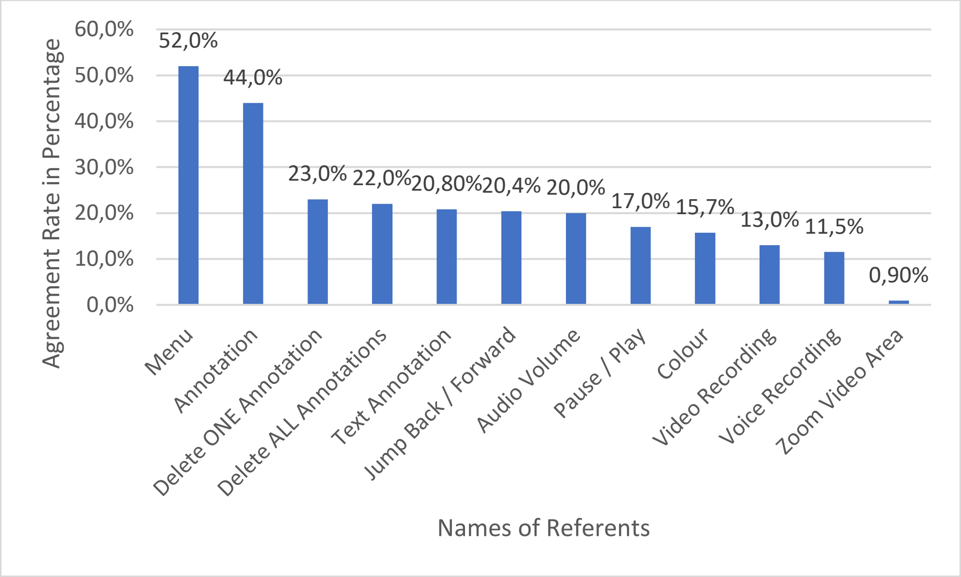


Figure 7: Agreement rates from elicitation study for tabletop interaction. Note: Referents are ordered on the horizontal axis in descending order of their agreement rates.

The agreement rates vary from 52% for opening and closing the menu - being the gesture most agreed upon - to 0,90 % for zooming into an area of the video - being the gesture least agreed upon – with an average AR of 16,3% and an SD of 13,2%. “Menu” and “Annotation” are the two gestures that have the highest AR. Both of them are close to double the AR of the third gesture “Delete ONE Annotation” with 23% and fall – according to Vatavu and Wobbrock [33] - into the range of “very high agreement” or “high agreement” respectively. “Delete ONE Annotation” till “Voice Recording” have an AR starting at the bottom half of the 20% and, ranging till the bottom half of the 10% area and can be classified as medium agreement. Finally, “Zoom Video Area” with 0,90% AR should be classified with low agreement.

A shortlist of gesture proposals was created to give an overview of the diversity of gestures proposed. Proposals range from UI elements - such as the burger menu or symbol selection (with a click of the respective symbol) - to symbolic hand poses - like the imitation of a camera lens/video frame -, and to motion gestures performed in mid-air - clap hands or holding a sponge. The shortlist is provided in Table 2.

Referent	AR	Most frequent	Second most frequent
Menu	0,52	Dragging in from top left corner to center	Burger Menu (UI element)
Annotation	0,44	Draw with index finger	Selection of symbols (UI element)
Delete ONE Annotation	0,23	Select element & drag out of area	Imitate cleaning sponge
Delete ALL Annotations	0,22	Big gesture swiping the table	Select all elements and drag them out of area
Text Annotation	0,208	Long click & digital keyboard pops up	Draw with finger (similar/identical with first suggestions for Annotation)
Jump Back / Forward in Video	0,204	Double tap left / right	Swipe left / right
Audio Volume	0,20	Volume bar slider: left-hand base and right-hand indicator	Diverging hands: the further apart, the louder
Pause / Play Video	0,17	Tap in the middle of video	Swipe left to right: Pause; Swipe right to left: play
Change Color for Annotation	0.157	Long press on annotation & select colour	Spread fingers
Video Recording	0,13	Imitating camera (lens)/video frame	Swipe in from one edge
Voice Recording	0,115	Clap hands	Imitate talking with hands
Zoom Video Area	0,009	Encircle area of interest with finger	Spread fingers over area of interest

Table 4: First and second most frequent gesture proposals for each referent

To better understand the gesture proposals, they are analyzed based on the taxonomies defined in sub-chapter 2.3.2. Figure 8 illustrates the observed percentages of gestures falling in each category. The dimensions are split into their respective categories and accumulate to a total of 100 %. For the *Nature* domain, there seems to be a balance between abstract and metaphoric gestures – each approx. 40% - with symbolic gestures being the least used category. Examples of these gesture categories would be “imitating camera (lens) / video frame” for symbolic, “imitate cleaning sponge” for metaphoric, and “spread fingers” for abstract gestures. For the *Structure* domain, the biggest category would be buttons with 32,40%. The second biggest category would be “Hand poses & motion” with 26,1%, followed by “Hand motion” with 22,5%, and “Hand poses” with 15,7%, respectively. The lowest category, with 3,30% would be mixed locales. Examples for these categories would be – in order of occurrence in the prior sentence: “click burger menu”, “hands parallel to the camera and slowly pulling away from each other”, “swipe with any hand”, and “imitating the clearing movement of a table by putting the flat hand in a 180° on the table and then ‘pushing’ everything from the table”. The third domain – *Complexity* – has an 89,8% portion of simple gestures e.g., “put the hand in the middle to ‘play/pause’ the video” and 10,20% of compound gestures such as: “the left hand is used for the base of the audio control and the right hand is used to define the volume. The further away the right hand from the base, the louder the volume should be”. For the fourth dimension – *Symmetry* – more than half (52,10%) of the gesture proposals focus on the “dominant unimanual” category, followed by the “Symmetric Bimanual” category (35,20%). The “Asymmetric Bimanual" category has a 9,30% share, and the final category – “Nondominant Unimanual” – has a 3,40% share. Examples of proposed gestures would be respectively: “imitating holding a pen and then drawing to make the annotation”, “clap to start and stop the audio recording”, “imitating camera (lens) / video frame”, and “long press with the left – non-dominant - hand to call the colour UI and then selecting with the right hand”.

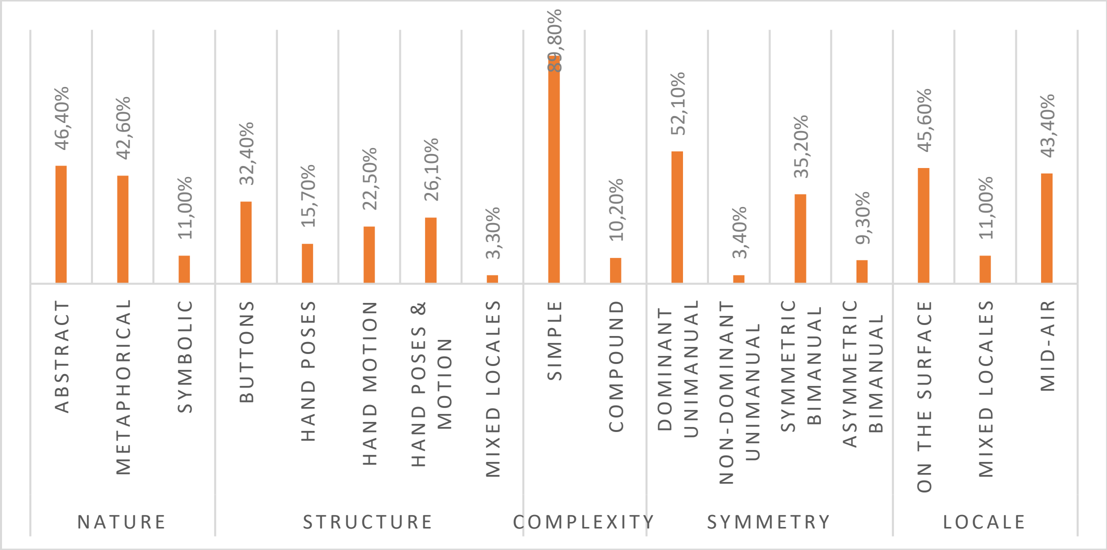


Figure 8: Observed percentages of gestures for each category of the chosen taxonomy

The final dimension would be Locale. The categories “On the surface” and “Mid-air” have a similar value of approx. 44,00%. “Mixed locales” on the other hand was set to 11%. Examples of the gestures would be “double tap on the left side of the video to jump back some seconds”, “putting index finger and thumb together twice, imitating talking”, and the audio control gesture.
The explanations used for the different domains are quotes – translated into English – from participants.

In summary, the gestures were categorized based on handedness, scale, and direction. A total of 112 distinct gestures were identified. Agreement rates varied, with opening and closing the menu having the highest agreement (52%) while zooming into a video area had the lowest (0.90%). A shortlist of proposed gestures showcased a diverse range, from UI elements to symbolic hand poses and motion gestures. The analysis revealed a balance between abstract and metaphoric gestures in the Nature domain, with symbolic gestures being less common. In the Structure domain, buttons were the most prevalent category, followed by hand poses and motion. Complexity primarily consisted of simple gestures (89.8%), while Symmetry leaned towards dominant unimanual gestures (52.10%). Locale categories "On the surface" and "Mid-air" were equally common (44%), while mixed locales made up 11%. The explanations in the different domains were provided by participants, offering valuable insights into the proposed gestures.

4.2 Gesture Dictionaries

Based on the analysis in section 4.1.1, two GDs were created. The design of the GDs is based on Gheran’s [18], [56] and Wobbrock’s [5] depiction of gestures for their elicitation studies. The first GD was created by using the AR, following the approach propagated by Villarreal-Narvaez et.al and Wobbrock et al [20], [57]. The second is also based on an approach by Wobbrock et al. [5], focusing on mental models, mentioned by the participants during the elicitation study. The second approach of a dictionary design was chosen because it seemed an interesting idea and was fostered by the remarks by Gheran et al. [18], [56] that users might need more experience with systems to get an understanding of their preferred gestures.

