**Title**: Point-cloud avatars to improve spatial communication in immersive collaborative virtual environments

**Authors**: Guillaume Gamelin1, Amine Chellali1*, Samia Cheikh1, Aylen Ricca1, Cedric Dumas2, Samir Otmane1

**Affiliation**s: [1]IBISC Laboratory, Univ Evry, Université Paris Sacalay, Evry, France

[2]LS2N, Institut Mines Telecom Atlantique, Nantes, France

**Acknowledgments**

The authors would like to thank all the volunteers that participated to the experimental study. The authors would like also to thank the Paris Ile-de-France Region (grant # 17002647) and Genopole for the financial support.

Corresponding author:

Dr. Amine CHELLALI

ORCID: 0000-0002-6143-5898

Amine.chellali@univ-evry.fr

Université d'Evry

40, Rue du Pelvoux,

Courcouronnes, 91020 Evry Cédex

France

Tel : +33 1 69 47 75 33

# Point-cloud avatars to improve spatial communication in immersive collaborative virtual environments

**Guillaume Gamelin[1], Amine Chellali[1*], Samia Cheikh[1], Aylen Ricca[1], Cedric Dumas[2], Samir Otmane[1]**

[1]IBISC Laboratory, Univ Evry, Université Paris Sacalay, Evry, France

[2] LS2N, Institut Mines Telecom Atlantique, Nantes, France

**Abstract**

Collaborative virtual environments allow remote users to work together in a shared 3D space. To take advantage of the possibilities offered by such systems, their design must allow the users to interact and communicate efficiently. One open question in this field concerns the avatar fidelity of remote partners. This can impact communication between the remote users, more particularly when performing collaborative spatial tasks. In this paper, we present an experimental study comparing the effects of two partner's avatars on collaboration during spatial tasks. The first avatar was based on a 2.5D streamed point-cloud and the second avatar was based on a 3D preconstructed avatar replicating the remote user movements. These avatars differ in their fidelity levels described through two components: visual and kinematic fidelity. The collaborative performance was evaluated through the efficacy of completing two spatial communication tasks, a pointing task and spatial guidance task. The results indicate that the streamed point-cloud avatar permitted a significant improvement of the collaborative performance for both tasks. The subjective evaluation suggests that these differences in performance can mainly be attributed to the higher kinematic fidelity of this representation as compared to the 3D preconstructed avatar representation. We conclude that, when designing spatial collaborative virtual environments, it is important to reach a high kinematic fidelity of the partner's representation while a moderate visual fidelity of this representation can suffice.

**Keywords: Partner's avatar, Kinematic fidelity, Collaborative virtual environments, spatial communication, Immersive virtual reality.**

## 1 Introduction

Collaborative virtual environments (CVEs) allow multiple remote users to work together in a shared virtual space [*1*]. They are used in different applications such as industrial design, entertainment or training [*2*, *3*]. These systems must support some features, such as multimodal communication and must provide cues on the partners' activities in order to improve the sense of copresence between the team members and ensure an effective collaboration [*4*].

In this context, one of the open research questions when designing such systems is the way the remote partners and their actions are visually represented in the CVE [*5*]. In this paper, we are particularly interested in the effects that the fidelity of this representation can have on spatial interactions in an immersive CVE. Indeed, in everyday life, operators use

different body parts to communicate spatial information to other people (e.g., pointing out an object of interest with the finger). They also use their awareness of their partner's position to adapt their verbal descriptions of spatial relationships between the surrounding objects (e.g., describing the relative position of an object according to the partner's position/perspective). Thus, the fidelity of the visual representation of a user's body in the virtual environment can impact the efficiency of spatial communication between partners and thus the effectiveness of their collaboration [*6*]. The objective of this work is to study the impact of the partner's representation fidelity on the way persons collaborate when performing spatial tasks. For that purpose, we have conducted a user study to compare two representations of the partner that differ in their levels of fidelity. This work is a first step towards defining design guidelines which take into consideration the fidelity of the partner's visual representation in order to enhance the effectiveness of spatial communication and collaboration in CVEs.

## 2 Related work

Our literature review is divided into three main categories. First, we examine how spatial communication is performed in face to face situations and explore the way it is currently supported in CVEs. Second, we review users' representations in CVEs. Finally, we investigated how the fidelity of the user's representation can be described.

### 2.1 Spatial communication in CVE

The ability of operators to collaborate efficiently when performing spatial tasks together depends strongly on their ability of locating their partners and the shared objects in their common workspace [*7*]. The activity of exchanging information to locate an object and/or a person is referred to as spatial communication. The verbal description of an object position depends on various factors such as the relative position of this object according to the other objects of the environment or according to the position and/or the perspective of the speaker and the listener. Usually, to make this verbal description, an operator (the speaker) uses the egocentric reference system [*8*]. In this reference system, actions and object positions are described by the speaker according to his/her own body/perspective (for example, "the object on my left/right"). In a face to face collaborative situation, the partners do not necessarily share the same viewpoint of the environment. However, an operator is usually aware of the partner's (listener's) gaze direction and field of view. Therefore, he/she can more or less easily use the listener's egocentric reference system by projecting him/herself into the partner's body and locating objects according to the listener's perspective (for example, "the object on your left/right"). This is referred to as spatial perspective-taking [*9*]. The speaker can also enrich a spatial verbal description by using additional nonverbal cues such as pointing gestures, gaze, and head orientation. This suggests that spatial communication depends on different factors directly related to the way the partners see each other including their movements and actions.

These spatial interactions can become difficult to perform in CVEs. Indeed, virtual environments change the users' perception of their surrounding space and often include very little cues on the partner's activity [*4*]. It is for instance more complicated to use the listener's egocentric system if the speaker is not correctly aware of the partner's perspective in the CVE. In addition, the use of nonverbal cues is usually limited compared to real world interactions [*10*]. Pointing an object can for instance become complicated to perform and the partner's viewpoint difficult to perceive depending on how the body and movements of the remote partner are faithfully reproduced in the CVE. This may result in communication issues when two partners perform together a spatial task. This highlights also the importance of designing the partner's

representation with an appropriate level of fidelity in order to provide the necessary cues to the speaker and the listener, thus reducing the spatial communication issues in CVEs. Again, these cues should convey information about the movements and actions of the partner.

In the literature, some studies have investigated spatial communication in CVEs [*4*, *11*, *9*, *12*]. Instead of focusing on the use of avatars, some of these studies have proposed new interaction metaphors to improve spatial communication. For instance, Chellali et al. [*4*] included a common spatial frame of reference in the CVE to encourage the partners to spontaneously use an exocentric reference frame (locating an object according to the other environment objects) instead of an egocentric one during spatial interactions. Their study has concluded that this visual aid can improve spatial communication but is less helpful for users that have difficulties to perform mental rotations. In fact, using the exocentric reference frame requires performing complex mental rotations to locate objects in space. Hindmarsh et al. [*11*] proposed to provide each operator with a feedback on the partner's view of the CVE, and objects present in his/her field of view. This helped the partner performing spatial descriptions of the environment and improved spatial communication [*11*, *13*]. However, adding separate windows in the user's view field may increase the cognitive load and slow down the task completion time [*14*]. More recently, the study of Pouliquen-Lardy et al. [*9*] investigated role distubtion (manipulator vs guide) during a collaborative spatial manipulation task and its effect on perspective-taking during spatial communcaition. The results show that during spatial interactions, the partners preffered using the manipulator's perspective rather than the guide's perspective. While this study used avatars to represent the remote partners, the avatars were not animated limiting thus their fidelity.

The recent developments in immersive VR technologies and the reduced cost of tracking devices offer new opportunities to include higher fidelity avatars at lower cost. Some studies have investigated the influence of the user's representation on spatial interactions but in the case of single user virtual environments [*15*, *16*]. The study of Ries et al. showed an improvement of distance estimations in the virtual environment for users equipped with a virtual avatar over those without a virtual representation. In addition, Mohler et al. have shown that an animated avatar outperforms a static avatar in distance estimation tasks [*16*]. Our goal is to build on this existing literature by exploring avatar-based approaches in the case of spatial collaboration tasks in a systematic way. More particularly, we are interested in investigating how the fidelity of movements of the remote partner's avatar can impact the collaborative performance during spatial tasks.

### 2.2 Users' representation in CVE

User's representation is a fundamental problem when designing immersive virtual environments, and is usually acquired using an avatar: a visual representation of the user inside the virtual environment [*17*]. Avatars provide a direct relationship between the natural movements of a user in the real world environment and the animation of his 3D representation in the virtual environment. This animation is usually performed through motion capture devices. The importance of the own body perception in immersive virtual environments has been investigated for some time with results suggesting a strong correlation between the feeling of presence and the degree of association of the user with his/her virtual body. This association is referred to as the sense of embodiment [*18*]. In CVEs, the perception of the partner's body is even more important because it gives crucial information on the partners' activity such as their position, identity, objects of interest, gestures, mood, and actions [*6*, *19*].

