Ensuring a Robust Multimodal Conversational User Interface During Maintenance Work

Christian Fleiner
christian.fleiner@fau.de
University of Erlangen-Nuremberg
Nuremberg, Germany

Till Riedel
till.riedel@kit.edu
Karlsruhe Institute of Technology
Karlsruhe, Germany

Michael Beigl
michael.beigl@kit.edu
Karlsruhe Institute of Technology
Karlsruhe, Germany

Marcel Ruoff
marcel.ruoff@kit.edu
Karlsruhe Institute of Technology
Karlsruhe, Germany

ABSTRACT

It has been shown that the provision of a conversational user interface proves beneficial in many domains. However, there are still many challenges when applied in production areas, e.g. as part of a virtual assistant to support workers in knowledge-intensive maintenance work. Regarding input modalities, touchscreens are failure-prone in wet environments and the quality of voice recognition is negatively affected by ambient noise. Augmenting a symmetric text- and voice-based user interface with gestural input poses a good solution to provide both efficiency and a robust communication. This paper contributes to this research area by providing results on the application of appropriate head and one-hand gestures during maintenance work. We conducted an elicitation study with 20 participants and present a gesture set as its outcome. To facilitate the gesture development and integration for application designers, a classification model for head gestures and one for one-hand gestures were developed. Additionally, a proof-of-concept regarding a multimodal conversational user interface with support of gestural input during maintenance work was demonstrated. It encompasses two usability testings with 18 participants in different realistic, but controlled settings: notebook repair (SUS: 82.1) and cutter head maintenance (SUS: 82.7).

CCS CONCEPTS

• **Human-centered computing** → **Gestural input**; *Usability testing*; *Mobile devices*; *Natural language interfaces*; • **Applied computing** → **Computer-assisted instruction**.

KEYWORDS

Assistance systems, conversational agent, elicitation study, industry 4.0, multimodal interface, task guidance, user-defined gestures

1 INTRODUCTION

Conversational User Interfaces (CUI) were successfully applied in many domains, e.g. customer support [16] or cultural heritage [4, 33]. While virtual assistants (aka conversational agents), which come with a CUI, significantly grew in popularity for personal use[1], industrial virtual assistants have failed to establish until today. With respect to virtual assistants which are categorized as a kind of artificial intelligence (AI), Hoffmann et al. noted that "not everything that is possible with mainstream AI is applicable in industry" [19]. This is due to the additional constraints and challenges, industrial environments are subject to. Regarding maintenance tasks, the implementation of a CUI is challenging, because maintenance work is often performed in noisy and dirty environments [40, 46]. This has an impact on the available interface options. Text-based input via mobile devices like smartphones or tablets is hindered, because capacitive-sensing touchscreens are failure-prone to water smears and dirt [7, 32]. Furthermore, voice commands and utterances as an input become unreliable when ambient noise interferes with the voice recognition [22, 29].

However, this does not mean that there is no support available. In spite of all the challenges, instructive assistance systems (IAS) are already applied in production areas to support operators in knowledge-intensive maintenance processes. While most of them use visual cues, current research begins to focus on systems with a CUI, too. For instance, a case study with a conversational agent (CA) was conducted for maintenance work using the digital twin of a machine to show the operator how to fix the issue [12]. The operator was limited to written text for communicating with the CA, while the CA replied via text or voice. Recently, Serras et al. presented results on their multimodal system with a CUI applied for maintenance work of a robotic gripper [30]. Here, visual and

[1] https://www.forbes.com/sites/ilkerkoksal/2020/03/10/the-sales-of-smart-speakers-skyrocketed/, last accessed on February 2nd, 2021

aural interaction capabilities were provided. For mitigating the already mentioned effects of ambient noise, devices with noise-canceling capabilities were used. Although both systems had a positive impact on their users, it can be argued that the presented multimodal interaction is not reliable enough for maintenance work environments.

As a consequence, we suggest that a multimodal CUI must support gestural input in the context of maintenance work. Gestures provide an alternative modality that is robust to noise and also provides support for situated interaction. In spite of several devices on the market that support gesture recognition, e.g. head-mounted displays [14, 31], not all types of gestures may be appropriate for maintenance work. Especially, ergonomic factors are neglected concerning maintenance tasks. The work is often performed at inaccessible locations that coerce workers into constraint postures [5] like kneeling which limits the variety of suitable gestures.

There is a research gap about appropriate gestures aligned to maintenance work. This paper aims to close this research gap by providing two classification models and one explicit gesture set for this context. Also in response to the provocation paper of Schaffer and Reithinger [28], this paper argues that a CUI which is applied for industrial maintenance must at least provide a symmetric multimodality [39] in form of text- and voice-based interaction, but greatly benefits from complementary gestural input.

Regarding the paper's structure, we first outline the key requirements of an IAS and related work. Afterward, we report results from a gesture elicitation study aimed at identifying appropriate gestures during maintenance work. Finally, the results of a preliminary study which is comprised of two usability testings are presented to demonstrate the proof-of-concept for a multimodal CUI with gestural support in the context of maintenance work.

2 RELATED WORK

2.1 Overview of IAS Research

The purpose of an IAS is to facilitate the decision-making process by the provision of information, instruction and guidance to the user. IAS are predominantly applied on production sites to aid workers in assembling products, which is a response to the trend of mass customization [18, 20, 43]. In contrast to augmented reality solutions [2, 25] , only few research was done for IAS related to a CUI regarding maintenance work. For instance, Haslwanter et al. [18] compared aural guidance (using text-to-speech) to the original text-based guidance. Another interesting prototype was introduced by Zheng et al. who implemented a head-mounted display with aural and gestural input for workflow guidance and instant messaging capabilities via smartwatch to connect with remote experts [45].

At first, it is vital to determine the aptitude of an IAS based on the overall characteristics of maintenance work to make them comparable. Beigl [6] identified four key requirements that must be met for the provision of digital instructions. We apply these requirements to the context of maintenance work. The first requirement is (1) *ubiquity* (everywhere accessible). This requires the instructions to be accessible on a mobile device because maintenance work must be mainly performed at the affected machine and cannot be moved to a dedicated working station. Secondly, the application must be (2) *easy to use* (short training phase, low mental effort). This relates to the need of a familiar interface and the appropriate chunking of instructions. Furthermore, the IAS must be (3) *unobtrusive to use* (does not distract operator from primary task). In consideration of physical maintenance work, it is expected that both hands are required to perform the maintenance task. Consequently, none or at most one hand should be temporally used for IAS interaction. Lastly, (4) *extra capabilities* should benefit the overall purpose of the system. In the case of Beigl's *ElectronicManual* [6], the provision of rich media like video clips and updatable resources were named as extra capabilities that paper-based solutions could not provide. We define the multimodal interaction for robust communication as first essential extra capability. Besides text- and voice-based interaction, we propose mid-air gestures as a complementary input modality. In order to satisfy the key requirement *unobtrusive to use*, temporary one-hand and head gestures qualify to be used during maintenance work. The provision of a CUI as a social component [15] to enhance the users' engagement, serves as second extra capability.

