AllTogether: Effect of Avatars in Mixed-Modality Conferencing Environments

Payod Panda
payod.panda@microsoft.com
Microsoft Research
Cambridge, UK

Molly Jane Nicholas
University of California Berkeley
Berkeley, California, USA

Mar Gonzalez-Franco
Kori Inkpen
Eyal Ofek
Ross Cutler
Ken Hinckley
Jaron Lanier
Microsoft Research
Redmond, Washington, USA

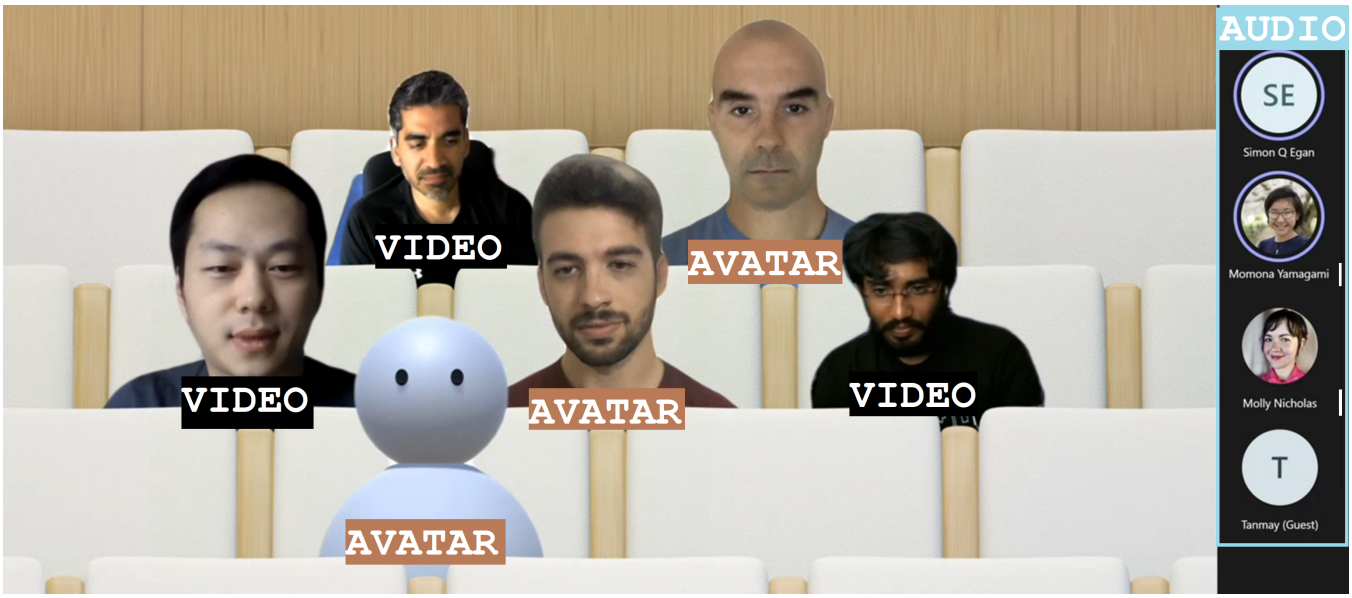

Figure 1: Avatars, video participants, and audio participants talking in a mixed-modality conferencing environment.

ABSTRACT

Visual representation in most video conferencing systems is a binary option between camera on and off. With such systems, voice-only participants might feel left-out, particularly in configurations that situate video participants in a shared virtual environment (e.g., Figure 1). Motivated by the results of a large-scale (n=1140) preliminary study indicating a perceived need for avatar-supported meeting attendance, we developed AllTogether, a system that provides voice-only participants with the option to be represented by an avatar in a call with other video participants. Past research has compared the effect of avatar representations with video and audio, but focused on single-modality calls (i.e. all participants represented as either avatars or video or audio only). We studied the use of our system across three conferencing sessions with 9 participants being represented by a mixture of avatar, video, or voice-only (no visual) representations to better understand users' perceptions and feelings of co-presence when being represented through these modalities. We found that the visual representation of self and others as well as body motion agency affected participants' feelings of co-presence and the level to which participants felt others were present in the video call respectively. Our results highlight the implications of visual realism and agency of control on users' perception of self and others. We propose avatars as a way to expand the binary choice of camera on and off to a spectrum of choices for the user, offer design implications for integrating avatars into video conferencing systems, and update the literature on users' avatar preferences.

CHIWORK '22, June 8–9, 2022, Durham, NH, USA

ACM ISBN 978-1-4503-9655-4/22/06.
https://doi.org/10.1145/3533406.3539658

CCS CONCEPTS

• **Human-centered computing** → **Collaborative and social computing**; **Empirical studies in HCI**.

KEYWORDS

avatar, video conferencing, together mode, copresence

ACM Reference Format:
Payod Panda, Molly Jane Nicholas, Mar Gonzalez-Franco, Kori Inkpen, Eyal Ofek, Ross Cutler, Ken Hinckley, and Jaron Lanier. 2022. AllTogether: Effect of Avatars in Mixed-Modality Conferencing Environments. In *2022 Symposium on Human-Computer Interaction for Work (CHIWORK '22), June 8–9, 2022, Durham, NH, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3533406.3539658

1 INTRODUCTION

Working from home has recently become commonplace, and many people have transitioned online for work. Video conferencing is a cornerstone of this work paradigm, resulting in a sharp increase in the adoption of video conferencing tools for day-to-day activities–over 300% for some products [1, 24].

New features have been introduced to help overcome challenges in current videoconferencing systems [4, 11]. For example, several solutions implement background subtraction techniques that allows participants to blur the background or replace it with a virtual background to maintain privacy. Some solutions like Together mode in Microsoft Teams utilize background subtraction and situate each participant in a shared virtual space (see Figure 2a), which has the potential to increase the sense of co-presence between peers.

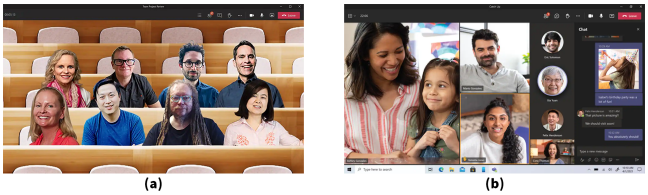

Figure 2: (a) Together mode, mimicking a physical social space, and (b) Standard grid view in most video conferencing applications today. (copyright Microsoft)

Enhanced presence systems like Together mode work well for users who have their video turned on. However, users who do not have video (either because of choice or necessity) cannot take advantage of the enhanced co-presence. Systems that only provide a binary switch between camera on and camera off constrain the choices available to such users for visual representation. This limits their ability to communicate non-verbal information via body language, such as head movements for agreement or disagreement. When users don't have access to non-verbal communication channels they can feel excluded from the social group [7].

