Computers & Graphics 82 (2019) 321–331

Contents lists available at ScienceDirect

Computers & Graphics

journal homepage: www.elsevier.com/locate/cag

Special Section on Cyberworlds 2018

2D, 3D or speech? A case study on which user interface is preferable for what kind of object interaction in immersive virtual reality

Daniel Hepperle[a,*], Yannick Weiß[b], Andreas Siess[a], Matthias Wölfel[a]

[a] Faculty of Computer Science and Business Information Systems, Karlsruhe University of Applied Sciences, Germany
[b] Faculty of Mathematics, Informatics and Statistics, Ludwig Maximilian University of Munich, Germany

ARTICLE INFO

Article history:
Received 14 November 2018
Revised 7 June 2019
Accepted 13 June 2019
Available online 18 June 2019

MSC:
68U35

Keywords:
Virtual Reality
Comparison of User Interfaces
Input Modalities
2D interface
3D interface
Speech interface

ABSTRACT

Recent developments in human machine interaction offer three principal different approaches to interact with 3D environments, namely: *2D overlays* using icons, *3D interfaces* resembling interactions of the real world and *speech interfaces* which matured in the last years and are becoming more and more popular in other context such as smartphones or smart homes. Faced with the task to select the best interaction strategy to interact with immersive environments one is left with best practice and literature. But neither offers a clear strategy and all the methods are used widely. In particular a consistent comparison providing insights on when to use what interface in immersive virtual environments is missing. In order to evaluate the relative strengths and weaknesses of each interface in relation to different tasks in immersive environments a quantitative user study has been conducted. Results showed significant differences on the interface performances according to different parameters: ease of learning (Speech and 2D are favored), uncomplicated handling (Speech), speed (Speech and 2D), overview (2D), fun (3D), comprehension (2D and 3D) and on how simple and efficient it is to handle text input (Speech).

1. Introduction

Finding good human machine interaction strategies is not an easy task. This is further complicated if new technologies such as immersive virtual environments come into play. Well known screen-interaction guidelines (i.e. for WIMP—Windows, Icons, Menus and Pointer or NUI—natural user interfaces) can only be transferred into virtual reality (VR) applications conditionally. This is mainly caused by an increased degree of freedom (DOF) in VR compared to 2D interfaces. Generally, a 2D interface has 2 DOFs (position on the 2 dimensions X and Y),[1] while a 3D interface supports 6 DOFs (position and orientation on the X,Y and Z-dimensions). Besides a simple change from 2D to 3D, the constant adaptation of the viewing point according to the head orientation provides a higher degree of immersion. Choosing the wrong interaction paradigms can therefore break the feeling of presence, while choosing a suitable one could even increase it. One might think that, instead of adapting screen-interaction guidelines, interaction strategies used in the physical world can simply be emulated. But once more, it is not as simple as it might look at first sight: The lack of haptic feedback or precision of the input device might draw a natural mapping impractical and alternative solutions are more promising. Besides technical limitations, design strategies might be required which overcome the limitations of the physical world by introducing techniques to enable, for instance, the selection or the manipulation of objects that might be out of reach otherwise; popular techniques in this respect are Go-Go [1], occlusion or ray-casting.

To make things even more complicated, popular VR gear comes along with a broad variety of possible implementation strategies by providing access to various sensory information: the handheld-controllers provide at least position, orientation, and switches while the head-mounted-display (HMD) provides at least position and orientation. Some of them include a microphone and speakers, some do not. Microsoft for example encourages users to use their speech interface (SI) in the new Windows Mixed Reality [2]. A popular extension is to mount a depth sensor, such as the LEAP-Motion, in front of the HMD to add free hand gestural input. In an earlier study, we came up with a first guideline that helps deciding when to use which interface [3]. These findings have been evaluated further within this work. Section 2 offers an exhaustive literature review on studies relevant to this research. This is followed by defining relevant tasks that are common for immersive VR environments in Section 3. We then describe the decision process for the respective interfaces used for this study in Section 4 and

* Corresponding author.
E-mail address: daniel.hepperle@hs-karlsruhe.de (D. Hepperle).
[1] In a 2D interface the orientation is usually not considered.

https://doi.org/10.1016/j.cag.2019.06.003
0097-8493/

explain the test preparation, the physical and virtual set up of our experiment and its procedure in Section 5. Section 6 shows and discusses the results gained from the user study. This is followed by a corresponding error analysis in Section 7 and a conclusion and outlook in Section 8.

In addition to the previous paper, the fastest interface — Speech[2]— is analyzed in more detail to discuss further improvements. We categorized the respective tasks that had to be fulfilled within the evaluation by their superordinate interaction-paradigms (see Section 5.4) to conclude further findings. We show that for the SI, there is a significant correlation between the number of steps needed to fulfill a task and the perceived overall experience of the application (See Section 6.7). Also, for a better understanding, the individual capabilities of each interface have been summarized. The interfaces then are discussed in regard to multimodal interaction and the resulting cognitive load as well as in regard to using avatars as a supplementary interface for virtual reality.

2. Literature review

Being left in an ocean of possibilities without good guidance one might first refer to the scientific literature to get some orientation. However, meaningful user-based studies that compare well established interaction methods in VR with each other, or with rather nontraditional methods, are hard to find.

Although many strategies are presented, they are not consistently compared [4,5]. Other publications highlight the basic rules for app designers to produce vis-a-vis artificial immersion [6]. In [7] different 3D object manipulation techniques, ranging from WIMP, to touch until mid-air interfaces are discussed. Poupyrev et al. [8] present results of a formal study that compares virtual hand and virtual pointer. Recent publications review interaction techniques in virtual environments with the focus on selection techniques without considering menus or semantic queries [9] or on more general interaction paradigms [10]. Recently Kaur and Yammiyavar [11] compared a gaze controlled 2D version with a 3D version of a mobile keyboard, stating a user preference for the 3D version, but the option of a SI is not considered in this discussion. A dated study shows an approach of combining all different interfaces in a cave automated virtual environment (CAVE), but does not investigate the individual performance of each interface [12]. Other work investigates speech-based interfaces to be used in VR applications, the literature however is already dated [13–17]. Speech recognition technologies have made tremendous progress in recent years, just like VR, which led us conclude that findings reported in the mentioned publications might not be valid for today's technology. Besides technological progress, todays users are also more used to speech interfaces, which contributes to a different evaluation performance. He et al. [18] address the problem of complex sentence formation, which a system can be unable to process, by introducing a highly standardized "instruction language", which can be understood more easily by language processing technology in clearly defined contexts. Lee [19] investigates single and multimodal speech and gesture interaction with 3D objects in augmented reality (AR), but not in VR. A direct transfer of the gained insights, even though head-mounted AR and VR is related to each other, might not be feasible for several reasons:

- Todays head-mounted AR devices offer smaller FOV.
- AR Interaction Systems have to take uncontrolled environmental conditions in consideration in real time (e.g. if physical space is limited, the AR System should suggest voice input, if you are in a library, it should not.).
- Most VR headsets ship with dedicated input devices that people would not carry around with them to interact in (semi) public space.
- People might be reluctant to use speech commands in public spaces.

While 2D and 3D interfaces are well described for VR and more recent comparisons can be found [20], speech-based interfaces lack recent investigations. This is particularly surprising since SIs seem to be omnipresent since the introduction of Apple's Siri and Amazon's Alexa.