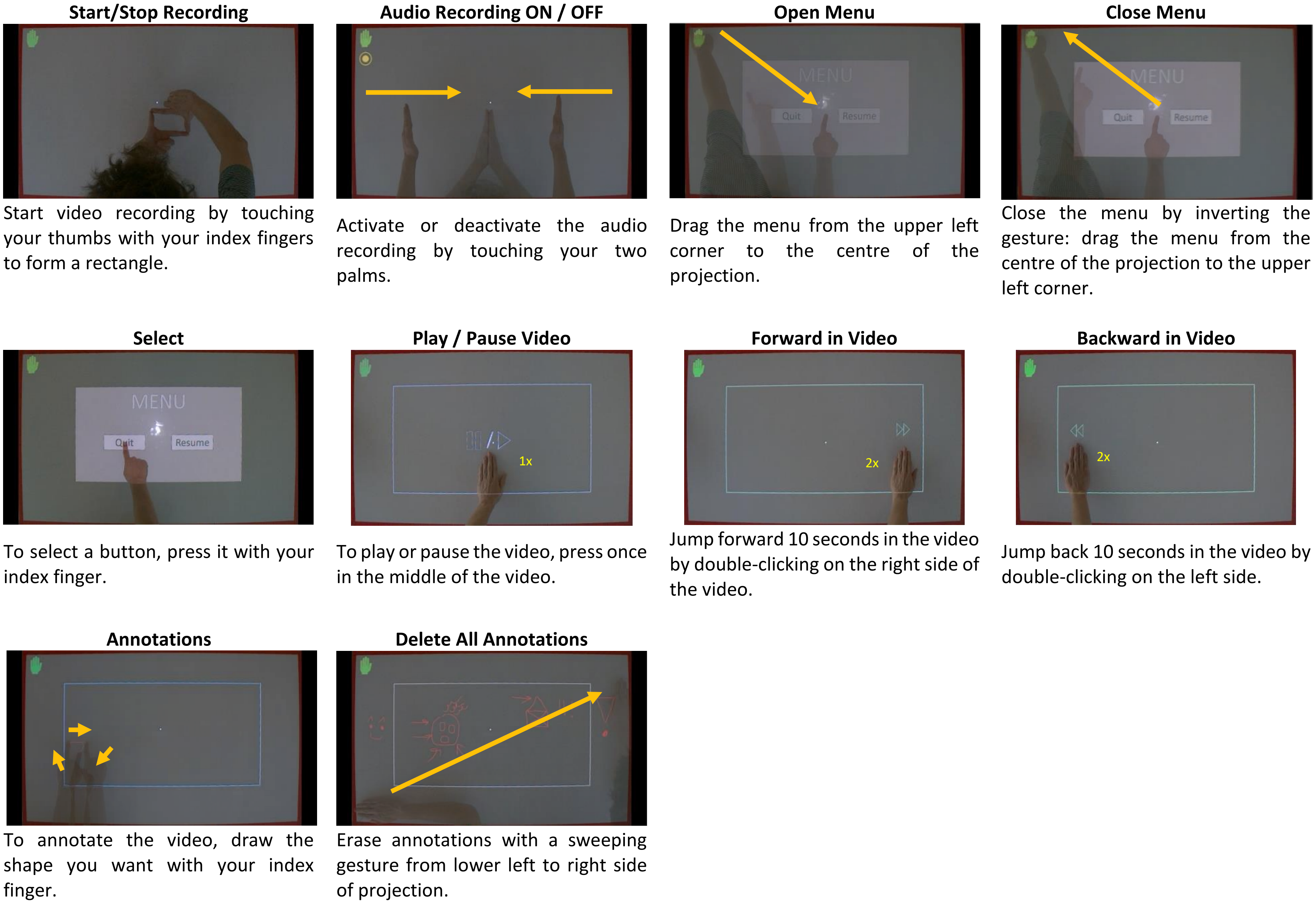
Not all the gestures evaluated by the elicitation study were integrated into the GDs, due to time reasons. Therefore, only the functionalities deemed as fundamental for the prototype were integrated. Gestures that were excluded are: (3) audio volume, (8) zoom video area, (10) change colour of annotation, (11) insert text to video, and (12) delete ONE annotation

4.2.1 Gesture Dictionary based on Agreement Rate

Inherently, no conflicts arose in cases where the same gesture was employed to execute distinct commands, as one gesture cannot lead to different results. For the implementation one adaptation of the gestures needed to be made though: for the video control (Play / Pause Video, Jump Forward / Backward in Video) the selection with the pointer was changed to a selection with the whole hand. This adaption was inspired by a suggestion from the elicitation study and according to [5], [58], changing the amount of fingers used for a gesture is non-essential for users.

Select, though only part of the elicitation study as a training instance, was added to the GD to navigate the main menu that has the functions of changing a scene and quitting the application. The two functions of the main menu were not implemented via gestures because they were perceived as too powerful – especially quitting the application – for a prototype and may cause a high level of frustration if executed accidentally. The chosen gesture for select is a click with one of the index fingers. The intention was to have a neutral gesture that is overall well-known.

Consequently, the AR-based gesture set – depicted in Figure 7 - is free of conflicts and encompasses 57.0% of all proposed gestures. The final result of the AR-based approach is a consistent set of user-defined gestures that contain 10 gestures, where 8 gestures are unimanual and 2 are bimanual.



Start/Stop Recording

Start video recording by touching your thumbs with your index fingers to form a rectangle.

Audio Recording ON / OFF

Activate or deactivate the audio recording by touching your two palms.

Open Menu

Drag the menu from the upper left corner to the centre of the projection.

Close Menu

Close the menu by inverting the gesture: drag the menu from the centre of the projection to the upper left corner.

Select

To select a button, press it with your index finger.

Play / Pause Video

To play or pause the video, press once in the middle of the video.

Forward in Video

Jump forward 10 seconds in the video by double-clicking on the right side of the video.

Backward in Video

Jump back 10 seconds in the video by double-clicking on the left side.

Annotations

To annotate the video, draw the shape you want with your index finger.

Delete All Annotations

Erase annotations with a sweeping gesture from lower left to right side of projection.

Figure 9: GD sheet and instructions based on AR

4.2.2 Gesture Dictionary based on Mental Models

Unlike the AR-based approach, this one is more based on the insights and thoughts of the designer. Solely based on the input received from the elicitation study, but still, more expert / design-driven decisions were made than the pure mathematical one by the AR. Therefore, there were also no conflicts in creating the second GD. As argued in the AR-based GD, the Select gesture was also added in this GD for the navigation of the main menu.

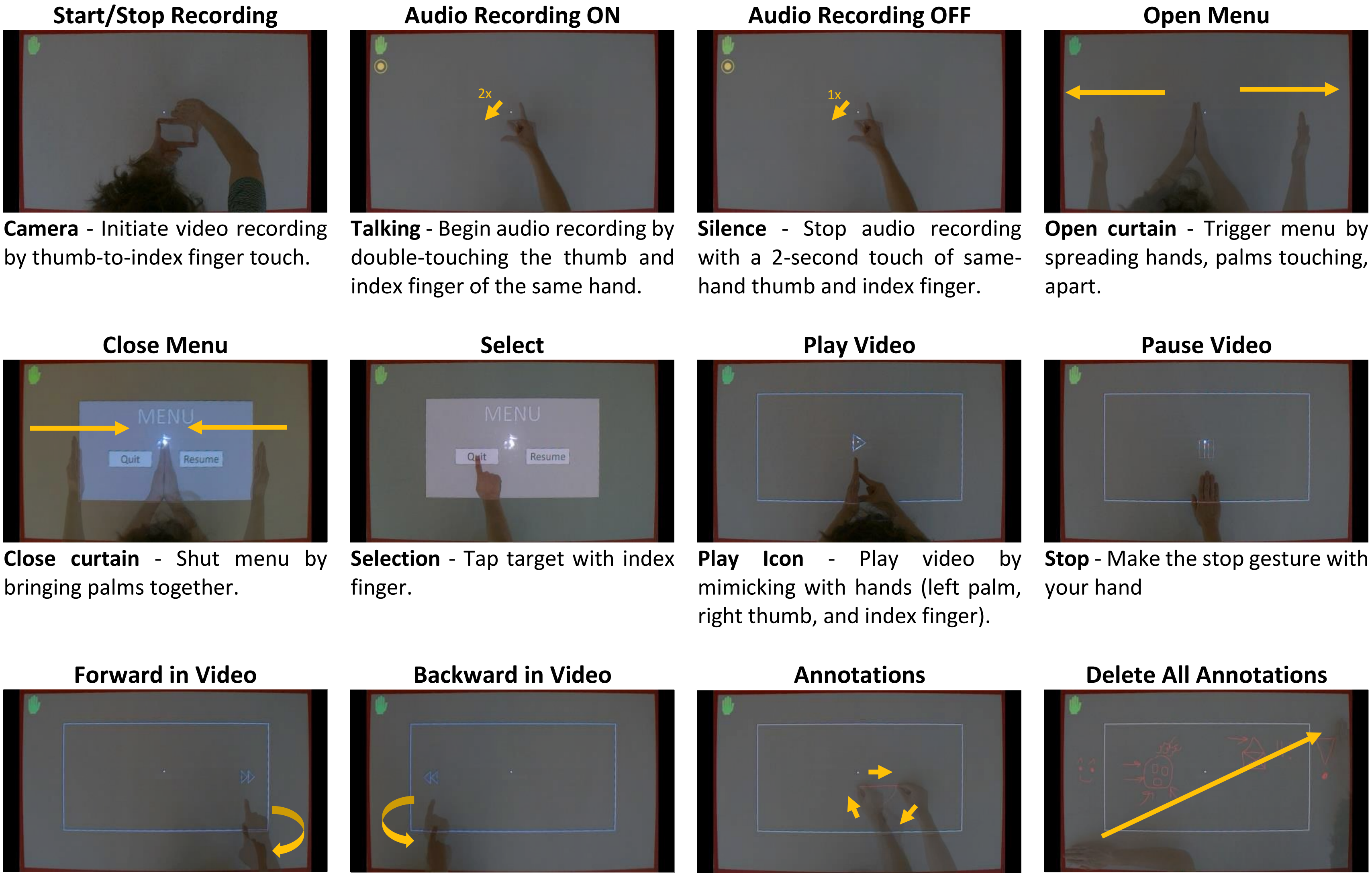
Table 5 includes a description of the gestures used for the mental model-based GD, including the quotes – mostly translated from German to English - and a description of the gesture as proposed by participants. The select gesture has been excluded because it is no analogy but still fits into the GD because of the underlying "touch screen mental model".