Several studies have investigated the effects of the presence and the realism of the partner's avatar in CVEs with results indicating that people feel higher levels of social presence when there is a visual representation available [*20*]. For instance, the study of Fribourg et al. has shown that being immersed with another person's avatar can improve the performance and task engagement of a CVE user, but does not influence one's sense of embodiment [*21*]. The study of Piumsomboon et al. has shown that the presence of a remote user's avatar improved the sense of social presence and the overall user experience in a mixed reality collaborative system [*22*]. Dodds et al. have shown that a self-animated avatar improves communication with a remote user within a CVE as compared to non-animated avatars and to avatars with a prerecorded listening behavior animation [*23*]. Steptoe et al. have shown that adding ocular behavior into the partner's avatar through gaze tracking improves social interactions between the collaborators [*6*]. The study of Garau et al. has shown a positive effect of the visual and behavioral realism of the partner's avatar on the perceived quality of communication in an immersive CVE [*24*]. By comparing two forms of the partner's visual representations (an abstract avatar Vs a preconstructed photo-realistic avatar) Latoschik et al. have shown a positive effect of the increased visual realism of the partner's avatar on the self-perception in a virtual environment [*25*]. The study of Cowell et al. indicates that incorporating trusting nonverbal cues in virtual characters increases the feeling of credibility and trustworthiness [*26*]. However, this study was conducted with users interacting with virtual characters and not with actual remote partners in a CVE. The study of Young et al. [*27*] compared the effects of a full-body representation of the partner to a hands-only representation when performing together a high-fiving gesture in a CVE. Their results show that the full-body representation of both users have improved the task performance. In the work of Economou et al. [*28*], different forms of user's representations in immersive CVEs were proposed. The authors have provided different guidelines for designing the partner's representation in CVEs. However, no systematic evaluation of the different proposed representations was conducted. In the work of Steptoe et al. [*29*], different users' representations were used to support acting rehearsal in a collaborative multimodal mixed reality environment. The system was informally evaluated by a director and two actors who performed together a remote rehearsal of a theatrical scene. The system was well appreciated by the users although the participants have commented that the inexpressivity of the actors' avatar may impact the rehearing via this system. Fairchild and al. presented a Mixed Reality system supporting the contextualization of nonverbal communication. The system was used in a simulation of Mars, within which the collaborators performed together exploration tasks and social interactions. To improve nonverbal communication, the authors used a real time 3D video reconstructed avatar. While this method permitted to improve nonverbal communication, it required the use of expensive tracking cameras [*30*]. The study of Roth et al. has shown that users can compensate for missing nonverbal behaviors when using non-visually realistic avatars to improve their social interactions by using other behavioral channels [*31*]. More recently, Wu et al. have compared a depth-sensor-based tracking to a controller-based partial-body tracking (using inverse kinematics) of the same 3D preconstructed avatar in a VR interview simulation [*32*]. Their experiment has shown that the depth-sensor-based tracking improves user experience, increases virtual body ownership and improves the users' rating of their own nonverbal behaviors. However, these avatars were used to represent the participants and not their partners. Cho et al. have shown that volumetric depth-sensor captured avatars increase users' social presence as compared to a 3D preconstructed mesh avatar and to a 2D video [*33*]. Yoon et al. have compared different appearances of the remote user's avatar ranging from a head & hands representation to a whole body representation [*34*]. They have also found that the whole body representation increases social presence. Finally, Regenbrecht et al. presented a voxel-based method to visually represent users in an interactive real-time environment [*35*]. Scenarios with persons collaborating in a 3D conferencing space were

presented by the authors highlighting the importance of nonverbal communication cues supported by this approach. However, only a subjective evaluation was conducted showing that this user's representation method is promising.

The previous review suggests the existence of a positive effect of the avatar presence and fidelity on social interactions in CVEs. In addition, nonverbal communication is better perceived when the behavioral fidelity of the avatars is increased. However, to our best knowledge, the impact of the fidelity of the partner's avatar on spatial communication in CVEs has not yet been systematically investigated.

### 2.3 Fidelity of the user's representation in CVE

To be able to assess the impact of different visual representations of the remote partner on collaboration in CVEs, it is important to identify the components that describe the avatar's fidelity. Fidelity is defined here as "the objective degree of accuracy with which real-world experiences are replicated in the virtual world" [*36*]. Two main components can be used to characterize the fidelity of the partner's avatar in a CVE:

#### 2.3.1 The visual fidelity

This component refers to the static appearance of the avatar. It can be defined as the degree of accuracy with which the different shapes of the represented user's body are reproduced in his/her virtual avatar. This component can be decomposed into three interdependent sub items.

*appropriateness of the morphology* which describes how close the selected morphology matches that of the real entity or how realistic the morphology is represented. For example, a gorilla is closer to a human shape and function than a fish to a human, or a humanoid avatar is more realistic than a human-like avatar with cartoonish appearances such as mice or dogs.

*photorealism of the virtual avatar* describes how detailed and clear the rendering of the virtual avatar is. This refers to the level of detail of the mesh and textures of the 3D model of the avatar. For instance, a photorealistic humanoid avatar has a higher resolution than a blurred humanoid avatar.

*Visual identity* defines how easy is the user able to recognize the identity of the remote partner by being exposed to his/her visual representation. This includes for instance the fidelity of the facial features (e.g., eye colors, nose shape) and body parts, the height, the weight, the clothing and the accessories. This component defines the similarity between a virtual avatar and a particular person allowing the user to identifying him/her [*37*]. In this case, a human-like avatar with cartoonish appearances but having the same haircut, and wearing glasses, cloths and accessories similar to those of a given person will have a higher fidelity level of the visual identity than a photorealistic avatar who does not resemble to this specific person. We exclude here any behavioral elements permitting to identify a person that are rather included in the next component. We exclude also all the non-visual elements such as the voice tone or haptic communication features.

#### 2.3.2 The kinematic fidelity

This component refers to the dynamic aspects of the avatar. It is defined as the degree of exactness with which the avatar replicates the movements of the represented user. It describes how the different movements (including position, velocity and acceleration) of all the visible body parts of the partner's body (arms, legs, fingers, head, eyes, facial expressions...)

are faithfully reproduced by the avatar. This includes both the deliberate movements used to communicate with the partner (e.g., pointing out an object with the arm or moving the lips to speak) and non-deliberate movements related to spontaneous human behavior (e.g., eye blinking and breathing). This is similar to the concepts of communicative realism and behavioral realism [*38*, *39*] proposed in the domain of social interaction. These concepts focus mainly on how realistic and natural does the visual representation behaves like an actual person which increases the social presence. We highlight here the importance of the movement precision which may play an important role during spatial communication. For instance, a pointing gesture performed by an avatar can be very realistic and natural. However, if it is not precise enough it may lead to misunderstandings between the partners.

The previous components can be used to analyze the degree of fidelity of a given representation of the partner in a CVE. In fact, each component has its specific requirements in order to reach a higher level of fidelity. However, depending on the collaborative task, not all the components need to reach a high level of fidelity in the same system to ensure an effective collaboration. Thus, combined with an analysis of the collaborative tasks to be supported by the CVE and its constraints, these concepts can help to determine the appropriate levels of fidelity for each component to correctly perform these tasks within the CVE.

In this paper, we focus on the impact of these fidelity components when designing the remote partner's avatar. In fact, the analysis of spatial interaction tasks in the real world and the identification of their constraints in CVE suggests the importance of the avatar fidelity for an efficient collaboration. More particularly, we explore how the movement precision can impact spatial interactions. To investigate this question, two different representations of the remote partner are compared regarding their fidelity levels. An experimental study, comparing the effects of these two representations on the collaborative performance when performing two spatial tasks in a CVE is also presented.

## 3 Materials and Methods

### 3.1 The avatars

Two types of visual representations (avatars) of the remote partner in an CVE are analyzed and compared:

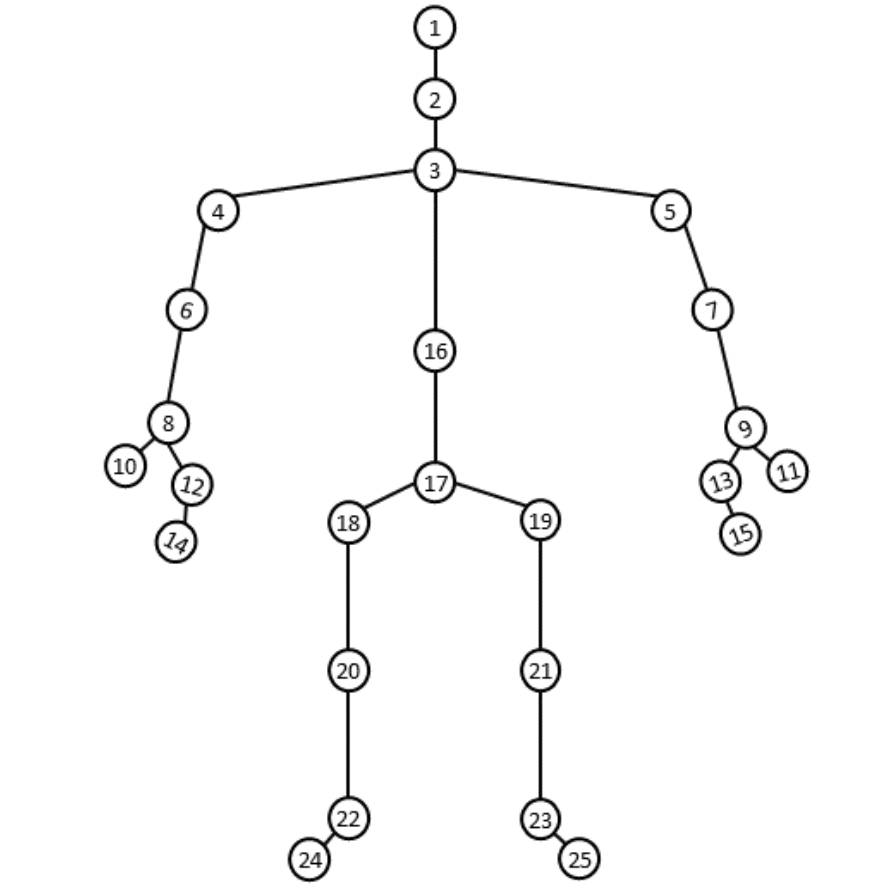
#### 3.1.1 The streamed point-cloud avatar

The first type of user's representation is based on a 2.5D point-based textured avatar built in real time through a depth sensor. This is a mixed reality representation [*40*] that allows to integrate the video of the remote user as a point-cloud in the virtual environment (Figure 1). This form of user's representation is similar to a video conferencing system. However, it offers the advantage of displaying the partner in 2.5D and in the appropriate location inside the CVE by mixing its video representation (without the surrounding real world artefacts) with the virtual environment. This representation is considered as a 2.5D avatar, because we only use one front-facing depth sensor to capture the user's video, thus not providing coverage of the rear-half of the body. It is though possible to combine several depth sensors to ensure that the partner is fully captured [*35*].