Already in 2000, Oviatt et al. [21] stated that besides a more effective natural language processing, dialogue processing, semantic recognition, error handling techniques and cognitive theory of multimodal interaction, multimodal interfaces still need to be addressed in more detail. 19 years later a lot of research in this direction has been conducted, however no clear statement on how to implement multimodal interfaces and which modality to use for what function range can be made. For real environments, Welch and Warren [22] came up with the term *modality appropriateness*, that says that our central nervous system adapts to the upcoming tasks in a statistically optimal manner. But is this true for VR as well? McGlashan [23] and Irawati [24] suggest the combination of different interfaces/modalities to further improve interaction and satisfaction of a virtual or augmented reality application.

3. Interaction tasks

In contrast to two-dimensional space providing 2 or maximum 3 DOF, three-dimensional space offers 6 DOF. While the increase of freedom doesn't seem much, it challenges the input device as well as the user [12]. To provide a more natural and intuitive interaction, manufacturers and researchers are working on all kind of interaction technologies that could add even more freedom (i.e. interfaces for hand tracking [25,26], haptic feedback [27–29], gesture recognition [30,31], limb recognition / full body tracking [32,33], etc.) to VR. But even for current interaction technologies, guidelines still need to be defined (e.g. a 3D gesture alphabet is still not defined). In order to compare different interfaces and their advantages as well as their disadvantages for particular tasks, the most common functions including freely modifiable parameters have to identified and classified. The distinguishable interaction tasks with their corresponding parameters, which we have used in our evaluation, are summarized in Table 1. Each task is briefly described next:

Table 1
Interaction tasks and their corresponding parameters as documented in literature.

Task	Parameter	Source
Selection	True, false	[4,34]
Manipulation	System state (on/off);	[4]
Position	Relative movement, absolute placement	[5]
Rotation	Relative orientation, absolute orientation	[5]
Creation	Generate objects, delete objects	[34]
Modification	Objects' appearance (e.g. color)	[34]
Text input	Character, word, phrase	[34]

- **Selection**
  This category is a fundamental prerequisite for most types of manipulation to follow. It is necessary to specify what object to refer to, before the actual action can be performed on it or with it.
- **Manipulation**
  This category includes those functions that are available after selection while only requiring just one simple interaction without any additional selection or specification of parameters, e.g.

[2] After error correction.

Table 2
Overview over similarities and differences between and individual capabilities of the evaluated interfaces. The values given only serve as descriptive indications for easier comparison or ordering and should not be seen as specific quantitative values.

	2D	3D	Speech
Perception	Visual	Visual	Audiovisual
Degrees of freedom	2 (3)	6	Vocabulary size
Non-sequential input possible	No	No	Yes
Object manipulation	Indirect	Direct	Indirect
Long range interaction	Yes	No	Yes
Interaction latency	Low	Low	High
Error susceptibility	Low	Low	High
Commonplace usage of UI	High	Low	Low

activating and deactivating functions of an object such as turning on/off the light of a lamp.

- **Position**
  These are manipulations that change the position of objects. They are performed through variable numeric values without the need of a distinct selection of possible options and can occur abruptly (e.g. teleportation) or continuously (e.g. dragging/carrying). Our test environment includes free positioning anywhere on the horizontal plane.
- **Rotation**
  This category includes manipulations that change the orientation of objects. These are also performed through arbitrary numeric values. In the test application, it is possible to rotate freely around the vertical axis.
- **Creation**
  This category is comprised of functions that change the number of active objects in the scene. They require a decision between certain predefined options on lists. For instance, the participants had the possibility of choosing a piece of furniture to spawn a new object inside the virtual environment.
- **Modification**
  These are functions, which change the appearance or properties of an object. An object's state can be modified through predefined options and requires a preceding selection of which object to affect. In the test environment, we allow the re-coloring of an object (by selection of the object and color or surface structure).
- **Text Input**
  This category includes the input of free texts for the independent and individual labeling of certain objects in a scene.

4. Description of the interfaces used

This chapter introduces the implementation of the 2D, 3D and Speech interfaces as used in our experiments. A general introduction to VR interfaces can be found in Jerald [4]. All interfaces make use of handheld controllers (HTC Vive Controller), with the only difference being the displayed image and the way of interaction with the controller itself. In addition, the microphone in the HTC Vive is used for speech input. While the SI only requires pressing a key on the controller (push-to-talk), the 2D and 3D interfaces depend on additional information for interactions, such as the position and orientation of the controllers. In addition to the detailed description, an overview of each interface's capabilities, as well as the differences and similarities between them is given in Table 2.

4.1. 2D user interface

The interaction method we implemented for the 2D interface is executed indirectly by pressing flat buttons with corresponding icons on a virtual two-dimensional surface in a 3D virtual environment. The implementation is realized in a manner as applied in various VR environments: A visible beam is cast from the virtually represented controller and adapts to its orientation and position (see Figure 1, left image). This beam can be utilized as a (mouse-) pointer by detecting the collision with the two-dimensional plane. Through pointing at controls—whose functions are indicated by commonly used symbols—and pressing the trigger button on the controller, objects can be manipulated, created, positioned and rotated. Selection in 2D is done by pointing on an element (i.e. a table) of the list. Tasks were given in such a way that you would not have to re-select the element again.

4.2. 3D user interface

To keep interaction as close as possible in a VR context to real life and to reflect the most common used strategy in 3D interaction in VR, we used simple hand interaction (direct representation of the controllers/hand in VR) instead of Go-Go, occlusion or ray-casting that would enable participants to reach objects that are not in their physically available range. For the 3D interface we required the user to collide with the object to interact with—a standard method often used in VR, which tries to emulate the physical interaction in reality. Objects colliding with the controller can be lifted by holding the trigger button on the respective controller and can be set down again by releasing the button. The data generated by the movement (velocity, acceleration) of the object is initially maintained when the object is released, to simulate inertia and allow objects to be thrown. Furthermore, tools (hammer and brush) are implemented that can be picked up and provide context-sensitive functions such as removing objects from the scene or recoloring them when in touch with other virtual objects (see Fig. 1, center image). If these interactions would require great force in reality, this is simulated through the controllers' ability for haptic feedback in form of vibrations varied in strength and duration. In addition, further keys of the controller are used to provide a comparable range of functions as the 2D interface (trackpad for teleportation; trigger for selection; grip button for manipulation e.g. switching the lamp on/off.[3] Since the user needs the ability to touch the object, it is necessary that movement within the virtual space is enabled.[4] In order for the participant to change position (teleport), he can press the HTC Vive controller's trackpad. Once pressed, the virtual controller sends out a visible light beam to point at the desired destination. Releasing the trackpad will teleport the participant to the destination selected. Objects attached to the controller are not detached when using teleportation.

4.3. Speech user interface

In contrast to the previous two interfaces the SI has to be activated by pressing the trigger button on the HTC Vive controller. This is required to prevent recognition of words which are uttered, but not meant as input to the speech system. After a short delay for the start of the input process, the user is able to say any arbitrary sentence in German.[5] The recognized words are displayed instantly into the user's field of view to enable direct feedback (see Fig. 1 right). When the user has finished his sentence, which is detected automatically by pausing to utter, the intention of the command is detected along with additional information or parameters. Now it is possible to trigger the desired action, depending on the

[3] For the HTC Vive controller's button naming conventions see the Vive user manual.

[4] The virtual space in our setup is larger than the physically available space, so regular movement within the VR space was not sufficient but, up to a specific distance, possible as well.

[5] The mother tongue of all tested participants.