We conducted a formative survey to understand why users might not want to use video cameras in conferencing sessions. We discovered factors like wanting to multitask, not wanting to show how they looked, and network bandwidth considerations (see Section 3.1). Additionally, we found that these users might use a realistic or non-realistic avatar in place of transmitting live video.

Based upon these findings, we designed AllTogether. Instead of relying on the binary camera on/off options, AllTogether offers a spectrum of choices from no visual fidelity (camera off) to full visual fidelity (camera on). The intermediate levels are offered by avatars of varying visual (cartoonish to realistic) and expressive fidelity (synthetic movement vs. tracked head movement). This gives the participant a safe way to have visual presence in the absence of webcam feed, while allowing them to choose the level of fidelity that they are comfortable with, and that is appropriate for the meeting.

We studied the use of AllTogether with 9 participants over three sessions in a *mixed modality* video conference–a conference call where all participants, irrespective of visual representation, appeared next to each other in a shared virtual space. We examined the effects of (1) being represented by and (2) interacting with avatars as well as behavioral responses of the different participants. Prior studies [20, 21] have compared various forms of representations (videos, avatars, audio), but focused on single-channel comparisons. These studies did not explore mixed configurations where multiple representations were present in the same session. Our research extends this earlier work to explore scenarios where all three representations are present, and explores how this impacts both users' perceptions and avatar preferences. We also wanted to understand how users felt about interacting with avatars, or being represented by an avatar, compared to video or audio-only representations.

We explore the self-reported effects of the avatar's visual, motion, and expressive fidelity on the user's experience. We discuss the participants' experiences from the OBSERVER and ACTOR perspectives (explained in Section 4), and expand upon:

- Presence and participation,
- Privacy and trust, and
- Non-verbal cues and body language.

We also share design implications for the integration of avatars in video conferencing solutions as we move to an increasingly digitized workplace, including the importance of agency and customization, thinking about forms of tracking input and body motion aside from cameras, and using graceful bandwidth degradation.

2 RELATED WORK

2.1 Avatars in Video Conferencing

The use and study of avatars for communication and socialization became popular with the advent of online virtual worlds such as Second Life [25] and World of Warcraft [27]. However, their use in video conferencing scenarios remains limited. Schroeder [36] reflected on this discrepancy and highlighted several fundamental differences in how people interact with video versus avatars.

Bente et al. [6] compared text-chat, audio, video conferencing, and avatars. Their results revealed significant differences between text-chat and the other real-time modalities but did not find differences between audio, video, or avatars. Later, Junuzovic et al. [21] directly compared video and avatars in a work video conferencing scenario. One of the key concerns about using avatars in the workplace is related to their appearance, such as feeling less professional [21], including concerns that their artificial nature could have a negative impact on trust or relatedness [6]. Inkpen and Sedlins's evaluation of avatar appearance for work settings revealed that while some avatars were not well received, in other instances users ranked avatars similarly to webcam photos, such as those that had a more formal, realistic appearance and did not feel "creepy" [20]. While many users preferred video, they still saw the potential for

avatars, particularly in terms of sociability. In our system, we allow the user to choose between different levels of visual fidelity, from abstract and playful to more realistic and serious, and discuss the implications.

Prior studies [20, 21] have compared various forms of representations (videos, avatars, audio), but focused on single-channel comparisons and did not explore mixed-modality configurations where multiple representations were present in the same session. Schroeder also noted that avatar and video offer two distinct choices in distinct environments but did not envision the possibility of combining these two modalities and how that could impact users' interactions. In this paper, we explore experiences in a mixed-modality configuration (see Section 4).

Finally, it is important to recognize that while prior work on avatar video conferencing is informative, circumstances around video conferencing in the workplace were radically different. Video conferencing was an occasional practice in contrast to today's extensive WFH scenarios. Users' acceptance and desire for rich forms of video-communication has likely evolved over time, with more exposure to new technologies for everyday communication. Additionally, the blurring of work and personal lives and requirements to be on video calls from home has changed users' desires for alternative representations to live video.

2.2 Implementing Avatars

Avatars are digital representations of users that may be abstract, cartoonish, or human-like. These avatars may represent the user (such as in a video call), or a character that the user controls (such as in a video game). Depending on the use case, avatars might be created manually by designers and character artists (e.g. for characters in animated movies), or use computer vision or other techniques to generate a digital 2D or 3D avatar in the likeness of a user's visual appearance [48]. Different implementations also allow for different levels of customization of the avatar, such as facial appearance, body type, and clothing and accessories.

Avatars can also be animated to allow users to express emotions and help users self-identify with the avatar [16, 44]. This is usually achieved by rigging a skeletal system [14, 15] which is moved by sensing and tracking the user. This can be achieved with cameras and computer vision, speech sentiment analysis, or animating visemes through audio input [14]. However, avatars don't need to have high fidelity or tracked animations. Even minimal animations such as idling facial animation can help participants self-identify with the avatar [16].

In this work, we compared users' perceptions and preferences of low-fidelity avatars against visually photorealistic avatars generated using an online service. Additionally, we compare rudimentary synthetic movement in avatars to tracked (through computer vision) and triggered (through mouse interactions) avatar movements and expressionism. We identify contexts for which high- or low-visual, motion, and expression fidelity might be appropriate for digital avatars in a conferencing environment.

2.3 Presence and co-presence

Presence is a multi-dimensional construct [23, 37, 38] that includes telepresence or spatial presence, co-presence, and social presence. Spatial presence or telepresence is the feeling of the user "being there" in a virtual environment. It indicates the degree to which a user feels like they are in a new environment where they feel physically immersed. Telepresence can be thought of as the user's sense of being in a mediated space, rather than where their body is physically located [41]. Social presence indicates a medium's capacity for supporting a sense of connection and attachment. Social presence tries to measure the medium's capability of developing interpersonal relationships, typically over an extended period of time [30].

Co-presence is used to describe the feeling of "being together" with other people in an environment [34, 35]. The sense of co-presence is high when a two-way perception between users in a medium exists–i.e., users feel that they can actively perceive others and that others can actively perceive them. Campos-Castillo & Hitlin introduce the term entrainment, and extend the definition of co-presence to mean the degree to which an actor perceives mutual entrainment (i.e., synchronization of attention, emotion, and behavior) with another actor [10]. Fox et al. [12] found that the perceived agency of an avatar (i.e., the feeling that the avatar is controlled by a real human) increased the feeling of co-presence. In contrast, Nowak & Biocca [29] found that agency had no effect, however a higher level of anthropomorphism resulted in reduced sense of co-presence. In this work, we investigate the effect that different avatar forms have on a user's subjective sense of co-presence through varying levels of agency (through motion fidelity) and anthropomorphism (through visual fidelity) in the avatar's design.