***Figure 1 : Two types of the partner's representation are compared : (left) streamed point-cloud avatar and (right) preconstructed and animated virtual avatar***

### 3.1.2 The preconstructed virtual avatar

The second type of user's representation is a 3D virtual preconstructed avatar. The avatar was modeled, rigged and skinned through a 3D modeling software (MakeHuman). This representation is then integrated into the virtual environment and animated in real time (Figure 1). The photorealistic textures of the face of the remote partner were scanned using a depth sensor. The avatar was then imported into the virtual environment and animated using the same depth sensor. For that purpose, the user's movements were captured through 25 key points and then used to animate in real time the skeleton of the preconstructed avatar (Figure 2).




| | | |
|---|---|---|
| 1. Head<br>2. Neck<br>3. Spine Shoulder<br>4. L. Shoulder<br>5. R. Shoulder | 6. L. Elbow<br>7. R. Elbow<br>8. L. Wrist<br>9. R. Wrist<br>10. L. Thumb | 11.R. Thumb<br>12. L. Hand<br>13R. Hand<br>14. L. Hand Tip<br>15R. Hand Tip |

| | |
|---|---|
| 16. Spine<br>17. Hip Center<br>18L. L. Hip<br>19. R. Hip<br>20. L Knee | 21. R. Knee<br>22. L. Ankle<br>23. R Ankle<br>24. L. Foot<br>25. R. Foot |

***Figure 2: 25 key points are used to animate the preconstructed avatar***

In the following section, the two representations are analyzed to determine their fidelity levels.

### 3.2 Avatars fidelity

To avoid any effect of the motion capture device on precision, we have used the same device which was placed in the exact same position for animating both avatars.

Both avatars have a high appropriateness of the morphology because they are photorealistic and anthropomorphic. The photorealism of the point-cloud representation has nevertheless a lower level of fidelity. When compared to the preconstructed virtual avatar representation, one can notice that some points representing the body shape of the user are invisible sometimes. We presume though that the human brain is able to reconstruct the full body representation when the avatar is moving. Moreover, only the front face of the partner is included in the virtual environment in this case. To avoid

any effect of the missing rear part in our experiment, the interaction scenario was designed so that the participants were always facing their partner, who was in turn facing the capture device during interactions in the virtual environment.

In addition, the streamed point-cloud representation has a higher level of kinematic fidelity. The user's movements are captured through an unlimited number of points that are combined to reconstruct the partner's representation in real time. On the other hand, only 25 key points of the skeleton of the user are captured and used to animate the preconstructed 3D representation. For instance, one can see the user's index finger pointing an object in the streamed point-cloud representation. On the other hand, the full hand is directed toward the object in the preconstructed virtual avatar because only the thumb and the hand tip are tracked (Figure 2). Moreover, it is possible to more or less distinguish the gaze direction using the streamed point-cloud representation while only the head direction is visible with the preconstructed virtual avatar.

Based on our analysis, we conclude that the two representations differ in their fidelity levels. The streamed point-cloud avatar has a higher level of kinematic fidelity than the preconstructed virtual avatar while the later has a higher level of visual fidelity than the former. In order to determine the effects of the fidelity difference on spatial interactions we have conducted the following user study.

### 3.3 Spatial tasks

In the following experiment, we chose to focus on two representative spatial communication tasks. The first one consists of an operator guiding verbally his/her partner avatar towards selecting and manipulating a virtual object of interest. This can permit to observe the operator's perspective-taking when he/she verbally describes the position of the object of interest according to the partner's position/point of view. Therefore, we will study the effects of the partner's avatar fidelity on the ability of the participants to successfully describe object positions to their partner according to this avatar.

The second spatial communication task consists of interpreting pointing gestures performed by a partner to designate an object of interest. This can show the importance of the avatar movement precision when performing pointing gestures. We will therefore study how successful will be the participants in interpreting a pointing gesture performed through the arms of the partner's avatar and therefore identifying the correct objects pointed out.

### 3.4 Research hypotheses

The previous analyses of the avatars fidelity and the collaborative spatial tasks give us indications on how each user avatar could impact the collaboration during these tasks. By comparing the two types of partner's avatar when performing these tasks, our main hypothesis is that the streamed point-cloud partner's avatar would improve spatial communication. This is expected because it has a higher level of kinematic fidelity and despite his lower level of visual fidelity. This will be investigated through the following working hypotheses:

H1: The streamed point-cloud avatar would help the user to verbally guide his/her partner towards selecting an object of interest. This is expected to decrease the guidance time, and to decrease the partner's errors rate when selecting the target objects in task 1.

H2: The streamed point-cloud avatar would facilitate the interpretation of pointing gestures performed by the partner to designate an object of interest. This is expected to decrease the time needed to correctly recognize the pointed rows and to decrease the recognition error rates in task 2.

H3: The better spatial communication expected with the streamed point-cloud avatar would increase the subjective feeling of presence, co-presence and social presence with this partner within the CVE. As suggested by previous research, a higher kinematic fidelity increases social presence [*24*, *33*, *34*, *35*].

H4: The streamed point-cloud avatar has a lower level of photorealism. Therefore, the participants would rate its visual realism lower than that of the preconstructed avatar.

H5: The streamed point-cloud avatar would decrease the perceived strangeness when looking at this avatar despite having a lower level of photorealism. Indeed, Slater and Steed [*41*] argue that the more photorealistic an avatar appears to the user the higher the expectation of behavioral realism. Recent studies suggest also an uncanny valley effect associated with photorealistic preconstructed avatars with low behavioral realism [*33*, *34*]. Therefore, the higher kinematic fidelity of the streamed point-cloud avatar is expected to increase its behavioral realism whose actions will be perceived as closer to an actual person than those of the preconstructed avatar.

H6: The better spatial communication expected with the streamed point-cloud partner's avatar would encourage the participants to use the partner's egocentric reference frame instead of an exocentric reference frame (relative position of objects). Indeed, perspective-taking would be facilitated with a higher kinematic fidelity.

### 3.5 Participants

Twenty-two participants (13 males, 9 females) from the local campus community (students and staff) were enrolled in this study (N=22). The mean age was 28.7 (SD = 9.5; min = 20; max = 50). All of them had normal or corrected to normal vision. All of them had a previous experience with video games (including smartphone games) with 16 of them playing video games frequently (at least once a week). All of them had also at least one previous experience with VR technologies with only 4 of them using VR devices frequently (at least once a week). The experimental protocol was approved by the institutional ethics committee of Université Paris Saclay (CER Paris Saclay) prior to enrolling any human subject. An informed written consent was also obtained from all the subjects involved in this study prior to their participation.

### 3.6 Experimental design

Similar to previous studies [*12*, *27*, *22*, *33*, *34*], and in order to provide a consistent behavior to the participants, all of them performed the two experimental tasks in collaboration with the same confederate (an experimenter). The study design was a between-subjects design with one experimental factor (avatar type) with two conditions: streamed point-cloud avatar (SPA) Vs preconstructed virtual avatar (PVA). Therefore, the only difference between the two conditions was the type of the avatar used to represent the partner in the virtual environment. The 22 participants were randomly assigned to one of the two groups (SPA group or PVA group) with 11 participants in each group. One female participant in the SPA group had to be excluded from the data set because she felt uncomfortable wearing the HMD and had to stop

the experimental session. At the end, the collected and analyzed data concerned 11 participants in the preconstructed virtual avatar condition (N=11) and 10 participants in the streamed point-cloud avatar condition (N=10).

## 3.7 Experimental setup

### 3.7.1 Experimental room

The participants and their confederate were collocated in the same experimental room. They were able to talk to each other (with restrictions, c.f. experimental tasks section) but were not able to (physically) see each other.

### 3.7.2 Collaborative virtual environment

The designed CVE consisted of a delimited space displayed at scale 1:1 for both users. In this environment, the participant was located on a small (2x2m) esplanade above the experimenter (Figure 3**Erreur ! Source du renvoi introuvable.**). Both users were facing each other and saw the virtual environment from a different perspective. They were allowed to walk only inside a delimited space. The walking space of the participant was delimited in the virtual environment by the edges of the esplanade. The participant was able to see the experimenter's avatar in the virtual environment (the type of the displayed avatar depended on the experimental condition, c.f. Experimental design section) but the experimenter did not see any avatar of the participant. The experimenter had a 4 (columns) x 2 (lines) virtual grid displayed in front of him. A similar grid was placed on the right hand side of the participant. Neither of the partners was able to see the other's grid. In addition, the experimenter's avatar was surrounded with 10 virtual white cubes (each having a volume of $0.2m^3$ and positioned 0.2m above the floor). The participant saw these same virtual cubes with 5 different colors (two cubes of each color). Finally, a sphere was placed on the left hand side of the participant. The color of this sphere indicated to the participant the color of the cubes of interest (c.f. experimental tasks section).