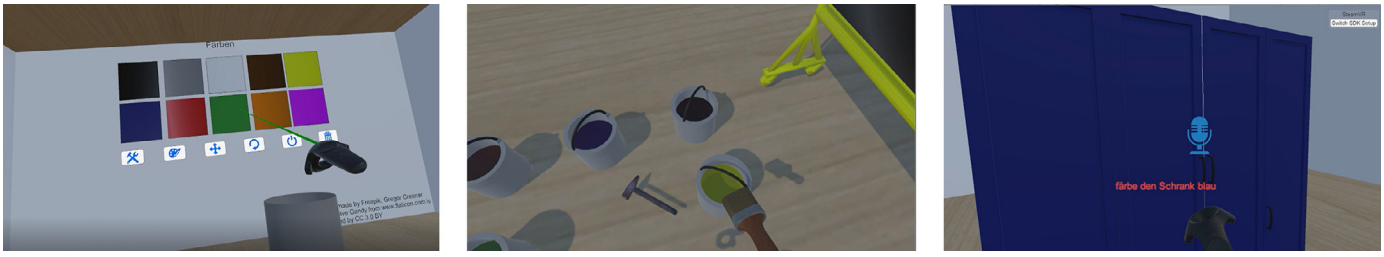

Fig. 1. Screenshots of the used input methods: 2D interface (left): selecting a color by pointing on one of the colored rectangles. 3D interface (middle): selecting a color by putting the brush into one of the buckets. Speech interface (right): selecting a color via voice command (The red text shows what sentence/word has been understood by the speech interface). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

required and optional parameters. In addition, the application repeats the executed action via vocal description. If no action can be assigned to the command the application generates feedback in the form of a request or an error message. Optional parameters that are not passed are either ignored or replaced by default values. If a necessary parameter is missing or context-dependent terms—e.g. directions or anaphors—are used, the application first tries to conclude the required information from the current context—e.g. the direction of the user's head or the objects' position/presence in the scene. If the application fails to do this, a specific demand for the missing parameter is generated, while the collected valid fragments are retained for the next input cycle. While speech user interfaces[6] are becoming more and more popular recently,[7] discoverability still is a big issue compared to other user interfaces. While in other interfaces—especially 2D interfaces—one almost instantly gets an idea of the scope of functions you can use, a SI usually lacks such an overview of functionality. In our case for instance, participants were asked to colorize a piece of furniture in a predefined color, so it was imaginable that it is possible to use that specific color. For the 2D and 3D interfaces it is instantly clear what colors one could use (see Fig. 1) while for the SI you only could guess which colors are available. On the other hand, it might be way easier to access a large amount of different colors via the SI while especially for the 3D interface it would quickly become more complex. The issue of discoverability was tackled in such a way, that we explained the basic concept of the interfaces while in training phase. We explained all commands that are not common in everyday interaction with other humans (i.e. "generate a chalkboard"). Commands like "move the cabinet right next to the TV" were explained briefly. So, we did not offer participants some kind of pre-printed vocabulary to learn by heart but helped them generate a fitting mental model [35] on how to interact with this interface.

5. Test setup

To compare different interfaces a test application is required, which allows different strategies to be implemented. Since we could not find a useful application fulfilling our requirements we had to develop our own solution. We decided to portray the scenario of a furnishing application, since this use case contains a broad variety of test-worthy functions that are categorized and described in Section 3. Each of the user interfaces are built on the same fundamental application, allowing all three interfaces to have the possibilities for an identical range of functions, as well as an effective and target-oriented usage.

5.1. Notes on the technical implementation

For creating the application, we used hardware and software tools frequently used within the VR community: standard PC with Microsoft Windows 10, Intel i7, Nvidia 1080 and an HTC Vive with hand controllers and *Unity3D*[8] as a development platform. To ensure that the investigated interfaces are reproducible and are reasonably comparable—e.g. in complexity and performance—to those in many other VR applications, we used common interaction techniques and used available and widely used toolkits. Within the game engine *Unity3D* we used *Valve's SteamVR*[9] and the open source *Virtual Reality Toolkit*.[10] To realize the SI, we communicated asynchronously with *Google Cloud Speech API*.[11] This allowed us to have a live display of what currently is recognized and also to see the final result of the speech recognition right before it is transferred to the dialog system *Dialogflow*.[12] To give verbal feedback speech synthesis was used, realized through the *SpeechSynthesizer* class within the.NET framework[13] and the standard voice included in Microsoft Windows 10.

5.2. Room setup

The test took part in a quiet room in which 3.5 × 2.5 m were used to move freely in the VR application. For safety reasons each side in the physical space had at least 1 m additional space. The virtual space was a rectangular room of 12 × 12 m containing a solid gray wall on each side, a wooden floor and ceiling. To mark the boundaries of the physical interaction space Valve's safety feature *Chaperone*[14] was used. Using a virtual space larger than the available physical space was intended to make the results derived from that design decision apply better to actual VR usage situations. A study conducted by Valve in 2015 (n = 99) shows that only 6% of the users had a VR play area larger than 4 × 4 m [36]. Testing actual real walking situations would not be expedient because of technical and physical limitations. The highest-selling VR games such as Fallout VR [37] also intend the users to somehow move in a virtual world that is larger than the typical available physical space.[15]

[6] Speech user interface and voice user interface are often used interchangeably. However, note, that the meaning of speech and voice is different: Speech recognition aims to strip out the personal differences to detect the words said. Voice recognition typically disregards the language and meaning to detect the physical person behind the speech.

[7] Microsoft, Amazon, Apple and Google all came up with their own speech assistant.

[8] See: https://unity3d.com/unity (accessed 2018-07-21).

[9] See: https://assetstore.unity.com/packages/templates/systems/steamvr-plugin-32647 (accessed 2018-07-21).

[10] See: https://github.com/thestonefox/VRTK/ (accessed 2018-07-21).

[11] See: https://cloud.google.com/speech/?hl=de (accessed 2018-04-18).

[12] See: https://dialogflow.com/ (accessed 2018-04-18).

[13] See: https://www.microsoft.com/net/ (accessed 2018-04-18).

[14] See: https://xinreality.com/wiki/Chaperone (accessed 2018-07-25).

[15] I.e. in Fallout VR the user can choose between teleport or using the controller's trackpad.

5.3. Test preparation

Since most participants are frequently using 2D interfaces but might have no experience with 3D and Speech interfaces, we asked the participants to get accustomed to the respective interfaces within the application. For a maximum of 5 min per interface (max. 15 min total) the participants could access all functions and got help to understand and learn all necessary features (e.g. which button to press, how to move, how to create something, how to write on the virtual chalkboard, etc.). The users were free to stop the learning period if they felt confident in using the system. Free learning and usage offered a comprehensive overview of the interfaces that the test participants were asked to evaluate in the subsequent questionnaire. Once the user acquired the necessary knowledge of the functions, the test was started by the test's supervisor manually.

