Effect of Facial Realism on Presence in Collaborative Virtual Environments

Investigating the Effect of Avatars with Eye and Mouth-Tracked Facial Expressions

Joshua B. Slik[1]

Supervisor(s): Prof. Dr. Ricardo Marroquim[1]**, Amir Zaidi**[1]

[1]**EEMCS, Delft University of Technology, The Netherlands**

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 24, 2024

Name of the student: Joshua Slik
Final project course: CSE3000 Research Project
Thesis committee: Prof. Dr. Ricardo Marroquim (responsible professor), Amir Zaidi (supervisor), Chirag Raman (examiner)

Abstract

Social uses of virtual reality, such as collaborative virtual environments (CVEs), are showing significant increases in general adoption. In these CVEs, it is desirable for users to feel a high level of presence, which can increase collaboration effectiveness. This study investigated the effect of facial realism on presence in CVEs. Facial realism was implemented through including eye- and mouth-tracked facial expressions in an avatar representation. A controlled within-dyad experiment was performed, consisting of a dyadic interaction between participants in a CVE. We found no significant difference between a Static Face condition and either Eye Tracked or Full Tracked (eye- and mouth) facial expression conditions. Some individual items on the questionnaire were significant or marginally significant, suggesting some positive impact on the assessment of partner reactions in the Full Tracked condition, compared to Static Face. Participants also reported a lower feeling of correspondence between virtual experiences and their physical body in the Full Tracked and Eye Tracked conditions. Some evidence has been found that this experiment warrants repetition with changes discussed in this paper, which may yield significant differences with a higher sample size.

1 Introduction

Advances in display technology and design evolution over the past decade has increased the access of consumers to extended reality (XR) technology. Whereas virtual reality (VR) and augmented reality (AR) were previously gadgets and areas of active research, in the last decade these XR technologies have started to be used by a wider audience, including availability to general consumers. More widespread adoption of XR has come with opportunities for people to socialise in collaborative virtual environments (CVEs). Currently, many social VR platforms are in various stages of availability and active development. Social VR platforms like VRChat, Rec Room, ChilloutVR, Resonite, and Meta Horizon Worlds are all currently being used by people to socialise, attend virtual concerts, or play games. All are also still under active development. Recently, these social VR platforms have been growing in the amount of users. For instance, looking at monthly average concurrent users and monthly peak concurrent users data on the Steam game marketplace [1], we see that VRChat's active user base triples from January 2020 (7980 average, 14 444 peak) to January 2024 (24 109 average, 51 321 peak) [2]. Across all VR platforms, not limited to Steam, VRChat's API reports around double the January 2024 numbers (51 099 average, 93 263 peak) [3].

In social VR applications, like CVEs, presence, the feeling of "being there" is an important aspect of immersion. It has been shown that higher presence correlates with collaboration effectiveness and task performance in CVEs [4]. Presence can be understood to consist of three types of presence: physical presence (i.e. virtual objects are experienced as physical objects), social presence (i.e. avatars of other humans are experienced as actual humans), and self-presence (i.e. the virtual self—the avatar—is experienced as the actual self). The distinction of presence consisting of these three factors is according to the model of presence proposed by Lee [5]. An important aspect of current research into CVEs is how to approach the design of avatars to increase the sense of presence.

One aspect of avatar design is behavioural realism. That is, the degree to which the avatar representation matches the user's movements. This research aims to further the research into behaviour realism of avatars in CVEs, specifically regarding facial realism. We will investigate to what extent facial-tracked eye and mouth expressions impact the three aspects of presence.

2 Related work

Previous research exists on the effect of various degrees of behavioural realism in upper body representation on the feelings of self-presence, social presence and interpersonal attraction [6]. Herrera *et al.* [6] found that in a dyadic interaction an avatar with floating tracked hands elicited a greater feeling of self-presence and social presence in users than a full-bodied avatar with arm movements inferred from the hand tracking.

However, more research into higher-fidelity avatar tracking is needed. In general, it has been found that increased levels of user-tracking have an effect on presence [7]. Herrera *et al.* [6] theorised the higher presence in the floating hands condition was due to a lack of behavioural realism: users could not fully control the arm of the avatar, thus the avatar was less behaviourally realistic. Therefore, only displaying those parts of the body which are tracked, and can be mapped to the virtual environment accurately, could increase presence.

Some social VR platforms afford higher-fidelity motion of avatars. Through adding tracking points on the ankles and waist, often referred to as Full Body Tracking (FBT), all limbs of the user would be tracked, as opposed to only head and hands. In such a setup, it is possible to infer the exact positions of all limbs of the user in greater detail. With these additional tracking points, the inferred limb positions can then be mapped to the avatar in VR space, more accurately representing the user's body. This method could be extended with tracking of knees, elbows, and chest to further increase the user's control of the avatar, and therefore the behavioural realism of the avatar. Such a setup, with all joints tracked, would afford a user control full control over all their virtual limbs, which could improve feelings of presence in full-bodied avatars.