Functionality	Gesture	Quote / Description
Start / Stop Recording	Putting index fingers and thumbs together, creating rectangular	"It is like creating a frame, where you can look through. Just like with a camera"
Audio Recording ON	Putting index finger and thumb, of one hand, together twice	"Like if you gesture someone is talking."
Audio Recording OFF	Putting index finger and thumb, on one hand, together once. Shortly lingering in this gesture.	"Similar to 'If you want someone to shut up'. Like 'just zip it'."
Open Menu	Putting your hands together and slowly move them apart.	"I imagine that my hands are curtains and I am slowly pulling the curtain open to see what is behind."
Close Menu	Inverse of Open Menu: Hands are put together	"When I am done with whatever is behind the curtains, I am just closing

		them. Because now I want to focus on what is in front of them.”
Play Video	Imitate the play symbol	“I just feel like showing the system that it should start playing.”
Pause Video	Stop motion	“Just like the stop sign you would make if someone comes too close to you.”
Forward / Backward in Video	Rotating one index finger (counter) clockwise	“Turning the clock forward or backwards.”
Annotations	Imitating holding a pen.	“I want to draw on it. So for me, holding a pen would make the most sense.”
Delete ALL Annotations	Big swipe from left to right	“Pushing everything off the table. Getting rid of it as quick as possible.”

Table 5: Description of the mental models for the second GD

Consequently, the mental model-based gesture set – depicted in Figure 10 - is free of conflicts and encompasses 16% of all proposed gestures. The final result of the mental model-based approach is a consistent set of user-defined gestures that contain 12 gestures, where 8 gestures are unimanual and 4 are bimanual.



Start/Stop Recording

Camera - Initiate video recording by thumb-to-index finger touch.

Audio Recording ON

Talking - Begin audio recording by double-touching the thumb and index finger of the same hand.

Audio Recording OFF

Silence - Stop audio recording with a 2-second touch of same-hand thumb and index finger.

Open Menu

Open curtain - Trigger menu by spreading hands, palms touching, apart.

Close Menu

Close curtain - Shut menu by bringing palms together.

Select

Selection - Tap target with index finger.

Play Video

Play Icon - Play video by mimicking with hands (left palm, right thumb, and index finger).

Pause Video

Stop - Make the stop gesture with your hand

Forward in Video

Fast forward time - Let the index finger of your right hand rotate clockwise.

Backward in Video

Rewind time - Rotate the index finger of your right hand counterclockwise.

Annotations

Writing - Imitate pen-holding, advance annotation with fingertips.

Delete All Annotations

Clear table - Sweep palm from bottom left to right side.

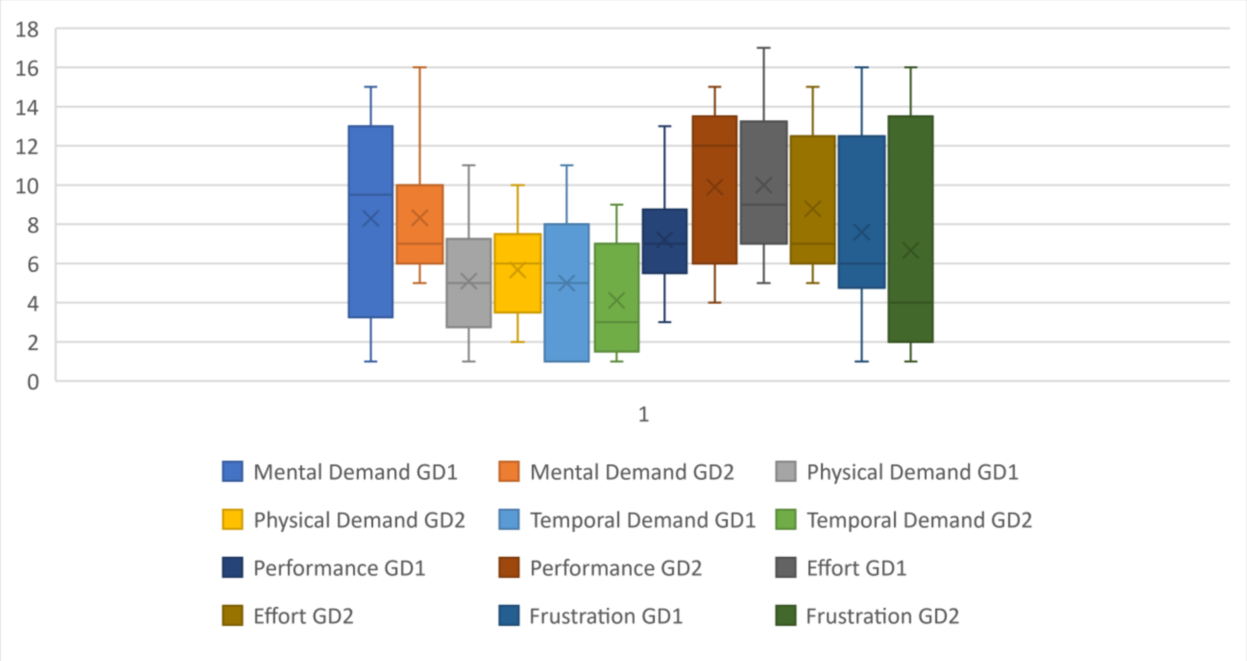
Figure 10: GD sheet and instructions based on MM

4.3 User Study

This section discusses the results of the user study. It starts with the analysis of the questionnaires, followed by the insights gained from the interviews. This leads to the presentation of the final GD, accompanied by design recommendations. For all the observed values a Jarque-Bera test was executed to assess whether dependent variables were normally distributed. Due to all of them being the case, parametric statistical tests such as the paired and unpaired t-test and Pearson correlations were used.

4.3.1 NASA-TLX

The boxplot in Graph 1 illustrates the results of the NASA-TLX questionnaire. As discussed in the literature review chapter, the metrics are analyzed separately. Following the two corresponding metrics – e.g.: Mental Demand GD1 and Mental Demand GD2 – will be discussed together.



Graph 1: Boxplot of the NASA-TLX results

For *Mental Demand,* the data for GD2 shows less variation than the one for GD1. The mean for each is at approx. 8,3 though the median for GD2 is lower (7) than for GD1 (9,5). *Physical Demand* is similar for both GDs rations, though GD1 has a slightly bigger Inter Quantile Ration (IQR) than GD2. The mean for GD1 is a bit lower (5,1) and nearly overlaps with the median (5) in comparison to the mean (5,6) and (6). The mean and median of *Temporal Demand* for GD1 would be 5, which is higher than the respective values for GD2: 4,1 and 3. The IQR for *Performance* for GD1 is significantly smaller than for GD2 which is also demonstrated by the mean (7,2 vs 9,8) and median (7 vs 12). For *Effort* the mean values are close (9,8 vs 10), though the median for GD1 is higher than for GD2 (12 vs. 9). The final metric – Frustration – shows that GD1 has a mean of 7,6 and a median of 6 vs a mean of 6,6 and median of 4 for GD2; Though, the IQR is bigger for GD2.

Following the boxplot, a paired and unpaired t-test and Pearson correlations were calculated for each metric. Table 6 summarizes the results and also shows the mean and variance of the respective category.

Category	Mean	Variance	Pearson Correlation	P-Value (one-tail)	P-Value (two-tail)
Mental Demand GD1	7,186153654	4,653420354	-0,031040489	0,353159748	0,706319496
Mental Demand GD2	7,787764581	3,333333333			
Physical Demand GD1	4,347129523	2,943920289	-0,512410092	0,402959823	0,805919646
Physical Demand GD2	5,120317173	2,357022604			
Temporal Demand GD1	4,136823785	3,235604395	-0,210498888	0,197313519	0,394627038
Temporal Demand GD2	3,093170706	2,845832994			
Performance GD1	7,038181032	2,753224821	-0,097730335	0,093959508	0,222437858
Performance GD2	8,994403304	3,899984172			
Effort GD1	9,716418835	3,559026084	-0,011930787	0,19521207	0,390424141
Effort GD2	8,148994236	3,48895966			
Frustration GD1	6,095233319	4,732342098	0,030228223	0,338563572	0,677964477
Frustration GD2	4,443860655	5,696002497			

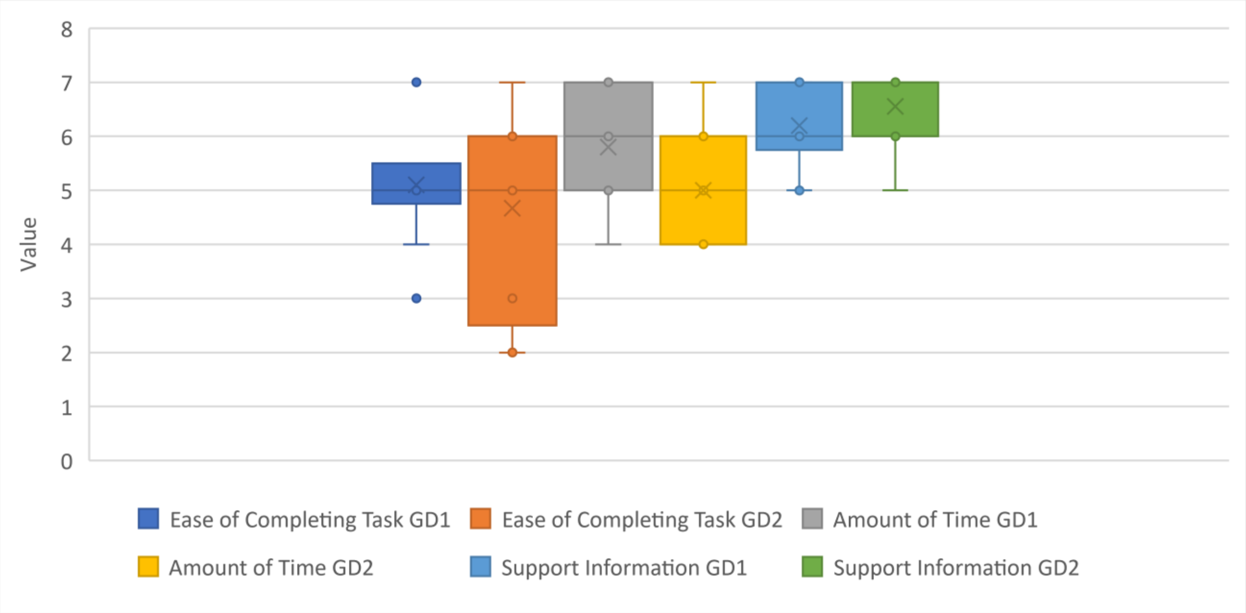
Table 6: Descriptive Analysis Parameter for NASA-TLX Categories

All the correlation values show little to no correlation except for the *Physical Demand* category with -0.51, which is also a rather low value. Regarding the p-value, none of them reached a significant value and therefore it is assumed that the H1 hypothesis – Dictionary 2 is better than Dictionary 2 – needs to be rejected.