These dimensions of presence have been studied widely, which can elicit different aspects of virtual environments and experiences [37]. Researchers have proposed several methods of measuring presence [46], primarily under two general categories of either subjective measures or objective corroborative measures. Subjective measures require a participant's judgement of their psychological state within the environment (for physical presence) or in relation to others (for social and co-presence). These include quantitative measures like presence questionnaires [41, 42, 49], continuous measurement [19], psychophysical measures [43, 45], as well as qualitative measures [13, 26, 32, 33] and subjective corroborative measures [17, 22]. On the other hand, objective measures utilize automatically or subconsciously generated user responses that might be correlated with measurable properties of the medium [18, 28, 31, 40]. Scholars agree that objective measures of presence are provide less depth in understanding the experience of the user, and measuring presence by asking users to describe their experience subjectively is typically more useful [23, 39]. In this paper, we try to understand our participants' subjective feelings of co-presence qualitatively.

3 METHODS

3.1 Formative Survey

After the shift to remote work due to the COVID-19 pandemic in March 2020, we conducted a 3-month survey with users of the Microsoft Teams software in order to understand video conferencing habits from typical users. This survey was presented at random to all users of the software within the USA, and asked respondents to provide data from their last video meeting. Responses to this

survey from 1140 users motivated further investigation of avatars as a viable form of visual representation in video calls.

Of the respondents, only 50% shared their video feed during their last meeting. When asked to reveal why they did not share their video, there were five common responses that accounted for 73% of the responses (users could only select one reason for this question):

1. "Other participant(s) were not sharing video" (24%);
2. "I didn't feel that sharing my video gave much value" (14%);
3. "I wanted to multitask during the meeting" (13%);
4. "I wanted to hide my visual appearance" (11%); and
5. "Not enough network bandwidth or poor network quality" (11%).

When asked what features would increase their use of video, the top five responses were (participants were allowed to select multiple choices):

1. "Video that uses much less bandwidth at the same video quality" (44%);
2. "Only share my video when I'm talking or unmuted" (39%);
3. "Low-light video enhancement" (23%);
4. "Eye gaze correction (preserve eye contact)" (21%); and
5. "Video avatar of you (either realistic or non-realistic)" (18%).

It is interesting to note that the top four responses are either quality or performance issues. This data motivated our study, indicating perceived need for avatar-supported meeting attendance.

3.2 General study design

In this study, we answer the following research questions (RQ):

- RQ1 How does representing some call participants as avatars affect the experience of participating in a group video call for avatar, audio, and video participants?
- RQ2.1 How does a participant's self-representation (audio / video / avatar) subjectively impact their feeling of being present with others?
- RQ2.2 What is the effect of other participants' representation (audio/video/avatar) on how present they are perceived to be in the call?

In order to understand these effects, we conducted a within-subjects qualitative study with a group of 9 participants. All 9 participants participated in all three sessions (one session per study condition for each participant).A within-subjects design was used to enable users to compare their experiences across conditions. Over the course of the study, each participant joined three group video sessions using Together mode in Microsoft Teams with 9 participants in each session. For each session, participants were either asked to use their webcam (VIDEO condition), to turn off their webcam but use their microphone (AUDIO condition) or choose an avatar from the provided choices (AVATAR condition). Participants in the AVATAR condition were shown how to use our avatar software. These users were assigned either a high or low motion fidelity (explained below). During the three sessions that each group participated in, the group discussed one of the following three topics: (1) a movie, (2) travel plans for after the pandemic, or (3) research in academia vs. industry.

3.2.1 Participants. For this study, we recruited nine participants (7 men, 2 women; 4 aged 18-30 years, 4 aged 30-40 years, 1 aged 40-50 years) from a large tech company. Some of the participants were acquainted with each other. While a randomized lab experiment might require complete strangers to reduce the interaction effects of prior acquaintance on dependent variables, prior acquaintance amongst call participants is a common occurrence in typical workplace video meetings. Additionally, being from the same workplace adds a level of entitativity (an individual's recognition of a social unit as a group), which has been found to be essential for any level of co-presence and social presence [7]. Recruiting participants from the same workplace adds ecological validity to our study.

For work-related calls, all of the study participants reported using video conferencing software at least once daily, with 6 participants reporting multiple daily usage. For personal calls, 8 of the 9 study participants reported using video conferencing software less than once a week, with 1 participant reporting not using video conferencing for personal needs at all. Participants reported sharing their video feed during personal calls (often: 2 participants, always: 7 participants) slightly more often than for work-related video calls (often: 7 participants, always: 2 participants).

3.2.2 Data collection. At the start of the study, each participant was asked to complete a demographic survey that included questions on their prior use of video conferencing software. In addition, all participants completed a pre-survey that asked about their avatar preferences for video conferencing, which asked if they would use avatars for video calling in personal or work settings.

On the days of the study sessions, all participants participated in a group video call moderated by a researcher for 15-20 minutes. After introducing the session topic, the researcher turned his camera and microphone off. The video-calls were recorded for later data analysis. After each video call, participants were invited to share open-ended feedback and thoughts on their experience through an online survey ("Please share any feedback or thoughts from the user study session today") (contributed to RQ1, RQ2.1, and RQ2.2).

The sessions were conducted on a Monday, Wednesday, and Friday. After completing the three study sessions, each participant filled out a post-survey on their avatar preferences. Participants were also invited to participate in a 1 hour focus group session which took place a week after the final study session (the Wednesday of the week after). Seven out of nine participants joined the focus group session. Focus groups can help understand participants' feelings of presence [13]. We asked open-ended questions, primarily around their experience of being represented by and interacting with other avatars. This kind of semi-structured approach has been shown to be suitable for exploring presence [26].

3.3 Data analysis

We performed an inductive thematic analysis [9] of the data collected during observation (video recordings of the sessions and the focus group). The data was transcribed using an online transcription service and cross-checked for accuracy. Non-verbal utterances were disregarded. We grounded the analysis in the collected data and started with open coding. For the first round of open coding we used descriptive labels, which we combined into categories in an axial coding step. Finally, through discussion amongst the researchers, we developed three semantic themes [9] which we expand upon from the perspective of the ACTOR or the OBSERVER in Section 4:

- Presence and participation,
- Privacy and trust, and
- Non-verbal cues and body language.

