

User-Defined Voice Commands, Display Interactions and Mid-Air Gestures for Smart Home Tasks

Fabian Hoffmann

University of Regensburg
Regensburg, Bavaria, Germany
fabian.hoffmann@stud.uni-regensburg.de

Miriam Ida Tyroller

University of Regensburg
Regensburg, Bavaria, Germany
miriam-ida.tyroller@stud.uni-regensburg.de

Felix Wende

University of Regensburg
Regensburg, Bavaria, Germany
felix.wende@stud.uni-regensburg.de

ABSTRACT

We present a study aimed to gain insight on users' perceptions and desires in the context of smart home interaction modalities. To achieve this, we conducted an elicitation study in which participants were asked to perform commands within a simulated smart home environment, facing three conditions: voice command, display interaction and mid-air gestures. Facing tasks of different areas in smart home that require user assistance, the participants suggested fitting commands and rated them on the grounds of goodness, ease, enjoyment and social acceptance, as well as their general preference of each modality. The collected measures allow us to present insights that can be used as possible future guidelines for smart home interaction modalities and future research in voice command, display interactions and mid-air gestures.

KEYWORDS

smart home, voice control, display control, mid-air gestures

ACM Reference Format:

Fabian Hoffmann, Miriam Ida Tyroller, and Felix Wende. 2018. User-Defined Voice Commands, Display Interactions and Mid-Air Gestures for Smart Home Tasks. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 7 pages. <https://doi.org/10.1145/1122445.1122456>

1 INTRODUCTION

Smart homes and its interaction modalities are widely spread ... Existing work focuses on creating smart home interaction modalities considered desirable and enjoyable to use for its users, as developed by Hagensby Jensen et al. [7]. ... This is caused by different obstacles, such as the interaction modalities being too expensive to implement or too complex to understand. Interacting with a smart home requires modalities that are easy to perform by its user, but also modalities that are easily coming to mind, with user enjoyment and social acceptance as variables also important to consider. Our goal is ...

2 APPROACH AND METHODOLOGY

We followed a similar approach as Dingler et al. [5] by showing and explaining different smart home tasks to the participants and subsequently asking them to propose a voice command, a display interaction and a mid-air gesture, to fulfil the specific tasks in their preferred way. All eleven tasks are listed in section 'Tasks'. A within-subject design was chosen, so every participant gave suggestions for every modality and task. We used a latin-square on the order of the interaction modalities to reduce sequence effects [4] and fatigue. The tasks were shown to the participants in random order. We took video recordings of all sessions. We also collected feedback from participants through questionnaires, on preferences of interaction modalities for a specific task and on goodness, ease, enjoyment and social acceptance of their suggestions. The study was conducted in German.

Interaction Modalities

We compared three different types of interaction modalities. The already commonly used voice and display control, as well as the existing but lacking development technique of mid-air gestures. Therefore, we were able to collect insights on the existing modalities and additionally gain a new set of mid-air gestures.

Tasks

The smart home market can be divided into six different categories [3]. Those are *home entertainment*, *smart household appliances*, *energy management*, *networking and control*, *comfort and light* and *building security*. We excluded the category *networking and control* for developing the tasks, because it does not include devices that can be controlled, but is rather the infrastructure of a smart home and would be responsible for the detection of performed commands. For all other categories we selected two common tasks [1] each, except for *building security* three because of its bigger market share. All categories with their assigned tasks are listed in table 1.

Participants

A total of 13 participants (7 female) took part in the study with an average age of 33.5 (SD = 15.1). We recruited the

Table 1: Categories with their assigned tasks

Category	Task
Home Entertainment	1. Increase the volume of the music. 2. Turn on the next TV channel.
Smart household appliances	3. Start multi-colored wash at 60 degree. 4. Turn off the oven.
Energy Management	5. Increase the room temperature. 6. Open the shutters.
Comfort and light	7. Turn on the light. 8. Dim the light.
Building security	9. Close the window. 10. Lock the front door. 11. Turn on the security camera.

participants through social networks and personal contacts. The participants were mostly students from different departments of the University of Regensburg and OTH Regensburg. All of them at least heard of smart homes before and are familiar with interaction through displays. According to the pre-questionnaire, ten participants are familiar with both voice control and display interaction to control other devices, but only one performed mid-air gestures for interaction yet. Seven participants own smart home devices like Google Home, Amazon Alexa, smart TVs or lamps and use them frequently. None of the participants owns a fully integrated smart home system.