2.2 Elicitation Studies

An elicitation study is like *Wizard of Oz* a low-fidelity prototyping technique that is especially "useful for open-ended, early-stage exploration" [9]. The objective of it is to design a symbol set that reduces training time for a specific interface due to its high guessability [41]. Hereby, a *symbol* denotes the means of input needed to execute a specific command (called *referent*). For instance, saying "turn on TV" may be a *symbol* for turning on the TV (*referent*). Within an elicitation study, multiple participants propose appropriate symbols for a set of actions (referents) related to a specific system [42]. In this case, symbols are conveyed via gestures. As a result, an appropriate gesture consensus set can be derived for the application or system in the proposed use case.

We identified three prior elicitation studies that concentrated on head or one-hand gestures. Dierk et al. conducted an elicitation study "to identify appropriate inputs to hat-worn technology" where various head and mid-air gestures were proposed by participants [13]. Zaiti et al. presented a collection of mid-air gestures to control a TV [44]. Another approach was taken by Ruiz et al. who focused on eliciting mid-air gestures with a smartphone [26]. Although these studies already identified several gestures that could be paired with referents that are also applicable for maintenance work, there were two reasons that made it imperative to conduct an elicitation study that would explicitly address a CUI regarding maintenance work support. The first reason was that the provision of instructions and the required navigation control concerning a CUI would require additional referents that were not considered in prior studies. Secondly, it was shown that the given conditions of a use case had a great impact on the participant's choice of input modality [13].

3 ELICITING USER-DEFINED GESTURES

We conducted a gesture elicitation study with 20 participants (16 male; 4 female) with a mean age of 40.0 years (SD: 19.4) to elicit user-defined gestures which are powerful enough to substitute utterances during maintenance work. Each participant received € 10 for their participation. Fifteen participants were right-handed, four left-handed and one ambidextrous. The sample was comprised of

two user groups. Ten participants were tech-savvy people aged 30 years or younger with limited working experience who were recruited on the university compound. The second sample addressed people aged over 30 with at least one year of working experience and were recruited by a cooperating manufacturer of eroding and milling machines.

3.1 Selection of Referents

From the three respective elicitation studies, we excerpted and edited the list of referents to fit the expected CUI of a multimodal IAS. For instance, the referent *copy* [13] was omitted because the application of editing functions is limited for task guidance. Next, we added referents that would be also beneficial when using an instant messaging platform:

- *Select 1-4. Option.* Contemporary messaging interfaces allow quickreplies which are a set of interface buttons that "suggest messages the user can send to the bot"[34].
- *Request Details.* The command provides additional options to retrieve complementary information that supports the overall understanding of an action.
- *Record Media* initializes any kind of media recording to document maintenance processes.
- *Request Summary* provides an outline of the current maintenance task's progress. The command is helpful for the operator to get the current status after a break or shift change.
- *Request Assistance.* If the documentation is incomplete, the option to escalate to an expert is essential. The IAS of Zheng et al. offered instant messaging capabilities to do so [45].
- *Switch View.* Regarding the structure of instant messaging conversations, media files are difficult to reselect. Therefore, we propose a command that alternates between the media library and the chat-view.

For the common hierarchy structure of manuals, we adapted Charwat's task hierarchy for the instruction navigation: [11]:

- *Task* defines the global objective and the purpose of the process.
- *Activity* is the container for a group of actions. Each activity ends with an achieved intermediary goal.
- *Action* is defined as an appropriate, memorable work unit.

In total, a list of 35 relevant referents was created to be used in this elicitation study (see table 1).

3.2 Experimental Setup and Procedure

Each participant was given headphones and a wristband for the right hand as artifacts to perform gestures. All referents were introduced via an instant messaging dialogue mock-up with an IAS to repair a kitchen product. Antagonistic referents like mute/unmute were presented together. For each referent, first a head gesture and then a hand gesture were proposed. We explained that only the motion was relevant for the head gesture detection. Furthermore, arm movement, finger movement and static postures were accepted as hand gestures. All proposals were subsequently rated on a 7-point Likert scale regarding ease of execution and referent fitness.

Table 1: List of referents presented to participants. Each referent can be seen as an user intent for conversational agents.

Category	Context	Referent
Action	Agent	Select 1st Option
		Select 2nd Option
		Select 3rd Option
		Select 4th Option
		Request Details
		Record Media
		Request Summary
		Request Assistance
	Application	Accept
		Decline
		Start Voice Command
		Cancel Voice Command
		Volume Up
		Volume Down
		Mute
		Unmute
	Media Control	Play
		Pause
		Stop
		Forward
		Rewind
	Call (Voice/Video)	Answer Call
		Ignore Call
		Hang Up
Navigation	Application	Switch View
	Chat-View	Next Action
		Previous Action
		Repeat Action
		Next Activity
		Previous Activity
		Cancel Task
	Tile-View	Panel Right
		Panel Left
		Panel Up
		Panel Down

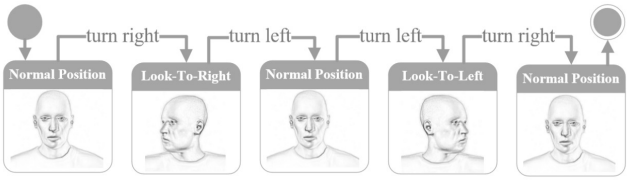


Figure 1: Exemplary state machine for the *head shake* gesture. Note that state *Normal Position* is not held when moving from right to left. Thus, there is no break of movement and both performed movements of *turn left* can be interpreted as one long movement of *turn left*. All sequences start and end in state *Normal Position*. In appendix A, a state machine is displayed to clarify the classification model.

Table 2: Dimensions with attributes for describing one-hand mid-air gestures. The attribute *None* of dimension *Movement Direction* is used to describe static gestures where any movement does not have impact on the gesture's meaning. Hand Shapes are illustrated in appendix B.

Type	Dimension	Attribute	Detail
Dynamic	Arm Participation	yes	main movement is done by (lower) arm
		no	main movement is done by hand wrist
	Type of Movement	linear	gesture with linear movement
		radial	radial movement (typical with ellbow joint)
		circle	radial gesture that returns to its initial position
	Movement Direction	none	no movement (for static gestures)
		backwards	backwards movement
		forwards	forwards movement
		right	rightward movement
		left	leftward movement
		up	upward movement
		down	downward movement
Static	Palm Orientation	palm-to-body	palm is facing body
		palm-inverse-to-body	back of the hand is facing body
		palm-right	palm is facing right side
		palm-left	palm is facing left side
		palm-up	palm is facing ceiling
		palm-down	palm is facing floor
	Hand Shape	ASL-1	ASL finger alphabet for number "1"
		ASL-2	ASL finger alphabet for number "2"
		ASL-3	ASL finger alphabet for number "3"
		ASL-4	ASL finger alphabet for number "4"
		ASL-A	ASL finger alphabet for letter "a"
		ASL-B	ASL finger alphabet for letter "b"
		ASL-C	ASL finger alphabet for letter "c"
		ASL-H	ASL finger alphabet for letter "h"
		ASL-O	ASL finger alphabet for letter "o"
		ASL-Q_close	ASL finger alphabet for letter "q" → index finger is touching thumb
		ASL-Q_open	index finger is touching thumb → ASL finger alphabet for letter "q"
		ASL-S	ASL finger alphabet for letter "s"
		ASL-Y	ASL finger alphabet for letter "y"
		ASL-Z	Hand shape for ASL finger alphabet for letter "z" without motion
		DGS-1	DGS finger alphabet for number "1"
		DGS-SCH	DGS finger alphabet for trigraph "sch"

Additionally, participants remarked if they would feel comfortable to use the gesture proposal in public as social acceptance influences the users' behavior [24, 38]. We encouraged all participants to share their thoughts about their gesture proposals (*Think-aloud*). We recorded audio and video during all sessions where each lasted approximately one hour.