5.4. Test procedure

We used a within group design where all interfaces have been tested by all participants. To reduce the learning effects, the order of the tested interfaces was chosen randomly for each participant. For each interface, the users had to complete the same 17 tasks. The following instructions were given to the users in order.[16] We added the task types classified in Section 3 in brackets (SM = Simple Manipulation, POS = Positioning and Rotating Objects, MOD = Create or Modify Objects Using Lists, TEX = Text Input):

1. We need a brown cupboard (MOD)
2. Move the cabinet to the wall ahead of you (POS)
3. Now we need a red lamp (MOD)
4. Place the lamp to the right of the cabinet (POS)
5. Create a TV set (MOD)
6. Rotate the TV set by 180° (POS)
7. Switch on the lamp (SM)
8. Destroy the cabinet (SM)
9. Move the TV set to the wall at your right (POS)[17]
10. Create a chalkboard (MOD)
11. Write the following word on the chalkboard: Hello (TEX)
12. Colorize the chalkboard in red. (MOD)
13. Write a sentence that contains at least 20 characters onto the chalkboard (TEX)
14. Switch off the lamp (SM)
15. Rotate the chalkboard by 90° clockwise (POS)
16. Destroy the chalkboard. (SM)
17. Move the Lamp to the TV (POS)

Once the procedure started, the upcoming task was written directly into the VR environment. Since the 2D interface was projected onto the walls, the instructions were placed in 3D space. For the other interfaces (3D and speech) it was written on the front facing wall. To fulfill the given tasks, the use of one or more functions (see Table 1) was required. Due to differences between the interfaces, the number of necessary interaction steps varies: The non-sequential SI can—but does not necessarily have to—lead to the desired objective through a smaller number of interaction steps. The user worked through each instruction in sequence and was given a new task once he has successfully accomplished the preceding one. Once the participant finished all tasks, they were asked to fill in an evaluation questionnaire in order to capture all subjective criteria. Fulfilling all 17 tasks for the three interfaces took around 20–30 min per participant plus the time to get accustomed to the interfaces (additional max. 15 min). In total it took each participant around one hour to learn the interactions, complete the tasks and fill out the questionnaire.

[16] The tasks were given in German and are accordingly translated into English for better understanding.

[17] Additionally an arrow that indicated the direction was shown.

5.5. Participants

The described test procedure was performed on 30 persons—20 male, 10 female—with an average age of 30 years (min. 18, max. 49, standard deviation 8.2). The majority of the test persons (66.7%) already had experience with VR systems and half of the participants (50.0%) already used at least one 3D input device. Even though two-thirds already experienced VR before, the time spent in VR was not very long: 10 persons had been using VR for less than 15 min, 3 persons had been using VR between 15 min and 1 h, and 7 persons had been using VR previously between 1 h and 10 h. None have used VR for more than 10 h. Almost one fourth (23.3%) are using digital speech assistants such as Apple Siri, Amazon Alexa or Google Assistant.

5.6. Evaluation data

For evaluation purposes we used two methods of data acquisition. While participants were executing the tasks in VR, the system automatically recorded the time it took a participant to fulfill the respective tasks and afterwards we asked the participants to fill in a questionnaire. The questionnaire was separated into three main parts: general information (age, gender—free text answers), existing knowledge about VR Systems (multiple choice answers) and specific questions about the respective interfaces (2D, 3D, Speech). For the specific questions a Likert Scale from 0 (strongly disagree) to 10 (strongly agree) was used for the subjective evaluation.

6. Results

This section presents measured results aggregated from the developed application and questionnaire. If no unit is specified, the values refer to the Likert Scale as described in Section 5.6. The tables present the mean values while the standard deviation is listed in round brackets. We carried out a non-parametric Friedman test on all results except task times. Here we conducted an ANOVA test with a Tukey Post-Hoc Test. For the non-parametric values, a Post-Hoc Friedman test using Fisher's least significant difference was carried out in addition. Statistical significance is assumed for p-values ≤ 0.05. These are marked with an asterisk. Groups marked with the same letter in square brackets are not significantly different.

6.1. Usability

In order to evaluate key usability factors, we queried perceived complexity and uniformity of the respective interfaces via questionnaire. In addition, the participants were asked to evaluate the overstrain induced by the number of input options, as well as the quality of feedback. Comparing the different interfaces, we see a trend regarding the mean values in favor of the SI (mean overall rating: 8.6)—this compares to 8.0 and 8.1 for 2D and 3D interfaces respectively (see Table 3). Performing a Friedman test on the different aspects of usability, uncomplicated handling as well as the aspect of an uncluttered interface does show a statistically significant result (see Table 3) in favor of the SI. Given the fact that Speech results in a significant difference for two factors and ranks well for the other aspects as well, Speech might be a good choice in regard to usability factors.

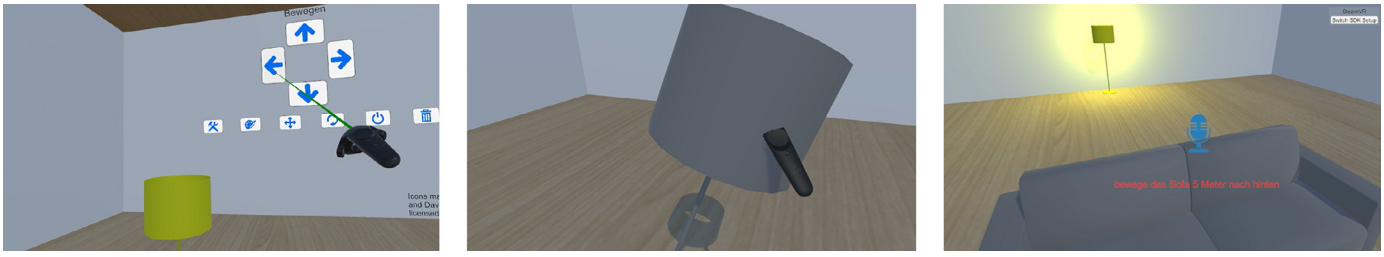

Fig. 2. Screenshots of the used input methods: 2D interface (left): moving an object via icon control. 3D interface (middle): moving an object via direct touch control (right): moving an object via speech command (The red text shows what sentence / word has been understood by the Speech interface. Translated from German it means: "Move the sofa 5 meters backwards"). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
Comparison of usability factors (questionnaire responses).

	2D	3D	Speech
Uncomplicated handling *	7.5 (1.9) [b]	7.6 (1.8) [b]	**8.7 (1.6) [a]**
Interface doesn't feel cluttered *	8.4 (1.6) [b]	8.8 (1.6) [ab]	**9.5 (0.7) [a]**
Input doesn't overextend	8.8 (1.4)	7.8 (2.7)	8.6 (1.9)
Consistent handling	8.5 (1.7)	8.9 (1.4)	9.0 (1.7)
Expectable functions	8.0 (1.9)	7.8 (1.7)	8.2 (1.5)
Satisfactory feedback	7.0 (2.7)	8.2 (1.9)	8.3 (2.1)

Table 4
Comparison of the individual perceived ease of learning (questionnaire responses).

	2D	3D	Speech
Easy to learn *	**8.7 (1.8) [a]**	7.8 (2.2) [b]	**9.1 (1.5) [a]**
Transferable to new tasks	8.8 (1.4)	8.4 (2.1)	8.8 (1.6)

6.2. Ease of learning

Concerning learnability, Table 4 shows comparable high values for the 2D and Speech interfaces, while the 3D input mode was considered much harder to learn. We observe a significant difference between the interface used and the ease of learning. Transferability (e.g. can the recently acquired skills easily be applied to other tasks) was ranked almost equally among all interfaces.