Overall, there might be some indicators of which GD might be a bit more interesting and intuitive to use for participants, but overall, there is no clear indication. For every metric the H1 had to be rejected indicating that the MM approach might not be better or worse than the AR approach.

4.3.2 ASQ

The boxplot in Graph 2 illustrates the results of the ASQ questionnaire. Similar to the NASA-TLX analysis, each category was analyzed separately. For the category of *Ease of Completing Task,* the IQR for GD2 is significantly bigger than for GD1. The median is for both 5 though the mean is slightly lower for GD2 (4,6) than for GD1 (5,1). For the second category – *Amount of Time* – median and mean for GD1 are respectively 6 and 5,8. Both values are higher than for GD2, where both values are 5. For the category Support Information, both GDs have similar values. The medians are for both 6 and the means would be 6,2 (GD1) and 6,5 (GD2).



Graph 2: Boxplot of the ASQ results

Table 7 shows the Pearson correlation and p-values for the ASQ questionnaire. Two of the three correlations show a slight tendency with Amount of Time having -0,40 and Support Information +0,50. Regarding the p-values, none reached the significant threshold of 0,05 and therefore the H1 hypothesis – Dictionary 2 is better than Dictionary 1 – was rejected.

Category	Mean	Variance	Pearson Correlation	P-Value (one-tail)	P-Value (two-tail)
Ease of Completing Task GD1	4,78397402	0,99380799	0,295803989	0,380640494	0,761280988
Ease of Completing Task GD2	4,251275279	1,763834207			
Amount of Time GD1	5,682080769	1,030402055	-0,409196604	0,077528821	0,244128682
Amount of Time GD2	4,894757363	1,054092553			
Support Information GD1	6,065561556	0,737027731	0,538027587	0,11479297	0,103786492
Support Information GD2	6,516051787	0,684934889			

Table 7: Descriptive Analysis Parameters for the ASQ Categories

Similarly to the NASA-TLX results, here the analysis of the questionnaire also doesn't lead to a clear result. Which fosters the need for the qualitative analysis of the interviews.

4.3.3 Interviews

This section is used to present the insights gained from the post-study interviews. They will be presented in two parts. At first, every question from GD1 will be discussed with the answers received. Followed by the replies to GD2. If replies overlapped, they were merged, to reduce the repetition of replies. This excludes questions three, four, and five - because it was important to know if a gesture was forgotten / liked / disliked multiple times, making it a potential candidate for exchange. Some of the following quotes were translated from German into English.

Gesture Dictionary One:

For the first question - *What was the most positive aspect of your experience?* – multiple participants stated that the interaction was fun and/or exciting: "*It was fun to use and very novel!*". It was also pointed out multiple times that the gestures felt novel and easy to comprehend: "Well-functioning gestures that feel efficient and intuitive". It was stated once that the form of interaction was interesting as an improvement for the prototype.

For the second question - *What was the most negative aspect of your experience?* – the most dominant feedback was that the system felt buggy: "I felt that it was a big buggy. I understand that it was a prototype, but gestural controls felt tough to use because it felt there was a lack of feedback when it was outside of the detection area”. As also stated in the quote the second most mentioned aspect was the lack of feedback of the system. It was desired that the system would add information to let the user know the current state of the gesture. The third most mentioned point was a feature request, stating that it was sometimes hard – especially while annotating – to stop the gesture.

The third question - *Could you please recount all the gestures you just learned and what functionality they have?* – revealed that all participants could recall all of the gestures and their respective functionalities. This leads to a coherence coefficient of 1,0.

For the fourth question - *Were there gestures you liked and if yes, what did you like about them?* – Table 8 was created to provide some overview. The most liked gesture – being mentioned five (5) times - of GD1 was “Play / Pause Video” often being referenced as YouTube control. “Audio Control” is in second place with three (3) mentions. “Jumping Forward / Backward” as well as “Annotation”, and the “Delete ALL Annotation” gesture have two (2) mentions. “Open Menu” and “Video Recording” are mentioned once (1) each.

Gesture	Amount of Time Referenced
Play / Pause Video	5
Audio Recording ON/OFF	3
Jumping Forward / Backward in Video	2
Annotations	2
Delete ALL Annotation	2
Open / Close Menu	1
Start / Stop Video Recording	1

Table 8: Most liked Gestures of GD1

Similarly to the fourth question, the fifth - *Were there gestures you did not like and if yes, what did you not like about them?* – will be presented with a table. Overall, “Delete ALL Annotation” was mentioned six (6) times as the least liked gesture. This is followed by “Open Menu” being mentioned two (2) times. The gestures “Annotation”, “Audio Control”, and “Selecting / Pointing” are mentioned once (1) respectively. For the last mentioned gesture – “Select” – it needs to be mentioned that the participant was heavily tattooed and the system seems to have problems recognizing tattooed hands. This will be further discussed in chapter 6 Conclusion & Future Work.

Gesture	Amount of Time Referenced
Delete ALL Annotation	6
Open / Close Menu	2
Annotations	1
Select	1
Audio Recording ON / OFF	1

Table 9: Most disliked Gestures of GD1

For the final question - *Any additional remarks you would like to make?* – it was mentioned again by some participants that the system needs to add additional feedback: "The system needs to show in what state the gesture recognition currently is". Additionally, it was pointed out that some gestures – especially "Delete ALL Annotations" – needed too much space for execution – from one side of the projection to the other, making it potentially uncomfortable. Another point mentioned was that two hand gestures are impossible to execute for people with disabilities. This should be considered in future designs. Two final remarks that were made are: "*Nice interaction, maybe more functions would be helpful such as change colours, zoom in & out*" and "*I can foresee this being used for education purposes if it was smoother to interact with*".

Gesture Dictionary Two:

For the first question - *What was the most positive aspect of your experience?* – multiple participants stated that they felt the gesture control was an intuitive form of interaction: "*This system is easy to understand and memorize. Signs are easily adaptable. It was fun*". The role of movement – probably the general role of moving one's body in this context - was also positively highlighted: "*The gestures are active. I have or rather can move my body to achieve something*". One participant also stated that this system might be interesting for the recording of DIY videos: "*Such a system would be nice for e.g.: YouTube videos / tutorials. DIY projects in general*".

For the second question - *What was the most negative aspect of your experience?* – the most dominant answer was that some gestures were not registered and needed to be executed multiple times: "My hand didn't get recognized properly with some gestures". The second most stated trouble was the obscuration of either the projection or hands by the upper body and/or hands: "My real hand hides the projected content" & "My head must have covered my hands which made it impossible for the system to interact, I assume".

Unlike in the first GD, the third question - *Could you please recount all the gestures you just learned and what functionality they have?* – for GD2 lead to participants forgetting five (5) gestures for seven (7) times in total. Table 10 shows that "Jump Forward & Backward" was forgotten by two (2) participants, which also happened with the "Select" gesture. "Annotation", "Menu" and "Recording" were not recollected once (1) each. This leads to a coherence coefficient of 0,73 as seen in the following calculation:

$$c_{GD2} = \frac{\frac{9}{10} + \frac{10}{10} + \frac{10}{10} + \frac{6}{10} + \frac{9}{10} + \frac{10}{10} + \frac{9}{10} + \frac{10}{10}}{10} = 0{,}73$$

Gesture	Amount of Time Forgotten
Jumping Forward / Backward in Video	2
Select	2
Annotations	1
Open / Close Menu	1
Start / Stop Recording	1

Table 10: Gestures not recollected by Participants and amount thereof for GD2

For the fourth question - *Were there gestures you liked and if yes, what did you like about them?* – Table 11 was created. Overall, the gestures for "Play" and "Main Menu" were liked the most with four (4) votes each. This is followed by "Pause" with three (3) votes. The gestures "Erase" and "Annotation" were mentioned twice (2), followed by the final gesture "Recording" mentioned once (1).

Gesture	Amount of Time Referenced
Play Video	4
Open / Close Menu	4
Pause Video	3
Delete ALL Annotations	2
Annotations	2
Start / Stop Recording	1

Table 11: Most liked gestures of GD2

Similarly, to the fourth question, the fifth - *Were there gestures you did not like and if yes, what did you not like about them?* – lead to the creation of a table – Table 12. The two most disliked gestures would be "Writing" and "Jump Forward / Backward" with three (3) mentions each.

“Audio Recording” was mentioned as the least-liked gesture twice (2). “Menu”, “Video Recording”, and “Play” were mentioned once (1) each.

Gesture	Amount of Time Referenced
Annotations	3
Jumping Forward / Backward in Video	3
Audio Recording ON / OFF	2
Open / Close Menu	1
Start / Stop Recording	1
Play	1

Table 12: Most disliked gestures of GD2

For the final question - *Any additional remarks you would like to make?* – some recommendations for improving the system were stated. Two comments addressed an improvement of the annotation approach. The first comment proposes to slow down the video while in annotation mode to allow a more precise annotation without needing to use the “Stop/Pause” gesture: “*You could slow down the video while annotating. Allowing me to not always having to pause and continue the video*”. The second comment proposes to stop the video entirely and zoom the area that is annotated, for more granular annotations on the video: “*I’d prefer it the video just stops when I start the annotation gesture and the area where I want to annotate zooms in. I can then draw more precisely. When I am done, stop the annotation gesture, the video should just continue*”. One participant stated that gesture control reminded her of sign language and that this could be also considered. The final remark that seems interesting was that the child of one of the participants tends to use the index finger for writing, which might lead to a more natural interaction for the annotation gesture: “*My child, when he wants to write, uses his index finger. Wouldn’t that be more natural than holding a pen?*”.

In conclusion, the analysis of the interview showed that overall participants were interested in the interaction form of GCs. The bugginess of the system was referenced as the primary cause of frustration. In regards to remembering gestures, GD1 scored 1,0 for the coherence coefficient; Whereas, GD2 scored 0,73.

4.4 Final Gesture Dictionary

After the above-stated analysis of the results – quantitative as well as qualitative – it was decided that the insights gained by the qualitative aspects are primarily used for the design of the final GD. This was done by looking at the gestures mentioned in questions three, four, and five and calculating an overall score. The score was calculated by gaining a point (+1) for being mentioned as the liked gesture and deducting a point (-1) if the gesture was not recalled by a participant during question 3 or explicitly stated as disliked in question five. Table 13 shows the calculation process for GD1 and Table 14 the process for GD2. It needs to be mentioned that this approach has not been discussed in any of the literature analyzed and is a technique created for this work.