3.3 Analysis and Classification Models

The collected data included pre-test questionnaires, videos, gesture proposal evaluations and post-test interview transcripts. Two participants could not perform gestures for all referents due to time constraints (only 22 referents and 18 referents respectively). In total, 1,340 gestures were collected. Hand gestures that scored below their head gesture counterparts were replaced by them. This was done to emphasize the head gestures' potential because head gestures were identified as less popular than mid-air gestures [13]. Proposals that scored less than four points (< *neutral*) in each ease or fitness were classified as "refusals". By including "refusal" as a symbol into our analysis, we were able to identify referents unfitted for gesture pairing. The final set of proposals consisted of 1,076 performed gestures and 264 "refusals".

We developed two classification models for gesture characterization using grounded theory. The classification models for head and hand gestures are based on *chunking* [8] to split gestures into their primitives. Regarding head gestures, we defined eleven movement vectors as primitives in alphabetical order: hold, push backwards, push forwards, quarter-circle clockwise, quarter-circle counter-clockwise, tilt down, tilt left, tilt right, tilt up, turn left and turn right. The vectors can be unified into a state machine (an example is given in figure 1).

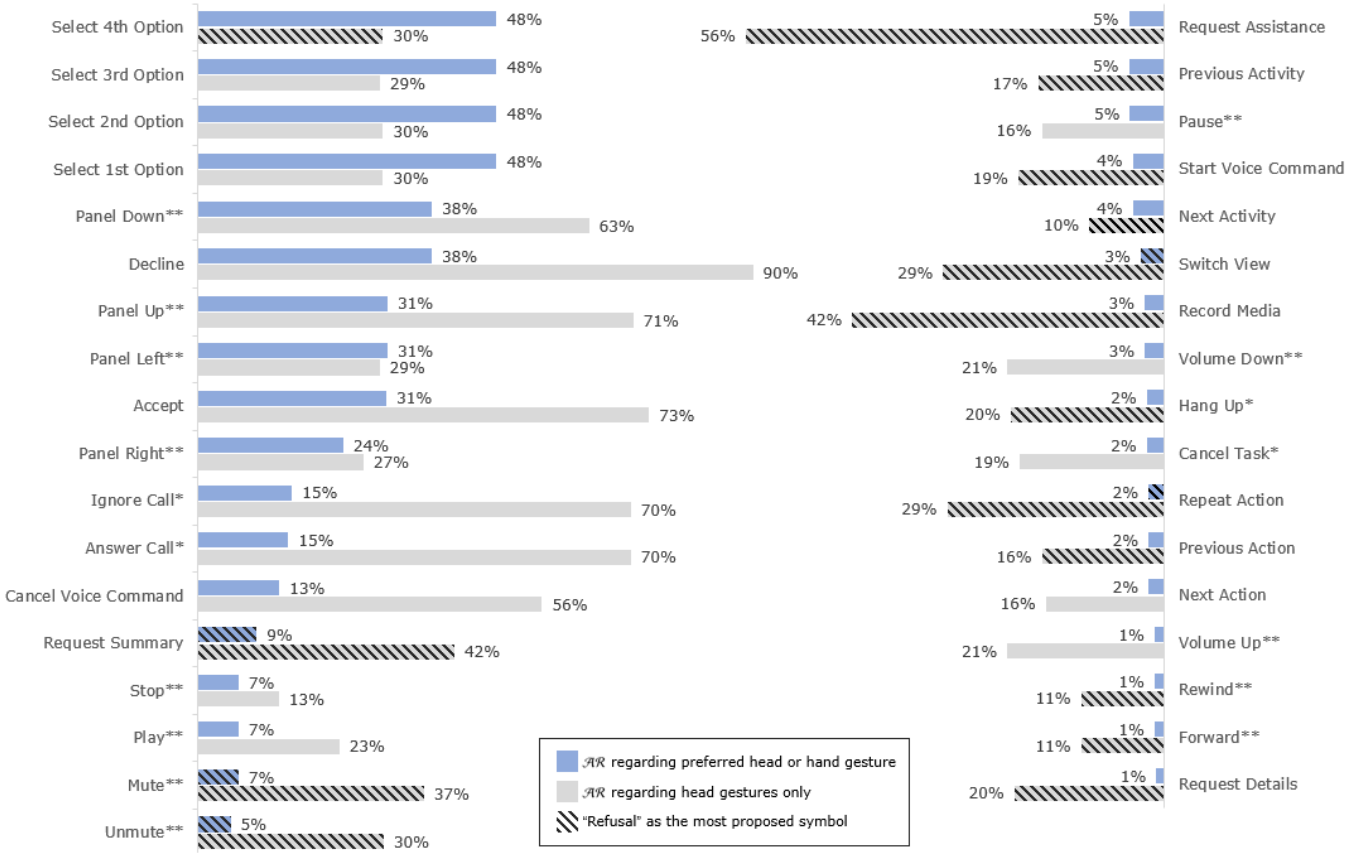


Figure 2: Overview of the agreement rates of all gestures. Referents with no asterisk had 20 proposals for each gesture type; referents with one asterisk had 19 and referents with two asterisks received only 18 proposals.

Gestures that could not be correctly captured via this state machine, were described via the action *draw* (e.g. drawing an "x" for the referent *cancel task*). The model is based on the design space class *head gesture* [13]. For one-hand gestures not only the movement, but also the hand's posture can be relevant for the gesture's meaning [35]. The resulting degrees of freedom were too large to build an appropriate state machine. We alternatively derived five dimensions based on two types [17, 35] to describe most gestures efficiently and provide the means for an easy replication: *arm participation, type of movement, movement direction* (dynamic) and *palm orientation, hand shape* (static). Hand postures of proposed gestures were matched with hand shapes of the finger alphabet of the American Sign Language (ASL) and the German Sign Language (DGS). The final classification model of one-hand gestures is displayed in table 2. The only gestures that we could not adequately describe by this model were drawing actions and snapping. We used keywords to identify these gestures.

3.4 Results

As the measurement for a consensus among the group, we applied Vatavu and Wobbrock's agreement rate $\mathcal{AR}$ [36]:

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

P is the set of all proposals for referent r with $|P|$ being r's amount of elicited proposals. P_i is a subset comprised of identical gesture proposals. If $\mathcal{AR}$ equals 0.0, every participant proposed a unique gesture. If $\mathcal{AR}$ equals 1.0, everybody proposed the same gesture. Vatavu and Wobbrock [36] proposed the following qualitative interpretations for $\mathcal{AR}$: *low agreement* ($\leq .100$), *medium agreement* ($.100 - .300$), *high agreement* ($.300 - .500$) and *very high agreement* ($> .500$). We calculated the $\mathcal{AR}$s and double-checked the results by using the AGreement Analysis Toolkit (AGATe v2.0)[2]. In total, 124 unique head gestures and 239 unique one-hand gestures were identified. The $\mathcal{AR}$s among head gestures ranged from 0.100 to 0.900 (⌀ 0.339). Among the preferred gestures, $\mathcal{AR}$s ranged from 0.011 to 0.484 (⌀ 0.150). The $\mathcal{AR}$s are illustrated in figure 2.