Further, little research has been conducted on the effect of eye and facial expressions on the feelings of physical presence, social presence, and self-presence. Oh *et al.* [8] found that mapping an enhanced smile on an avatar increases affect and social presence in dyadic interactions, though found no significant difference between mapping an enhanced or authentic smile and only mapping a mouth's open-closed state. To the best of our knowledge, no research has been conducted into comparing the effect of different levels of facial realism of an avatar on presence.

3 Present Study

This paper aims primarily to further the work of Herrera *et al.* [6], by researching if their findings of behavioural realism extend to facial realism. We hypothesise that a higher level of behavioural realism in facial expressions could improve presence. This will be the focus of this paper: the effect of eye and facial expressions, achieved through eye and mouth tracking, on the different aspects of presence in a collaborative virtual environment (CVE). The research question this paper will answer is the following:

> *To what extent does facial realism of an avatar, operationalized through adding eye and mouth tracking, impact the feeling of presence in a CVE, when compared to an avatar with only head and hands tracking?*

We will investigate the broad concept of presence in terms of the three types of presence described by Lee [5]. The main question will be answered through the perspective of the following sub-questions:

- *To what extent does facial realism, as above, affect physical presence?*
- *To what extent does facial realism, as above, affect social presence?*
- *To what extent does facial realism, as above, affect self-presence?*

This paper is especially interested in social presence, as it is the most clearly applicable factor of presence related to social VR applications. We hypothesise that social presence is positively affected by higher degrees of facial realism. Further, we hypothesise that physical presence will not be affected either way, and self-presence will be slightly positively affected by higher realism.

These questions will be answered by conducting a controlled within-subject experiment where participants will have a dyadic interaction while represented by an avatar with differing levels of facial-motion fidelity. The research question will be answered by statistical analysis of questionnaire results provided by the participants after these interactions.

4 Method

We will research if facial realism has an effect on presence in CVEs by performing a controlled experiment of participants in a dyadic interaction. Facial realism will be achieved through eye- and mouth-tracked facial expressions, that are accurately mapped from tracking data to implemented facial expressions on an avatar mesh. This section will go into detail on all aspects of the controlled experiment.

4.1 Participants

A total of 14 participants, split into 7 dyads, were recruited from Dutch university and vocational university students, mainly from personal network and academic acquaintances. All students and supervisors in the research group of the author were excluded from participating.

4.2 Materials and Apparatus

In the experiment, participants interacted in a VR environment using a head-mounted display (HMD) supporting eye and facial tracking, and a pair of hand controllers supporting binary extend/close tracking for thumb, index finger and middle finger. The HMDs used were two headsets of the model HP Reverb G2 Omnicept Edition (2021) with second generation Windows Mixed Reality hand controllers. The HMD had a per-eye resolution of 2160 × 2160 pixels, with a display refresh rate of 90 Hz. The HMD used inside-out camera-based tracking for its own position and rotation, as well as the hand controllers' position and rotation. The virtual bounds of the HMDs, represented by semi-transparent walls, were configured to not overlap. The bounds measured 2 m wide × 1.5 m long.

The HMDs were equipped with included infrared face cameras with a resolution of 400 × 400 pixels. One HMD's face camera was unresponsive, it was replaced with an under-the-HMD mounted Logitech C920 webcam. The webcam's 640 × 480 pixels image was horizontally cropped to 480 × 480 pixels from both sides equally, then scaled down to 400 × 400 pixels, and finally desaturated, to match the original HMD face camera as close as possible.

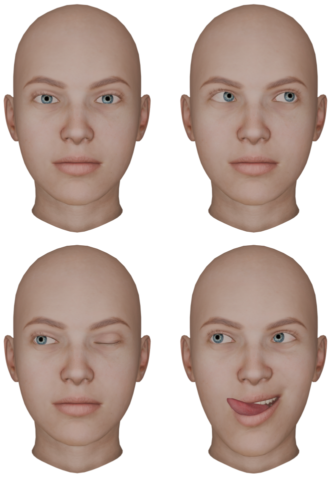
The open-source project VRCFaceTracking (VRCFT) [9], which the Unified Expressions standard is a part of, was used to send tracking events to the social VR platform. Inside VRCFT, the module VRCFTOmniceptModule [10] was installed to connect to the HMD and capture its eye tracking data, and the module VRCFT-Babble [11] was used to receive events sent by Project Babble [12], an application that processes the HMD face camera through image recognition models to yield a value between 0 and 1 for many facial expressions to represent how strongly any shape is currently expressed. Version information for all used software is provided in their respective references, and licenses can be found in Section 8.2.

The social VR platform the interactions took place in was VRChat, in a publicly available virtual world called "The Black Cat," resembling a café environment. The environment was set up before the experiment to place the participants' avatars facing away from each other, to allow eye and mouth tracking data to settle when applying and removing the HMDs. This was done so participants would not see each other until HMDs were fully applied and tracking was accurate, and therefore would not see inaccurate or unnatural facial expressions.