Gesture	Liked Gesture Value	Disliked Gesture Value	Overall Score
Play / Pause Video	+ 5	0	+ 5
Audio Recording ON / OFF	+ 3	- 1	+ 2
Jumping Forward / Backward in Video	+ 2	0	+ 2
Annotations	+ 2	- 1	+ 1
ON / OFF Recording	+ 1	- 1	0
Open / Close Menu	+ 1	- 2	- 1
Select	0	- 1	- 1
Delete ALL Annotations	+ 2	- 6	- 4

Table 13: Score calculation of the GD1 gestures

Gesture	Liked Gesture Value	Forgotten Gesture Value	Disliked Gesture Value	Overall Score
Play Video	+ 4	0	- 1	+ 3
Pause Video	+ 3	0	0	+ 3
Delete ALL Annotations	+ 2	0	0	+ 2
Open / Close Menu	+ 4	- 1	- 1	+ 2
Start / Stop Recording	+ 1	- 1	- 1	- 1
Annotations	+ 2	- 1	- 3	- 2
Audio Recording ON / OFF	0	0	- 2	- 2
Select	0	- 2	0	- 2
Jump Forward / Backward in Video	0	- 2	- 3	- 5

Table 14: Score calculation of the GD2 gestures

The overall scores of the different GD are then compared to define which gestures were preferred over the other resulting in Table 15:

Gesture	Result GD1	Result GD2	Final Decision
Audio Recording ON / OFF	+ 2	- 2	GD1
Annotations	+ 1	- 2	GD1
Delete ALL Annotations	- 4	+ 2	GD1 / GD2
Jumping Forward / Backward in Video	+ 2	- 5	GD1
Open / Close Menu	- 1	+ 2	GD2
Play /Pause Video	+ 5	+ 3 / +3	GD1
Select	- 1	- 2	GD1 / GD2
Start / Stop Recording	0	- 1	GD1 /GD2

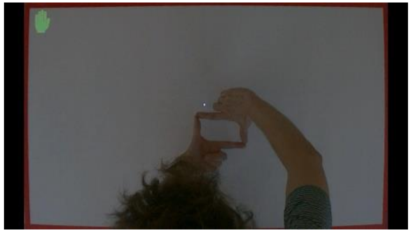
Table 15: Overall score comparison of the gestures

There is one conflict in the final gestures presented in Table 15: “Audio Recording ON / OFF” from GD1 shares the same gesture as “Open / Close Menu” from GD2. It was decided to give preference to the “Open / Close Menu” function in terms of the gesture, because the overall score of “Result GD1 + Result GD2” was +1, compared to 0 for “Audio Recording ON / OFF”. This results in the function of “Audio Recording ON / OFF” getting the gesture from GD2.

Another point of discussion would be the Play / Pause Video function: Though for GD2 each gesture achieved +3 as being liked, it was decided to give GD1 (though only +5) higher priority because it is believed that the overall mention of one gesture being liked 5 times, rather than two gestures being liked 3 times individually, is more difficult to achieve.

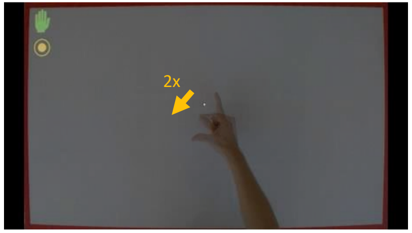
With these decisions, the final GD can be found in Table 16.

Start/Stop Recording



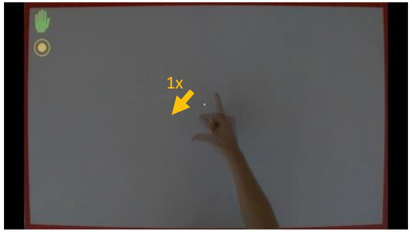
Camera - Initiate video recording by thumb-to-index finger touch.

Audio Recording ON



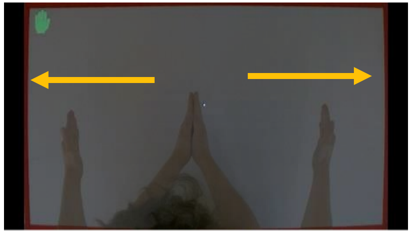
Talking - Begin audio recording by double-touching thumb and index finger of the same hand.

Audio Recording OFF



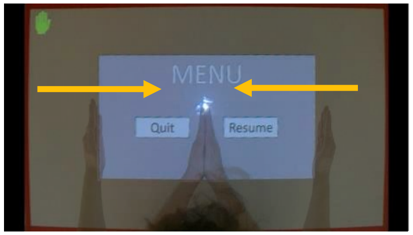
Silence - Stop audio recording with a 2-second touch of same-hand thumb and index finger.

Open Menu



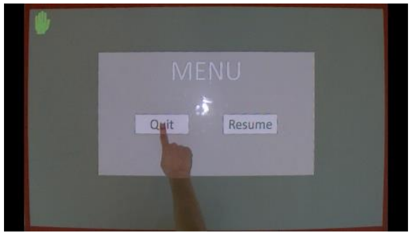
Open curtain - Trigger menu by spreading hands, palms touching, apart.

Close Menu



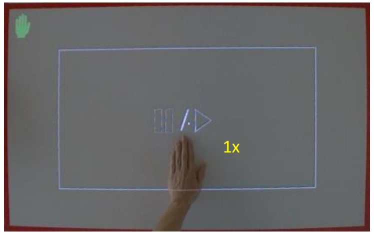
Close curtain - Shut menu by bringing palms together.

Select



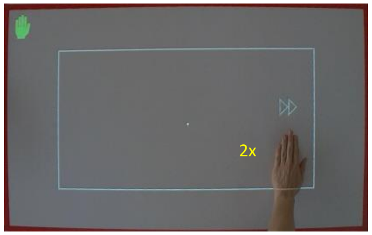
Selection - Tap target with index finger.

Play / Pause Video



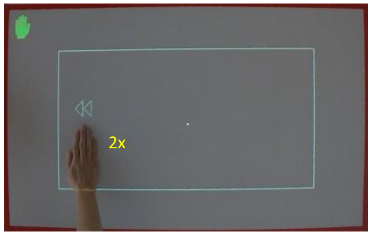
Touch - To play or pause the video, press once in the middle of the video.

Forward in Video



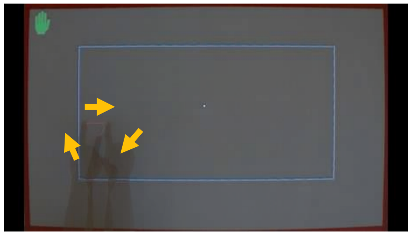
Touch - Jump forward 10 seconds in the video by double-clicking on the right side of the video.

Backward in Video



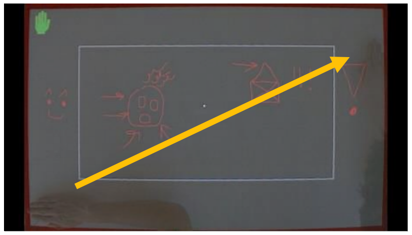
Touch - Jump back 10 seconds in the video by double-clicking on the left side.

Annotations



Finger Paint - To annotate the video, draw the shape you want with your index finger.

Delete All Annotations



Clear table - Sweep palm from bottom left to right side.

Table 16: Final Gesture Dictionary

4.5 Design Recommendation

Based on the results, observations, and user feedback, four design recommendations were deduced. The first two recommendations focus on the visual feedback of the system, the third on the possible interactions, whereas the fourth is a recommendation for using user studies for data gathering. While these implications should hold relevance for the general user population, though due to the specific sociodemographic region of the participants, there might be a bias.

4.5.1 Visual Feedback

Users expressed a heightened interest in the visibility of the status of gesture recognition. Notably, gestures that did not immediately reflect users' actions, like the “Delete All Annotation” or “Open / Close Menu” gesture from the second GD, resulted in less control and precision compared to gestures that closely mirrored their movements such as “Select” – excluding one participant who was heavily tattooed. Participants stated that it might be helpful to know if the system changed into a certain mode, such as a Delete or Annotation mode. Additionally, users suggested augmenting their hands with a digitalized version such as a digital skeleton – as shown in Figure 6 – to aid them in understanding how the system currently perceives their hands. This design recommendation has already been adopted by optical see-through technologies such as the Microsoft Hololens 2[23] and Occulus Meta Quest 3[24] and could serve as an impetus for improvement.

Both of these design recommendations seem to correlate with the 10 usability heuristics by Nielsen [59]. The first is a reference to the heuristic “Visibility of the system status” whereas the second can be seen as a reference to “Match between the system and the real world”.

4.5.2 Multiple Input Modalities

Given the amount of UI and click interactions proposed by participants of the first study, it is proposed that two modes of interaction should be provided for the users: UI and gesture. UI interaction – which would include the click gesture - for the users who are new to the system or use it infrequently and the more sophisticated gestures – presented in the final GD - for expert

[23] https://learn.microsoft.com/en-us/windows/mixed-reality/mrtk-unity/mrtk2/features/input/hand-tracking?view=mrtkunity-2022-05

[24] https://developer.oculus.com/meta-quest-3/

users. This aligns with the second golden rule by Schneiderman that frequent users should be enabled to use shortcuts.

4.5.3 Data Collection

Datasets are an essential part of DNN model training. Unfortunately, especially when the use case is an interaction niche, data may not always be available. Both studies that were conducted in this thesis are possible approaches to gathering data. Especially during the user study in the process of gesture learning, participants repeated gestures multiple times; Furthermore, participants can be encouraged to comment on the gestures while learning them. Therefore, it is recommended to integrate and video record learning tasks for user studies focused on gesture evaluation. This would allow for gathering data that can then further be used to model training and / or sharing it with the community. Latter would help the community, foster insights, and allow for the development of new models that could be used for new studies, as stated by Nebeling et al. [32] and Gheran et al. [56].

4.6 Final Model & Video Database

To enable replication and extension of the results of this work, the study material including the source code, video material, and model of the GDs is publicly available at: github.com/DPHofer-HAII/gesture-control-for-tabletop-interaction.

In summary, the analysis of the second study included quantitative and qualitative data. The quantitative data originates from the two questionnaires, which resulted in having no statistical significance and therefore being not used in the design of the final GD. The qualitative data on the other hand resulted in insights indicating that some gestures were preferred by participants more than others and vice versa. The final GD is a cumulation of gestures used in the first and second GD. In addition to the final GD, four design recommendations – clustered into three categories – are proposed. The first two propagate the use of visual feedback for users to understand the state of the system. The third recommendation states that two interaction forms – UI and gesture – should be provided. The final recommendation states that training segments should be integrated into user studies for data gathering.