[2]https://depts.washington.edu/acelab/proj/dollar/agate.html, last accessed on January 6th, 2021

Table 3: Overview of the user-defined gesture set which is eligible to substitute text- or voice-based interaction : (A) consensus (B) choice-based (C) authors' recommendation (D) omitted. The referents' names can be directly interpreted as the user intents. Note that only the best gesture for each referent is displayed and the choice of the gesture type for a referent is not exclusive.

Category-Context	Referent	Gesture Type	Characteristic	Origin
Action-Agent	Select 1st Option	hand	(no, linear, none, palm-inverse-to-body, ASL-1)	A
	Select 2nd Option	hand	(no, linear, none, palm-inverse-to-body, ASL-2	A
	Select 3rd Option	hand	(no, linear, none, palm-inverse-to-body, ASL-3)	A
	Select 4th Option	hand	(no, linear, none, palm-inverse-to-body, ASL-4)	A
	Request Details	hand	(yes, circle, (up, forwards), palm-to-body, DGS-SCH)	C
	Record Media	hand	(no, linear, none, palm-inverse-to-body, ASL-O)	C
	Request Summary	head	(4x quarter-circle clockwise)	C
	Request Assistance	hand	(yes, radial, (up, backwards), palm-to-body, ASL-Y)	C
Action-Application	Accept	head	(tilt down, tilt up)	A
	Decline	head	(turn right, turn left, turn left, turn right)	A
	Start Voice Command	hand	Snap	C
	Cancel Voice Command	head	(turn right, turn left, turn left, turn right)	A
	Volume Up	head	(tilt up, hold, tilt down)	A
	Volume Down	head	(tilt down, hold, tilt up)	A
	Mute	hand	(no, radial, (left, down), palm-right, DGS-1)	C
	Unmute	hand	(no, radial, (right, up), palm-left, DGS-1)	C
Action-Media Control	Play	hand	(yes, linear, forwards, palm-down, ASL-Z)	B
	Pause	hand	(yes, linear, forwards, palm-inverse-to-body, ASL-B)	B
	Stop	hand	(yes, linear, right, palm-down, ASL-B)	C
	Forward	hand	(yes, circle, (up, right), palm-down, ASL-S)	B
	Rewind	hand	(yes, circle, (up, left), palm-down, ASL-S)	B
Action-Call	Answer Call	head	(tilt down, tilt up)	A
	Ignore Call	head	(turn right, turn left, turn left, turn right)	A
	Hang Up	head	(turn right, turn left, turn left, turn right)	C
Navigation-Application	Switch View	hand	(yes, radial, (up, right), palm-down, ASL-Z)	C
Navigation-Chat-View	Next Action	head	(tilt down, tilt up)	A
	Previous Action	hand	(yes, linear, left, palm-left, ASL-B)	C
	Repeat Action	head	(4x quarter-circle counter-clockwise)	C
	Next Activity	-	-	D
	Previous Activity	-	-	D
	Cancel Task	head	(turn right, turn left, turn left, turn right)	A
Navigation-Tile-View	Panel Right	hand	(no, linear, right, palm-left, ASL-B)	A
	Panel Left	hand	(no, linear, left, palm-left, ASL-B)	A
	Panel Up	hand	(no, linear, up, palm-down, ASL-B)	A
	Panel Down	hand	(no, linear, down, palm-down, ASL-B)	A

The results of the elicitation study indicated that only nods and head shakes were relevant head gestures. In contrast to the work of Dierk et al. [13], participants stated that they would always prefer utterances before using any kind of gesture. While there was no objection to perform any proposed hand gesture in public, participants felt uncomfortable to perform head gestures with a transition sequence of five movements or longer in public.

In spite of the higher $\mathcal{AR}$s for head gestures compared to hand gestures, participants preferred hand gestures over head gestures. When participants had the option to choose between hand or head gestures, 82% (550) of the proposals were hand gestures and 18% (120) were head gestures. This issue was already addressed by Dierk et al. [13]. There, participants named the lack of expression power of head gestures as one reason for that. The lack of expression power compared to hand gestures could be caused by the smaller degree of freedom.

Due to the lack of expression power, participants could not always propose head gestures that they perceived as positive or neutral. This effect could only be mildly observed for hand gesture proposals. If there is a high consensus regarding a referent with "refusal" as most proposed symbol, it is likely that there is no eligible symbol within the respective design space that would be accepted by the user group. In this study, we set an $\mathcal{AR}$ of .300 as threshold to classify this kind of referent. Accordingly, it should be avoided to pair the referents *mute*, *unmute*, *record media*, *request assistance*, *request summary*, and *select 4th option* with head gestures.

3.5 Developing a User-Defined Gesture Set

We classified all gestures with a medium or higher agreement rate as an appropriate substitute to text- or voice-based interaction, because it is likely that users would expect this gesture to initialize the respective referent. Vice versa, it could be irritating for users if the referent is paired with another gesture. Therefore, a gesture with an acceptable agreement rate should be reserved for this referent.

The results of the elicitation study have displayed that there is only a small consensus regarding media and volume control. However, both are essential for a smooth interaction with an aural output interface. Thus, we decided to conduct an additional choice-based elicitation study covering the best three proposals for each referent from the first elicitation study. For this study, 14 industrial freshman workers (11 male; 3 female) were recruited. The mean age was 17.1 years (SD: 0.83). Here, only gestures with a very high agreement rate were reserved for the respective referent. This applied for referents *Play* ($\mathcal{AR} = 0.527$), *Pause* ($\mathcal{AR} = 0.615$), *Forward* ($\mathcal{AR} = 0.615$), and *Rewind* ($\mathcal{AR} = 0.527$).

Referents with only a low agreement rate gestures indicate that participants do not have any expectations yet. Those referents could be paired with unbiased, novel gestures to enrich the overall usability of the system. To provide a complete gesture set for the list of referents, we paired the remaining referents with promising gestures that were proposed in the elicitation study as the authors' recommendation. The referents concerning activities and actions had similar proposals. As the activity referents are not essential for the guidance navigation, we omitted to pair them with a gesture to prevent any confusions. The complete user-defined gesture set is shown in table 3.

4 PRELIMINARY STUDY

As a second contribution, we conducted two usability testings within realistic scenarios (notebook repair, laser cutter head maintenance). It is important to distress that this study's purpose was solely to demonstrate the proof-of-concept of a multimodal CUI with gestural input support in the area of maintenance work, but not to evaluate the proposed user-defined gesture set. Therefore, we implemented a CA running on an instant messaging platform as a basic prototype. Similar to Aromaa et al. [2], the prototype was a multi-device system. The core device was a smartphone running the instant messaging client and provided the user with visual instructions. Headphones conveyed aural information and the user could use the integrated microphone for utterances. Head gestures were detected via camera using the Viola-Jones object detection framework for face detection [37] and the Lucas-Kanade method [21]. A gesture wristband [1] was used to detect hand gestures where a simplified recognition model was applied covering only the dynamic dimensions *Type of Movement* and *Movement Direction*. The prototype supported the referents *Request Details*, *Mute*, *Unmute*, *Next Action*, *Previous Action*, *Repeat Action* and *Cancel Task*. *Request Details* was split into four commands to reduce the navigation overhead for this simple prototype: *request short/long description*, *request video* and *request result description* (see figure 3).