6.3. Rating of manipulation modes

Since the main interaction task in this study was built around the scenario of a furnishing application, the manipulation of objects in the virtual environment is of major importance (see Fig. 2 for the different movement options). Hence the participants were asked to evaluate the position and rotation of objects in terms of individuality[18] and precision for each interface. They were also asked to rate the input modes' capabilities regarding overview and comprehension when manipulating objects using predefined lists. Selecting items from a list is one of the key features used for the 2D and the 3D environment and was necessary for all tasks marked with "MOD" (see Section 5.4). Since these options have to be translated for non-expert users to be interpreted in a questionnaire, the mapping cannot be done one to one. So, the results for selection and list-usage preferences are found in Table 5 "selection and manipulation". Here questions like "the user interface offers a good overview for all selection possibilities" and "lists offer comprehensible symbols, terms and abbreviations" were asked to refer to lists and selection possibilities (see Fig. 3). For the free text input, participants were asked to rate the criteria efficiency and simplicity. The results—listed in Table 5—show an advantage of the 2D interface when a clear and comprehensible presentation of the choices is demanded. The 3D and Speech interfaces showed a deficit in overview, but the 3D interface was fairly better understood through the realistic interaction with the objects. The SI provided, by missing graphic representation, the worst overview and the worst understanding about the choices. The SI also performs poorly when position and rotation of objects was demanded: the participants criticized the lack of precision and individuality when compared to the 3D input mode and especially when the 2D interface is put into consideration.

Table 5
Comprehensive rating of the respective interaction/manipulation modes. Average questionnaire responses using a Likert Scale from 0 (=strongly disagree) to 10 (strongly agree).

	2D	3D	Speech
Selection & manipulation			
Overview *	**8.9 (1.1) [a]**	7.5 (2.2) [b]	7.4 (2.4) [b]
Comprehension *	**8.3 (1.6) [a]**	**8.4 (1.5) [a]**	7.3 (2.3) [b]
Position & rotation			
Individuality	6.6 (2.3)	6.6 (2.7)	5.2 (3.7)
Precision	7.2 (1.9	6.6 (2.6)	6.2 (3.6)
Text input			
Efficiency *	6.1 (2.7) [b]	–	**8.6 (2.2) [a]**
Simplicity *	6.8 (2.8) [b]	–	**9.2 (1.2) [a]**

Table 6
Measured average time the participants needed to complete a task for the respective interface. The Speech‡ column refers to the mean times the participants would have needed if the Speech interface would have performed without errors.

	2D	3D	Speech	*Speech‡*
Overall *	**11.9 s** [a]	22.4 s [b]	16.9 s [c]	*10.9 s*
previous Speech interface experience				
No *	**12.0 s** [a]	23.5 s [b]	18.0 s [c]	*11.1 s*
Yes *	**11.7 s** [b]	18.8 s [a]	13.1 s [b]	*10.3 s*
previous Virtual Reality experience				
No *	**11.7 s** [b]	22.5 s [a]	16.5 s [b]	*9.7 s*
< 1h *	**11.1 s** [a]	22.0 s [b]	17.7 s [c]	*11.7 s*
Yes	**12.7 s**	20.7 s	15.9 s	*11.1 s*

6.4. Measured time

During the task the participants' required time for each of the three interfaces was measured. Table 6 shows the average time participants needed to fulfill the different tasks.

A one-way ANOVA and a Tukey Post-Hoc Test were conducted to compare the effect of the three interface types in regards to time needed. For all conditions, except previous VR experience, a significant difference between the interfaces was observed.[19]

[18] Individuality in this context describes the amount of freedom the user had in choosing the exact position/orientation of the objects (could be influenced e.g. by DOF, constraints and snapping-techniques).

[19] The column Speech‡ is not statistically comparable with the other columns.

Fig. 3. Screenshots of the different selection modes: 2D interface (left): flat overview of the possible selections. 3D interface (middle): objects distributed circularly (right): no selection possibilities available. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Looking at the overall mean times a clear chronological order can be given without considering error-rates: 1. 2D, 2. Speech and 3. 3D. Without accounting common speech related errors, the SI would have outperformed the other interfaces. Table 6 also indicates a significant correlation between the interface used and the respective time requirements. Furthermore, there was a noticeable statistical correlation between prior experience with speech assistants and measured mean times.

When participants are separated according to their previous experience with SIs we observe two effects: prior experience with speech input has a drastic impact on the time to solve the tasks on 3D and Speech interfaces, yet 2D time requirements remained unchanged. While it is obvious that previous experience in Speech interfaces helps in that particular use case, the drastic reduction in 3D interfaces' mean times comes as a surprise. However, the rather obvious assumption that this effect is due to a higher technical affinity of this group could not be confirmed (Pearson's $r = 0.1$).

It should also be noted, that the times measured include system processing time which—in case of 2D and 3D interfaces—can be neglected. On average the SI needed two turns to successfully complete one task and the processing of one request needed 1.1 s (standard deviation = 0.4 s). Thus, on average the processing required 2.2 s. This states that the SI, in an ideal world, without errors and processing time, would take only 8.7 s (avg.) and thus, would by far outperform any other interface. Having in mind that SIs are becoming more popular and sophisticated as well, error rates will drop and at some point, be near to error free.

6.5. Aspects of immersion and presence

Immersion and presence are difficult parameters to define and therefore to determine reliably. According to Slater and Wilbur [38], presence is seen as a conscious state or sense of actually being inside the virtual world. This illusion of presence is very subjective and can't be measured directly. In order to avoid further expanding the already very substantial questionnaire even further, we decided not to use pre-designed presence questionnaires (e.g. Witmer and Singer or iGroup Presence Questionnaire[20]), but rather to query only some narrowed aspects of presence like usability, intuitive operation and effects on the experience in general.

Intuitive usability is one of the key elements of interfaces [39]. For this reason, we were interested how users rated the three presented concepts in terms of naturalness and intuitive access. Therefore, the participants were questioned to evaluate each interface respectively. When referring to Table 8 it can be stated that Speech and 3D was rated significantly more natural and intuitive than 2D. Since effectiveness is not the only goal to achieve, we also asked the users whether the input methods presented had a positive or negative effect on their perceived overall experience in VR [4]. With this question we wanted to determine whether there are any differences between the interfaces, which on the one hand react very efficiently to user input, but on the other hand contribute to a break in experience—one of the most relevant key factors for a broad range of applications in VR. It can be stated that 3D as well as Speech excel in this dimension, while the tested 2D interface was rated less convincing. Due to the arbitrary nature of text inputs [34] at first, they seem inappropriate for immersive environments [40]. Whether text input generally leads to a loss of immersion is currently still under discussion [34,41]. Text input in VR therefore seems still quite inconclusive regarding its consequences on experience and immersion hence we questioned this aspect directly: The participants were asked to rate whether the text input method lowered their perceived VR experience. In this regard we can state that there is also a significant difference between Speech and 2D (see Table 8), where Speech was rated the best in maintaining a constant VR experience.

Table 7
Overview of the steps it took all participants to fulfill a creation / colorizing, positioning, switching something On/Off/deleting - task. (see Section 5.4 for respective tasks). The standard deviation is listed in brackets behind the respective mean values. Note that Sub-tasks like selection is not considered here.

Category (Task Number)	Median	Average (StDev)
Create / Colorize (1,3,5,10,12)	1	1.46 (0.92)
Position / Rotate (2,4,6,9,15,17)	1	2.91 (4.22)
Turn On / Off; Delete (7,8,14,16)	1	1.41 (0.43)

6.6. Overall experience

At last, the overall experience with the interfaces was evaluated by the participants. Table 9 shows, that the SI induced the test participants' greatest overall satisfaction. However, the observed differences are not significant according to the conducted Friedman test. The 3D interface was also rated fairly well, while the 2D input mode was rated the worst. The high ratings assigned to all interfaces demonstrate that all interfaces were satisfactory and capable of fulfilling the requirements of the interface. What also can be stated is the significant correlation between the perceived fun and the interface used (see Table 9): Here the 3D input mode achieved the best rating, although the participants criticized that this particular interface was the hardest to learn (see Table 4) and was prone to overextension (see Table 3).