5 Discussion

This section engages in a reflection on the research questions, their approach, and the ensuing results. It is outlined by the sub-research questions, beginning with the collection of gestures and concluding with recommendations and limitations that effected the work presented in this master’s thesis.

5.1 What gestures are intuitively associated with video recording and player functionalities in a tabletop projection setting?

In retrospect, employing additional metrics, such as participants' creativity levels as proposed by Gheran et al. [18], [56], could have enriched the analysis of gesture proposals. This measure, gauging participants' creativity, holds promise in excluding suggestions deemed insufficiently innovative. This could have been particularly beneficial for participants inclined towards web-centric interactions, who, despite being advised to diversify their proposals, persisted in suggesting the integration of UI elements. With the exclusion of participants, it would have caused a need to conduct the elicitation study longer, trying to achieve the threshold of 20 participants. On the other hand, not excluding participants lead to the design recommendation of multiple forms of interaction. While the proposition of participant exclusion is based on an assumption, validating it would necessitate a replication of the study with the incorporation of a creativity score. It might even be possible that more UI-based interaction would be proposed, which leads to the issue of replicability.

Recent work by Gheran et al. [56] entailed a reexamination of their prior study [18], revealing that only half of the gestures proposed in their second study align with those from the first. This observation aligns with the thoughts expressed by Nebeling et al. [32] and Gheran et al. [56] themselves, emphasizing the community's challenge in replicating results.

Furthermore, Gheran et al. additionally underscore the need for further investigation into the optimal number of participants in elicitation studies. Meaning that there might be an increase in the amount of needed participants to gain more sophisticated insights. This might become rather tricky because gaining 20 participants was already a rather exhausting experience. The need for more participants might become a problem, especially when a study may not be able to provide an attractive incentive for potential participants.

In terms of analysis, the taxonomy proposed by Gheran et al. [18] was chosen, because it is a combination of three alternative taxonomies and uses more general dimensions. It is believed that with the use of a different taxonomies such as Wobbrock et al. [5], Ruiz et al. [22], and Piumsomboon et al. [21], a different result of the gesture clustering and henceforth the AR would be achieved. But this would need further investigation.

The creation of the two GDs was an interesting approach. Creating the first GD, based on the AR, was rather straightforward. The creation of the second GD on the other hand took more time. Analyzing the descriptions of the signs proposed by the participants and understanding their mental model, which includes the chance of potentially misinterpreting some meanings. Furthermore, as the results of the user study have shown, not all of the gestures of the first GD were preferred over the ones by the second GD. This could potentially be an indication that both approaches combined – and also the inclusion of system designers – might be a more beneficial approach.

Regarding gestures, the zooming function might need some additional discussion. Contrary to assumptions that the divergence / moving apart of two fingers for zooming might become the dominant proposal, it is the gesture with the least agreement; Meaning, that the most variation was proposed by the participants in the elicitation study. This outcome correlates with the use of taxonomy for gesture clustering and may change if the taxonomy proposed by Wobbrock et al. [5] is used. Nonetheless, it was interesting to see two basic patterns for zooming crystalized: movement from the centre outwards and encircling of the area of interest. The former entails a combination of gestures where fingers move from the centre of the area that should be zoomed outward. Latter with different combinations of gestures that encircled the area of interest, though the form was not always a circle but could also be a square / rectangle.

Nevertheless, it is highly recommended to undertake replications, given that environmental variables, such as spatial arrangements and location, have been shown to influence participant engagement in gesture elicitation studies. Additionally, the specific setting and audience context can exert a notable impact on the social acceptability of gestures, particularly in public spaces.

5.2 Given the elicitation study outcome, what AI approach can serve as a proof of concept for implementing gesture recognition models?

The implementation of the prototype proved to be a substantial undertaking, taking approximately two weeks of implementation. Unfortunately, due to these constraints, some initially proposed gestures did not make it into the second study, a development that is met with regret. Incidentally, some of the gestures / functionalities left out were mentioned by participants. This could be interpreted as a sign that the functions should be addressed in a follow-up study.

Several technical limitations persist. As noted in the literature review, issues concerning the obscuration of projected images remain prevalent. Additionally, a tradeoff between the visibility of projection and hand detection necessitates careful consideration; Therefore, the balance between optimal hand detection and image visibility remains a challenge.

Moreover, the technology for Hand Gesture Recognition (HGR) is still evolving and it is believed to not yet be considered mature. Numerous variables can influence such a system, making specific use cases accessible primarily to technically proficient individuals.

5.3 How can participant preferences for gestures be evaluated using two AI models – one informed by the Agreement Rate and the other by Mental Models?

Overall, the conduction of the second study was rather experimental, meaning that few scientific publications could be used for guidance. As stated before, follow-up studies of elicitation studies are rather rare and depending on the goal might vary greatly. Nonetheless, it is believed that the output of this work is a contribution to the scientific community.

The use of the coherence score, as well as the “calculation” of the final GD, is also experimental. The former is deducted by a study format executed by Peshkova et al. [35] but thought to be aligned with one of the questions from the post-study interview. The “calculation” of the final GD was created to demonstrate an analytical way to determine the gestures and to avoid making the impression of deciding on gestures via gut feeling. Whether the final outcome would be superior to any of the other GDs would need to be researched via a usability study with well-implemented systems; Hence, subject to future research.

Regarding the creation of the final GD, it was a good decision to include quantitative and qualitative data. The questionnaires were used because it was believed they might give some insights into the preferences of users, which, unfortunately, seems to not truly be the case. There was no statistical significance in any of the results leading to the use of qualitative insights. These on the other hand allowed some interesting insights. Showing that some gestures were clearly preferred over others and that a dictionary with more different gestures – GD 2 – potentially led to forgetting some gestures and with that, functionalities of the system. The forgetting of gestures is especially problematic in a setting where there are little to no clues about the respective functionalities. However one could argue that there should be multiple ways for the execution of functionalities, and not just gestures. This could be part of a follow-up study.

Given the specific setup of the prototype, a notable observation was that many participants exhibited a preference for gestures aligned with touch interfaces. This finding sheds light on the influence of interface design on user gesture preferences and highlights the importance of tailoring interactions to user expectations, habits, but also level of expertise.

5.4 What insights and recommendations can be derived from developing and evaluating AI-driven gesture recognition models in tabletop projection settings for future projects?

A critical insight emerging from this study is the enhancement of system feedback. Participants expressed the desire to not only be informed about the successful detection of hands – which the system currently provides – but also the current state / mode of the system, as well as detailed information regarding the hand positions perceived by the system via e.g.: a digital skeleton. This insight leads into the realms of eXplainable and human-centred AI and could be a future research endeavour.

It is also worth mentioning that all participants in this study were right-handed, introducing a bias in the dataset that warrants attention in future research. While efforts were made to implement a non-discriminatory model, this aspect may still pose challenges, because AI models can only become as good as their training data.

Tattoos were found to significantly impact the detection rate of the model. It became evident that the basic model provided by Google may have limited data with tattooed hands. This limitation merits further consideration.

A potential format for comparing the two – or potentially even the three - Gesture Dictionaries could have been a Wizard of Oz setup, as discussed by Nebeling et al. [32]. However, due to resource constraints, this approach was not feasible for this study. Though arguably the Wizard of Oz approach might fit a more comparative – or even usability - study rather than one that also aims to collect as much data as possible.

The discussion section reflects the research approach and findings. It is structured according to the sub-research question and commences by examining the insights gained for gesture proposal collection and subsequently delves into broader considerations regarding the research's applicability, generalizability, and limitations. The Elicitation Study could benefit from additional metrics, particularly in evaluating participants' creativity levels, as proposed by Gheran et al. [14]. The potential influence of users' familiarity with specific technologies, such as web interfaces or the Windows operating system, might be a crucial aspect to consider. Replication challenges within the community, as noted by Gheran et al. [56], underscore the need for further investigation into optimal participant numbers for elicitation studies. Technical limitations persist, notably issues related to image projection obscuration and the tradeoff between hand detection and image visibility. Additionally, the evolving nature of Hand Gesture Recognition technology calls for ongoing refinement. The study's right-handed participant bias and the impact of tattoos on the model's detection rate highlight areas for future consideration. While a Wizard of Oz setup could offer a valuable format for comparing Gesture Dictionaries, resource constraints prevented its implementation in this study. These insights collectively underscore the complex landscape of gestural interfaces and it is hoped that they will help for future research.

6 Conclusion & Future Work

This work has discussed an approach to establish gestures and integrate the corresponding controls for a projector-based tabletop interaction system. Starting with an examination of the prototype, which serves as the foundational element of this interaction concept, this work contributes to the existing academic discourse.

Early endeavours employing video camera-based input modalities in tandem with tabletop projections unveiled persistent challenges, notably the occlusion of the projection by the user's body and the pronounced influence of ambient light on hand detection rates. Intriguingly, the distinctive use case of this project prompted to create novel gestures. Notably, the absence of adequate open-source data for training a DNN led to the conduct of two studies.

The first study – the elicitation study - focused on collecting signs / gestures with participants, which resulted in crafting two distinct gesture dictionaries. One dictionary is based on the AR, whereas the second is on MM. The ensuing second study served a dual purpose: discerning the favoured gestures from these dictionaries and accruing essential data to train the final DNN model.

For the models that were used in the user study, a combination of UI elements and transfer learning was applied. It needs to be pointed out that this approach was enough to conduct the study, but is in no regard a substitute for a thoroughly trained model.

The final output of this work is the following: (1) a gesture dictionary for the given context, (2) a method to rate gestures according to user preferences, (3) an open-source video and AI model database and (4) design recommendations and best practices for such a project.

In regards to future work, multiple points could be addressed: (1) usability study of the final GD, (2) implementation of the left-out gestures / functionalities, (3) adding additional functions, (4) allowing for text input, (5) using the collected gestures in different settings, (6) applying new analysis concepts for elicitation studies, (7) conduct a study on the premise of learnability rather than guessability, (8) developing a system that allows multiple approaches to execute a function, and (9) visual feedback for system status.

As a next step, the most important task should be a usability study. This study should provide important insights into how the final DNN model addresses the needs of a diverse user base, ensuring its effectiveness, accessibility, and user-friendliness.

Moreover, attention must be directed towards gestures identified in the initial study but not advanced to the subsequent user study. Functionalities such as "controlling audio volume," "altering the colour of annotation," and "deleting selected annotations," were elicited but ultimately deemed unviable within the temporal constraints. Attending to these aspects should enhance the usability of the system.