For the evaluation of the prototype, the System Usability Scale (SUS) was applied. We concentrated on measuring the perceived usability to display the users' interest and the acceptance of such a CUI.

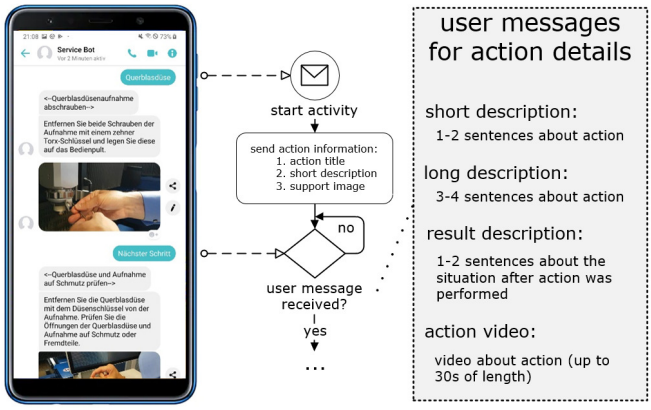


Figure 3: Instant messaging client of prototype showing exemplary instructions from the industrial use case (left) and its conversation flow (middle). Besides navigation-based commands, the participants had four options to request additional information of the current action which are related to referent *request details* (right). An exemplary conversation flow instance of an activity is shown in appendix C.

4.1 Domestic Use Case: Notebook Repair

We recruited seven male engineering students (M: 27.4; SD: 2.1) who already had basic experience in computer repair and thus, could provide qualitative feedback. Each student had to perform six repair tasks which were comprised of the disassembly and the replacement of a component, e.g. the cooling fan. The participation was rewarded with € 10. The tasks were performed in sequence to uphold the realistic scenario.

The prototype achieved a mean of 82.1 (mark *A* [27]; SEM: 4.6) on the SUS. None of the participants performed any head gestures after the training phase. In the post-test interview, one participant added that "head gestures were not necessary for this scenario because the hands could be freed easily by putting tools or components back on the table". Five of seven students preferred aural input instead of gestures because utterances were easier to perform. One of them emphasized that he "could work on the notebook while listening to aural guidance" which helped him to locate components more easily. In contrast, one participant felt pressured by the prototype to complete the tasks in comparison to the available paper-based manual. When asked where the participants could imagine the application of such a system, participants named car repair, maintenance of electronic devices and the assembly of furniture as potential scenarios.

4.2 Industrial Use Case: Laser Cutter Head

We conducted a second usability testing within an industrial environment. Eleven freshman workers (M: 16.9; SD: 0.8) from the choice-based elicitation study performed maintenance tasks related to the cutter head of a 2d laser cutting machine. All participants

had none or limited hands-on experience with this type of machine. Four maintenance tasks and one training task were selected which were part of the maintenance routine in case of cutting problems. All tasks were hardware-related and did not require any interaction with the control panel of the machine. Therefore, participants did not need any prior machine-related training. The machine was located on a shop floor with representative ambient noise.

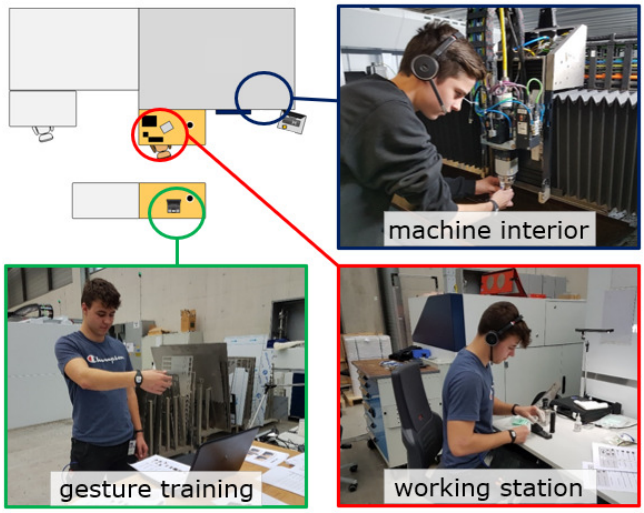


Figure 4: Setup for the industrial use case: cutter head maintenance. First, participants became familiar with the smart wristband by interacting with a gesture-controlled media player (bottom-left). Three maintenance tasks were performed within the machine (top-right) and one task was completed on the respective working station (bottom-right).

In this testing, we omitted the option for head gestures via camera due to the company's privacy regulations. The setup is illustrated in figure 4. The prototype scored a similar mean of 82.7 (mark *A*; SEM: 2.8) compared to the domestic use case. In the post-test questionnaire, eight participants marked the voice recognition as reliable, whereas only three participants were satisfied with the reliability of the gesture recognition. The stated benefit of utterances was hands-free operation and gestures could be used as "an alternative in noisy environments". Similar to the domestic use case, participants named car repair and maintenance of electronic devices as potential scenarios for the application of such a system.

5 DISCUSSION

Classification Models. Prior elicitation studies described gestures with keywords, short sentences or exemplary visualizations [3, 13, 26, 42]. This could lead to ambiguity and misunderstandings on how to perform a described gesture. We understand the provision of the classification models as an improvement for the quality of gesture replication, although few gestures like snapping could not be classified. In several elicitation studies, taxonomies were provided to classify the design space that application designers work with. Dierk et al. [13] already proposed a taxonomy of head-based gestures. However, application designers must find their own approach to identify gestures within this design space. The presented classification models facilitate this process for application designers by providing building blocks on how to design gestures. Besides the design of gestures, the classification models could be easily implemented into an image processing classifier for gesture detection. We also expect users to better understand the gesture detection process through these classification models. Lastly, we want to remark that the application of these models could be insufficient for other application areas. Therefore, we encourage follow researchers to test and to improve the proposed classification models.

User-Defined Gesture Set. We developed a gesture set for 33 referents that is eligible to substitute text- or voice-based interaction of a CUI in the context of maintenance work. The gesture set can be used as a reference by application developers to add gestural input to their CUI. The gesture set's application is not limited to the context of maintenance work because many referents are generic for any system, e.g. *accept*, and were also applicable for surface [42] or ubiquitous computing [10].

It is important to note that the agreement rate only points out gestures that are shared among many for a referent. Novel gestures that could enrich the gesture set remain unnoticed. Therefore, the presented gesture set is comprised of gesture proposals with a medium agreement or higher and novel gestures that are promising, but only received a low agreement rate. Hand gestures were more popular than head gestures because the design space of head gestures is smaller. The argumentation of the small design space is aggravated by the concerns for public use when head gestures were longer than four movements. Also, participants were quickly exhausted by the continuous use of head gestures. Therefore, we recommend that developers should focus first on the provision of hand gestures and implement head gestures only as an auxiliary option. However, even hand gesture proposals scored low for some referents. We want to refer to Mignot et al. [23] whose work indicated that speech and gesture had a complementary relationship because participants preferred gestures for simple actions, while speech was used for more abstract actions within their research. This behavior could also be observed in the elicitation study, e.g. for referent *request summary*.