3.4 Avatar software and setup

For the study, we built custom client-side avatar software in Unity3D game engine that interfaces with Together mode in Microsoft Teams. Together mode was chosen because it offered a shared virtual space that could be inhabited by video participants, avatars, and audio participants at the same time. Each participant that was assigned the avatar condition for the session ran this software on their machine. Using this software, participants could choose an avatar to represent themselves and switch between different avatars on the fly. The system allows for the avatar to either be controlled through motion capture from a webcam, or to generate artificial movements. In addition, it allows users to trigger emotions expressed through facial and body movements selected on a point and click interface (Figure 3). Our software runs the avatar in the background and we use an open-source Windows DirectShow filter, Unity Capture [2], to broadcast frames from our software as a virtual camera. The participant then chooses this virtual camera in Microsoft Teams.

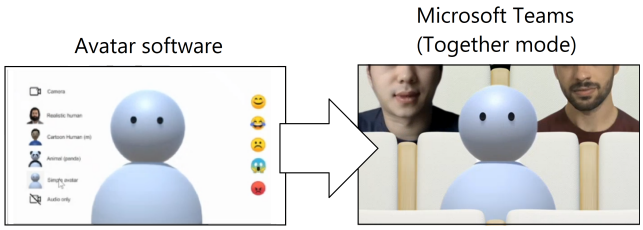

Figure 3: The Avatar software lets the user select various avatars and trigger emotions in Microsoft Teams.

Figure 4: Visual fidelity of various avatars in the study. The greyed out avatars were not offered as choices to the participants.

3.4.1 Avatar visual fidelity. The avatar choices given to the participants had different levels of visual fidelity, representing different levels of realism (Figure 4). The high visual fidelity avatars were visually photorealistic, with a strong resemblance to the participant. These were generated from the participant's photograph using an online service. For the remaining avatars, their resemblance to humans receded as the avatars became more abstract. These avatars were created by the authors in Unity3D using simple geometric shapes like spheres, capsules, and boxes. Four levels of fidelity were provided to the participants: 1) realistic human; 2) cartoon human (male or female); 3) animal (panda); and 4) simple generic avatar. Participants were allowed to choose their desired level of fidelity.

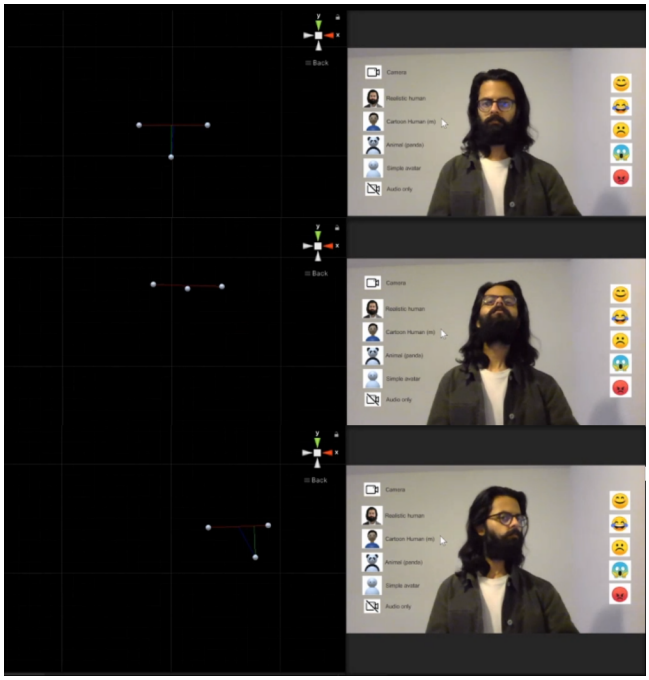

Figure 5: The system senses the location of the nose and the two eyes, and the simple algorithm calculates the approximate pose based on the relationship between their locations.

3.4.2 Avatar motion fidelity. Two levels of body motion were also implemented for the avatars. Avatar head motion was either mapped to the participant's head movements sensed through webcams for high motion fidelity, or was synthetically generated by the software. For the synthetic motion, the avatars randomly looked in different directions at irregular intervals, creating low motion fidelity.

The motion fidelity of the avatar was assigned to each user, with half using head-movement and the other half using synthetic motion. For the head-mapped motion, we used the OpenPose library [3]. OpenPose can detect key landmarks on the user's face on a 2D plane. We implemented a simple algorithm to estimate the 3D head pose from the 2D location of the landmarks (Figure 5). This mode required access to a camera device.

3.4.3 Avatar expressions. The avatar software allows users to trigger basic emotions through a point-and-click interface. The expressions available to the users are anger, laughter, sadness, smile, and surprise (Figure 6). These were selected by the research team based on which expressions we would like to have access to for both personal and workplace video calls. In addition to the triggered expressions, the avatars also had idling animations like eye blinks. Additionally, all avatars reacted to the audio from the microphone–including lip-sync for avatars with lips, and other visual indicators (e.g., a scaling head) for low visual fidelity avatars without a mouth (e.g. see Figure 6).

Figure 6: Three of the expressions offered by AllTogether. L to R: laughter, surprise, anger.

4 FINDINGS AND DISCUSSION

Here we share and discuss key themes from the thematic analysis. For clarity, we additionally indicate if an insight addresses the participant's experience of interacting with (OBSERVER) or being represented as (ACTOR) audio/video/avatar.

4.1 Presence and participation

ACTOR. Participants felt more present in the call when they had a visual representation (AVATAR or VIDEO conditions) than when they did not (AUDIO condition). This might be especially true in a mixed-modality conferencing system like Together mode, where call participants are viewed as part of a "group" that is together in a virtual shared environment, building entitativity for visually present participants [7]. Additionally, AUDIO participants commented that they found it hard to participate in the conversation with no visual representation. They felt more involved in the call with other participants and found it easier to contribute in both AVATAR or VIDEO conditions compared with the AUDIO-only condition. These findings confirm that a visual representation can help participants feel more present in a social environment [34, 35].

> *Participant 2*: Yeah, I definitely felt that [participating] was a hierarchy [with] video, it was really easy. [With] Avatar, it was [more] difficult than video. And [with] audio [it] was almost impossible.

OBSERVER. On average, participants indicated that it was easy to forget that audio-only participants were present in the call. This could be a nested recurrent problem–feeling less present due to no visual representation could lead audio participants to not engage in the call as much as participants with visual representation. From the OBSERVER's perspective, audio participants are neither present in the visual field, nor are they present in active conversation as much as participants with visual representations. This lowers entrainment (see [10]) and harms co-presence. These issues were addressed by avatars irrespective of their visual representation, with visually realistic avatars faring better than non-realistic representations because of their identifiability. While avatars helped mark the participants in an audience better than audio-only, some degree of avatar personalization could help speakers identify audience members being represented by generic avatars better. Participants speculated that they might feel "listened to" and reduce speaker anxiety by seeing avatars in the audience rather than empty seats (audio-only).