Avatar

A single avatar was used to represent the participants regardless of their appearance. The final mesh and rig of the avatar is based on 'CC Character Base 3' by Reallusion [13], modified and used within allowances of the provided licence. On this mesh, the eye and face expressions were implemented through shape keys (sometimes called blend shapes). These are variables that have a floating point value between 0 and 1, where 0 means 'shape not expressed' and 1 meaning 'shape maximally expressed'. One defines a basis shape on the mesh for which vertex positions are stored as normal for any mesh. All shape keys then store some translation of vertices from the basis mesh to the shape key definition. By linear interpolation, vertices can then move between the basis shape (value 0) and the maximally expressed shape key (value 1). Multiple shape keys can receive values larger than 0, and these shapes are then combined by multiple linear interpolations, yielding the possibility for complex expressions by combination of simple expressions. The expression shape keys were implemented following the Unified Expressions standard [14], which provides for translation of vendor-specific tracking standards to one encapsulating standard. This allowed for animation of the mesh to be driven by tracking events received from many commercially available HMDs, which can aid future reproducibility. Defined shape keys included individual eye rotation, eyelid closure, and pupil dilation for eye expressions, as well as mouth open, smile, frown, and puckering of the lips for facial expressions. Some example expressions are shown in Figure 1, while the full list of implemented shapes is available in Appendix C, and the avatar itself in Appendix A

Figure 1 Various facial expressions. In reading order: basis shape; looking left and upward with 4 shapes; looking right and left eye closed with 2 shapes; complex facial expression of 12 underlying shapes.



4.3 Design

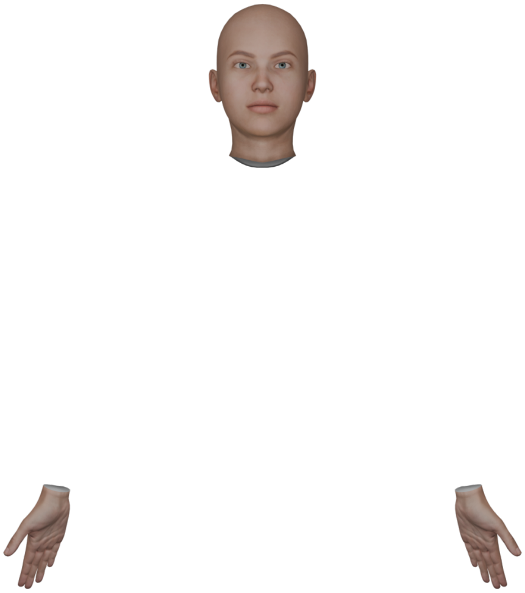
The study used a within-dyads design with three conditions: Static Face, Eye Tracked, and Full Tracked. Both participants in the dyad were represented with the same type of avatar within each condition. In all conditions, the avatar consisted of a floating head and hands, as seen in Figure 2. The head position and rotation of the participant were mapped to the avatar's head. The hand controller position and rotation, as well as the three binary states of finger tracking, were mapped to the floating hands. Participants could control the thumb to extend and close, the index finger was able to extend and close, and the middle finger binary position was mapped to extending and closing the remaining three fingers. In total, eight hand shapes were supported through combining these binary finger tracking features.

In the first condition, Static Face, participants embodied an avatar with no face tracking at all. In the second condition, Eye Tracked, participants embodied an avatar that supported granular eyelid closure and X/Y rotation for individual eyes, as well as average tracked pupil dilation affecting the avatar's joint-eye dilation. In the final condition, Full Tracked, participants embodied an avatar supporting a vast amount of mouth shapes in addition to all previously mentioned eye tracking from the Eye Tracked condition. The order of the conditions was counterbalanced to account for possible learning effects for six total condition orderings.

4.4 Procedure

Participants were invited to the same room and selected into the condition order with the fewest recorded dyads up to that point, with random selection breaking ties.

Figure 2 Avatar representation in base pose.



Before commencing with the experiment proper, the participants were presented with the HMDs and explained how to adjust them to their head, and aided in this process as required until both were comfortable. Then, the eye tracking of the headset was calibrated using the HMD vendor's provided tool, consisting of adjustments to the HMD's top strap, interpupilary distance, and a short game of following a virtual sphere with their eyes.

After calibrating, participants were asked to remove their HMDs, and were informed that they would be playing a game of 20 Questions in a virtual environment resembling a café, where they could move around a little, but were asked to stay within the virtual bounds represented by semi-transparent walls. Upon entering the CVE, they would turn around and see their partner's avatar in front of them, at which point they would commence the game.

The choice of the game 20 Questions was based on prior research into task performance and presence in CVEs also using the 20 Questions game to structure their interaction [15, 16, 6]. Further, considering the three conditions and required setup time, a verbal game that is simple and quickly explainable was desirable. The rules of the game 20 Questions were explained to all participants following a written script. In the CVE, one participant chosen at random would start as the guessing participant by trying to identify an object through asking "yes" or "no" questions to their partner, who was shown the name of the object by the researcher just before the game commenced. The researcher kept track of the amount of questions asked, and informed the participants of the current count every five elapsed questions, and counted down the final five questions individually. The game would end when the guessing participant correctly guessed the word, or when they asked 20 questions regardless of a correct guess at the end. At that point, the researcher would record the time elapsed for that round, and the game was repeated with switched roles, with a different word given to the participant previously guessing, now non-guessing. The given words were always 'apple', 'boat', 'computer', 'pen', 'bicycle', and 'book', in that order across the rounds and conditions.