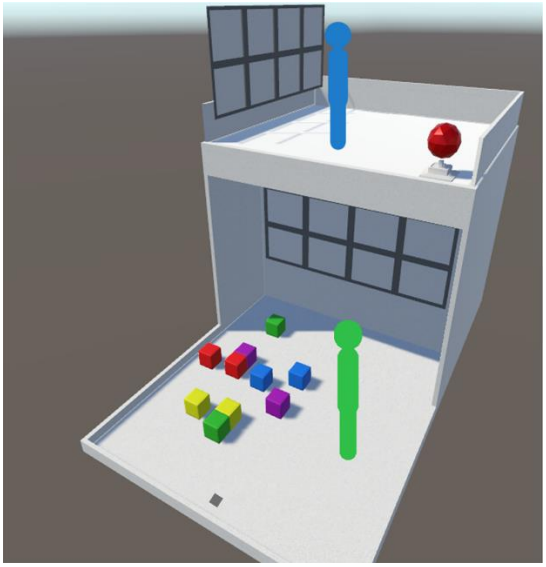


***Figure 3 : Overview of the collaborative virtual environment: the participant (in blue) was placed upstream the experimenter (in green)***

### 3.7.3 Apparatus

The participants were immersed in the virtual environment using a VR headset (HTC Vive; Figure 4**Erreur ! Source du renvoi introuvable.**). It has a resolution of 1080 x 1200px and a refresh rate of 90fps for each eye. The participants used

also two Vive Controllers to interact with the virtual environment and were in a direct voice contact with their confederate (the experimenter). The experimenter was immersed through a 4-sided CAVE (2x2x2.5m) and used a PlayStation Move type controller to interact with the virtual environment (Figure 4**Erreur ! Source du renvoi introuvable.**). Head and hand tracking of the experimenter were performed using 4 OptiTrack Prime 13 cameras. We have preferred using the CAVE as the experimenter's display in order to facilitate his body tracking using the Kinect for the avatar animation and point-cloud streaming. Yet, both displays are highly immersive to ensure an immersion and interaction symmetry between the partners.

***Figure 4: Experimental setup (left) the participant wearing the HMD (right) the experimenter inside the CAVE***

The HTC vive workstation has an Intel I7 processor, 16GB RAM, Nvidia GeForce GTX 1080 graphics card under Microsoft Windows 10. The CAVE workstation has an Intel I7 processor, 16GB RAM, and an Nvidia Quadro K5000 graphics card under Microsoft Windows 7. It uses also 4 full HD projectors.

For both conditions, a Microsoft Kinect for Windows V2 was used and connected to the HTC vive workstation using the Kinect Adapter for Windows [1]. This sensor has a depth resolution of 512 x 424px with a FoV of 70.6° x 60° and a framerate of 30fps. Its color camera has a resolution of 1920 x 1080px with a FoV of 84.1° x 53.8° and a framerate of 30fps. Its operative measuring range is from 0.5m to 4.5m. After some preliminary tests, the sensor was placed in the same exact position for both conditions: on top of the front screen of the CAVE, 2.3m above the floor, facing downwards at about 35°). This position permitted to capture the stream from the same perspective as the participants and ensured that all body parts of the experimenter were always inside the FoV of the sensor during the experiment.



[1] https://developer.microsoft.com/en-us/windows/kinect/

The motion capture and animation of the preconstructed virtual avatar were performed using the Kinect sensor . For that purpose, the device extracts a standard video stream and a depth map and tracks the user's skeleton using 25 key points (Figure 2).

The colored point-cloud was generated using the same Kinect sensor using a method similar to that proposed by Nahon et al. [*40*]. The point-cloud was directly extracted from the two video streams of the Kinect (depth and color Data). The data was then streamed to the localhost client application in order to reconstruct the user's avatar and display it in real time inside the virtual environment.

Both avatars were rendered on the HMD at a framerate of 30fps. For both conditions, a calibration phase was necessary before each experimental session to calculate the projection matrix between the Kinect sensor and the OptiTrack reference systems. This ensured the experimenter's avatar was correctly located within the virtual environment.

#### 3.7.4 Software

The virtual environment was modeled using Blender and exported into the Unity 2018 applications implemented with C#. They were installed on both workstations which were connected through a local network. The network communication between the two Unity applications was provided by the Unity Multiplayer and Networking API[2]. The CAVE display and interactions were implemented using the MiddleVR for Unity plugin[3] (only on the experimenter's side). The HTC Vive integration was provided using the SteamVR plugin for Unity[4] (only on the participants' side). The motion capture data for the preconstructed virtual avatar was integrated into the Unity application using the MS-SDK V2 and the Kinect V2 Unity package[5] (only on the participants' side).

### 3.8 Experimental tasks

During the experiment, the participants were asked to perform together with their confederate, an experimental trial based on two successive collaborative tasks. The two experimental tasks were designed based on the two spatial communication scenarios discussed above (c.f. spatial tasks section).

#### 3.8.1 Task 1

The first task carried out by the participants (Figure 5) consisted in verbally guiding the experimenter towards selecting two cubes among the 10 cubes surrounding him (2 cubes to select and 8 distractors). The initial position of each cube was random and was changed at the beginning of each trial. They were also positioned 0.2m above the floor to ensure the experimenter's interaction device is correctly tracked when manipulating them. The cubes were displayed to the participant with 5 different colors (two green, two red, two yellow, two blue and two purple cubes; Figure 3). On the other hand, the experimenter only saw these cubes in a white neutral color (Figure 6). To ensure consistency across all participants, the experimenter's starting position was always the same. This position was marked out on the floor using a virtual grey square (in the middle, 1.8m from the front wall of the CAVE, Figure 4). At the beginning of each trial, the



[2] https://docs.unity3d.com/Manual/UNet.html
[3] https://www.middlevr.com/middlevr-for-unity/
[4] https://assetstore.unity.com/packages/tools/integration/steamvr-plugin-32647
[5] https://rfilkov.com/2014/08/01/kinect-v2-with-ms-sdk/

color of the two cubes to be selected was displayed to the participants on the sphere placed on their left hand side (not visible by the experimenter). Then, they had to identify the two corresponding cubes among those surrounding the experimenter and to verbally guiding him to select them (once at a time). Following these verbal instructions, the experimenter had to physically move towards the cubes and to select them using his virtual hand collocated with his interaction device. It is to be noted that the only verbalizations used by the experimenter to interact with the participant during this task were the use of deictic expressions (this/that) when selecting the cubes. Once selected, the cube was moved towards a validation area (represented by a green point-cloud sphere; Figure 6). If the cube was correctly selected, it disappeared and a "success" sound was displayed for both users, otherwise it came back to its original position and a "failure" sound was displayed for both users (so that no further verbal interactions are needed). The first task ended when the second cube was correctly selected and moved to the validation area.

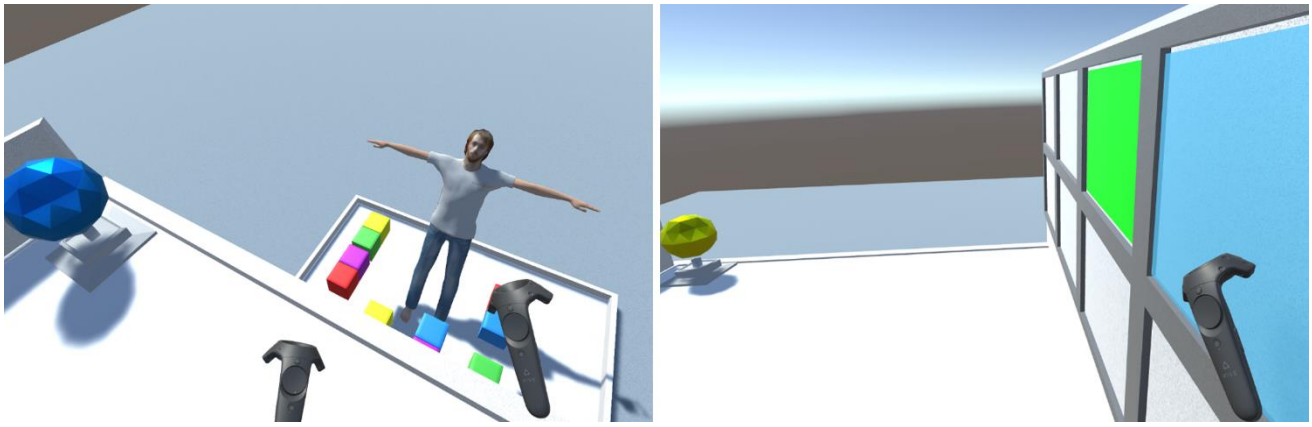


***Figure 5: Participant's views of the collaborative environment: (left) task 1 He/she saw the colored cubes and the color of the sphere indicated the color of the cubes to be selected. (right) task 2: He/she used the interaction device to select a row on the grid.***

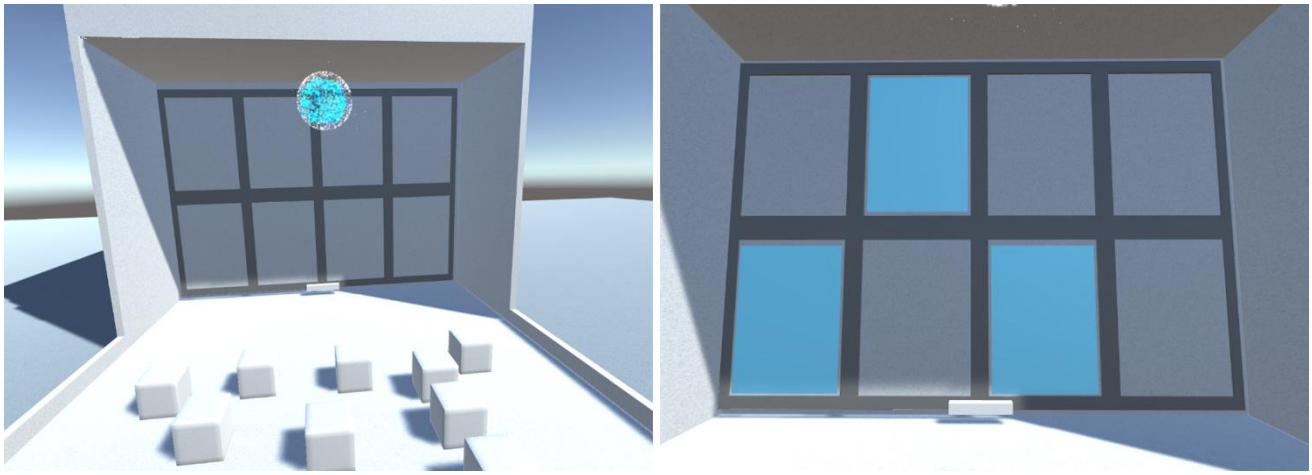


***Figure 6: Experimenter's views of the collaborative environment. (left) task 1: he saw the cubes in white neutral color. (right) task 2: he saw the 3 rows pattern and had to target the middle of each row during the pointing gesture.***