Preliminary Study. The objective of the two usability testings was to get first results concerning a multimodal CUI that supported head and one-hand gestures. The prototype was well received by the participants which was reflected by the SUS. Head gestures were not performed after the training phase in the first testing and were omitted for the second one. One reason for that was that actions, where both hands were required, rarely occurred. Participants acknowledged gestures as an appropriate alternative to utterances. However, the prototype acted only as a proof-of-concept and the sample size was too small to represent the population. Lastly, the participants stated that the application of such a system could be beneficial for car repair and maintenance work of electronic devices. We interpret this as a reason to conduct further research in this area.

6 CONCLUSION

This paper contributes to the definition process of an appropriate standard for a conversational user interface (CUI) in the context of maintenance work by exploring user-defined gestural input for the communication with a conversational agent (CA) as an instructive assistance system (IAS). We focused on head and one-hand gestures to align to the special constraints that arise from maintenance work environments. Two classification models were derived from the 1,340 gesture proposals that were collected from the conducted elicitation study with 20 participants. A user-defined gesture set was presented to complement text- and voice-based interaction with a CA concerning 33 different user intents. In 82% of the cases, participants preferred to perform a hand gesture instead of a head gesture. In capturing gestures for this study, we have gained insights into the mental models of non-technical users that are valuable for application designers who want to enrich their CUI with gestural input. We want to stress that the application of the classification models and the gesture set is not limited to maintenance work, but is applicable for other contexts regarding multimodal CUI. As a second contribution, we provided results on two usability testings with 18 participants in total using an IAS prototype with a gesture-enabled, multimodal CUI for digital guidance during maintenance work. These studies demonstrated the proof-of-concept of such a CUI and shall encourage further research in this area.

ACKNOWLEDGMENTS

This research was partially funded by the BMBF Software Campus project "IntEnseChoI" (FKZ 01/S17042) in cooperation with the TRUMPF Group. We also want to thank the exeron GmbH for their contribution in this project. We also want to thank Markus Scholz and Alexander Mädche for their insights on this matter.

REFERENCES

[1] Christoph Amma, Marcus Georgi, Tomt Lenz, and Fabian Winnen. 2016. Kinemic Wave: A Mobile Freehand Gesture And Text-Entry System. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '16)*. ACM, 3639–3642. https://doi.org/10.1145/2851581.2890263

[2] Susanna Aromaa, Antti Väätänen, Iina Aaltonen, and Tomi Heimonen. 2015. A Model for Gathering and Sharing Knowledge in Maintenance Work. In *Proceedings of the European Conference on Cognitive Ergonomics 2015 (ECCE '15)*. Association for Computing Machinery, Article 28, 8 pages. https://doi.org/10.1145/2788412.2788442

[3] Ilhan Aslan, Tabea Schmidt, Jens Woehrle, Lukas Vogel, and Elisabeth André. 2018. Pen + Mid-Air Gestures: Eliciting Contextual Gestures. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction* (Boulder, CO, USA) *(ICMI '18)*. Association for Computing Machinery, New York, NY, USA, 135–144. https://doi.org/10.1145/3242969.3242979

[4] Fabricio Barth, Heloisa Candello, Paulo Cavalin, and Claudio Pinhanez. 2020. Intentions, Meanings, and Whys: Designing Content for Voice-Based Conversational Museum Guides. In *Proceedings of the 2nd Conference on Conversational User Interfaces* (Bilbao, Spain) *(CUI '20)*. Association for Computing Machinery, New York, NY, USA, Article 8, 8 pages. https://doi.org/10.1145/3405755.3406128

[5] BAuA - Gemeinsames Ministerialblatt 2014. *Tätigkeiten mit wesentlich erhöhten körperlichen Belastungen mit Gesundheitsgefährdungen für das Muskel-Skelett-System*. Retrieved April 9, 2020 from https://www.baua.de/DE/Angebote/Rechtstexte-und-Technische-Regeln/Regelwerk/AMR/pdf/AMR-13-2.pdf?__blob=publicationFile&v=2

[6] Michael Beigl. 1999. ElectronicManual: Helping Users with Ubiquitous Access. In *Proceedings of the HCI International '99 (the 8th International Conference on Human-Computer Interaction) on Human-Computer Interaction: Communication, Cooperation, and Application Design*, Vol. 2. L. Erlbaum Associates Inc., Hillsdale, NJ, USA, 246–250. http://dl.acm.org/citation.cfm?id=647944.743470

[7] Mudit Ratana Bhalla and Anand Vardhan Bhalla. 2010. Comparative study of various touchscreen technologies. *International Journal of Computer Applications* 6, 8 (2010), 12–18.

[8] William Buxton. 1995. Chunking and phrasing and the design of human-computer dialogues. In *Readings in Human–Computer Interaction*, Ronald M. Baecker, Jonathan Grudin, William A.S. Buxton, and Saul Greenberg (Eds.). Morgan Kaufmann, 494 – 499. https://doi.org/10.1016/B978-0-08-051574-8.50051-0

[9] Julia Cambre and Chinmay Kulkarni. 2020. Methods and Tools for Prototyping Voice Interfaces. In *Proceedings of the 2nd Conference on Conversational User Interfaces* (Bilbao, Spain) *(CUI '20)*. Association for Computing Machinery, New York, NY, USA, Article 43, 4 pages. https://doi.org/10.1145/3405755.3406148

[10] Edwin Chan, Teddy Seyed, Wolfgang Stuerzlinger, Xing-Dong Yang, and Frank Maurer. 2016. User Elicitation on Single-hand Microgestures. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, 3403–3414. https://doi.org/10.1145/2858036.2858589

[11] Hans Jürgen Charwat. 1994. *Lexikon der Mensch-Maschine-Kommunikation* (2nd ed.). Oldenbourg, München. 32 pages.

[12] André da Silva Barbosa, Felipe Pinheiro Silva, Lucas Rafael dos Santos Crestani, and Rodrigo Bueno Otto. 2018. Virtual assistant to real time training on industrial environment. In *Transdisciplinary Engineering Methods for Social Innovation of Industry 4.0: Proceedings of the 25th ISPE Inc. International Conference on Transdisciplinary Engineering, July 3–6, 2018*, Vol. 7. IOS Press, 33. https://doi.org/10.3233/978-1-61499-898-3-33

[13] Christine Dierk, Scott Carter, Patrick Chiu, Tony Dunnigan, and Don Kimber. 2019. Use Your Head! Exploring Interaction Modalities for Hat Technologies. In *Proceedings of the 2019 on Designing Interactive Systems Conference (DIS '19)*. ACM, 1033–1045. https://doi.org/10.1145/3322276.3322356

[14] Markus Funk, Thomas Kosch, and Albrecht Schmidt. 2016. Interactive Worker Assistance: Comparing the Effects of In-situ Projection, Head-mounted Displays, Tablet, and Paper Instructions. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '16)*. ACM, 934–939. https://doi.org/10.1145/2971648.2971706

[15] Ulrich Gnewuch, Stefan Morana, Marc Adam, and Alexander Maedche. 2018. Faster is Not Always Better: Understanding the Effect of Dynamic Response Delays in Human-Chatbot Interaction. In *26th European Conference on Information Systems*.

[16] Ulrich Gnewuch, Stefan Morana, and Alexander Maedche. 2017. Towards Designing Cooperative and Social Conversational Agents for Customer Service. In *Proceedings of the 38th International Conference on Information Systems (ICIS), Seoul, ROK, December 10-13, 2017. Research-in-Progress Papers*. AIS eLibrary (AISeL).