After receiving these instructions, participants were asked to put on their HMDs and were given their hand controllers. A few checks were carried out: participants were asked if the HMDs felt comfortable, and the researcher would observe the view of both participants mirrored on the desktop monitors to ensure tracking was operating as expected for that condition. After these checks, they were instructed to turn around and commence the game.

After each interaction, HMDs were removed and the participants filled out a questionnaire on paper, during which the researcher switched the avatars to the next condition in the selected order, and the experiment was repeated for all conditions, starting with giving one of the participants a new word for the 20 Questions game.

4.5 Measures

Presence was measured by a questionnaire with three subscales respective to the different factors of presence. Subscales were constructed by adapting items from the Multimodal Presence Scale (MPS) [17], which was developed according to the Lee model of presence [5], and the Herrera *et al.* [6] scales. For every subscale participants were asked to what extent they agree with these statements on a 7 point Likert-type scale (1 = *Strongly Disagree*, 7 = *Strongly Agree*). The full questionnaire is provided in Appendix B. Reliability of the full questionnaire was excellent, *Cronbach's alpha* = .948, 95% CI [0.921, 0.964].

Physical presence

Physical presence was measured using the 5 items from the physical presence subscale in the MPS [17]. Scale reliability was good, *Cronbach's alpha* = .89, 95% CI [0.826, 0.929].

Social presence

Physical presence was measured by items adapted from the social presence subscale in the MPS [17] and items from Herrera *et al.* [6]'s social presence questionnaire. Scale reliability was very good, *Cronbach's alpha* = .90, 95% CI [0.837, 0.938].

Self-presence

Physical presence was measured by items adapted from the self-presence subscale in the MPS [17] and items from Herrera *et al.* [6]'s self-presence questionnaire. Scale reliability was very good, *Cronbach's alpha* = .91, 95% CI [0.846, 0.937].

5 Results

The paper questionnaires were transcribed by hand into a CSV file containing a dyad identifier, an identifier for the participant in a dyad (either participant 1 or 2), a condition and condition order identifier, and 17 question answers on a scale of 1 to 7. This resulted in a total of 42 observations consisting of 7 dyads with 2 participants each, and 3 different conditions per participant. This data was processed in Python with the `numpy` and `pandas` packages to calculate the presence scores. Subscale scores were calculated by the mean of their respective questions, which can be viewed in Appendix B. The total presence score was calculated by the mean of all questions. In addition, a unique identifier for each person (from 0 to 13) was inserted into the dataframe for use in analysis.

The means and standard deviation of all dependent variables are summarized by condition in Table 1. On average, the Static Face interactions took the shortest ($M = 6.28$ min, $SD = 2.41$), followed by the Full Tracked interactions ($M = 7.84$ min, $SD = 2.85$), and the Eye Tracked interactions taking the longest ($M = 7.89$ min, $SD = 3.32$).

Table 1 Means and Standard Deviations for Dependent Variables by Condition

	Static Face	Eye Tracked	Full Tracked
Measure	$M \pm SD$	$M \pm SD$	$M \pm SD$
Physical presence	4.43 ± 1.13	4.43 ± 1.39	4.44 ± 1.12
Social presence	4.36 ± 1.50	4.49 ± 1.35	4.65 ± 1.21
Self-presence	3.39 ± 1.24	3.37 ± 1.50	3.20 ± 1.00
Total of presence subscales	4.09 ± 1.22	4.14 ± 1.31	4.16 ± 1.01

Note: M = mean, SD = (sample) standard deviation

5.1 Analysis Method

The experimental setup requiring dyads results in the data being unavoidably dependent on the dyad. Since the study was also within-user, with every participant providing three data points, the data is also dependent on the individual participants in the dyads. This dependence violates the assumption of independent data required for standard analysis of variance (ANOVA) or linear regression analysis, so a Linear Mixed Model (LMM) analysis was carried out, which allows for compensation for these random effects. The double dependence, first on the dyads, second the participants in each dyad, required a nested LMM analysis with one level of nesting (dyad → participant). Since the study was not carried out in multiple regions, the dyad is our highest grouping, and no regional grouping factor needs to be considered.

The LMM analysis was carried out in R using the `lme4` and `lmerTest` packages. The model was constructed using the following definition:

```
model ← lmer(dv ~ treatment + (1|dyad_id/person_id
    ), REML = TRUE, data = data)
```

Listing 1 LMM model definition. *Note:* `dv` = dependent variable

As Listing 1 demonstrates, the model was constructed to explain variations in a dependent variable `dv`, based on variations in a fixed effect `treatment` (the scenario identifier: Static Face, Eye Tracked, Full Tracked). Because of the dependence on dyads and specific participants in dyads, a nested random effect `(1|dyad_id/person_id)` was included in the model. The model will account for dyads having an unknown random effect (`dyad_id`), and specific participants nested in the dyads having an unknown random effect (`person_id`). The found regression coefficients and significance scores are summarized in Table 2, grouped first by dependent variable, and second by effect of the Eye Tracked and Full Tracked conditions when compared to the Static Face condition.