4.2 Privacy and trust

ACTOR. Participants felt that embodying a non-realistic avatar was valuable for how it protected an individual's privacy. Specifically, there was an imbalance in vulnerability that participants felt while being represented as video, realistic avatar, and non-realistic avatar. Participants felt "exposed" and as though they were "sharing more" as VIDEO or realistic avatars than non-realistic avatars.

> *Participant 3*: having the non-realistic avatar was nice from the user [ACTOR] perspective, but from the group [OBSERVER] perspective, it was a bit more awkward. If I'm in the meeting with video and other people are with the non-realistic avatars, I found that socially a bit stifling.

OBSERVER. Participants expressed concern that it might be hard to trust that audience members using avatars were paying attention in a meeting. However, using realistic avatars might mitigate this to some extent, in line with Inkpen and Sedlins's finding of users tending to trust realistic avatars more, particularly in workplace settings:

> *Participant 3*: I suppose the more [visually] realistic it was the more I was willing to believe that the person was actually there and engaged even if they weren't. Whereas like if they were a non-realistic avatar I was just kind of like, oh, it's like audio only, I have no idea if they're really paying [attention].

A combination of high visual fidelity and good tracking increased the likelihood that participants trusted the avatars from an OBSERVER perspective. Poor tracking was particularly problematic for participants from the ACTOR perspective–since the user has full knowledge of the baseline truth (i.e. how the participant is actually moving), seeing a discrepancy in how the avatar representing them is moving can cause dissonance. This finding confirms the importance of good tracking algorithms that have been presented in past work [15, 44].

OBSERVER. Prior acquaintance between call participants lowered the feeling of trust for avatars with poor simulated tracking even further. For example, P5 knew P4 well, and was familiar with his natural body movements. P5 found that P4's synthetic avatar body movements made her disregard his realistic visual representation entirely.

> *Participant 5*: It was enough for my brain to be like, "The visual information is incorrect. You should disregard it" and I basically completely focused on the audio. Which was so surprising to me, that it had that effect.

4.3 Non-verbal cues and body language

ACTOR. Participants felt that representing their head and limb movement without facial expressions was enough for conveying their presence in the meeting while maintaining privacy. They valued the use of simple expressions such as nodding, which were a rich social signal that helped both ACTOR and OBSERVER feel that they were actively participating in the conversation. Even simple body and limb tracking animating an avatar can increase agency, which increases the feeling of co-presence [8, 12].

> *Participant 1*: They're not really trying to mimic facial expression[s]. [Instead, they do] the default tracking

> of where I'm looking, embodying what I'm doing in the meeting.

Body motion can be used to mediate turn-taking [47], which points to the importance of accurate tracked body motions. At the same time, random, larger movements in a naive synthetic body motion algorithm (like in AllTogether) can interrupt understanding the flow of the conversation for both ACTOR and OBSERVER.

> *Participant 5*: One thing that I like to do when I start talking is I really like to use my body to draw attention.
>
> *Participant 7*: [The avatar's random motion] really bummed me out both as a listener [OBSERVER]– to see who was talking and see where the conversation flow was – but also for me [as an ACTOR], managing my own reactions.

OBSERVER. The simplified appearance of non-realistic avatars drew more attention to the avatar's body motion. Viewing others' emotional expression was perceived to be less important than natural body movements which implicitly encode levels of attention, focus, and interest. This information was lacking for the audio call, a distinct disadvantage. A low visual and expressive fidelity avatar with high motion fidelity could lead to higher levels of presence, which confirms Nowak and Biocca's findings [29].

> *Participant 9*: [By seeing head movements] I can see who's paying attention and who is kind of lost or who is agreeing with me and who is disagreeing with me. So I think that's more important.

ACTOR. Additionally, while previous studies have shown that facial expressions can help users self-identify with avatars [16], participants expressed lack of trust in facial emotion tracking systems, and concern about their emotional state being conveyed incorrectly to others. They also surmised that incorrectly tracking facial emotions could introduce "creepiness" for both ACTOR and OBSERVER.

> *Participant 4*: When you try to mimic facial expressions, you start entering the uncanny valley, but with just the head tracking, I think it's great, because even with just head tracking, you're able to convey a lot of information.

5 DESIGN IMPLICATIONS FOR AVATAR INTEGRATION IN VIDEO CONFERENCING

Participants reported that they enjoyed talking with a realistic avatar (OBSERVER lens) (e.g. Figure 7), and the higher visual fidelity helped participants accept that they were talking to a real person. They also mentioned that being represented by an avatar did not bother them either (ACTOR lens). While both low- and high-fidelity avatars have their pros and cons for various contexts, we offer a few insights that should be kept in mind for integrating avatars in future video conferencing systems.

5.1 Avatar choice and customization

Giving users the agency to choose between avatars of different visual fidelity and the option to customize these avatars for personalization and easier identification has merit.

Avatars with different levels of fidelity can be useful in different circumstances, like a non-realistic, simple avatar in a large video call as an audience member, or a realistic avatar for a team meeting at work. Combined with the different options for enabling avatar movement (body tracking, synthetic motion), we can create a rich ecology of choices for call participants. Each of these choices will have pros and cons (eg. higher trust of tracked avatars (from the observer's perspective), but higher freedom to do other work during the call with avatars with synthetic motion (for the actor)). A conferencing system could potentially suggest different avatar visual and tracking fidelity for the user based on the context (e.g. low fidelity avatars for large meetings, high fidelity for small meetings, video for individual meetings).

A participant represented by a more realistic avatar was easier to identify, accepted by the other participants, and regarded as more reliable. However, study participants that sought privacy through avatars felt exposed by using realistic avatars. While it was harder to distinguish low-fidelity avatars from each other, offering personalization for such avatars could aid in differentiating between participants while retaining user privacy.

Figure 7: An avatar with high visual fidelity.

5.2 Alternate forms of input tracking

In our study, avatars either had synthetic motion, or motion tracked through the camera. Participants noted the benefits of having tracked motion, saying that it makes the avatars feel more natural. However, there were advantages of offering synthetic motion–e.g. being visually present in the call while not being tied to the physical location of the computer. A good synthetic motion model that does not require direct sensing of the users may address the needs of realism and privacy. However we suspect that a naive implementation of such a synthetic motion algorithm might lower presence [12].