6.7. Speech interface - steps

Even though it took participants only 1 Step in Median to fulfill a task when using the SI (see Table 7), it still took participants almost 11 seconds to fulfill a task on average using the SI (see Table 6). Given the fact that SIs theoretically offer an almost endless amount of possible inputs, it could mean that the

[20] See: http://www.igroup.org/pq/ipq/index.php (accessed 2019-04-03).

Table 8
Comparison of different parameters linked to perceived presence.

	2D	3D	Speech
Natural & intuitive input *	6.9 (2.4) [b]	**7.9 (1.7) [a]**	**8.1 (1.4) [a]**
Interface does not lower VR experience *	7.9 (2.4) [b]	**9.3 (0.9) [a]**	**9.1 (1.4) [a]**
Text input does not lower VR experience *	7.3 (3.0) [b]	–	**9.0 (1.7) [a]**

Table 9
Comparison of different parameters.

	2D	3D	Speech
Fun *	7.3 (2.8) [b]	**8.9 (2.1) [a]**	8.1 (2.0) [b]
Professional	7.2 (2.2)	7.2 (2.4)	7.7 (2.2)
Efficiency	7.5 (2.0)	7.2 (2.0)	8.1 (2.0)
Overall experience	7.4 (2.1)	7.6 (1.9)	8.3 (1.5)

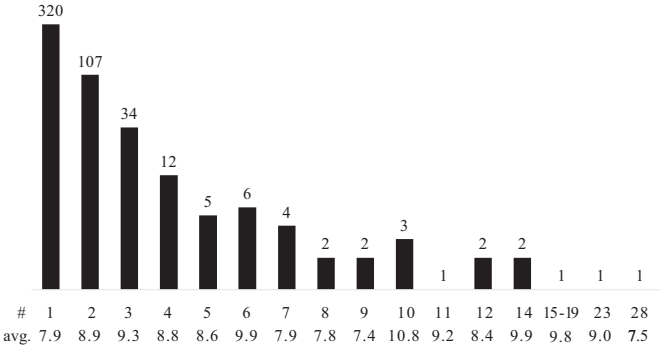

Fig. 4. Overview of how often it took users to fulfill a task with the respective numbers of steps for the Speech interface (320 times it took the users 1 speech command). The avg. row shows the average duration it took per step (Standard Deviation: 1.05). A logarithmic scale is used to better display the huge difference in results.

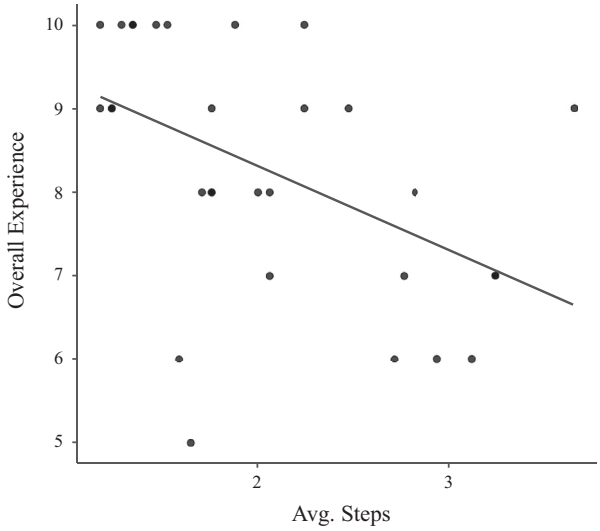

Fig. 5. Correlation Plot between overall experience and the average number of steps it took a participant to fulfill a task. ($p \leq 0.05$; Pearsons r −0.479).

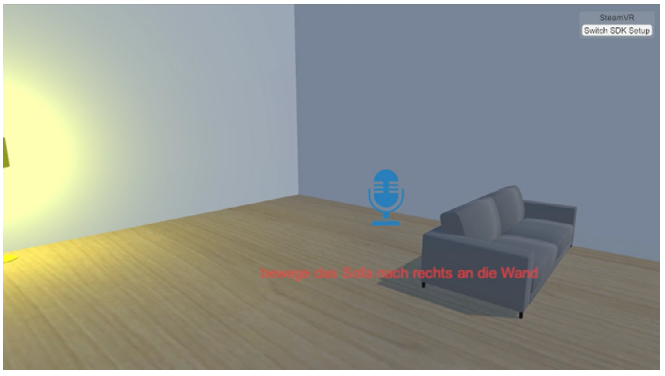

Fig. 6. Different way how participants used the Speech interface. The red text is written in German and means “move the sofa to the wall at my right.” (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

resulting cognitive load for constructing a command is quite high. Fig. 2 (right image) for example shows that users sometimes simulated the 2D interface moving objects bit by bit instead of making use of the given constraints such as the walls as done in Fig. 6. This is also shown in Table 7: Although it was not necessary, some participants tended to position elements stepwise instead of putting the current context into consideration. Also, the SI does not explicitly offer an *undo* option like the arrow in the opposite direction for the 2D user interface or the naturalness of the movement when using a 3D interface (see Fig. 2). This biases the steps towards a higher number for fulfilling a task as seen in Fig. 4 tremendously, which directly reflects into the perceived overall experience as shown as correlation plot in Fig. 5. We observed a statistical relevant correlation ($p \leq 0.05$; Pearsons r: −0.479) between the needed number of steps and the perceived overall experience, saying that the experience suffers from an additional number of steps that are needed to fulfill the task. This interestingly is not reflected in the SI error rate in regard to the overall experience. Here we did not observe a statistical correlation.

7. Error analysis

Each type of interface is more or less error prone. Errors can be caused by poor interface design (ambiguous icon or test, small interaction elements, hidden information) or technical restriction (low precision, classification errors, parsing errors). The findings presented here are based on quantitative analysis as well as observed errors by the test supervisors.

7.1. 2D interface

For the 2D interface there has been observed one deficiency for the user interface: since the selected piece of furniture did not properly indicate that it was selected, for instance by a glowing outline, it was observed that participants sometimes (∼3%) had problems accessing the objects even after learning it in the preparation session.

7.2. 3D interface

One of the main problems the test’s supervisor could observe for the 3D interface was that some users (∼5%) had problems using the teleport function while holding an object in the other hand. Since—in case of acceleration etc.—we use regular physics in VR, a collision between two objects will cause an acceleration

of the respective object in hand and therefore will move into an unexpected place. The scenario where this happened was, when participants were asked to place furniture right next to a wall. For instance, they teleported themselves too close to a wall. Also, some participants accidentally deleted objects while teleporting and holding the hammer in their hand.[21] During the testing process of the 3D user interface, most participants also showed difficulties using the Grip-Button of the HTC Vive controller, which could have contributed to lower user satisfaction. Due to its location on the side of the controller, it had to be activated by pressing the lower palm of the hand against it. This was probably intended to respond to a strong grip of the user's hand, but instead resulted in participants needing additional mental and physical effort finding the button and building enough pressure to press it. This could be clearly observed during testing despite the Grip-Button only being used to switch objects on and off, which is not demanded as often as the other used buttons and only required the button to be pressed once and not to be held down.

7.3. Speech interface

In contrast to 2D and 3D, the SI showed a significant higher number of malfunctions caused by usage errors, parsing errors of the dialog system and by speech recognition errors. On average, errors occurred in about 17% of the interactions, while the error rate dropped to only 7% for participants with experience in SIs. Inexperienced participants, on the other hand, had an error rate of about 20%. The occurrence of errors while using the SI also had an impact on the average rating the participants were asked to give in the questionnaire. Weiss et al. investigate further, in how far the *expectable functions* and *overall experience* vs. *speech error* correlate [3].