### 3.8.2 Task 2

The second task consisted in identifying a 3-rows pattern. The 3 corresponding rows had to be selected among the 8 rows of the virtual grid. Once the task 1 ended, the experimenter had to move back to his initial position and then 3 rows were displayed in blue (Figure 6) on the grid in front of him (not visible by the participant). He then had to use only pointing gestures with his right arm to indicate to the participant the corresponding rows to select on his/her own grid (one row at a

time). The experimenter had to point-out each row while verbally announcing the corresponding row number starting from the bottom left one and ending with the top right one. No additional verbal instructions were given by the experimenter. For consistency across all participants, the experimenter had to always target the middle of the pointed row with his index finger. He was also always standing at the starting position marked out on the floor by a small virtual grey square (Figure 4Figure 4). The participant was allowed to physically move back and forth to observe the pointing gesture and then select the corresponding row on his/her own grid. Once the participant's hand was in contact with the grid, its color turned to blue indicating that it can be selected. He/she was then able to use the vive controller trigger button to confirm the selection. Once a row was selected, its color turned to green (Figure 5). Once a row was selected, the participant had to ask the experimenter to point out the next row. Once the participant had selected the third row, a success/failure sound was displayed for both users and a green/red color was displayed during 2 seconds on both grids depending on the success/failure of the 3-row pattern selection. In case of failure, the process was repeated until the correct pattern was entered indicating the end of the experimental trial.

The experimental trial (including both tasks) was then repeated 5 times for each participant. The participants were instructed to perform the two tasks as quickly as possible and to limit the number of errors.

### 3.9 Experimental procedure

The experimental procedure started with signing the consent form and filling in a demographics questionnaire. The participants performed then a mental rotation test to evaluate their spatial abilities [*42*]. This test is used to control for any individual differences regarding the spatial abilities of the participants. After that, the participants wore the HMD and started the training and familiarization phase. In this phase, the participants were immersed inside a virtual environment similar to the experimental one. They were allowed to navigate inside this environment and to test the rows selection. They were also allowed to observe environment from the experimenter's perspective in order to have a better understanding of the tasks to be performed together. However, no avatar was displayed at this time. During this phase, audio instructions were displayed to explain the experimental tasks and the interactions to the participants. They were able to repeat the instructions until they felt correctly understanding them. Once the tutorial was finished, the actual experimental session started. The participants were explicitly asked to stop the experiment if there was any sign of discomfort. Once the participants had performed the five experimental trials, they removed the HMD and were asked to fill in a post experimental questionnaire before leaving the experimental room.

### 3.10 Measurements and data analyses

The collaborative performance was evaluated through the completion time and the errors rate of the two experimental tasks. The time calculation for the first task started once the color of the target cubes was displayed on the participant's sphere and ended once the second cube was correctly selected by the experimenter. The time calculation for the second task started when the pattern was displayed on the experimenter's grid and ended when the correct pattern was entered by the participants on their own grid. The number of the incorrectly selected cubes and incorrectly entered patterns was used to measure the errors rate. All the data was recorder automatically on text files by the Unity networked applications. For each participant, an average score was then calculated over the five trials for the four variables.

In addition to objective measurements, we have collected subjective data to evaluate the feeling of presence, co-presence and social presence of the participants with their partner. The questions (Q1-Q11) are inspired from questionnaires used in peer-reviewed international publications [*43*, *25*] and were translated into French (authors' own translation). We have also proposed other questions to serve the purpose of our own study (Q12-Q15). All the questions were rated on a 5-point Likert scale. The items are available in Table 1.

***Table 1: Items of the post experimental questionnaire***

| ID | Question |
|---|---|
| Q1 | To what extent did you feel immersed in the environment you Saw? N -T |
| Q2 | I tried to create a sense of closeness between us N C |
| Q3 | My partner tried to create a sense of closeness between us N C |
| Q4 | To what extent was this like you were in the same room with your partner? N S |
| Q5 | To what extent was this like a face-to-face meeting? N S |
| Q6 | I felt isolated from my partner |
| Q7 | My actions were strongly conditioned by the instructions of my partner |
| Q8 | The actions of my partner were strongly conditioned by my instructions |
| Q9 | I felt my partner tried to help me |
| Q10 | I tried to help my partner |
| Q11 | I felt understood by my partner |
| Q12 | The virtual body of my partner helped me describing the position of the objects to grasp |
| Q13 | The virtual body of my partner helped me to enter the right pattern |
| Q14 | It was easy to locate my partner in the virtual environment |
| Q15 | It was easy to interpret the pointing gestures of my partner |

We have also asked the participants to rate the "visual realism" of their partner's avatar on a 3-points scale (not realistic, moderately realistic or very realistic). They were also asked to indicate whether they felt any "strangeness" looking at any body parts of the experimenter's avatar (yes/no answer regarding the strangeness of the face, hair, torso, arms, hands, legs, and feet). Finally, they were asked to indicate the reference frame they have mainly used when describing objects position to their partner (one choice among: partner's full body position, partner's arms position, tor he relative positions of the virtual cubes).

We have used a confidence level of 95% for all our statistical analyses. As a consequence, a result is considered significant when $p < 0.05$. All data analyses were performed using SPSS v.21.0 (IBM Corp., Armonk, NY, USA) with the appropriate statistical tests.

## 4 Results

In this section we first present the statistical tests used to analyze the data and then report both descriptive and inferential statistical analyses of the collected data.

### 4.1 Statistical tests

The first step of data analyses consisted in checking the normality assumption of the collected data to determine whether parametric tests can be used. For that purpose, the Shapiro-Wilk test was run on the Vandenberg mental rotation test (VMRT) data, the completion time data, and the errors rate variables. The results indicate that only the VRMT and the errors rate for task 2 variables follow a normal distribution. Therefore, the non-parametric Mann-Whitney U was used for comparing the means for all the other dependent variables. Moreover, the Levene's test shows that the equality of variances is not assumed for the errors rate for task 2 variable. Therefore, the Welch's parametric t test was used to compare the means for this variable. We have also used the Spearman's non parametric correlation test to investigate whether the VMRT scores have a significant correlation with the different dependent variables for both conditions. In addition, the non-parametric Mann-Whitney U test was used to compare the mean scores of the subjective questionnaire data (ordinal data). Finally, the non-parametric Chi-square ($\chi^2$) test of independence was used for reporting the relationship between the nominal variables ("visual realism", "observed strangeness" and "reference frame") and the avatar type.

### 4.2 Mental rotation test

The mean VRMT score was 16.20 (SD = 4.41) for the streamed point-cloud (SPA) condition and 14.81 (SD = 3.28) for the preconstructed virtual avatar (PVA) condition. The student's independent samples t-test shows no significant difference in VMRT scores between the two groups [$t_{(19)}$=0.819, p=0.42, d=0.35]. The Spearman's correlation tests show no significant correlation between the VMRT scores and the completion time for task 1, the completion time for task 2, the errors rate for task 1 and the errors rate for task 2 (Table 2).

***Table 2 : correlation tests for the VMRT scores***

| Variable | SPA condition | | PVA condition | |
|---|---|---|---|---|
| | Spearman's r | P-value | Spearman's r | P-value |
| Completion time task 1 | 0.091 | 0.802 | -0.023 | 0.947 |
| Completion time task 2 | 0.006 | 0.987 | 0.391 | 0.235 |
| Errors rate task 1 | 0.524 | 0.120 | -0.139 | 0.683 |
| Errors rate task 1 | 0.251 | 0.485 | 0.412 | 0.207 |

### 4.3 Completion times

The non-parametric Mann-Whitney U tests show a significant effect of the avatar type on the mean completion times of task 2 [U=22, p=0.02, r =0.25]. No significant effect of the avatar type was found for the mean completion time of task 1 [U=33, p= 0.132, r = 0.11]. In both tasks, the participants performed the tasks faster in the SPA condition (19% less time for task 1 and 56% less time for task 2; Figure 7).

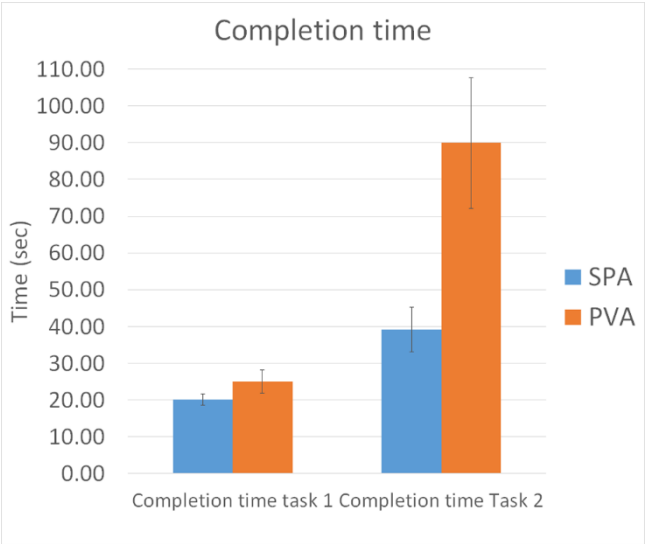



***Figure 7: Mean completion times for task 1 and task 2 (error bars represent the standard error)***

### 4.4 Error rates

The non-parametric Mann-Whitney U test shows a significant effect of the avatar type on the errors rate for task 1 [U=23.5, p=0.011, r=0.31]. The Welch's t-test shows a significant effect of the avatar type on the errors rate for task 2 [$t_{(8.84)}$=10.82, p=0.013, d =1.30]. The participants performed both tasks with fewer errors in the SPA condition than in the PVA condition (90% less errors for task 1 and 81% less errors in task 2; Figure 8).