Apparatus

The study was carried out in a quiet room. The different tasks were illustrated through pictures, which showed the state before and after issuing the command. Mid-air gestures, voice commands and comments of the participants were recorded by a mounted camera. Display interaction was documented through a sketch on paper. The whole setup is shown in Figure 1. None of the interaction modalities were actually implemented.

Procedure

Before starting the session, the participants were asked to fill out a consent form and a demographic questionnaire. Then

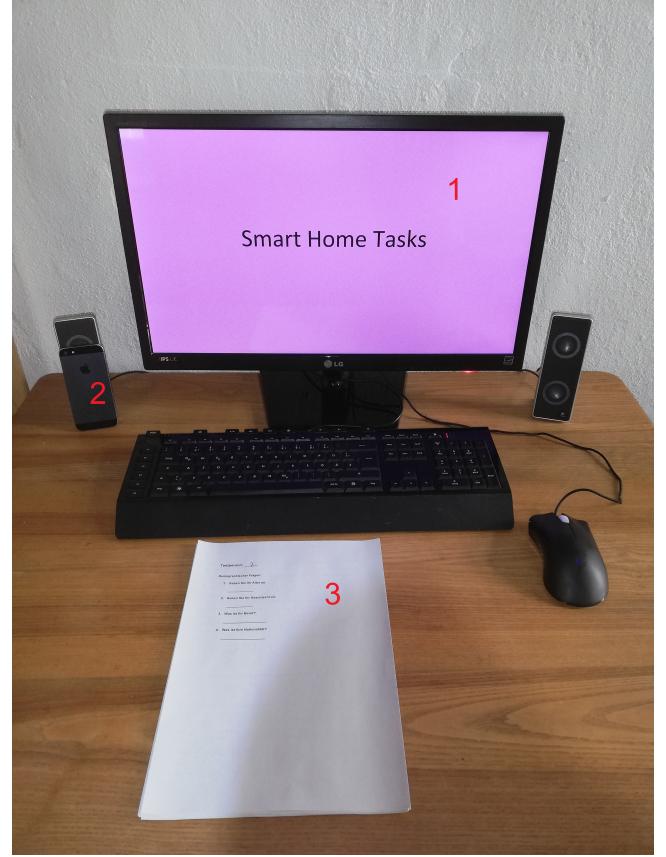


Figure 1: Setup with monitor to illustrate smart home tasks (1), camera for video recording (2) and the questionnaires (3)

they had to fill out a questionnaire in terms of their previous knowledge and usage of smart home devices and the three interaction modalities. After that the tasks were presented to the participants in a random order. At first all tasks had to be fulfilled with a single interaction modality, then with the second and after that with the remaining modality. Additionally to the illustration through pictures, the tasks were explained verbally. The participants were allowed to talk, move and interact with a display in any way they wanted and were encouraged to explain their choices in a thinking-aloud approach. After each task the participants rated their specific suggestion on goodness, ease, enjoyment and social acceptance on four 7-point Likert scales. When all tasks were finished with each interaction modality the participants rated the three interaction modalities for each on 7-point Likert scales, on how good each modality is to perform the specific task. They were asked to do this independently of their own suggestions. At the end a semi-structured interview was conducted to explore the motivation of the participants for each choice and allow them to rate the different interaction

modalities under the aspects of efficiency, simplicity, naturality, desirability and enjoyment. This is based on a similar approach in the elicitation study on foot gestures by Felberbaum et al. [6]. The study took about an hour, for which the participants were compensated with sweets.

3 RESULTS

With nine participants and eleven tasks, we collected for each of the three interaction modalities 143 suggestions and in total $13 * 11 * 3 = 429$. Our results include the video recording, taxonomies for each interaction modality, user-defined sets of voice commands, display interactions and gestures, subjective ratings of the sets, qualitative observations and an assessment on the modalities for each task.

Classification of Voice Commands

The participants suggested 43 unique voice commands. The authors manually classified each voice command along five dimensions: *nature*, *form*, *flow*, *context* and *complexity*. Within each dimension are multiple categories, shown in Table 2. We adopted the dimensions from Wobbrock *et al.* [10] and Ruiz *et al.* [8] and adapted them to voice commands.