[17] Shuangping Gon, Huajuan Ma, Yihang Wan, and Anran Xu. 2019. Machine Learning in Human-Computer Nonverbal Communication. In *NeuroManagement and Intelligent Computing Method on Multimodal Interaction* (Suzhou, China) *(ICMI '19)*. Association for Computing Machinery, New York, NY, USA, Article 4, 7 pages. https://doi.org/10.1145/3357160.3357670

[18] Jean D. Hallewell Haslwanter, Michael Heiml, and Josef Wolfartsberger. 2019. Lost in Translation: Machine Translation and Text-to-speech in Industry 4.0. In *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments (PETRA '19)*. ACM, 333–342. https://doi.org/10.1145/3316782.3322746

[19] Martin W. Hoffmann, Rainer Drath, and Christopher Ganz. 2021. Proposal for requirements on industrial AI solutions. In *Machine Learning for Cyber Physical Systems*, Jürgen Beyerer, Alexander Maier, and Oliver Niggemann (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 63–72.

[20] Thomas Kosch, Yomna Abdelrahman, Markus Funk, and Albrecht Schmidt. 2017. One Size Does Not Fit All: Challenges of Providing Interactive Worker Assistance in Industrial Settings. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers (UbiComp '17)*. ACM, 1006–1011. https://doi.org/10.1145/3123024.3124395

[21] Bruce D. Lucas and Takeo Kanade. 1981. An Iterative Image Registration Technique with an Application to Stereo Vision. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 2* (Vancouver, BC, Canada) *(IJCAI'81)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 674–679.

[22] Akhil Mathur, Anton Isopoussu, Fahim Kawsar, Robert Smith, Nicholas D. Lane, and Nadia Berthouze. 2018. On Robustness of Cloud Speech APIs: An Early Characterization. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers* (Singapore, Singapore) *(UbiComp '18)*. Association for Computing Machinery, New York, NY, USA, 1409–1413. https://doi.org/10.1145/3267305.3267505

[23] Christophe Mignot, Claude Valot, and Noëlle Carbonell. 1993. An Experimental Study of Future &Ldquo;Natural&Rdquo; Multimodal Human-computer Interaction. In *INTERACT '93 and CHI '93 Conference Companion on Human Factors in Computing Systems (CHI '93)*. ACM, 67–68. https://doi.org/10.1145/259964.260075

[24] Calkin S. Montero, Jason Alexander, Mark T. Marshall, and Sriram Subramanian. 2010. Would You Do That? Understanding Social Acceptance of Gestural Interfaces. In *Proceedings of the 12th International Conference on Human Computer Interaction with Mobile Devices and Services* (Lisbon, Portugal) *(MobileHCI '10)*. Association for Computing Machinery, New York, NY, USA, 275–278. https://doi.org/10.1145/1851600.1851647

[25] Nils Petersen and Didier Stricker. 2012. Learning task structure from video examples for workflow tracking and authoring. In *2012 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. 237–246. https://doi.org/10.1109/ISMAR.2012.6402562

[26] Jaime Ruiz, Yang Li, and Edward Lank. 2011. User-defined Motion Gestures for Mobile Interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, 197–206. https://doi.org/10.1145/1978942.1978971

[27] Jeff Sauro and James R. Lewis. 2012. Chapter 8 - Standardized Usability Questionnaires. In *Quantifying the User Experience*, Jeff Sauro and James R. Lewis (Eds.). Morgan Kaufmann, 185–240. https://doi.org/10.1016/B978-0-12-384968-7.00008-4

[28] Stefan Schaffer and Norbert Reithinger. 2019. Conversation is Multimodal: Thus Conversational User Interfaces Should Be as Well *(CUI '19)*. Association for Computing Machinery, New York, NY, USA, Article 12, 3 pages. https://doi.org/10.1145/3342775.3342801

[29] Benedikt Schmidt, Reuben Borrison, Andrew Cohen, Marcel Dix, Marco Gärtler, Martin Hollender, Benjamin Klöpper, Sylvia Maczey, and Shunmuga Siddharthan. 2018. Industrial Virtual Assistants: Challenges and Opportunities. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers* (Singapore, Singapore) *(UbiComp '18)*. Association for Computing Machinery, New York, NY, USA, 794–801. https://doi.org/10.1145/3267305.3274131

[30] Manex Serras, Laura García-Sardiña, Bruno Simões, Hugo Álvarez, and Jon Arambarri. 2020. Dialogue Enhanced Extended Reality: Interactive System for the Operator 4.0. *Applied Sciences* 10, 11 (2020). https://doi.org/10.3390/app10113960

[31] A. Syberfeldt, O. Danielsson, and P. Gustavsson. 2017. Augmented Reality Smart Glasses in the Smart Factory: Product Evaluation Guidelines and Review of Available Products. *IEEE Access* 5 (2017), 9118–9130. https://doi.org/10.1109/ACCESS.2017.2703952

[32] Ying-Chao Tung, Mayank Goel, Isaac Zinda, and Jacob O. Wobbrock. 2018. RainCheck: Overcoming Capacitive Interference Caused by Rainwater on Smartphones. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction* (Boulder, CO, USA) *(ICMI '18)*. Association for Computing Machinery, New York, NY, USA, 464–471. https://doi.org/10.1145/3242969.3243028

[33] Angeliki Tzouganatou. 2018. Can Heritage Bots Thrive? Toward Future Engagement in Cultural Heritage. *Advances in Archaeological Practice* 6, 4 (2018), 377–383. https://doi.org/10.1017/aap.2018.32

[34] Francisco A. M. Valério, Tatiane G. Guimarães, Raquel O. Prates, and Heloisa Candello. 2017. Here's What I Can Do: Chatbots' Strategies to Convey Their Features to Users. In *Proceedings of the XVI Brazilian Symposium on Human Factors in Computing Systems* (Joinville, Brazil) *(IHC 2017)*. Association for Computing Machinery, New York, NY, USA, Article 28, 10 pages. https://doi.org/10.1145/3160504.3160544

[35] Radu-Daniel Vatavu and Stefan Pentiuc. 2008. Multi-Level Representation of Gesture as Command for Human Computer Interaction. *Computing and Informatics* 27 (01 2008), 837–851.