The implementation we tested in this paper does not allow the user to have tracked motion without being physically present in front of a camera. However, there are be other ways of tracking user's movements. For example, headphones with a built-in Inertial Measurement Unit (IMU) could be used to track a user's head movements, which can enable the avatar to mimic the participant's motions. These augmented headphones could allow for tracked movements even while the user might be physically located elsewhere.

These forms of input tracking relate back to one of the reasons for users turning their video off: "I wanted to multitask during the meeting" was chosen as a reason by 13% of the respondents in our formative survey (Section 3.1). Avatars that allow such users to do

other tasks, such as laundry or house chores, could allow them to have visual presence in large group calls while enabling them to accomplish other tasks.

5.3 Graceful bandwidth degradation

One of the major reasons that users turn off their video cameras is the network bandwidth that video sharing demands (Section 3.1). Using avatars can be an advantage to people without access to a high-speed internet connection. Since avatar movements are computationally controlled, the avatars can be implemented in the cloud for server-mediated video conferencing solutions like Microsoft Teams, or individually on the client side for peer-to-peer (P2P) video conferencing like Skype [5]. This reduces the amount of data that a user has to upload in order to communicate. For synthetic motion the user only has to upload audio, and for tracked motion it is sufficient to transfer movement data only which can be used to compute the avatar's representation on the server or on the client side. Avatars can also be beneficial for graceful bandwidth degradation. Compression and downsampling the video stream are common practice in video conferencing solutions. For video conferencing, avatars can provide most of the benefits of full visual fidelity (video, high bandwidth usage) at a bandwidth cost similar to audio-only transmission. When the user's network speed drops, the video conferencing system can automatically and gracefully transition from video to avatar [16].

6 LIMITATIONS

The implementation of the avatar software had a few shortcomings. We used OpenPose [3] for head tracking which required high performance computers. This, combined with running the study during the COVID-19 global pandemic, limited us in a few ways.

First, we could not use advanced tracking, and were limited to utilizing just a few landmarks to get a respectable framerate for real-time tracking on moderate performance computers (in our case, three landmark points–two eyes and nose tip: Figure 5).

Second, the high performance requirement limited the number of participants we could recruit. Because of the remote study requirement, the participants needed to run the software on their home computers (which needed to meet a certain performance threshold). This limited our sample size.

Our sample does not represent a true random sample from the population of users that use video conferencing tools for work and personal communication. However, for a constructivist approach to qualitative data analysis, random sampling is not an ideal–richness of data is. That said, our study only included two female participants, introducing a potential gender bias.

7 FUTURE WORK

This paper's contribution focuses on understanding the perceived experience of users as they use avatars in a mixed-modality video conferencing environment. The themes that emerged from our analysis highlighted the tension between the use of low- and high-fidelity avatars, particularly from the perspective of trust and privacy.

There are multiple facets to the construct of trust in this context, including (1) trusting the system as an ACTOR: the *actions* and body motion that one's avatar performs on screen are indicative of the physical actions and body motion that the user is performing and (2) trusting somebody else's avatar as OBSERVER: trusting that the avatar's *appearance* is indicative of the participant's actual appearance.

While generally high visual fidelity avatars were trusted by observers, a higher motion fidelity was more important from the actor perspective. Our participants indicated that they valued body motion tracking more than facial expressions. However, while this might increase "mutual behaviour" [10], the lack of any emotional expression would reduce the degree of "mutual emotion", an important aspect of developing co-presence [10]. This might be mitigated with untracked emotional expressions, e.g. triggered by a point-and-click interface in our prototype. However, future work should study in more depth the effects of avatar expressionism vs. body motion on mutual entrainment and co-presence.

Low-visual fidelity avatars offered higher privacy, and so might be favorable from the actor perspective. Future work needs to explore how to navigate this ambiguous space of user trust and privacy, and how to best balance between the two through visual and motion fidelity of avatars.

Finally, as we move toward an increasingly digitized workplace and the metaverse, future work must explore the ramifications of using avatars across systems where 2D desktop users could meet and collaborate with 3D VR users.

8 CONCLUSION

Current video conferencing systems typically offer users a binary choice for visual representation–to either turn their cameras on or off. Users with access to a camera benefit from having rich representations in the meeting. However, those who don't have a video feed either by necessity or by choice might feel excluded from the social group. This is especially true for configurations where the participants with a video feed are situated in a shared virtual space, like in Together mode (Figure 2(a)).

Our formative survey (Section 3.1) showed numerous instances when participants did not have video on during a meeting, for a variety of reasons including: needing to multitask, wanting to hide their visual appearance, and network quality issues. However, our results revealed challenges for voice-only participants. Results from our user study demonstrate a potential for avatars to fill the gap between audio-only and video representation, including enhancing a participant's sense of feeling present as well as their sense of other participants being present in the call. We also provide insights on mixed-modality conferencing environments where meetings can include both avatars and video representations in the same virtual space (Figure 1).

Participants in our study generally preferred avatars to audio-only, which was comparable to their preference for video for these dimensions:

- the feeling of being present in the call along with others (when self is represented as avatar),
- the feeling of other being present in the call (when the other is represented as avatar),
- being able to be seen in a group call and identified as a participant,

- being trusted by a speaker (this was only true when the avatar is perceived to be mimicking user's movement),
- being able to initiate non-verbal communication, for instance through head movement for tracked avatars and emotion elicitation.

While participants generally preferred video over avatars, there were some scenarios where they preferred avatars over video:

- Avatars have the potential to allow the participant to be visually present in the meeting while not being physically present at the computer. This was only true for avatars that do not use a sensing device attached to the user's computer in order to track their movements (like a webcam). This could be done with avatars with synthetic motion (implemented in AllTogether), or avatars that use other forms of input tracking like headphones (not implemented in AllTogether). This addresses one of the reasons suggested for not using the webcam video in our formative survey ("multitasking preference", Section 3.1).
- Avatars allow the participant to be present while hiding their visual appearance (completely, in the case of non-realistic avatars, and partially, in the case of realistic avatars). This addresses another of the reasons for not using webcam in our formative survey ("I wanted to hide my visual appearance", Section 3.1).

There are also technical benefits to using avatars in terms of bandwidth usage, enabling transmission that can utilize very low network bandwidth–addressing another user concern from the formative survey. This can also support graceful bandwidth degradation akin to automatic downsampling as used in video streaming services.

When integrating avatars into video conferencing solutions, we advise designers to offer a range of avatar choices to the users, consider using alternate forms of input tracking for body motion, and evaluate the value of graceful bandwidth degradation for their application.