Taxonomy of Voice Commands. The *nature* dimension comprises *action* voice commands which state the action to perform. An example of this type of voice command is saying "increase temperature". *State* voice commands describe the desired condition of a device. For example, a *state* voice command is "cameras on" to start camera surveillance.

The *form* dimension describes how much words are used in the voice command and if they have the structure of a full sentence. A *single word* command can be "next" to get to the next TV channel. *Two words* voice commands mostly consist out of the mentioning of the device to be controlled and an action or state. *More words* commands are similar to *two words* but use additional filler words. Finally, voice commands that are correct sentences were classified with the category *sentence*.

The *flow* dimension categorizes the voice commands, if response of a device occurs after or while the user acts. A voice command is *discrete*, when a device performs the command after the participant stopped talking. A *continuous* voice command would be starting an action with a command and stop the ongoing action with another command.

The *context* dimension describes, if the voice command requires a specific context or can be performed independently. For example saying "turn off" to turn off the oven is *in-context*, whereas "oven off" is considered *no-context*.

The *complexity* dimension describes if the voice command consists out of a single or a composition of more voice commands. A *compound* voice command can be decomposed into *simple* voice commands.

Table 2: Taxonomy of voice commands for smart home tasks

Taxonomy of voice commands		
Nature	Action	Voice command states the action to perform
	State	Voice command describes the desired condition
Form	Single word	Voice command consists out of a single word
	Two words	Voice command consists out of two words
	More words	Voice command consists out of more words without sentence structure
	Sentence	Voice command uses sentence structure
Flow	Discrete Continuous	Response occurs <i>after</i> the user acts Response occurs <i>while</i> the user acts
Context	In-context	Voice command requires specific context
	No-context	Voice command does not require specific context
Complexity	Simple	Voice command consists of a single voice command
	Compound	Voice command can be decomposed into simple voice commands

User-defined Voice Command Set

We collected a total of 143 voice commands, which we used to create a user-defined voice command set for our specified tasks. For each task, we grouped identical voice commands together and the group with largest size was chosen to be the representative voice command for this corresponding task. To evaluate the degree of consensus among the participants, we computed the *agreement score* A_t (Equation 1), as proposed by Vatavu and Wobbrock [9], for each task.

$$A_t = \frac{|P_t|}{|P_t| - 1} \sum_{P_i \subseteq P_t} \left(\frac{|P_i|}{|P_t|} \right)^2 - \frac{1}{|P_t| - 1} \quad (1)$$

In equation 1, t is a task in the set of all tasks T , P_t is the set of suggested voice commands for t and P_i is a subset of identical voice commands from P_t . As an example of calculation of an agreement score, consider the task *increase the volume of the music*. The task has four groups of identical voice commands with a size of 7, 4, 1 and 1. Therefore the agreement score for *increase the volume of the music* is:

$$A = \frac{13}{12} \left(\left(\frac{7}{13} \right)^2 + \left(\frac{4}{13} \right)^2 + \left(\frac{1}{13} \right)^2 + \left(\frac{1}{13} \right)^2 \right) - \frac{1}{12} = 0.346 \quad (2)$$

Figure 2 illustrates the agreement scores for each task using voice commands. Participants had the least agreement on commands for the tasks *start multi-colored wash at 60 degree*

Table 3: User defined voice command set with the two most common options for each task

Task	German	English	Frequency
1	lauter Musik lauter	louder music louder	53.8 % 30.8 %
2	Nächster Kanal weiter	next channel next	46.2 % 23.1 %
3	Buntwäsche 60 Grad	colored laundry 60 de- grees	30.8 %
	Starte Buntwäsche 60 Grad	start colored laundry 60 degrees	15.4 %
4	Ofen aus ausschalten	oven off turn off	84.6 % 7.7 %
5	Wärmer Raumtemperatur erhöhen	warmer increase room temper- ature	30.8 % 30.8 %
6	Rollladen öffnen Rollladen auf	open roller shutter roller shutter up	69.2 % 23.1 %
7	Licht an	light on	100.0 %
8	Licht dimmen Licht dunkler	dim light light darker	30.8 % 15.4 %
9	Fenster schließen Fenster zu	close window window closed	53.8 % 46.2 %
10	Haustür absperren Tür zu	lock front door door closed	61.5 % 23.1 %
11	Kamera an Kamera einschalten	camera on switch on camera	61.5 % 23.1 %

(Task 3) and *increase the room temperature* (Task 5). This is attributable to the complexity of the tasks. Since the study was conducted in German, we also provide the original version of the voice commands to prevent losses during translation. Table 3 shows the most common and second most common voice commands and their frequency according to the different tasks.