Another problem observed several times (about 5–10% of the positioning tasks) was that when a participant went through an error in regard to the SI, for example the interface misinterpreted a rather complex input (i.e. move the table to the left of the lamp), they lowered the complexity of their input (i.e. move the table one meter to the left), whereas the interface only got one word wrong. For some participants, this even affected the subsequent tasks, leading them to break up the task into a sequence of operations, as is required with the other tested interfaces. This could mean that people (perhaps especially those with lower previous experience in SIs) misunderstood the way a SI works and tried to accommodate by adapting their speech pattern to what they thought was easier for computers to comprehend—Table 7 shows this effect for positioning tasks. To fulfill these tasks via the 2D interface you had to press a direction button several times (see Fig. 2). We deduced two possible problems based on our observations: Firstly, users tend to misinterpret the capabilities of SIs and secondly, they might be afraid to hastily speak out a speech command, because of a missing *undo* function. While the participants were changing positions of the virtual objects through the SI, another interesting pattern of errors or misinterpretations could be observed. Participants often had different—or rather completely contrary—interpretations of specific directional terms and commands. While terms used for moving an object to the left or the right were easily understood to be taken in the context of the user's left and right respectively, front and back were used inversely by some participants. This could be explained by the nature of speech in this specific situation (object positioning inside a room): Since most furniture have a distinctive front- and backside and mostly face away from the walls towards the human, terms like forward and backward can be interpreted both from the point of view of the user or of the object itself, resulting in completely reversed meanings (e.g. pushing the object "forward towards the wall" by your point of view, or pushing it "backwards against the wall" by the object's perspective). The 2D interface had clear directional clues through arrows and the 3D interface just relied on the position of the user and his hand, leaving the SI isolated with this issue. The viewing perspective when uttering the commands seemed to vary depending on the user but also on the object's proximity to a wall—the closer it got, the more likely a user was to invert the directional call. These two factors can't accurately be compensated by the SI, which means that the cognitive workload of figuring out with directional term to use is ultimately put on the user.

[21] The hammer is used for creating and deleting objects.

Table 10
Preference of the interface according to different parameters. + positive, ○ neutral, − negative. Note: Recommendation in regard to previous publication has changed since we use Friedman-test instead of ANOVA here. Value in brackets show the old recommendations.

	2D	3D	Speech
Position & rotation	+	○	−
Selection & manipulation	+	○	−
Text input	−	−	+
Ease of learning	+ (○)	○	+
Fun	−	+	○
Uncomplicated handling	○	○	+
Interface experience	−	+	+
Natural & intuitive input	− (○)	+	+
Time	+	−	○/+[a]

[a] The Speech interface would bear significantly better average times if the amount of errors could be reduced. See Table 6 for details.

8. Conclusion

We evaluated a 2D, a 3D and a SI to complete predefined furnishing tasks in an immersive VR environment to get insights about the different interaction paradigms. Even though we are well aware that our findings cannot answer all questions and are biased towards our specific implementations and design decisions, we hope, that they can serve as an indicator for those looking for support when building an interface in VR and as foundation for future research in this area. We believe that variations within a 2D, 3D or speech interface cause a smaller change in the discussed categories than the decision between the interfaces. Our results show that the SI performs significantly better than the 2D or the 3D interface or both in 8 categories. 2D and 3D only are advantageous in 4 categories. In order to summarize our findings, we derived a brief overview with guidelines from the results presented in this paper. Table 10 resumes which interface could be advantageous for a particular interaction task or design goal. Although, it has to be stated that in our conducted test a majority of participants were quite satisfied with all presented input technologies in general. Therefore, the table should be understood in the sense that the recommendations are always expressed in relation to the lowest measured value. Even if this value on its own could be categorized as rather satisfactory if regarded independently.

Furthermore—especially in case of the Speech and 3D interface—the favored interface depends not only on the interaction task, but also on the user's previous experience and prejudice. Our research points out, that there is a statistically significant correlation between the users' experience in SIs in regard to their individual performance. Also, it took participants with experience in using a VR System on average 2.7 s (12%) less time to perform a task (see Table 6). Another important finding in regard to SIs is, that it is crucial to minimize the steps it takes a user to fulfill a task. The perceived overall experience of

the application suffers when the user cannot solve a task with a small number of steps (see Fig. 5). After all, the designer or developer has to decide or investigate what focus is important for the application and user group; for instances:

- if fun and perceived presence is the most important, a 3D interface is recommended,
- if there are a lot of objects to be manipulated fast and accurate a 2D interface might be the best choice, and
- if the interface should be easy to learn and a lot of text needs to be entered, speech might be the best choice.

While our findings hopefully transfer to many applications, designers/developers are encouraged to evaluate their solutions and to see if our findings hold for their application.

What still needs to be evaluated further in this context is the variation of our findings within a given interaction paradigm (e.g how Go-Go compares to Simple Hand) and if switching between the different modalities in a too frequent manner (e.g. changing the interface from task to task) would raise the participants cognitive load in such a way that sometimes it might be better to stick with one of the interfaces. Also, in terms of production costs / price-performance ratio, it might be more profitable to stick with only one modality. Other factors that might be of interest when choosing the fitting interface are parameters like inspiration: In how far might a better overview of the possible options contribute to a higher degree of inspiration? For example, when implementing a VR furnishing application, people possibly would like to choose between pre-given objects that come with a 2D interface. Another question that still needs to be answered is, to what extend these findings can be transferred into other virtual environment like CAVES or Mixed Reality.

After all, looking through the results indicates a clear tendency towards implementing SIs in immersive VR applications. Not only is it the fastest interface,[22] it also was perceived the best solution in 8 cases. Still it seems as for some tasks a combination of the interfaces might be a better choice. For example, a supervisor - apprentice situation might be a fitting situation that illustrates the challenge. Sometimes it is just faster to do something on your own (2D & 3D interface) than to ask someone for help (SI). This leads to another interesting discussion: Would it be easier for users to create their mental model for the SI if the interface was represented by a virtual avatar? For example, having some kind of a robot assistant might frame the possible commands he can complete based on previous experiences (It might be more obvious for users to ask the robot to bring them something than just saying “create a sofa” into to void.). We observed that the results of mimicking natural interaction methods (Here: Speech) with current state of the art technology already can be seen as a very promising approach and that it quite possibly will become a (future) standard. Especially in combination with other relevant developments in this research-area, such as full-body or emotion tracking, continued research on how to successfully implement those technologies is of major importance.

Declaration of Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

[22] After error correction.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.cag.2019.06.003.

References

[1] Poupyrev I, Billinghurst M, Weghorst S, Ichikawa T. The go-go interaction technique: non-linear mapping for direct manipulation in vr. In: Proceedings of the 9th annual ACM symposium on user interface software and technology. ACM; 1996. p. 79–80.

[2] Microsoft. Use speech in windows mixed reality. https://support.microsoft.com/en-us/help/4041322/windows-10-speech-in-windows-mixed-reality; 2018. [Online; accessed 03. April 2018].

[3] Weiß Y, Hepperle D, Sieß A, Wölfel M. What user interface to use for virtual reality? 2D, 3D or speech-a user study. In: International conference on cyberworlds. IEEE; 2018. p. 50–7.

[4] Jerald J. The VR book: human-centered design for virtual reality. Morgan & Claypool; 2015.

[5] LaViola Jr JJ, Kruijff E, McMahan RP, Bowman D, Poupyrev IP. 3D user interfaces: theory and practice. Addison-Wesley Professional; 2017.