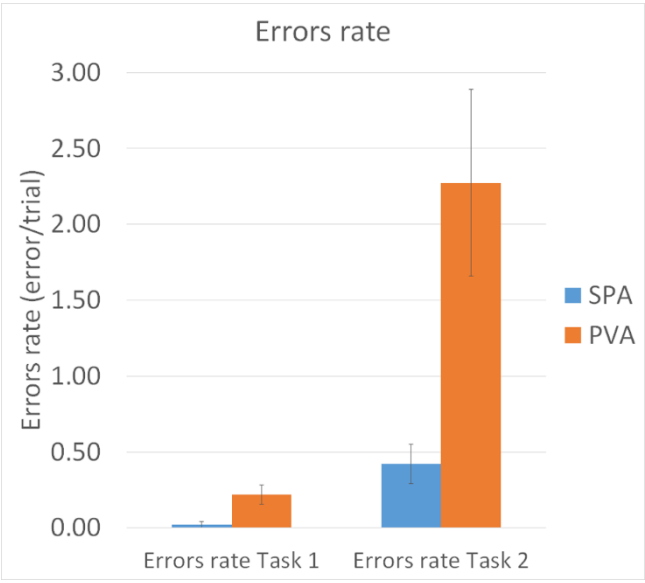



***Figure 8: Mean error rates for task 1 and task 2 (error bars represent the standard error)***

### 4.5 Subjective measurements

The non-parametric Mann-Whitney U tests show a significant effect of the avatar type on the participant's mean scores for questions Q1, Q8, Q9, Q12, Q13, Q15 (Table 3). No significant effects were found for the other questions (Table 3).

For questions Q1, Q8, Q12, Q13 and Q15, the mean scores were significantly higher in the SPA condition whereas for question Q9, the mean scores were significantly higher in the PVA condition.

***Table 3: descriptive and statistical analyses for the questionnaires***

| | Mean scores (SD) | Mean scores (SD) | | | |
|---|---|---|---|---|---|
| | SPA condition | PVA condition | Mann-Whitney U | P-value | Effect size (r) |
| Q1 | 4.8 (0.42) | 4.27 (0.65) | 30.000 | .043 | 0.19 |
| Q2 | 3.6 (0.69) | 3.64 (0.50) | 51.000 | .750 | 0.00 |
| Q3 | 3.3 (1.05) | 3.55 (0.69) | 49.500 | .669 | 0.01 |
| Q4 | 4.3 (0.48) | 4.64 (0.67) | 35.000 | .111 | 0.12 |
| Q5 | 2.9 (1.19) | 3.09 (1.04) | 50.000 | .715 | 0.01 |
| Q6 | 2.1 (0.73) | 1.82 (0.98) | 41.500 | .311 | 0.05 |
| Q7 | 4.3 (0.82) | 3.91 (0.83) | 40.500 | .269 | 0.06 |
| Q8 | 4.5 (0.84) | 3.55 (1.13) | 23.000 | .017 | 0.27 |
| Q9 | 2.5 (1.26) | 4.09 (0.83) | 16.500 | .004 | 0.40 |
| Q10 | 4.0 (1.24) | 4.18 (0.75) | 55.000 | 1.000 | 0.00 |
| Q11 | 4.3 (1.05) | 4.73 (0.47) | 45.000 | .397 | 0.03 |
| Q12 | 4.7 (0.48) | 3.64 (1.36) | 23.500 | .014 | 0.29 |
| Q13 | 4.4 (0.84) | 3.36 (1.12) | 25.000 | .027 | 0.23 |
| Q14 | 4.6 (0.69) | 4.36 (0.50) | 40.000 | .231 | 0.07 |
| Q15 | 3.6 (0.84) | 2.73 (0.65) | 21.000 | .009 | 0.32 |

In addition, the results show that 8 participants (80%) in the SPA condition rated the partner's avatar as moderately realistic, while only 2 of them (20%) found it very realistic. On the other hand, 7 participants (63.6%) rated the partner's avatar as moderately realistic in the PVA condition, 3 of them (27.3 %) found it very realistic, and only one of them (9.1%) found not realistic. The non-parametric Chi-square ($\chi^2$) test of independence for the nominal variable "visual realism of the avatar" indicates a non-significant relationship between the avatar type and the perceived avatar realism [$\chi^2(2) = 1.22$, $p = 0.54$, Cramér's $V = 0.24$].

Moreover, 5 participants (45.5%) in the PVA condition indicated that they felt some discomfort (strangeness) looking at the arms of the partner's avatar, 4 of them (36.4%) looking at its head and face and 2 of them (18.2%) did not feel any discomfort looking at any body part of the avatar. On the other hand, only 2 of the participants (20%) in the SPA condition indicated that they felt some discomfort looking at the head and face of the partner's avatar while 8 of them (80%) did not feel any discomfort looking at any body part of the partner's avatar. The non-parametric Chi-square ($\chi^2$) test of independence for the nominal variable "strangeness of body parts" shows a significant relationship between the avatar type and the perceived strangeness of the avatar's body parts [$\chi^2(2) = 9.24$, $p = 0.01$, Cramér's $V = 0.66$].

Finally, 9 participants (90%) in the SPA condition indicated having used the full-body position of their partner's avatar as the main reference frame when describing the cubes position to their partner while only one of them (10%) reported using the position of the avatar's arms. On the other hand, only 5 of the participants (45.5%) in the PVA condition indicated having used the full-body position of their partner's avatar as the main reference frame while 6 of them (54.5% ) reported using the position of the avatar's arms. No participant has reported using the relative position of the virtual cubes as the main reference frame. The non-parametric Chi-square ($\chi^2$) test of independence for the nominal variable "main reference frame" shows a significant relationship between the avatar type and the reference frame used by the participants [$\chi^2(2) = 4.67$, $p = 0.03$, Cramér's $V = 0.47$].

## 5 Discussion

In this section we discuss the results of our experimental study and their implications on the design of CVE.

### 5.1 Mental rotation tests

The results indicate no significant difference between the mental rotation test scores of the two experimental groups and no significant correlation between these scores and the objective dependent variables. As a consequence, we consider that both groups were comparable in terms of mental rotation abilities. It is to be noted that although not reported in this manuscript, none of the gender, gaming experience and VR experience control variables have shown a significant effect on the dependent variables.

### 5.2 Verbal guidance task

The first spatial communication aspect that we have investigated in our experiment was whether the participants were able to correctly guide their confederate towards an object of interest based of the avatar type.

The results indicate that the use of the streamed point-cloud avatar to represent the remote confederate has significantly improved the verbal guidance in task 1. Indeed, the errors rate was significantly lower in this condition. However, the completion time was not significantly improved when using the point-cloud avatar. This can be explained by the simplicity of this task. Indeed, the average completion time for this task was relatively low with 20.13 seconds (SD = 4.72) for the streamed point-cloud avatar condition and 25.00 seconds (SD = 10.25) for the preconstructed virtual avatar condition. The time increase in the preconstructed avatar condition although not significant, can be attributed to the increase in the errors rate. This suggests some difficulties of the participants when verbally guiding their confederate towards selecting the right cubes when the preconstructed avatar was used.

The subjective evaluation supports this finding and provide more indications for interpreting the increase in the errors rate. Indeed, the participants felt that their instructions had significantly more impact on the actions of their confederate when the point-cloud avatar was used (question Q8). This suggests that they found their verbal instructions to more useful to their confederate and helped him to find the right targets. Surprisingly, no significant difference was observed neither for the question regarding the feeling of being understood by the partner (Q11) nor for that regarding the need of helping the partner (Q10). This also may be attributed to the simplicity of the task and the fact that all the participants succeeded in completing it in a relatively short time.

While there was no significant difference regarding the localization of the partner within the virtual environment, the results indicate that the participants felt that the point-cloud avatar helped them better describe the position of the target cubes (question Q12). This suggests that both types of avatars were useful to localize the partner inside the virtual environment. However, the avatar with the higher level of kinematic fidelity was even more useful when verbally describing the objects position. The reported preferred reference frame provides more indications on the difficulties encountered by the participants in the preconstructed avatar condition. In fact, 54% of them reported using the arms of the partner's avatar as the main reference frame. This can be explained by the fact that when verbally guided, the experimenter had to select the cubes by physically moving his body and his arm towards them. The participants used then the real-time body and arm movements as a feedback from the confederate to their verbal guidance. When the kinematic fidelity of the arm movements was low (preconstructed avatar condition), the participants may have had some difficulties to determine whether the experimenter was about to grasp the correct object. Therefore, they had to direct their attention towards the hand and arm movements instead of the whole body movement. This is supported by the increased number of participants (45%) reporting feeling some strangeness looking at the avatar's arms suggesting that the participants found the arm movements in this condition less precise. On the other hand, the higher arms kinematic fidelity in the point-cloud avatar condition encouraged the participants (90% of them) to use the whole avatar body as a frame of reference with a higher success rate in this case. Moreover, none of them have reported any strangeness looking at the avatar's arms. H5 is then validated for task 1. This further supports the fact that a higher kinematic fidelity is more useful during the guidance task.