[36] Radu-Daniel Vatavu and Jacob O. Wobbrock. 2015. 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 (CHI '15)*. ACM, 1325–1334. https://doi.org/10.1145/2702123.2702223

[37] Paul Viola and Michael Jones. 2001. Rapid Object Detection Using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001.* 511–518. https://doi.org/10.1109/CVPR.2001.990517

[38] Sruthi Viswanathan, Fabien Guillot, and Antonietta Maria Grasso. 2020. What is Natural? Challenges and Opportunities for Conversational Recommender Systems. In *Proceedings of the 2nd Conference on Conversational User Interfaces* (Bilbao, Spain) *(CUI '20)*. Association for Computing Machinery, New York, NY, USA, Article 40, 4 pages. https://doi.org/10.1145/3405755.3406174

[39] Wolfgang Wahlster. 2006. *Dialogue Systems Go Multimodal: The SmartKom Experience*. Springer Berlin Heidelberg, Berlin, Heidelberg, 3–27. https://doi.org/10.1007/3-540-36678-4_1

[40] Carsten Wittenberg. 2008. Is multimedia always the solution for human-machine interfaces? - a case study in the service maintenance domain. In *2008 15th International Conference on Systems, Signals and Image Processing*. 393–396. https://doi.org/10.1109/IWSSIP.2008.4604449

[41] Jacob O. Wobbrock, Htet Htet Aung, Brandon Rothrock, and Brad A. Myers. 2005. Maximizing the Guessability of Symbolic Input. In *CHI '05 Extended Abstracts on Human Factors in Computing Systems (CHI EA '05)*. ACM, 1869–1872. https://doi.org/10.1145/1056808.1057043

[42] Jacob O. Wobbrock, Meredith Ringel Morris, and Andrew D. Wilson. 2009. User-defined Gestures for Surface Computing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)*. ACM, 1083–1092. https://doi.org/10.1145/1518701.1518866

[43] Josef Wolfartsberger, Jean D. Hallewell Haslwanter, Roman Froschauer, René Lindorfer, Mario Jungwirth, and Doris Wahlmüller. 2018. Industrial Perspectives on Assistive Systems for Manual Assembly Tasks. In *Proceedings of the 11th PErvasive Technologies Related to Assistive Environments Conference (PETRA '18)*. ACM, 289–291. https://doi.org/10.1145/3197768.3201552

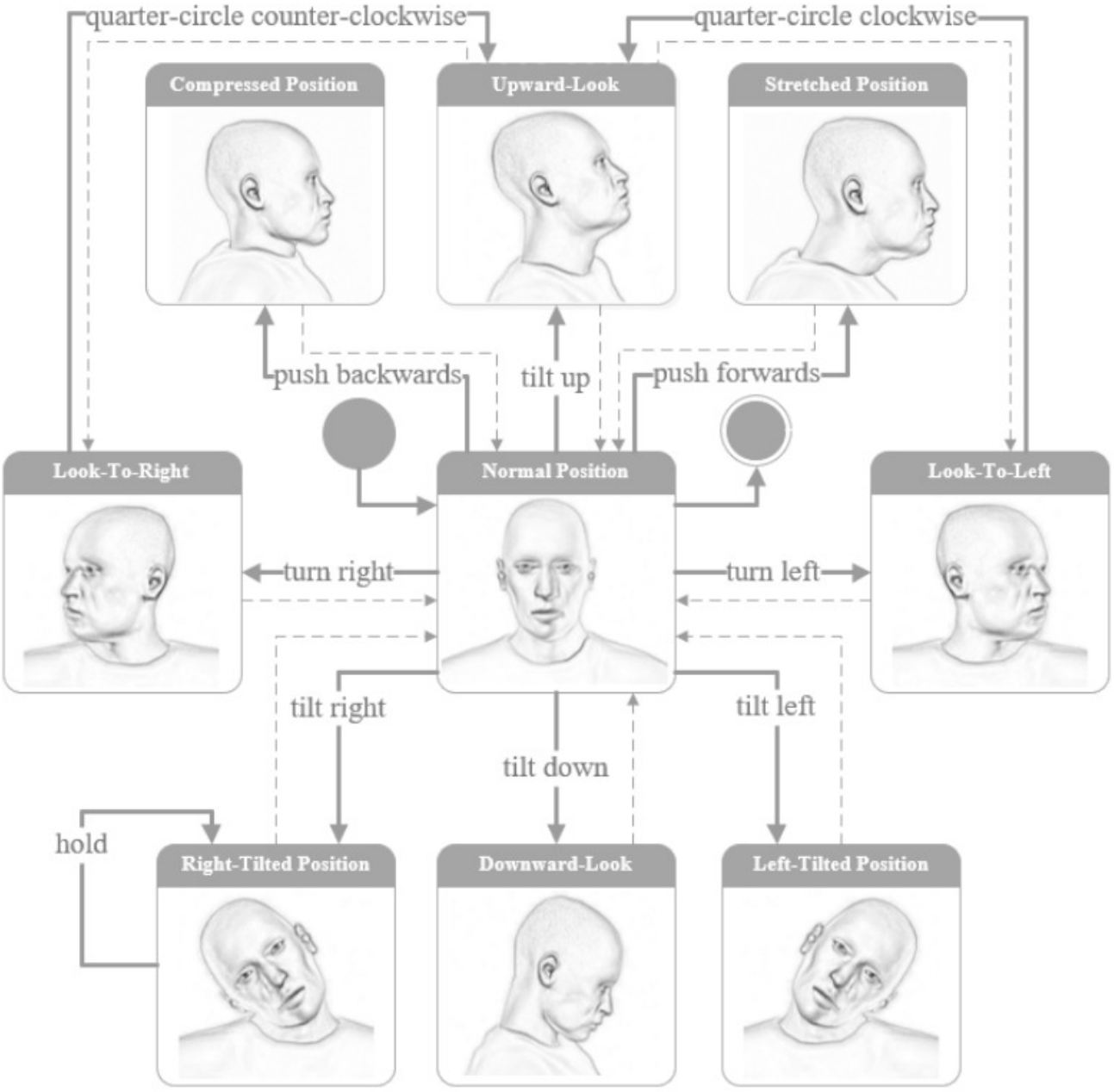
[44] Ionut-Alexandru Zaiti, Stefan-Gheorghe Pentiuc, and Radu-Daniel Vatavu. 2015. On free-hand TV control: experimental results on user-elicited gestures with Leap Motion. *Personal and Ubiquitous Computing* 19, 5 (2015), 821–838. https://doi.org/10.1007/s00779-015-0863-y

[45] Xianjun Sam Zheng, Patrik Matos da Silva, Cedric Foucault, Siddharth Dasari, Meng Yuan, and Stuart Goose. 2015. Wearable Solution for Industrial Maintenance. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '15)*. ACM, 311–314. https://doi.org/10.1145/2702613.2725442

[46] Jens Ziegler, Sebastian Heinze, and Leon Urbas. 2015. The potential of smartwatches to support mobile industrial maintenance tasks. In *2015 IEEE 20th Conference on Emerging Technologies Factory Automation (ETFA)*. 1–7.

A EXEMPLARY STATE MACHINE FOR CHUNKING HEAD GESTURES

The presented state machine is optimized for clarity and does not include all possible states. Each edge is representing a movement vector and is only described once in the figure to preserve clarity. Inverse vectors can be identified by the figure's symmetric characteristic. For instance, the inverse vector of movement *tilt up* is *tilt down*. Movement *hold* is applicable for all states and describes an intentional stagnancy of movement. The head model was created using MakeHuman[3].



[3]http://www.makehumancommunity.org/, last accessed on January 30th, 2021

B HAND SHAPE ILLUSTRATIONS

Illustrations of the proposed hand shapes which originated from the American Sign Language (ASL) and the German Sign Language (DGS).

ASL-1 ASL-2 ASL-3 ASL-4

ASL-A ASL-B ASL-C ASL-H

ASL-O ASL-Q_open ASL-Q_close ASL-S

ASL-Y ASL-Z DGS-1 DGS-SCH

C EXEMPLARY CONVERSATION FLOW INSTANCE

The instance shows the task guidance for replacing the notebook's heat sink assembly that was part of the first usability testing. It illustrates the multimodal dialogue. Note that this example does only show a subset of the referents that were available at both usability testings. The instructions were originally in German, but were translated into English for a better accessibility.