In summary, avatars show promise as an addition to the choices for visual representation offered to video conferencing participants. As society continues to transition to extensive remote and hybrid work scenarios, advancements in the ways users can be represented and engage with each other are critical, and avatars offer a plausible way forward.

REFERENCES

[1] 2020. A Message to Our Users - Zoom Blog. https://blog.zoom.us/a-message-to-our-users/.

[2] 2020. GitHub - schellingb/UnityCapture: Streams Unity rendered output to other Windows applications as virtual capture device. https://github.com/schellingb/UnityCapture.

[3] 2020. OpenPose: Real-time multi-person keypoint detection library for body, face, hands, and foot estimation. https://github.com/CMU-Perceptual-Computing-Lab/openpose.

[4] 2020. 'Zoom fatigue' is taxing the brain. Here's why that happens. https://www.nationalgeographic.com/science/2020/04/coronavirus-zoom-fatigue-is-taxing-the-brain-here-is-why-that-happens/.

[5] Salman A Baset and Henning G Schulzrinne. 2006. An analysis of the Skype peer-to-peer internet telephony protocol. In *Proceedings - IEEE INFOCOM*. https://doi.org/10.1109/INFOCOM.2006.312

[6] Gary Bente, Sabine Rüggenberg, Nicole C. Krämer, and Felix Eschenburg. 2008. Avatar-Mediated Networking: Increasing Social Presence and Interpersonal Trust in Net-Based Collaborations. *Human Communication Research* 34, 2 (04 2008), 287–318. https://doi.org/10.1111/j.1468-2958.2008.00322.x arXiv:https://academic.oup.com/hcr/article-pdf/34/2/287/22325251/jhumcom0287.pdf

[7] Anita L. Blanchard, Leann E. Caudill, and Lisa Slattery Walker. 2018. Developing an entitativity measure and distinguishing it from antecedents and outcomes within online and face-to-face groups:. *https://doi.org/10.1177/1368430217743577* 23 (1 2018), 91–108. Issue 1. https://doi.org/10.1177/1368430217743577

[8] Kristopher J Blom. 2007. On Affordances and Agency as Explanatory Factors of Presence. *Psychnology Journal* (2007). http://www.psychnology.org

[9] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. https://doi.org/10.1191/1478088706qp063oa

[10] Celeste Campos-Castillo and Steven Hitlin. 2013. Copresence: Revisiting a Building Block for Social Interaction Theories. *Sociological Theory* 31 (2013), 168–192. Issue 2. https://doi.org/10.1177/0735275113489811

[11] Liz Fosslien and Mollie West Duffy. 2020. How to combat zoom fatigue. *Harvard Business Review* (2020).

[12] Jesse Fox, Sun Joo (Grace) Ahn, Joris H. Janssen, Leo Yeykelis, Kathryn Y. Segovia, and Jeremy N. Bailenson. 2015. Avatars Versus Agents: A Meta-Analysis Quantifying the Effect of Agency on Social Influence. *Human–Computer Interaction* 30, 5 (2015), 401–432. https://doi.org/10.1080/07370024.2014.921494 arXiv:https://doi.org/10.1080/07370024.2014.921494

[13] Jonathan Freeman and Steve E. Avons. 2000. Focus group exploration of presence through advanced broadcast services. In *Electronic Imaging*.

[14] Mar Gonzalez-Franco, Zelia Egan, Matthew Peachey, Angus Antley, Tanmay Randhavane, Payod Panda, Yaying Zhang, Cheng Yao Wang, Derek F. Reilly, Tabitha C Peck, Andrea Stevenson Won, Anthony Steed, and Eyal Ofek. 2020. MoveBox: Democratizing MoCap for the Microsoft Rocketbox Avatar Library. In *2020 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*. IEEE, 91–98. https://doi.org/10.1109/AIVR50618.2020.00026

[15] M Gonzalez-Franco, E Ofek, Y Pan, A Antley, A Steed, B Spanlang, A Maselli, D Banakou, N Pelechano, S Orts-Escolano, V Orvalho, L Trutoiu, M Wojcik, MV Sanchez-Vives, J Bailenson, M Slater, and J Lanier. 2020. The Rocketbox library and the utility of freely available rigged avatars for procedural virtual humans and embodiment. *Frontiers in Virtual Reality* (2020).

[16] Mar Gonzalez-Franco, Anthony Steed, Steve Hoogendyk, and Eyal Ofek. 2020. Using Facial Animation to Increase the Enfacement Illusion and Avatar Self-Identification. *IEEE Transactions on Visualization and Computer Graphics* 26, 5 (2020), 2023–2029.

[17] Hunter G. Hoffman, K.C. Hullfish, and Suzy Houston. 1995. Virtual-reality monitoring. *Proceedings Virtual Reality Annual International Symposium '95* (1995), 48–54.

[18] Wijnand Ijsselsteijn. 2004. *Presence in Depth*.

[19] Wijnand Ijsselsteijn, Huib De Ridder, Roelof Hamberg, Don Bouwhuis, and Jonathan Freeman. 1998. Perceived depth and the feeling of presence in 3DTV. *Displays* 18 (5 1998), 207–214. Issue 4. https://doi.org/10.1016/S0141-9382(98)00022-5

[20] Kori M. Inkpen and Mara Sedlins. 2011. Me and My Avatar: Exploring Users' Comfort with Avatars for Workplace Communication. In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work* (Hangzhou, China) *(CSCW '11)*. Association for Computing Machinery, New York, NY, USA, 383–386. https://doi.org/10.1145/1958824.1958883

[21] Sasa Junuzovic, Kori Inkpen, John Tang, Mara Sedlins, and Kristie Fisher. 2012. To see or not to see: A study comparing four-way avatar, video, and audio conferencing for work. In *GROUP'12 - Proceedings of the ACM 2012 International Conference on Support Group Work*. 31–34. https://doi.org/10.1145/2389176.2389181

[22] Robert S. Kennedy, Norman E. Lane, Kevin S. Berbaum, and Lilienthal Mg. 1993. Simulator Sickness Questionnaire: An enhanced method for quantifying simulator sickness. *The International Journal of Aviation Psychology* 3 (1993), 203–220.