Classification of Display Interaction

The participants suggested 61 unique display interactions. Similar to the voice commands we manually classified each display interaction along three dimensions for graphical user interface (GUI) elements (*form, elements, flow*) and for touch gestures along four dimensions (*form, nature, binding, flow*). Within each dimension are multiple categories, shown in Table 4 for GUI elements and for touch gestures in Table 5. The dimensions and categories for GUI elements were inspired by the work from Wobbrock *et al.* [10] and Ruiz *et al.* [8]. As taxonomy for touch gestures we used Wobbrock *et al.* [10] taxonomy of surface gestures, which is displayed in Table 5.

Table 4: Taxonomy of display interactions for smart home tasks (GUI elements)

Taxonomy of display interactions (GUI elements)		
Form	Direct Action	Single interaction that directly leads to the action
	Selection & Confirmation	Selection of the action and starting through another element
Elements	Single clickables	The GUI includes one or more single clickables (button, checkbox, etc.)
	Slider	The GUI includes one or more sliders
	More words	Voice command consists out of more words without sentence structure
	Rotation	The GUI includes one or more rotational elements
	Text & number entry	The GUI includes one or more options to enter text or numbers
Flow	Symbolic	The GUI includes one or more special symbolic elements
	Discrete	Response occurs <i>after</i> the user acts
	Continuous	Response occurs <i>while</i> the user acts

Table 5: Taxonomy of display interactions for smart home tasks (touch gestures)

Taxonomy of display interactions (touch gestures)		
Nature	Symbolic	Gesture visually depicts a symbol
	Physical	Gesture acts physically on objects
	Metaphorical	Gesture indicates a metaphor
	Abstract	Gesture-referent mapping is arbitrary
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 and path with one Finger
Binding	Object-centric	Location defined with respect to object features
	World-dependent	Location defined with respect to world features
	World-independent	Location can ignore world features
	Mixed dependencies	World-independent plus another
Flow	Discrete	Response occurs <i>after</i> the user acts
	Continuous	Response occurs <i>while</i> the user acts

Taxonomy of Display Interaction (GUI Elements): ToDo: Beschreibung Taxonomy wie bei voice

Taxonomy of Display Interaction (Touch Gestures): ToDo: Beschreibung Taxonomy wie bei voice