Concurrently, consideration should be given to additional functionalities that were not initially part of the elicitation study but hold significance for the establishment of a gesture-centric interaction system. These include transitions to the subsequent scene and the possibility to exit the application. Integration of these functionalities will give the users control over basic concepts, commonly used in applications. Additionally, the concept

The concept of Word Gesture Keyboards presents an intriguing way for future research. This approach allows users to input words using gestures, a technique commonly employed in touch-based devices. Incorporating this concept into mid-air gesture interactions may offer an intuitive and efficient means of interaction. Exploring the work of Benoit et al. [60] on word gesture keyboards for mid-air gestures could provide valuable insights and inspire further advancements in this domain.

The utilization of gestures in alternative setups or concepts, such as a "web on the wall," holds promise for future exploration. This avenue of investigation could open new possibilities for intuitive interactions in diverse technological environments. By adapting gestures to unconventional contexts, innovative ways to enhance user experiences and usability may be uncovered.

Tsandilas et al. [61] introduce an intriguing approach for elicitation study analysis, inspired by machine learning-based cross-validation. This method estimates the guessability error of sign-to-referent mappings with increasing sample sizes. While potentially valuable for evaluating the data collected in this study, it was regrettably discovered after the completion of this work. The implementation of such an approach in future studies could offer insights into the optimal participant sample size required for robust results.

To take a completely different approach, this work could be replicated with learnability in mind, rather than guessability. It would be interesting to see the differences and if one approach, in a rather long-term mentality, would be more beneficial for users or not.

A natural next step should be to create a system that allows the user to use different approaches for the execution of functions. This might be a combination of gestural, vocal, and UI-based interactions.

Finally, based on the feedback received from study participants - and as discussed in the section 4.5.1 Visual Feedback – additional system information should be added to the system.

In following up on this future work, the development of this work holds the potential not only to refine the existing interaction paradigm but to potentially even set a new standard for immersive and user-driven technology. Nonetheless, these mentioned tasks seem to be difficult to undertake.

7 Bibliography

[1] R. A. Bolt, '"Put-that-there": Voice and gesture at the graphics interface', in *Proceedings of the 7th annual conference on Computer graphics and interactive techniques - SIGGRAPH '80*, Seattle, Washington, United States: ACM Press, 1980, pp. 262–270. doi: 10.1145/800250.807503.

[2] R.-D. Vatavu and J. O. Wobbrock, 'Clarifying Agreement Calculations and Analysis for End-User Elicitation Studies', *ACM Trans. Comput.-Hum. Interact.*, vol. 29, no. 1, pp. 1–70, Feb. 2022, doi: 10.1145/3476101.

[3] D. Saffer, *Designing gestural interfaces*, 1st ed. Beijing ; Cambridge: O'Reilly, 2009.

[4] D. Wigdor and D. Wixon, *Brave NUI world: designing natural user interfaces for touch and gesture*. Burlington, Mass: Morgan Kaufmann, 2011.

[5] J. O. Wobbrock, M. R. Morris, and A. D. Wilson, 'User-defined gestures for surface computing', in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Boston MA USA: ACM, Apr. 2009, pp. 1083–1092. doi: 10.1145/1518701.1518866.

[6] M. W. Krueger, T. Gionfriddo, and K. Hinrichsen, 'VIDEOPLACE---an artificial reality', in *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '85*, San Francisco, California, United States: ACM Press, 1985, pp. 35–40. doi: 10.1145/317456.317463.

[7] P. Wellner, 'The DigitalDesk calculator: tangible manipulation on a desk top display', in *Proceedings of the 4th annual ACM symposium on User interface software and technology - UIST '91*, Hilton Head, South Carolina, United States: ACM Press, 1991, pp. 27–33. doi: 10.1145/120782.120785.

[8] W. Ju, R. Hurwitz, T. Judd, and B. Lee, 'CounterActive: an interactive cookbook for the kitchen counter', in *CHI '01 Extended Abstracts on Human Factors in Computing Systems*, Seattle Washington: ACM, Mar. 2001, pp. 269–270. doi: 10.1145/634067.634227.

[9] P. Dietz and D. Leigh, 'DiamondTouch: a multi-user touch technology', in *Proceedings of the 14th annual ACM symposium on User interface software and technology*, Orlando Florida: ACM, Nov. 2001, pp. 219–226. doi: 10.1145/502348.502389.

[10] L. Bonanni and C. Lee, 'The kitchen as a graphical user interface', *Digit. Creat.*, vol. 16, no. 2, pp. 110–114, Jan. 2005, doi: 10.1080/14626260500173096.

[11] C. Harrison, H. Benko, and A. D. Wilson, 'OmniTouch: wearable multitouch interaction everywhere', in *Proceedings of the 24th annual ACM symposium on User interface software and technology*, Santa Barbara California USA: ACM, Oct. 2011, pp. 441–450. doi: 10.1145/2047196.2047255.

[12] S. Murugappan, Vinayak, N. Elmqvist, and K. Ramani, 'Extended multitouch: recovering touch posture and differentiating users using a depth camera', in *Proceedings of the 25th annual ACM symposium on User interface software and technology*, Cambridge Massachusetts USA: ACM, Oct. 2012, pp. 487–496. doi: 10.1145/2380116.2380177.

[13] Y.-Q. Chen, C.-F. Chang, and P.-C. Su, ‘A tabletop lecture recording system based on gesture control’, in *2015 IEEE International Conference on Consumer Electronics - Taiwan*, Taipei, Taiwan: IEEE, Jun. 2015, pp. 372–373. doi: 10.1109/ICCE-TW.2015.7216950.

[14] S. Mitra and T. Acharya, ‘Gesture Recognition: A Survey’, *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.*, vol. 37, no. 3, pp. 311–324, May 2007, doi: 10.1109/TSMCC.2007.893280.

[15] D. Efron, ‘Gesture and environment.’, 1941.

[16] D. McNeill, *Gesture and Thought*. University of Chicago Press, 2005. doi: 10.7208/chicago/9780226514642.001.0001.

[17] J. C. Tang, ‘Findings from observational studies of collaborative work’, *Int. J. Man-Mach. Stud.*, vol. 34, no. 2, pp. 143–160, Feb. 1991, doi: 10.1016/0020-7373(91)90039-A.

[18] B.-F. Gheran, J. Vanderdonckt, and R.-D. Vatavu, ‘Gestures for Smart Rings: Empirical Results, Insights, and Design Implications’, in *Proceedings of the 2018 Designing Interactive Systems Conference*, Hong Kong China: ACM, Jun. 2018, pp. 623–635. doi: 10.1145/3196709.3196741.

[19] G. A. Lee, J. Wong, H. S. Park, J. S. Choi, C. J. Park, and M. Billinghurst, ‘User Defined Gestures for Augmented Virtual Mirrors: A Guessability Study’, in *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, in CHI EA ’15. New York, NY, USA: Association for Computing Machinery, Apr. 2015, pp. 959–964. doi: 10.1145/2702613.2732747.

[20] S. Villarreal-Narvaez, J. Vanderdonckt, R.-D. Vatavu, and J. O. Wobbrock, ‘A Systematic Review of Gesture Elicitation Studies: What Can We Learn from 216 Studies?’, in *Proceedings of the 2020 ACM Designing Interactive Systems Conference*, Eindhoven Netherlands: ACM, Jul. 2020, pp. 855–872. doi: 10.1145/3357236.3395511.

[21] T. Piumsomboon, A. Clark, M. Billinghurst, and A. Cockburn, ‘User-Defined Gestures for Augmented Reality’, in *Human-Computer Interaction – INTERACT 2013*, vol. 8118, P. Kotzé, G. Marsden, G. Lindgaard, J. Wesson, and M. Winckler, Eds., in Lecture Notes in Computer Science, vol. 8118. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 282–299. doi: 10.1007/978-3-642-40480-1_18.

[22] J. Ruiz, Y. Li, and E. Lank, ‘User-defined motion gestures for mobile interaction’, in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Vancouver BC Canada: ACM, May 2011, pp. 197–206. doi: 10.1145/1978942.1978971.

[23] R.-D. Vatavu, ‘User-defined gestures for free-hand TV control’, in *Proceedings of the 10th European Conference on Interactive TV and Video*, Berlin Germany: ACM, Jul. 2012, pp. 45–48. doi: 10.1145/2325616.2325626.

[24] R.-D. Vatavu and I.-A. Zaiti, ‘Leap gestures for TV: insights from an elicitation study’, in *Proceedings of the ACM International Conference on Interactive Experiences for TV and Online Video*, Newcastle Upon Tyne United Kingdom: ACM, Jun. 2014, pp. 131–138. doi: 10.1145/2602299.2602316.

[25] M. D. Good, J. A. Whiteside, D. R. Wixon, and S. J. Jones, ‘Building a user-derived interface’, *Commun. ACM*, vol. 27, no. 10, pp. 1032–1043, Oct. 1984, doi: 10.1145/358274.358284.

[26] G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais, ‘The vocabulary problem in human-system communication’, *Commun. ACM*, vol. 30, no. 11, pp. 964–971, Nov. 1987, doi: 10.1145/32206.32212.

[27] G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais, ‘Statistical Semantics: Analysis of the Potential Performance of Keyword Information Systems’, in *Human Factors in Computer Systems*, USA: Ablex Publishing Corp., 1984, pp. 187–242.

[28] S. Wiedenbeck, ‘The use of icons and labels in an end user application program: An empirical study of learning and retention’, *Behav. Inf. Technol.*, vol. 18, no. 2, pp. 68–82, Jan. 1999, doi: 10.1080/014492999119129.

[29] J. Epps, S. Lichman, and M. Wu, ‘A study of hand shape use in tabletop gesture interaction’, in *CHI ’06 Extended Abstracts on Human Factors in Computing Systems*, Montréal Québec Canada: ACM, Apr. 2006, pp. 748–753. doi: 10.1145/1125451.1125601.

[30] C. Mignot, C. Valot, and N. Carbonell, ‘An experimental study of future “natural” multimodal human-computer interaction’, in *INTERACT’93 and CHI’93 Conference Companion on Human Factors in Computing Systems*, 1993, pp. 67–68.

[31] S. Robbe, ‘An empirical study of speech and gesture interaction: Toward the definition of ergonomic design guidelines’, in *CHI 98 Conference Summary on Human Factors in Computing Systems*, 1998, pp. 349–350.