[23] Matthew Lombard, Theresa B Ditton, and Lisa Weinstein. 2009. Measuring Presence: The Temple Presence Inventory. *Proceeding of Presence 2009 : the 12th International Workshop on Presence* (2009), 1–14. http://www.temple.edu/ispr/prev_conferences/proceedings/2009/Lombard_et_al.pdf

[24] Marketwatch. 2020. Zoom, Microsoft Teams usage are rocketing during coronavirus pandemic, new data show. Retrieved Sept 14, 2020 from https://www.marketwatch.com/story/zoom-microsoft-cloud-usage-are-rocketing-during-coronavirus-pandemic-new-data-show-2020-03-30

[25] Rosa Mikeal Martey and Mia Consalvo. 2011. Performing the Looking-Glass Self: Avatar Appearance and Group Identity in Second Life. *Popular Communication* 9, 3 (2011), 165–180. https://doi.org/10.1080/15405702.2011.583830 arXiv:https://doi.org/10.1080/15405702.2011.583830

[26] Craig D. Murray, Paul Arnold, and Ben Thornton. 2000. Presence Accompanying Induced Hearing Loss: Implications for Immersive Virtual Environments. *Presence: Teleoperators & Virtual Environments* 9 (2000), 137–148.

[27] Bonnie Nardi and Justin Harris. 2010. *Strangers and Friends: Collaborative Play in World of Warcraft*. Springer Netherlands, Dordrecht, 395–410. https://doi.org/10.1007/978-1-4020-9789-8_24

[28] Sarah Nichols, Clovissa Haldane, and John R. Wilson. 2000. Measurement of presence and its consequences in virtual environments. *Int. J. Hum. Comput.*

Stud. 52 (2000), 471–491.

[29] Kristine L. Nowak and Frank Biocca. 2003. The Effect of the Agency and Anthropomorphism on users' Sense of Telepresence, Copresence, and Social Presence in Virtual Environments. In *Presence: Teleoperators and Virtual Environments*, Vol. 12. MIT Press 238 Main St., Suite 500, Cambridge, MA 02142-1046 USA journals-info@mit.edu, 481–494. https://doi.org/10.1162/105474603322761289

[30] Edwin B. Parker, John Short, Ederyn Williams, and Bruce Christie. 1978. The Social Psychology of Telecommunications. *Contemporary Sociology* 7, 1 (1978), 32. https://doi.org/10.2307/2065899

[31] Jerrold D Prothero, Donald E Parker, Thomas A Furness III, and Maxwell J Wells. 1995. Towards a Robust, Quantitative Measure for Presence. In *Conference on Experimental Analysis and Measurement of Situational Awareness*. 359–366. http://www.hitl.washington.edu/publications/p-95-8/

[32] Xavier Rétaux. 2003. Presence in the environment: theories, methodologies and applications to video games. *PsychNology J.* 1 (2003), 283–309.

[33] Liam Rourke, Liam Rourke, Terry Anderson, D Randy Garrison, and Walter Archer. 1999. Assessing Social Presence in Asynchronous Text-based Computer Conferencing. *The Journal of Distance Education / Revue de l'ducation Distance* 14 (1999), 50–71. Issue 2.

[34] Louis Schneider and Erving. Goffman. 1964. Behavior in Public Places: Notes on the Social Organization of Gatherings. *American Sociological Review* 29, 3 (1964), 427. https://doi.org/10.2307/2091496

[35] Ralph Schroeder. 2002. *Social Interaction in Virtual Enviroments: Key Issues, Common Themes, and a Framework for Research.* Springer-Verlag, Berlin, Heidelberg, 1–18.

[36] Ralph Schroeder. 2011. *Comparing Avatar and Video Representations.* Springer London, London, 235–251. https://doi.org/10.1007/978-0-85729-361-9_12

[37] Thomas Schubert, Frank Friedmann, and Holger Regenbrecht. 2001. The experience of presence: Factor analytic insights. *Presence: Teleoperators and Virtual Environments* 10, 3 (2001), 266–281. https://doi.org/10.1162/105474601300343603

[38] Thomas W. Schubert. 2009. A New Conception of Spatial Presence: Once Again, with Feeling. *Communication Theory* 19, 2 (may 2009), 161–187. https://doi.org/10.1111/j.1468-2885.2009.01340.x

[39] Thomas B. Sheridan. 1992. Musings on Telepresence and Virtual Presence. *Presence: Teleoperators and Virtual Environments* 1 (1 1992), 120–126. Issue 1. https://doi.org/10.1162/PRES.1992.1.1.120

[40] Mel Slater, Andrea Brogni, and Anthony Steed. 2003. Physiological Responses to Breaks in Presence: A Pilot Study. *Presence* (2003).

[41] Mel Slater, Amela Sadagic, Martin Usoh, and Ralph Schroeder. 2000. Small-group behavior in a virtual and real environment: A comparative study. *Presence: Teleoperators and Virtual Environments* 9, 1 (mar 2000), 37–51. https://doi.org/10.1162/105474600566600

[42] Mel Slater, Anthony Steed, John McCarthy, and Francesco Maringelli. 1998. The influence of body movement on subjective presence in virtual environments. *Human Factors* 40, 3 (sep 1998), 469–477. https://doi.org/10.1518/001872098779591368

[43] Michael P. Snow and Robert C. Williges. 1998. Empirical models based on free-modulus magnitude estimation of perceived presence in virtual environments. *Human factors* 40 (1998), 386–402. Issue 3. https://doi.org/10.1518/001872098779591395

[44] Bernhard Spanlang, Jean-Marie Normand, David Borland, Konstantina Kilteni, Elias Giannopoulos, Ausiàs Pomés, Mar González-Franco, Daniel Perez-Marcos, Jorge Arroyo-Palacios, Xavi Navarro Muncunill, et al. 2014. How to build an embodiment lab: achieving body representation illusions in virtual reality. *Frontiers in Robotics and AI* 1 (2014), 9.

[45] Kay M. Stanney and Robert S. Kennedy. 1998. Aftereffects from Virtual Environment Exposure: How Long do They Last?:. *http://dx.doi.org/10.1177/154193129804202103* 2 (nov 1998), 1476–1480. https://doi.org/10.1177/154193129804202103

[46] Joy van Baren and Wijnand IJsselsteijn. 2004. Measuring Presence: A Guide to Current Measurement Approaches. *OmniPres Project* (2004).

[47] Marla B. Wadsworth and Anita L. Blanchard. 2015. Influence tactics in virtual teams. *Computers in Human Behavior* 44 (3 2015), 386–393. https://doi.org/10.1016/J.CHB.2014.11.026

[48] Stephan Wenninger, Jascha Achenbach, Andrea Bartl, Marc Erich Latoschik, and Mario Botsch. 2020. Realistic Virtual Humans from Smartphone Videos. In *Proceedings of the ACM Symposium on Virtual Reality Software and Technology, VRST*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3385956.3418940

[49] Bob G. Witmer and Michael J. Singer. 1998. Measuring presence in virtual environments: A presence questionnaire. *Presence: Teleoperators and Virtual Environments* 7, 3 (1998), 225–240. https://doi.org/10.1162/105474698565686