User-Defined Display Interaction Set

ToDo: Beschreibung + Tabelle set wie voice

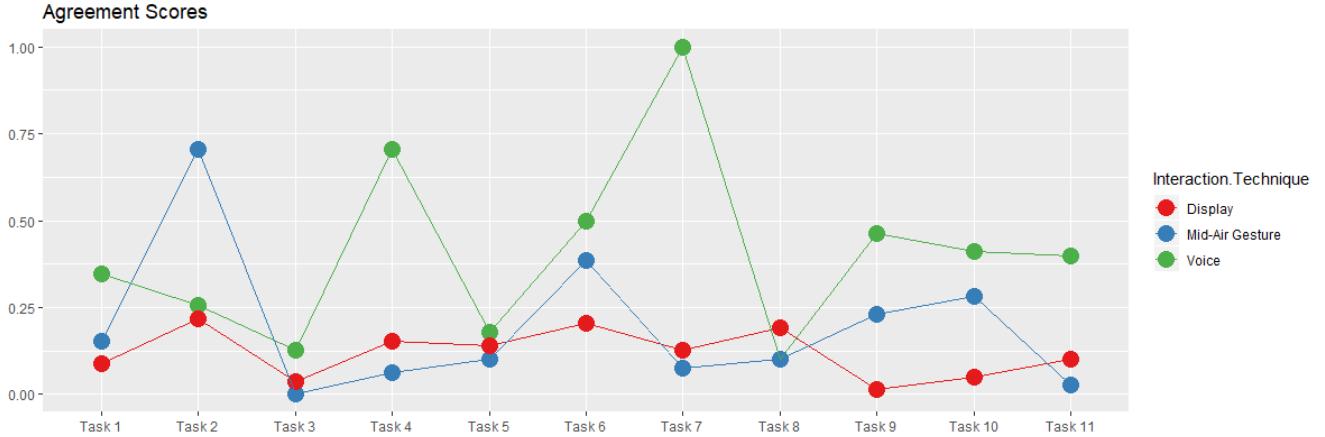


Figure 2: Agreement scores for each task with all interaction modalities

Classification of Mid-Air Gestures

The participants suggested 55 unique mid-air gestures. Once again we manually classified each mid-air gesture, this time along eight dimensions. The dimensions are *nature*, *flow*, *context*, *interaction*, *dimension*, *position*, *movement* and *complexity*. Within each dimension are multiple categories, shown in Table 6. We adopted the dimensions *nature* (small adjustment at category *physical*), *flow*, *context* and *complexity* and their corresponding categories from Wobbrock *et al.* [10] and Ruiz *et al.* [8]. *Dimension* from Ruiz *et al.* [8] was adapted to our needs. *Interaction* was inspired by Dingler *et al.* [5] and we further extended the taxonomy by the two dimensions *position* and *movement*.

Taxonomy of Mid-Air Gestures: The *nature* dimension includes *symbolic* mid-air gestures, which visually depict symbols. An example for that is drawing a "X" into the air to turn off the oven. *Physical* mid-air gestures imitate real physical actions like locking a door with a key, by rotating the wrist with a fist like hand position. A *metaphorical* gesture may be wrapping the arms around yourself with rubbing, to indicate you are freezing and want to raise the temperature in the room. Mid-air gestures that did not fit into any of these three categories are classified as *abstract*.

The *flow* dimension is the same as in the other taxonomies and categorizes the mid-air gestures, if response of a device occurs after or while the user acts. A mid-air gesture is *discrete*, when the response occurs after the movement, and *continuous*, when the response occurs during the movement.

The *context* dimension describes, if the mid-air gesture requires a specific context or can be performed independently. For example making a horizontal hand movement to change the TV channel is *in-context*, whereas pointing at the

TV and then performing the hand movement is considered *no-context*.

The *interaction* dimension simply describes if the participant used only one hand or both hands for their suggested mid-air gesture.

The *dimension* of a mid-air gesture is used to describe how many axis are involved in the movement of the hand. Some gestures, like just rotating the wrist happen along only a single axis. The translation of a hand or a rotational motion from the wrist are considered *tri-axis*. The combination of those two movements is classified as *six-axis*. The movement of fingers was ignored in this dimension and is described more in the dimension *movement*.

The *position* dimension describes the finger position at the beginning of the mid-air gesture. The difference between *flat hand* and *open hand* is, that the fingers are together at *flat hand* and spread at *open hand*.

The *movement* dimension simply states if the participant includes relevant finger movement (*movement*) or no finger movement (*no movement*) into his mid-air gesture.

The *complexity* dimension describes, just like voice commands, if the mid-air gesture consists out of a single or a composition of more mid-air gestures. A *compound* mid-air gesture can be decomposed into *simple* gestures.

User-Defined Mid-Air Gesture Set

We collected 142 mid-air gestures, because one participant could not think of a mid-air gesture for starting the multi-colored wash at 60 degree. Then we followed the procedure as with voice commands and display interactions. We grouped the identical mid-air gestures together, chose the group with the largest size as representative gesture for the corresponding task and computed the *agreement score* (Equation 1) for each task. The latter are also shown in Figure 2