It is to be noted that in both conditions, all the participants reported using mainly the partner's egocentric reference frame (either his arm or his whole body) instead of an exocentric one (the virtual cubes for instance), or their own egocentric reference frame. H6 is then rejected. This indicates that the presence of either avatar encouraged them for taking the perspective of their partner during the verbal guidance. This is consistent with the literature that suggests that when guiding a partner during a spatial task, operators are more likely to take the manipulator's perspective in order to reduce their partner's mental load [*9, 44*] which can be interpreted as a sign of a better collaboration.

To summarize, the findings suggest that hypothesis H1 is partially verified. The use of the streamed point-cloud avatar has improved the localization of the partner and has more particularly helped the participants to have a more precise feedback on the partner's actions. These cues on the partner's activities facilitated spatial communication during the guidance task which is consistent with the literature [*22*]. The results of the subjective evaluation suggest that this may be attributed mainly to the higher kinematic fidelity of this avatar. Finally, the presence of the partner's avatar, regardless the level of its kinematic fidelity encourages the perspective taking, which also improves spatial communication.

### 5.3 Interpreting pointing gestures of the partner

The second aspect of spatial communication investigated in our study was whether the participants were able to correctly find the virtual objects pointed out by their partner through his virtual arm.

The results indicate that the collaborative performance during this task was significantly improved when the confederate was represented using the point cloud avatar. In fact, both the errors rate and the completion time were significantly lower in this condition. Again, the time increase in the preconstructed avatar condition is mainly due to the increase in the errors

rate which was costlier in this task. The increase in the errors rate can be related to the difficulties encountered by the participants when interpreting the pointing gestures of their confederate.

The subjective evaluation supports also this finding. The participants found the point-cloud avatar to be more helpful to infer the right pattern (question Q13) and facilitated the understanding of the experimenter's pointing gestures (question Q15). This suggests that the higher kinematic fidelity associated with the point cloud avatar permitted a better interpretation of the pointing gestures. On the other hand, the participants in the preconstructed virtual avatar condition felt that their partner tried more often to help them than the participants in the other condition (question Q9). This reflects their difficulties to interpret the gestures of the experimenter. This is associated with an increased feeling of strangeness looking at the preconstructed virtual avatar's arms suggesting a lower precision of the arm movements. Therefore, H5 is also validated for task 2. This further suggests that the lower kinematic fidelity resulted in an increased difficulty to interpret the pointing gestures.

Some of the participants commented after the experiment that the orientation of the preconstructed avatar's arm had disturbed them, more particularly during the pointing gestures. We suspect this to be due to the direction of the pointing gestures performed toward the grid. In fact, the grid was displayed for the experimenter in the front screen of the CAVE which means that his pointing gestures were directed towards the Kinect sensor. In this case, the right elbow, right wrist, right hand, right thumb and right hand tip key points were almost aligned on the same vector (directed toward the sensor) causing a possible occlusion issue. This may have altered the tracking of the hand/arm position during the pointing gesture and resulted in imprecise movements as observed in previous studies using the Kinect [*32*]. This in turn, may have made the interpretation of the pointing direction difficult for these participants. This may also be applied, but with a lesser extent to task 1 when the experimenter was selecting the cubes. However, additional measurements on the quality of the tracking in these conditions are needed to confirm this hypothesis.

These results validate our second hypothesis H2 that the use of the streamed point-cloud avatar helps the users in correctly interpreting the spatial information transmitted by the partner.

Finally, there was no difference regarding the feeling of participants that their actions were strongly dependent on their partner's instructions (question Q7). This can be explained by the fact that the participants in both conditions felt that, regardless the results of their actions (success or failure), these actions were dependent on their own interpretation of the pointing gestures performed by the confederate.

### 5.4 Additional findings

In addition to the previous findings, the results of the present study indicate a higher sense of presence and immersion (question Q1) when the streamed point-cloud avatar was used. This can be explained by the fact that a better collaborative performance in the virtual environment had a positive impact on the feeling of immersion and presence of the participants.

Moreover, there was no significant difference in the responses to the questions regarding the copresence and social presence (questions Q2-Q6). This can be explained by the fact that the participants and their partner were physically in the same room and thus did not feel interacting with him remotely. Therefore, the hypothesis H3 is partially validated.

Regarding the visual realism, the results indicate no significant difference between the two conditions with most participants rating both avatars as moderately realistic. H4 is then rejected. This can be explained by the fact the lower kinematic fidelity associated with the preconstructed avatar led to decrease the perceived visual realism of this avatar despite its higher photorealism level. This also suggests that a moderate visual realism can be sufficient to correctly accomplish collaborative spatial tasks when the level of kinematic fidelity is high enough. However, further investigations are needed to better understand the relationship between visual and kinematic fidelity and the impact of each component on spatial interactions.

## 6 Conclusions

In this section we summarize the contributions of this work and the implications for the design of spatial collaborative virtual environments, the current limitations and directions for future work.

### 6.1 Design implications

In this work, we have discussed two components that characterize the fidelity of the partners' avatar in immersive CVEs. We have also presented an experimental study to compare the effect of two types of partner's avatar which differ in their fidelity levels, on performing two spatial tasks in a CVE. Our general hypothesis was that the point cloud avatar which has a higher kinematic fidelity and a lower visual fidelity would improve the spatial collaborative performance and co-presence with the remote partner when verbally guiding the partner and when interpreting the partner's pointing gestures.

The results of our study confirm that the use of this avatar significantly improves the collaborative performance and leads to a better perception of the partner's actions when carrying out the two spatial tasks. While previous studies have suggested that the avatar facial anomalies are more disturbing than body motion errors to convey emotional content [*45*], our findings suggest on the contrary that the fidelity of the body movements of the avatar is more important in spatial collaborative tasks. Other researchers suggested also that photorealism is less important than behavioral fidelity [*46*]. However, we argue that this needs to be linked with the constraints collaborative task to be performed.

Our recommendation for the design of CVE supporting spatial tasks is to increase the level of kinematic fidelity of the partner's avatar in order to improve spatial communication and thus the collaborative performance. The presence of an animated avatar with a moderate visual realism is important to facilitate the localization of the partner. Moreover, a higher kinematic fidelity of the arm and hand movements are more particularly important for correctly communicating spatial information and for giving a more efficient real-time feedback on the partner's activities and nonverbal behavior during spatial tasks.

### 6.2 Limitations and future work

There are some limitations to the present study. First, we have used a depth sensor to animate the preconstructed virtual avatar. There are currently several technologies and techniques offering better motion capture options [*27*]. Using those options would have increased the kinematic fidelity of the preconstructed avatar. We have chosen though to use the Kinect for animating both avatars for several reasons. First, it is a simple-to-use and an inexpensive option. In addition, it permits to avoid any bias related to the device precision and calibration issues, which can differ from one device to another. Finally, it does not require placing additional specific markers on the user's body which may also be a bias for

such a study. Indeed, when one want to capture natural behavior of users, it helps if they are not encumbered by excessive markers and restrictions on movements [*47*]. In future studies, we plan though to compare different motion capture technologies to determine the most suited for increasing the kinematic fidelity.

The second limitation is also related to the use of the Kinect sensor. In fact, we have used only one front-facing sensor to capture the user's video, thus not providing coverage of the rear-half of the partner's body in the streamed point-cloud condition. A single Kinect can achieve a partial 3D reconstruction of the user in the plane it is pointing towards. Therefore, the experimenter had to always face the device so that the participants were able to see his front face. Again, we have decided to use only one sensor for consistency across the two experimental conditions. Our software offers currently the possibility to combine several depth sensors to ensure the user is fully captured. In future studies, we can use this setup in order to give more freedom of movement to the user.

The two compared avatars differ both in their levels of kinematic and visual fidelity. Although the subjective measurements suggest that the difference in performance is mainly attributed to the difference in kinematic fidelity, the effect of the visual fidelity difference is still unclear. In future experiments, we plan to isolate each component to better understand how each of them can affect the communication and collaboration in spatial collaborative tasks.

The verbal guidance task was too easy to perform. As a consequence, the difference in performance between the two experimental conditions was very low (although significant for the errors rate). In the future, we could use more challenging tasks by adding more distractors and by reducing the size of the objects to be selected. This may further demonstrate the advantage of increasing the avatar's kinematic fidelity when performing spatial guidance tasks.

Our experiment included a small number of participants. This resulted in collected data not following a normal distribution. This can be a limitation for the generalization of the results. Nevertheless, the use of non-parametric tests permitted to show significant effects in our experiment. In the future, we plan to run the experiment with a larger sample size for a better generalization of the results.

In our future work, we plan also to propose metrics permitting to determine more objectively the fidelity levels of each component of the avatar fidelity, such as tracking data to evaluate the kinematic fidelity. We plan also to evaluate other forms of partner's representations and other types of collaborative tasks. The results of this research will permit extracting design guidelines for virtual environments that support collaborative work between remote partners. An example of a possible application of these systems is the remote training of a surgical team in a virtual operating room.

## 7 Conflict of Interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

## 8 Author Contributions

Guillaume Gamelin, Amine Chellali and Aylen Ricca contributed to the design of the application and the user study; Guillaume Gamelin, Amine Chellali and Samia Cheikh conducted the literature review, Guillaume Gamelin developed the CVE prototype and integrated all software/hardware components; Cedric Dumas developed the point-cloud streaming server, Amine Chellali and Samia Cheikh performed the data analysis; Guillaume Gamelin and Amine Chellali wrote the first draft of the manuscript. All authors contributed to manuscript revision, read and approved the submitted version.

## 9 Funding

This work was supported by the Paris Ile-de-France Region (grant # 17002647). AR received a Ph.D. grant from the University of Evry. We also acknowledge support from Genopole.