[6] Fuchs P. The challenges and risks of democratization of VR-AR. Virtual Reality Augmented Real 2018:289–301.

[7] Mendes D, Caputo F, Giachetti A, Ferreira A, Jorge J. A survey on 3D virtual object manipulation: from the desktop to immersive virtual environments. In: Computer graphics forum. Wiley Online Library; 2018.

[8] Poupyrev I, Ichikawa T, Weghorst S, Billinghurst M. Egocentric object manipulation in virtual environments: empirical evaluation of interaction techniques. In: Computer Graphics Forum, 17. Wiley Online Library; 1998. p. 41–52.

[9] Argelaguet F, Andujar C. A survey of 3D object selection techniques for virtual environments. Comput Graph 2013;37(3):121–36.

[10] Jankowski J, Hachet M. Advances in interaction with 3D environments. In: Computer graphics forum, 1,34. Wiley Online Library; 2015. p. 152–90.

[11] Kaur A, Yammiyavar PG. A comparative study of 2D and 3D mobile keypad user interaction preferences in virtual reality graphic user interfaces. In: Proceedings of the 23rd ACM symposium on virtual reality software and technology. New York, NY, USA: ACM; 2017. p. 42:1–42:2. ISBN 978-1-4503-5548-3. doi:10.1145/3139131.3141221.

[12] Cohen P, McGee D, Oviatt S, Wu L, Clow J, King R, et al. Multimodal interaction for 2D and 3D environments [virtual reality]. IEEE Computer Graphics and Applications 1999;19(4):10–13.

[13] Cohen PR, Dalrymple M, Moran DB, Pereira F, Sullivan JW. Synergistic use of direct manipulation and natural language. In: ACM SIGCHI bulletin, SI, 20. ACM; 1989. p. 227–33.

[14] McGlashan S, Axling T. A speech interface to virtual environments. In: Proc., international workshop on speech and computers; 1996.

[15] Wauchope K, Everett S, Perzanowski D, Marsh E. Natural language in four spatial interfaces. In: Proceedings of the fifth conference on applied natural language processing. Association for Computational Linguistics; 1997. p. 8–11.

[16] Cheyer A, Julia L. Spoken language and multimodal applications for electronic realties. Springer-Verlag London Ltd, Virtual Reality 1999;3:1–15.

[17] Dorozhkin DV, Vance JM. Implementing speech recognition in virtual reality. In: ASME 2002 international design engineering technical conferences and computers and information in engineering conference. American Society of Mechanical Engineers; 2002. p. 61–5.

[18] He Z, Lv C, Peng D, Yu D. A speech recognition-based interaction approach applying to immersive virtual maintenance simulation. In: 2017 second international conference on reliability systems engineering (ICRSE); 2017. p. 1–5.

[19] Lee M. Multimodal speech-gesture interaction with 3D objects in augmented reality environments; 2010. PhD Dissertation University of Canterbury Department of Computer Science and Software Engineering.

[20] Kyritsis M, Gulliver SR, Morar S, Stevens R. Issues and benefits of using 3D interfaces: visual and verbal tasks. In: Proceedings of the fifth international conference on management of emergent digital EcoSystems, MEDES ’13. New York, NY, USA: ACM; 2013. p. 241–5.

[21] Oviatt S, Cohen P, Wu L, Duncan L, Suhm B, Bers J, et al. Designing the user interface for multimodal speech and pen-based gesture applications: State-of-the-art systems and future research directions. Hum–Comput Interact 2000;15(4):263–322.

[22] Welch RB, Warren DH. Immediate perceptual response to intersensory discrepancy. Psychological Bulletin 1980;88(3):638–67.

[23] Mcglashan S. Speech interfaces to virtual reality. In: Proceedings of 2nd international workshop on military applications of synthetic environments and virtual reality; 1995.

[24] Irawati S, Green S, Billinghurst M, Duenser A, Ko H. An evaluation of an augmented reality multimodal interface using speech and paddle gestures. In: Advances in Artificial Reality and Tele-Existence. Springer Berlin Heidelberg; 2006. p. 272–83.

[25] VR Gluv Corporation. Feeling is believing. 2018. https://vrgluv.com/; [Online; accessed 8. Nov. 2018].

[26] Leap Motion Corporation. Reach into virtual reality with your bare hands. 2018. https://www.leapmotion.com/; [Online; accessed 8. Nov. 2018].

[27] Hepperle D, Wölfel M. Do you feel what you see? Multimodal perception in virtual reality. In: Proceedings of the 23rd ACM symposium on virtual

reality software and technology, VRST '17. New York, NY, USA: ACM; 2017. p. 56:1–56:2.
[28] Siu AF, Gonzalez EJ, Yuan S, Ginsberg JB, Follmer S. shapeshift: 2d spatial manipulation and self-actuation of tabletop shape displays for tangible and haptic interaction. In: Proceedings of the 2018 CHI conference on human factors in computing systems, CHI '18. New York, NY, USA: ACM; 2018. p. 291:1–291:13. ISBN 978-1-4503-5620-6.
[29] Simeone A, Velloso E, Gellersen H. Substitutional reality: Using the physical environment to design virtual reality experiences. In: Proceedings of the 33rd annual ACM conference on human factors in computing systems, CHI '15. New York, NY, USA: ACM; 2015. p. 3307–16.
[30] Wölfel M. Kinetic Space - 3D Gestenerkennung für Dich und Mich. Konturen 2012;32.
[31] Microsoft. Project gesture. 2018. https://labs.cognitive.microsoft.com/en-us/project-gesture; [Online; accessed 03. April 2018].
[32] Ruch WF, Platt T, Hofmann J, Niewiadomski R, Urbain J, Mancini M, et al. Gelotophobia and the challenges of implementing laughter into virtual agents interactions. Front Hum Neurosci 2014;8.
[33] Optitrack Corp.. Low latency, wide area vr tracking for caves and hmds. 2018. [Online; accessed 8. Nov. 2018]; https://optitrack.com/applications/.
[34] Sherman W. Understanding virtual reality: interface, application, and design. San Francisco, CA: Morgan Kaufmann; 2003. ISBN 978-1-55860-353-0.
[35] Corbett E, Weber A. What can i say? addressing user experience challenges of a mobile voice user interface for accessibility. In: Proceedings of the 18th international conference on human-computer interaction with mobile devices and services, MobileHCI '16. New York, NY, USA: ACM; 2016. p. 72–82. ISBN 978-1-4503-4408-1.
[36] Valve. room scale vr survey. 2018. https://steamcommunity.com/app/250820/discussions/0/483366528921496674/; [Online; accessed 03. April 2018].
[37] Lang B. Valve reveals 2017s top selling vr games on steam. 2018. https://www.roadtovr.com/valve-reveals-steams-top-selling-vr-games-in-2017/; [Online; accessed 03. April 2018].
[38] Slater M, Wilbur S. A framework for immersive virtual environments five: speculations on the role of presence in virtual environments. Presence 1997;6(6):603–16.
[39] Bærentsen KB. Intuitive user interfaces. Scand J Inf Syst 2000;12(1):4.
[40] Bowman DA, Rhoton CJ, Pinho MS. Text input techniques for immersive virtual environments: an empirical comparison. Proc Hum Factors ErgonomSoc Ann Meeting 2002;46(26):2154–8.
[41] Speicher M, Feit AM, Ziegler P, Krüger A. Selection-based text entry in virtual reality. In: Proceedings of the 2018 CHI conference on human factors in computing systems. ACM; 2018. p. 647.