5.2 Presence scores

As is apparent from Table 2, no significant difference was observed between the Static Face condition and either the

Table 2 Dependent Variables Linear Mixed Model Coefficients and Significance Scores

Parameter	β	SE	t	p
Physical presence				
(Intercept)	4.429	0.326	13.570	5×10^{-11}
Eye Tracked	-5×10^{-15}	0.231	-2×10^{-14}	1.000
Full Tracked	0.014	0.231	0.062	0.951
Social presence				
(Intercept)	4.357	0.363	12.018	7×10^{-10}
Eye Tracked	0.133	0.236	0.562	0.579
Full Tracked	0.296	0.236	1.253	0.221
Self-presence				
(Intercept)	3.386	0.338	10.013	1×10^{-8}
Eye Tracked	−0.014	0.217	−0.066	0.948
Full Tracked	−0.186	0.217	−0.855	0.400

Note: β = Estimate, SE = standard error, t = t-score, p = p-value

Table 3 Significance Scores for Individual Questionnaire Items

Q	Static Face	Eye Tracked		Full Tracked	
	$M \pm SD$	$M \pm SD$	t	$M \pm SD$	t
Q1	4.21 ± 1.63	4.21 ± 1.63	0.00	4.36 ± 1.34	0.43
Q2	4.86 ± 1.29	4.79 ± 1.48	−0.24	4.93 ± 1.59	0.24
Q3	4.07 ± 1.64	4.21 ± 1.89	0.40	4.14 ± 1.17	0.20
Q4	5.00 ± 1.24	4.86 ± 1.35	−0.51	5.00 ± 1.18	0.00
Q5	4.00 ± 1.24	4.07 ± 1.69	0.22	3.79 ± 1.37	−0.66
Q6	4.14 ± 1.70	4.50 ± 1.09	0.92	4.71 ± 1.49	1.47
Q7	3.86 ± 1.75	3.86 ± 1.51	0.00	4.71 ± 1.59	1.88$^{\bullet}$
Q8	3.93 ± 1.77	4.50 ± 1.74	1.25	4.50 ± 1.09	1.25
Q9	5.00 ± 1.75	5.07 ± 1.90	0.20	5.36 ± 1.34	1.02
Q10	4.86 ± 1.61	4.93 ± 1.54	0.22	5.07 ± 1.69	0.66
Q11	4.50 ± 1.99	4.00 ± 1.88	−1.84$^{\bullet}$	3.86 ± 1.66	−2.36*
Q12	4.21 ± 2.04	4.57 ± 2.03	1.19	4.36 ± 2.06	0.48
Q13	3.29 ± 1.38	3.71 ± 2.02	1.40	3.29 ± 1.64	0.00
Q14	3.07 ± 1.44	3.29 ± 1.73	0.65	3.07 ± 1.07	0.00
Q15	4.43 ± 1.65	4.21 ± 2.12	−0.59	4.36 ± 1.39	−0.20
Q16	3.07 ± 1.49	3.00 ± 1.62	−0.22	2.79 ± 1.42	−0.88
Q17	3.07 ± 1.44	2.64 ± 1.45	−1.87$^{\bullet}$	2.50 ± 1.02	−2.50*

Note: M = mean, SD = (sample) standard deviation, t = t-score
$^{*} = 0.01 < p < 0.05$, $^{\bullet} = p < 0.1$

Eye Tracked or Full Tracked conditions in any of the three presence measures. The highest observed non-significance is that for social presence when comparing Static Face to Full Tracked ($p = 0.221$).

5.3 Individual Questions

After no significant difference was found on any subscale score, individual questionnaire items were analysed more closely. Their individual means, standard deviation, and significance scores are summarized in Table 3. From the individual items, there are two questions with a significant difference, and a further three with a marginally significant difference. A significant difference was found between the Static Face and the Full Tracked conditions, with lower ratings from participants in the Full Tracked condition for Q11 ($p = 0.026$) and Q17 ($p = 0.056$). Further, a marginally significant difference was observed between the same conditions, with participants in the Full Tracked condition scoring Q7 higher ($p = 0.071$). The final marginally significant differences were observed between the Static Face and Eye Tracked conditions, with the participants in the Eye Tracked condition rating Q11 ($p = 0.078$) and Q17 ($p = 0.072$) lower.

5.4 Correlation among Dependent Variables

Dependent variables were analysed for their correlation. This analysis was carried out in R using the package `rmcorr` to correct for the repeated measures among participants.

Analysis shows that physical presence was significantly and positively correlated with social presence ($r = .57$, $p < 0.01$). Additionally, social presence and self-presence were significantly and positively correlated ($r = .48$, $p < 0.01$). The correlation between physical presence and self-presence was especially strongly positive and significant ($r = .70$, $p < 0.0001$).