## 10 Ethics statement

The user study was approved by the University of Paris Saclay Ethics Committee (#CER-Paris-Saclay-2018-024). An informed written consent was also obtained from all the subjects involved in this study prior to their participation.

## References

[1] Churchill EF, Snowdon DN, Munro AJ (2001) Collaborative Virtual Environments: Digital Places and Spaces for Interaction, Springer Verlag

[2] Chellali A, Jourdan F, Dumas C (2013) VR4D: an immersive and collaborative experience to improve the interior design process. Joint Virtual Reality Conference of EGVE and EuroVR (JVRC), pp 61-65

[3] Chellali A, Dumas C, Milleville-Pennel I (2011) Influences of Haptic Communication on a Shared Manual Task in a Collaborative Virtual Environment. Interacting With Computers 23(4): 317-328

[4] Chellali A, Milleville-Pennel I, Dumas C (2012) Influence of Contextual Objects on Spatial Interactions and viewpoints sharing in Virtual Environments. Virtual Reality 17(1): 1-15

[5] Steed A, Schroeder R (2015) Collaboration in immersive and non-immersive virtual environments Immersed in Media, Springer, pp 263-282

[6] Steptoe W, Steed A, Rovira A, Rae J (2010) Lie tracking: social presence, truth and deception in avatar-mediated telecommunication. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp 1039-1048

[7] Luff P, Hidmarsh J, Heath C (2000) Workplace Studies: Recovering Work Practiceand Informing System Design. Cambidge University Press

[8] Bridgeman B (1999) Separate representations of visual space for perception and visually guided behavior

[9] Pouliquen-Lardy L, Milleville-Pennel I, Guillaume F, Mars F (2016) Remote collaboration in virtual reality: asymmetrical effects of task distribution on spatial processing and mental workload. Virtual Reality 20: 213-220

[10] Peña Pérez Negrón A, Rangel Bernal NE, Lara López G (2015) Nonverbal interaction contextualized in collaborative virtual environments. Journal on Multimodal User Interfaces 9: 253-260

[11] Hindmarsh J, Fraser M, Heath C, Benford S, Greenhalagh C (1998) Fragmented interaction: Establishing mutual orientation in virtual environments. In Proceedings of the ACM 1998 Conference on Computer-Supported Cooperative Work: 217-226

[12] Chen W, Clavel C, Férey N, Bourdot P (2014) Perceptual Conflicts in a Multi-Stereoscopic Immersive Virtual Environment: Case Study on Face-to-Face Interaction through an Avatar. PRESENCE: Teleoperators and Virtual Environments 23: 410-429

[13] Spante M, Schroeder R, Axelsson AS (2004) How Putting Yourself into the Other Person´s Virtual Shoes Enhances Collaboration, Valencia, pp 190-196

[14] Gaver WW, Sellen A, Heath C, Luff P (1993) ONE IS NOT ENOUGH: MULTIPLE VIEWS IN A MEDIA SPACE. In Proceedings of INTERCHI: 335-341

[15] Ries B, Interrante V, Kaeding M, Anderson L (2008) The Effect of Self-embodiment on Distance Perception in Immersive Virtual Environments. Proceedings of the 2008 ACM Symposium on Virtual Reality Software and Technology, New York, NY, USA, pp 167-170

[16] Mohler BJ, Creem-Regehr SH, Thompson WB, Bülthoff HH (2010) The Effect of Viewing a Self-avatar on Distance Judgments in an Hmd-based Virtual Environment. Presence: Teleoper. Virtual Environ. 19: 230-242

[17] Benford S, Bowers J, Fahlen LE, Greenhalgh C, Mariani J, Rodden T (1995) Networked virtual reality and cooperative work. Presence: Teleoperators and Virtual Environments 4(4): 364–386

[18] Kilteni K, Groten R, Slater M (2012) The sense of embodiment in virtual reality. Presence: Teleoperators and Virtual Environments 21: 373-387

[19] Pan Y, Steed A (2017) The impact of self-avatars on trust and collaboration in shared virtual environments. PLOS ONE 12: 1-20

[20] Oh CS, Bailenson JN, Welch GF (2018) A Systematic Review of Social Presence: Definition, Antecedents, and Implications. Frontiers in Robotics and AI 5: 114

[21] Fribourg R, Argelaguet F, Hoyet L, Lécuyer A (2018) Studying the sense of embodiment in VR shared experiences. IEEE Virtual Reality and 3D User Interfaces, pp 1-8

[22] Piumsomboon T, Lee GA, Hart JD, Ens B, Lindeman RW, Thomas BH, Billinghurst M (2018) Mini-Me: an adaptive avatar for mixed reality remote collaboration. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, New York, NY, USA, pp 1-13

[23] Dodds TJ, Mohler BJ, Bülthoff HH (2011) Talk to the Virtual Hands: Self-Animated Avatars Improve Communication in Head-Mounted Display Virtual Environments. PLOS ONE 6: 1-12

[24] Garau M, Slater M, Vinayagamoorthy V, Brogni A, Steed A, Sasse MA (2003) The impact of avatar realism and eye gaze control on perceived quality of communication in a shared immersive virtual environment. Proceedings of the SIGCHI conference on Human factors in computing systems, pp 529-536

[25] Latoschik ME, Roth D, Gall D, Achenbach J, Waltemate T, Botsch M (2017) The effect of avatar realism in immersive social virtual realities. Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology, pp 1-10

[26] Cowell AJ, Stanney KM (2005) Manipulation of non-verbal interaction style and demographic embodiment to increase anthropomorphic computer character credibility. International journal of human-computer studies 62: 281-306

[27] Young MK, Rieser JJ, Bodenheimer B (2015) Dyadic Interactions with Avatars in Immersive Virtual Environments: High Fiving. Proceedings of the ACM SIGGRAPH Symposium on Applied Perception, New York, NY, USA, pp 119-126

[28] Economou D, Doumanis I, Argyriou L, Georgalas N (2017) User experience evaluation of human representation in collaborative virtual environments. Personal and Ubiquitous Computing 21: 989-1001

[29] Steptoe W, Normand J, Oyekoya O, Pece F, Giannopoulos E, Tecchia F, Steed A, Weyrich T, Kautz J, Slater M (2012) Acting Rehearsal in Collaborative Multimodal Mixed Reality Environments. Presence 21: 406-422

[30] Fairchild AJ, Campion SP, Garcia AS, Wolff R, Fernando T, Roberts DJ (2016) A mixed reality telepresence system for collaborative space operation. IEEE Transactions on Circuits and Systems for Video Technology 27: 814–827

[31] Roth D, Lugrin JL, Galakhov D, Hofmann A, Bente G, Latoschik ME, Fuhrmann A (2016) Avatar realism and social interaction quality in virtual reality. Virtual Reality (VR), 2016 IEEE, pp 277-278

[32] Wu Y, Wang Y, Jung S, Hoermann S, Lindeman RW (2019) Exploring the use of a robust depth-sensor-based avatar control system and its effects on communication behaviors. 25th ACM Symposium on Virtual Reality Software and Technology, pp 1–9

[33] Cho S, Kim S, Lee J, Ahn J, Han J (2020) Effects of volumetric capture avatars on social presence in immersivevirtual environments. 2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)

[34] Yoon B, Kim H, Lee GA, Billinghurst M, Woo W (2019) The effect of avatar appearance on social presence in an augmented reality remote collaboration. 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), pp 547-556

[35] Regenbrecht H, Park N, Ott C, Mills S, Cook M, Langlotz T (2019) Preaching voxels: an alternative approach to mixed reality. Frontiers in ICT 6: 7

[36] Gerathewohl SJ (1969) Fidelity of simulation and transfer of training: a review of the problem, Department of Transportation, Federal Aviation Administration, Office of Aviation Medicine

[37] Benford S, Greenhalgh C, Bowers J, Snowdon S, Fahlén L (1995) User Embodiment in Collaborative Virtual Environments. In Proceedings of CHI'95

[38] Blascovich J (2002) The Social Life of Avatars. In: Schroeder R (ed), Springer-Verlag, Berlin, pp 127-145

[39] Harris H, Bailenson JN, Nielsen A, Yee N (2009) The Evolution of Social Behavior over Time in Second Life. Presence: Teleoper. Virtual Environ. 18: 434-448

[40] Nahon D, Subileau G, Capel B (2015) "Never Blind in VR" enhancing the virtual reality headset experience with augmented virtuality. Virtual Reality (VR), 2015 IEEE, pp 347-348

[41] Slater M, Steed A (2002) Meeting people virtually: experiments in shared virtual environments The social life of avatars, Springer, pp 146–171

[42] Vandenberg SG, Kuse AR (1978) Mental Rotations, a Group Test of Three-Dimensional Spatial Visualization. Perceptual and Motor Skills 47: 599-604

[43] Nowak KL, Biocca F (2003) The Effect of the Agency and Anthropomorphism of Users' Sense of Telepresence, Copresence, and Social Presence in Virtual Environments. Presence: Teleoper. Virtual Environ. 12: 481-494

[44] Roger M, Knutsen D, Bonnardel N, Le Bigot L (2013) Landmark Frames of Reference in Interactive Route Description Tasks. Applied Cognitive Psychology 27: 497-504

[45] Hodgins J, Jörg S, O'Sullivan C, Park SI, Mahler M (2010) The Saliency of Anomalies in Animated Human Characters. ACM Transactions on Applied Perception 7: 1-14

[46] Blascovich J (2002) Social influence within immersive virtual environments The social life of avatars, Springer, pp 127–145

[47] Roberts DJ, Fairchild AJ, Campion SP, O'Hare J, Moore CM, Aspin R, Duckworth T, Gasparello P, Tecchia F (2015) withyou—an experimental end-to-end telepresence system using video-based reconstruction. IEEE Journal of Selected Topics in Signal Processing 9: 562–574