Table 6: Taxonomy of mid-air gestures for smart home tasks

Taxonomy of mid-air gestures for smart home tasks		
Nature	Symbolic Physical Metaphorical Abstract	Gesture visually depicts a symbol Gesture imitates a physical action Gesture indicates a metaphor Gesture-referent mapping is arbitrary
Flow	Discrete Continuous	Response occurs <i>after</i> the user acts Response occurs <i>while</i> the user acts
Context	In-context No-context	Gesture requires specific context Gesture does not require specific context
Interaction	One hand Two hands	Gesture was performed with one hand Gesture was performed with two hands
Dimension	Single-Axis Tri-Axis Six-Axis	Motion occurs around a single axis from Motion involves either translational motion from hand or rotational motion from wrist, not both Motion involves translational motion from hand and rotational motion from wrist
Position	Flat hand Open hand Closed hand Single finger Two fingers More fingers	Gesture started with flat hand Gesture started with open hand Gesture started with closed hand (fist) Gesture started with a single outstretched finger Gesture started with two outstretched fingers Gesture started with three or four outstretched fingers
Movement	No movement Movement	No change in finger position Change in finger position
Complexity	Simple Compound	Gesture consists of a single gesture Gesture can be decomposed into simple gestures

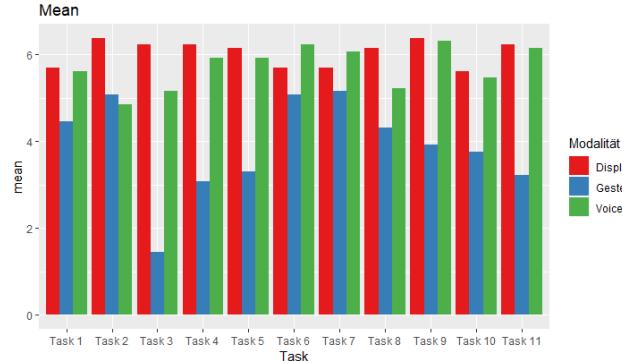
(blue). Task 3 (*start multi-colored wash at 60 degree*) has an *agreement score* of 0.0 because there were no two identical mid-air gestures. Compared to the other two modalities the participants had the highest agreement on Task 2 (*turn on the next TV channel*) with an *agreement score* of 0.705. ToDo: Beschreibung + Tabelle set wie voice

Comparing the Modalities

PLACEHOLDER Figure 3

Qualitative Results

The qualitative data obtained through the semi-structured interview was analysed by splitting the participants opinions into six categories, obtained through an open-coding approach carried out by the three researchers individually.

**Figure 3: Modalities rated by the participants, for each task on a scale from 1="not fitting at all" to 7="very fitting"**

These are evaluation of the experiment, voice control, display control and mid-air gesture control, as well as possible mixtures of interaction modalities and further suggested ones.

The participants mainly thought of the experiment as interesting and innovative. Participants stated it was "entertaining and interesting to test out new things I've never done before".

Looking into general opinions on the different interaction modalities, the participants voice mainly favourable opinions about voice control. They stated voice control is "clear and unambiguous, like commanding". Display control was regarded as easy to control and intuitive, "it's the most universal and comfortable to use", while also classing it as time-consuming, stating it's "tedious, always having to hold a display, like a smart phone, in your hands". On the other hand, participants were not as convinced by mid-air gesture control, as is mirrored in the collected quantitative data. They stated it's "complicated to use" and "only intuitive if all gestures are the same", explaining that "gestures you use seldom, you completely forget after like a month". The surveillance aspect also valued into those opinions, with participants stating that "I don't want to be under constant surveillance, I wouldn't implement such an interaction technique", which also influenced their view on voice control, even though not as much, as participants specified that "voice control is not as bad as video monitoring, though it would be better if you could turn it off".

When asked about the possibilities of mixing two or three of the suggested interaction modalities in one smart home environment, participants were generally in favour of it, stating "combining like two, like voice and display, would be the most convenient". Participants were also asked about other interaction modalities different from voice, gesture and display control that come to their mind, resulting in interesting approaches for possible future work such as "I'd

like haptic feedback, such as stomping once on the floor to activate my underfloor heating" or "I don't want to talk too much to my smart home, I want it to know my demands autonomously".

4 DISCUSSION

PLACEHOLDER

5 FUTURE WORK

The research offers a groundwork for future work in the surroundings of smart home development. Additionally, it also provides the base of producing even more insights on the examined interaction modalities through extending and improving the existing study. To achieve this, future work can build on the qualitative research that showed participants stating they mostly prefer display and voice control, with a possible combination of these two interaction modalities. Further work needs to focus on this approach, while also testing a combination of interaction modalities within a setting including more actual implemented parts of a smart home system. Although simulations offer research that is faster to develop and perform, they can not guarantee with an absolute certainty users would act the same way in a real smart home environment. To achieve a higher transferability and generalizability of the research, the number of participants and population to draw from needs to be significantly extended.