6 Discussion

This study aimed to examine to what extent facial realism has an effect on physical presence, social presence, and self-presence during a dyadic interaction inside a collaborative virtual environment (CVE). We were unable to find any significant differences between the Static Face condition and either of the Eye Tracked and Full Tracked conditions on any of the tested presence subscales.

Interestingly, for two individual items in the questionnaire a significant difference was observed, and a further three showed a marginally significant difference. One of the significant findings is for participants in the Full Tracked condition reporting lower agreement with Q17 than participants in the Static Face condition, as well as marginally significant lower agreement in the Eye Tracked condition compared to Static Face. This question was stated as follows "*Q17: When something happened to my avatar, I felt like it was happening to me.*" This is an interesting result, as the interaction did not require anything physically happening to the avatar of a user, other than the fully-verbal game. From observation during the experiments, some dyads did include virtual touching in their interaction. Participants were physically separated, so no physical contact ever took place, but that did not preclude some participants to perform a gestured 'virtual' high-five in celebration after guessing a word correctly, or other gestured 'virtual' contact like holding their partner's hand, shaking hands, extending and touching fingers. An emerging phenomenon in VR is *phantom touch*, or *phantom touch illusion*, or *phantom tactile sensation*. This phenomenon is described as a sensation of tingling, heat, or pressure on parts of the user's body that appear to be touched by themselves or other users in VR, in the absence of any physical stimulation [18, 19, 20]. This sensation extends to invisible (inferred) parts of a user's virtual limbs [19], such as the invisible arms of the avatar this study used. Although research into this phenomenon is in a very early stage, this finding may indicate the phantom touch sensation is lessened or suppressed in a dyadic interaction when both users are represented by avatars with eye- and mouth-tracked facial expressions.

The last marginally significant finding was for participants in the Full Tracked condition, with eye- and mouth-tracked facial expressions, reporting higher agreement with Q7 than participants in the Static Face condition. Item 7 on the questionnaire was as follows: "*Q7: I felt like I was able to assess my partner's reactions to what I said.*" This indicates that there is some marginally significant perceived benefit in assessing the emotional state of the dyadic partner from accurately mapped eye and mouth expressions in CVEs. It is likely this reported increase in emotional assessment is due to the increased affordance in nonverbal communication of the participants.

In general, when reviewing the means of presence scores across conditions in Table 1, it is clear that physical presence is not affected by the three types of avatar representation in this study ($M_{Static} = 4.43$, $M_{Eye} = 4.43$, $M_{Full} = 4.44$). This finding matches our hypothesis of no significant effect on physical presence. For social presence and self-presence, a difference in mean scores is visible, increasing in the case of social presence and decreasing in the case of self-presence as facial realism increases, but our LMM analysis did not find significance.

Given our findings of a marginally significant difference in Q7, we looked at the means of all social presence items more closely. As reported, Q11 has a significant negative shift for participants in the Full Tracked condition, and marginally significant in the Eye Tracked condition, both compared to Static Face. This question, "*Q11: I felt like during the simulation there were times where the computer interface seemed to disappear, and I was working directly with another person.*", was the longest statement on the questionnaire. Anecdotally, this question was reported to be hard to understand, with multiple participants asking the meaning of the question during the experiment (which was not answered by the researcher to prevent bias). One participant noted that they only understood the meaning of the question in the third round, and asked to change their earlier answers (which was denied). Moreover, multiple dyads reported the mouth expressions to be "unnerving", "unnatural" or "uncanny". This may explain the significant result for Q11, as the computer interface was more clearly present due to the avatars being unnerving. Multiple participants reported feeling uncomfortable looking at each other during this condition. Various effects may be the cause of this. From later analysis, the shape key responsible for the jaw movements, opening and closing the mouth, was set to a likely unnaturally widely opened mouth at the extreme value. During the experiment it was also clear that the facial tracking was slower than anticipated, with mouth movements often lagging more than a second behind speech of the participant. Question 11 is the only question whose mean is lower across both tracked conditions, with all other social presence items having higher observed means for both tracked conditions. If Q11 was not part of the questionnaire and is removed from the social presence score, the significance of the Full Tracked condition compared to Static Face does increase to marginally significant (originally $p = 0.392$, Q11 removed $t = 1.82$, $p = 0.08$). We are not considering the social presence score with this removed question 11 to be definitively indicative of a positive social presence increase, however. To conclude that, more research is necessary, preferably with a higher sample size, and revisions to the experiment. We will advocate this in later sections.

This study cannot reject the null hypothesis of facial realism having no effect on the feeling of presence in CVEs. Given the above marginally significant findings, a repeated experiment with a larger sample size is desirable.

6.1 Recommended changes

This study's dyadic interaction was a structured one, consisting of participants playing the 20 Questions game. This game is widely used in previous research as a simple structured interaction. From our observation of the participants as they played this game, however, many participants looked away from their partner during large parts of the interaction, especially as question-askers needed to think of their next question. Given these observations, we would recommend future research into avatar representations to consider an alternatively structured interaction that incentivises participants to look at their partners, as the differences between conditions in this research was fully contained to only the face of the avatars. Participants can easily miss these changes, or they may have a reduced effect if participants look away from their partner often. Anecdotally, some participants reported not noticing any change between the conditions.