[32] M. Nebeling, A. Huber, D. Ott, and M. C. Norrie, ‘Web on the Wall Reloaded: Implementation, Replication and Refinement of User-Defined Interaction Sets’, in *Proceedings of the Ninth ACM International Conference on Interactive Tabletops and Surfaces*, Dresden Germany: ACM, Nov. 2014, pp. 15–24. doi: 10.1145/2669485.2669497.

[33] R.-D. Vatavu and J. O. Wobbrock, ‘Formalizing Agreement Analysis for Elicitation Studies: New Measures, Significance Test, and Toolkit’, in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, Seoul Republic of Korea: ACM, Apr. 2015, pp. 1325–1334. doi: 10.1145/2702123.2702223.

[34] E. Peshkova and M. Hitz, ‘Coherence evaluation of input vocabularies to enhance usability and user experience’, in *Proceedings of the ACM SIGCHI Symposium on Engineering Interactive Computing Systems*, Lisbon Portugal: ACM, Jun. 2017, pp. 15–20. doi: 10.1145/3102113.3102118.

[35] E. Peshkova, M. Hitz, D. Ahlstrom, R. W. Alexandrowicz, and A. Kopper, ‘Exploring intuitiveness of metaphor-based gestures for UAV navigation’, in *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, Lisbon: IEEE, Aug. 2017, pp. 175–182. doi: 10.1109/ROMAN.2017.8172298.

[36] M. Hitz, E. Königstorfer, and E. Peshkova, ‘Exploring Cognitive Load of Single and Mixed Mental Models Gesture Sets for UAV Navigation’, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:198332850

[37] S. G. Hart and L. E. Staveland, ‘Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research’, in *Advances in Psychology*, vol. 52, Elsevier, 1988, pp. 139–183. doi: 10.1016/S0166-4115(08)62386-9.

[38] J. R. Lewis, ‘Psychometric evaluation of an after-scenario questionnaire for computer usability studies: the ASQ’, *ACM Sigchi Bull.*, vol. 23, no. 1, pp. 78–81, 1991.

[39] R. Francese, I. Passero, and G. Tortora, ‘Wiimote and Kinect: gestural user interfaces add a natural third dimension to HCI’, in *Proceedings of the International Working Conference on Advanced Visual Interfaces*, Capri Island Italy: ACM, May 2012, pp. 116–123. doi: 10.1145/2254556.2254580.

[40] M. L. Bolton, E. Biltekoff, and L. Humphrey, ‘The Mathematical Meaninglessness of the NASA Task Load Index: A Level of Measurement Analysis’, *IEEE Trans. Hum.-Mach. Syst.*, vol. 53, no. 3, pp. 590–599, Jun. 2023, doi: 10.1109/THMS.2023.3263482.

[41] M. Oudah, A. Al-Naji, and J. Chahl, ‘Hand Gesture Recognition Based on Computer Vision: A Review of Techniques’, *J. Imaging*, vol. 6, no. 8, p. 73, Jul. 2020, doi: 10.3390/jimaging6080073.

[42] L. Guo, Z. Lu, and L. Yao, ‘Human-Machine Interaction Sensing Technology Based on Hand Gesture Recognition: A Review’, *IEEE Trans. Hum.-Mach. Syst.*, vol. 51, no. 4, pp. 300–309, Aug. 2021, doi: 10.1109/THMS.2021.3086003.

[43] L. Fiorini *et al.*, ‘Daily Gesture Recognition During Human-Robot Interaction Combining Vision and Wearable Systems’, *IEEE Sens. J.*, vol. 21, no. 20, pp. 23568–23577, Oct. 2021, doi: 10.1109/JSEN.2021.3108011.

[44] F. Cavallo *et al.*, ‘Preliminary evaluation of SensHand V1 in assessing motor skills performance in Parkinson disease’, in *2013 IEEE 13th International Conference on Rehabilitation Robotics (ICORR)*, Seattle, WA: IEEE, Jun. 2013, pp. 1–6. doi: 10.1109/ICORR.2013.6650466.

[45] K. B. Shaik, P. Ganesan, V. Kalist, B. S. Sathish, and J. M. M. Jenitha, ‘Comparative Study of Skin Color Detection and Segmentation in HSV and YCbCr Color Space’, *Procedia Comput. Sci.*, vol. 57, pp. 41–48, 2015, doi: 10.1016/j.procs.2015.07.362.

[46] Ganesan P and V. Rajini, ‘YIQ color space based satellite image segmentation using modified FCM clustering and histogram equalization’, in *2014 International Conference on Advances in Electrical Engineering (ICAEE)*, Vellore, India: IEEE, Jan. 2014, pp. 1–5. doi: 10.1109/ICAEE.2014.6838440.

[47] O. Saman and L. Stanciu, ‘Image Processing Algorithm for Appearance-Based Gesture Recognition’, in *2019 23rd International Conference on System Theory, Control and Computing (ICSTCC)*, Sinaia, Romania: IEEE, Oct. 2019, pp. 681–684. doi: 10.1109/ICSTCC.2019.8885888.

[48] Y. Fang, K. Wang, J. Cheng, and H. Lu, ‘A Real-Time Hand Gesture Recognition Method’, in *Multimedia and Expo, 2007 IEEE International Conference on*, Beijing, China: IEEE, Jul. 2007, pp. 995–998. doi: 10.1109/ICME.2007.4284820.

[49] J. Molina, J. A. Pajuelo, and J. M. Martínez, ‘Real-time Motion-based Hand Gestures Recognition from Time-of-Flight Video’, *J. Signal Process. Syst.*, vol. 86, no. 1, pp. 17–25, Jan. 2017, doi: 10.1007/s11265-015-1090-5.

[50] D. Konstantinidis, K. Dimitropoulos, and P. Daras, ‘SIGN LANGUAGE RECOGNITION BASED ON HAND AND BODY SKELETAL DATA’, in *2018 - 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON)*, 2018, pp. 1–4. doi: 10.1109/3DTV.2018.8478467.

[51] W. Qi, S. E. Ovur, Z. Li, A. Marzullo, and R. Song, ‘Multi-Sensor Guided Hand Gesture Recognition for a Teleoperated Robot Using a Recurrent Neural Network’, *IEEE Robot. Autom. Lett.*, vol. 6, no. 3, pp. 6039–6045, Jul. 2021, doi: 10.1109/LRA.2021.3089999.

[52] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[53] C. Lugaresi *et al.*, ‘MediaPipe: A Framework for Perceiving and Processing Reality’, in *Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR) 2019*, 2019. [Online]. Available: https://mixedreality.cs.cornell.edu/s/NewTitle_May1_MediaPipe_CVPR_CV4ARVR_Workshop_2019.pdf

[54] N. Mohamed, M. B. Mustafa, and N. Jomhari, ‘A Review of the Hand Gesture Recognition System: Current Progress and Future Directions’, *IEEE Access*, vol. 9, pp. 157422–157436, 2021, doi: 10.1109/ACCESS.2021.3129650.

[55] P. Garg, N. Aggarwal, and S. Sofat, ‘Vision Based Hand Gesture Recognition’, *World Acad. Sci. Eng. Technol. Int. J. Comput. Electr. Autom. Control Inf. Eng.*, vol. 3, pp. 186–191, 2009.

[56] B.-F. Gheran, R.-D. Vatavu, and J. Vanderdonckt, ‘New Insights into User-Defined Smart Ring Gestures with Implications for Gesture Elicitation Studies’, in *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg Germany: ACM, Apr. 2023, pp. 1–8. doi: 10.1145/3544549.3585590.

[57] R.-D. Vatavu and J. O. Wobbrock, ‘Clarifying Agreement Calculations and Analysis for End-User Elicitation Studies’, *ACM Trans. Comput.-Hum. Interact.*, vol. 29, no. 1, pp. 1–70, Feb. 2022, doi: 10.1145/3476101.

[58] N. Beringer, ‘Evoking Gestures in SmartKom - Design of the Graphical User Interface’, in *Gesture and Sign Language in Human-Computer Interaction*, vol. 2298, I. Wachsmuth and T. Sowa, Eds., in Lecture Notes in Computer Science, vol. 2298. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2002, pp. 228–240. doi: 10.1007/3-540-47873-6_25.

[59] J. Nielsen, ‘Enhancing the explanatory power of usability heuristics’, in *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, 1994, pp. 152–158.

[60] G. Benoit, G. M. Poor, and A. Jude, ‘Bimanual Word Gesture Keyboards for Mid-air Gestures’, in *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, Denver Colorado USA: ACM, May 2017, pp. 1500–1507. doi: 10.1145/3027063.3053137.

[61] T. Tsandilas and P. Dragicevic, ‘Gesture Elicitation as a Computational Optimization Problem’, in *CHI Conference on Human Factors in Computing Systems*, New Orleans LA USA: ACM, Apr. 2022, pp. 1–13. doi: 10.1145/3491102.3501942.

8 Appendix

8.1 Questionnaire

Study: Hand Gesture Interaction

Demographic Information

- What gender do you identify as?

  - Female
  - Male
  - Diverse
  - Other: __________
  - Prefer not to say

- What is your age?

  - __________
  - Prefer not to answer

- What is your level of experience with Gesture Interaction Technology?

  - None
  - Novice
  - Occasional
  - Regular use (1 – 3 time per week)
  - Frequent (more than 3 times per week)
  - Daily
  - Prefer not to answer

- If yes, which technologies are you using?

  - Hololens
  - Oculus
  - Leap Motion
  - Kinect
  - Laptop / Tablet
  - Smart Phone / Watch
  - Sonstiges: __________

Questionnaire

1. Mental Demand — How mentally demanding was the task?

□ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □

Very Low — Very High

2. Physical Demand — How physically demanding was the task?

□ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □

Very Low — Very High

3. Temporal Demand — How hurried or rushed was the pace of the task?

□ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □

Very Low — Very High

4. Performance — How successful were you in accomplishing what you were asked to do?

□ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □

Perfect — Failure

5. Effort — How hard did you have to work to accomplish your level of performance?

□ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □

Very Low — Very High

6. Frustration — How insecure, discouraged, irritated, stressed, and annoyed were you?

□ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □

Very Low — Very High

7. Overall, I am satisfied with the ease of completing this task

□	□	□	□	□	□	□
1	2	3	4	5	6	7

8. Overall, I am satisfied with the amount of time it took to complete this task

□	□	□	□	□	□	□
1	2	3	4	5	6	7

9. Overall, I am satisfied with the support information (documentation) when completing this task

□	□	□	□	□	□	□
1	2	3	4	5	6	7