Leaving the approach to interact with user's short-time desires leads to us to future work centering around smart home-user interaction for long-time settings. Many of the participants stated their wish to interact with the smart home as little as possible. Instead, they preferred a scenario of setting their needs and demands at an early point in the implementation and rely on the smart home system working autonomous from that moment on. This could lead to a drastic change in participants' viewpoints on goodness, ease and social acceptance of interaction modalities. Performing a command only once compared to multiple times could lead to more lenient opinions on complex voice commands or mid-air gestures, or highly detailed structures of display commands.

Safety issues mentioned in the qualitative research concerning the need to monitor users' in order to implement voice and mid-air gesture control can also lead to further research. In the time of highly demanded data security and privacy, smart homes can not pose as threats to this. Future work can focus on studies developing interaction modalities that are free from such concerns or improve existing modalities to no longer rely on constant user supervision. For this, it's necessary to collaborate with experts of other fields.

6 CONCLUSION

PLACEHOLDER

REFERENCES

- [1] [n.d.]. Home - SmartHome Hilfe. <https://service.startsmarthome.de/de/>
- [2] [n.d.]. Infografik: Smart Home: Was spricht dafür, was dagegen? <https://de.statista.com/infografik/15254/argumente-fuer-und-gegen-die-nutzung-von-smart-home-produkten/>
- [3] [n.d.]. Smart Home - weltweit | Statista Marktprognose. <https://de.statista.com/outlook/279/100/smart-home/weltweit#market-revenue>
- [4] 2017. Latin Square Design: Definition and Balanced Latin Square Algorithm. <https://www.statisticshowto.datasciencecentral.com/latin-square-design/>
- [5] Tilman Dingler, Rufat Rzayev, Alireza Sahami Shirazi, and Niels Henze. 2018. Designing Consistent Gestures Across Device Types. In *Engage with CHI*, Regan Mandryk and Mark Hancock (Eds.). The Association for Computing Machinery, New York, New York, 1–12. <https://doi.org/10.1145/3173574.3173993>
- [6] Yasmin Felberbaum and Joel Lanir. 2018. Better Understanding of Foot Gestures. In *Engage with CHI*, Regan Mandryk and Mark Hancock (Eds.). The Association for Computing Machinery, New York, New York, 1–12. <https://doi.org/10.1145/3173574.3173908>
- [7] Rikke Hagensby Jensen, Yolande Strengers, Jesper Kjeldskov, Larissa Nicholls, and Mikael B. Skov. 2018. Designing the Desirable Smart Home. In *Engage with CHI*, Regan Mandryk and Mark Hancock (Eds.). The Association for Computing Machinery, New York, New York, 1–14. <https://doi.org/10.1145/3173574.3173578>
- [8] Jaime Ruiz, Yang Li, and Edward Lank. 2011. User-defined motion gestures for mobile interaction. In *Conference proceedings and extended abstracts / the 29th Annual CHI Conference on Human Factors in Computing Systems*, Desney Tan, Geraldine Fitzpatrick, Carl Gutwin, Bo Begole, and Wendy A. Kellogg (Eds.). ACM, New York, NY, 197. <https://doi.org/10.1145/1978942.1978971>
- [9] Radu-Daniel Vatavu and Jacob O. Wobbrock. 2015. Formalizing Agreement Analysis for Elicitation Studies. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*, Bo Begole, Jinwoo Kim, Kori Inkpen, and Woon-tack Woo (Eds.). ACM Press, New York, New York, USA, 1325–1334. <https://doi.org/10.1145/2702123.2702223>
- [10] Jacob O. Wobbrock, Meredith Ringel Morris, and Andrew D. Wilson. 2009. User-defined gestures for surface computing. In *CHI 2009 - digital life, new world*, Dan R. Olsen, Richard B. Arthur, Ken Hinckley, Meredith Ringel Morris, Scott Hudson, and Saul Greenberg (Eds.). ACM, New York, NY, 1083. <https://doi.org/10.1145/1518701.1518866>