A further recommended change is to use a different facial tracking approach. The eye tracking was reliable and quick, but the image recognition-based facial tracking was noticeably slower than real-time facial expressions, and was not accurate and detailed enough to facilitate correct animation of all the 67 implemented facial expressions. The experiment might be repeated with a marker-based or camera with depth-sensor (RGB-D) based approach, which may yield more accurate tracking.

6.2 Limitations

One clear limitation is the low amount of participants for this study. This small sample size should be acknowledged when drawing conclusions from the results. This research should be considered a pilot study, and this experiment, with consideration of suggested changes in Section 6.1, could be repeated with a larger sample size. A quick analysis of a hypothetical higher sample size was carried out by creating new datasets consisting of 70 dyads. These dyads were picked by random selection with replacement from the original dataset. To all mean presence scores, normally distributed noise was added, with the normal distribution centered around 0 and with a standard deviation of the respective dependent variables. The resulting presence scores with added noise were then clamped to the range $[1, 7]$. Dyad and person identifiers were updated to remain unique, as expected by the LMM nested analysis. This way, 4000 datasets were generated, with an unchanged standard deviation on the random variables (due to the added noise). Of these 4000 simulated datasets, 1839 yielded a significant difference between the Static Face and Full Tracked conditions on social presence scores, with 592 also yielding a significant difference between Static Face and Eye Tracked for social presence. The used to generate these datasets, as well as the analysis code, is provided in the same data analysis repository available in Appendix A. This finding of artificially generated larger sample size datasets supports a larger sample size may still yield significant results, and repetition of the experiment with suggested changes is desired.

A second limitation is the discrepancy in HMD camera setup. It is possible that the HMD with the replaced camera has a different level of fidelity for the mouth tracking.

Care was taken to equalise the output of the webcam, but there were differences in refresh rate and picture quality, not to mention that it is a visible-spectrum camera, not an infrared camera. The differing fidelity could bias the perceived social presence of the dyadic partner of the participant using this HMD, as the face tracking might be more accurate. It could also bias the participant's own perceived self-presence in the same way. Moreover, because of the positioning of the camera, this HMD was slightly more forward-heavy, which could result in a less comfortable fit of the HMD. This can in turn impact presence by introducing or reinforce a non-immersive factor and remind the participant they are wearing an HMD and are in a virtual environment.

A further limitation is that this study was unable to arrange separate physical rooms for the participants. Ideally, the dyadic interactions would have taken place with the participants in separate rooms, communicating through microphones, headphones, and whatever extent of non-verbal communication the different conditions allow. The fact that the participants could see each other prior to the experiment, and during switching of conditions, could positively bias their social presence. Although this bias should exist across all conditions, and will therefore not affect the significance of findings, it may shift social presence scores higher than actual-world conditions, where people using social VR are unlikely to be in the same room together.

7 Conclusions and Future Work

This study aimed to research the effect of facial realism on the three aspects of presence: physical presence, social presence, and self-presence [5]. A within-subject controlled experiment was conducted between dyads in a collaborative virtual environment (CVE) experiencing three different conditions of facial realism: Static Face, Eye Tracked including mapped eye expressions, and Full Tracked including mapped mouth expressions in addition to all Eye Tracked expressions. The aspects of presence were measured by questionnaire. We were unable to find any significant differences between the Static Face condition and either of the Eye Tracked and Full Tracked conditions on any of the tested presence subscales. Three agreement-based statements were individually significant or marginally significant. Two significant findings are participants reporting lower agreement in the Full Tracked condition, compared to Static Face, with "*Q11: I felt like during the simulation there were times where the computer interface seemed to disappear, and I was working directly with another person.*" and "*Q17: When something happened to my avatar, I felt like it was happening to me.*". For these same questions, a marginally significant finding was lower agreement being reported by Eye Tracked participants compared to Static Face. One question was marginally significant more positive, with participants in the Full Tracked condition reporting higher agreement with "*Q7: I felt like I was able to assess my partner's reactions to what I said.*", compared to Static Face.

7.1 Future work

As stated in Section 6.1 and Section 6.2, future research could repeat the experiment with consideration for suggested changes like a larger sample size, or perform a similar experiment into different degrees of facial realism. Unexplored aspects of this experiment are facial features like eyebrow and nose expressions, or viseme mapping, for example through face tracking or speech data. The HMDs used by this study did not support these kinds of facial expressions, but some commercially available HMDs do include support for tracking these features.

Further research could be conducted into lower degrees of facial realism that may convey emotions better. As Oh *et al.* [8] found for artificially enhancing a user's smile improving social presence, this may also be the case for a variety of facial expressions. Instead of implementing many small expressions that are tracked individually to combine into a final facial shape like the avatar of this research had, one may implement a limited set of clear expressions and enhanced versions of those expressions and measure presence of those conditions compared to a baseline. A further option is enhancing these expressions with algorithm-run simulations, like comparing user-tracked eye expressions to non-tracked simulated eye expressions, or a combination of the two, as well as the same possibility for facial expressions. Tracking data may show, for example, a user holding a certain face, where the animation algorithm would vary that shape slightly and introduce micro-expressions to account for low resolution tracking data.

8 Responsible Research

No large language models were used by the author at any point, for any purpose, as part of this research.

The author of this paper is aware of the Netherlands Code of Conduct for Research Integrity (NCCRI, 2018) and followed its guidance to the best of their ability. All results from data analysis were reported honestly, all performed data analysis was reported on in the paper, especially all analysis that yielded insignificant results. No data points were discarded, and would have been reported on if any kind of outlier removal was performed. The next subsections go further in depth regarding the human research ethics procedure in Section 8.1, and what measures were taken to ensure this research is reproducible in Section 8.2.

One day after submission, an error in the data was discovered. Two participants seem to have switched their paper questionnaires for one condition only, which results in a big outlier as the subjective differences between participants were big. Upon discovery of this error, immediate steps were taken to update this paper, as advised by the NCCRI pillar of responsibility for research findings. With this error removed, some significance values increased, and more findings turned out to be significant or marginally significant. Since you are reading this, you have read the updated paper.

8.1 Human Research Ethics

To safeguard all participants and their provided research data (closed-form questionnaire answers), the full Delft University of Technology Human Research Ethics Committee (HREC) procedure was followed. HREC was informed of the research and approved it on ethical grounds. Before an experiment commenced, the participants read an opening statement that informed them of their right to withdraw consent at any time during the experiment, which would guarantee no further data to be collected and any prior collected data to be destroyed. They were also informed exactly what data would be collected, how it would be stored, how it would be used, and the risks of VR research (chance of nausea, headaches, dizziness,) and that the researcher present would help them take their HMD off if at any point they wanted to exit the CVE.

To safeguard research data, the data was stored on paper and digitally on removable media (USB drive) at all times, and uploaded to Delft University of Technology archive storage after data analysis was completed. At no point was research data uploaded to an online storage provider.

8.2 Reproducibility and Open Science

Ensuring reproducibility was a large focus in the writing of this paper. As this research combines many existing software packages with assets constructed by the author specifically for this research (the avatar), the implementation of the experimental setup is reasonably complex. To this end, in line with Delft University of Technology guidelines for Open Science, all assets and code produced for this research are published publicly under a CC-BY licence. The modified avatar implementing the Unified Expressions standard, the Unity project that implements the VRCFT template on this avatar mesh, and all data analysis code are made available in Appendix A.

All references to existing software and standards are accompanied by version numbers in the references, and every step of the experimental setup is described in high detail. With this information, and the openly published assets and code, anyone that wants to reproduce the exact experimental setup is able to do so.

Both the anonymous data obtained from the study, and the code used for data analysis are also published publicly. The most important parts of data analysis specifics, such as the LMM definition and used R packages, are described in the paper for the purpose of transparency of the findings. Both the data and analysis code can be found in Appendix A, licensed CC-BY.

Acknowledgements

The author would like to thank VRChat for granting rights to upload the test avatars on their platform before their account trust rank policy would normally grant upload rights.

Further, the author wants to thank Cassandra Visser for her aid in testing the facial-tracking setup, and to thank Eline Hilbers for providing useful reading material regarding linear mixed model analysis.

Licenses

Used software ‘VRCFaceTracking’ (VRCFT) [9] is freely available under the Apache-2.0 licence. Modules ‘VRCFT-Babble’ [11] and ‘VRCFTOmniceptModule’ [10] for VRCFT are freely available under the MIT licence.

Used software ‘Project Babble’ [12] is freely available under the Apache-2.0 licence.

Used Unity prefab for receiving tracking events and mapping to shape keys ‘VRCFaceTracking-Template’ by Adjerry91 [21] is freely available under the MIT licence. Modifications to this template are provided in the Unity project in Appendix A.

Used base mesh and rig ‘CC Character Base 3’ by Reallusion [13] is freely available under the Reallusion Character Creator Base Model Licence. Modified version created for this study is provided in Appendix A.

References

Articles and Conference Papers

[4] A. Bayro, Y. Ghasemi, and H. Jeong, “Subjective and objective analyses of collaboration and co-presence in a virtual reality remote environment,” in *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, Christchurch, New Zealand: IEEE, Mar. 2022, pp. 485–487, ISBN: 978-1-66548-402-2. DOI: 10.1109/VRW55335.2022.00108.

[5] K. M. Lee, “Presence, explicated,” *Communication Theory*, vol. 14, no. 1, pp. 27–50, Feb. 2004, ISSN: 1050-3293, 1468-2885. DOI: 10.1111/j.1468-2885.2004.tb00302.x.

[6] F. Herrera, S. Y. Oh, and J. N. Bailenson, “Effect of behavioral realism on social interactions inside collaborative virtual environments,” *Presence: Teleoperators and Virtual Environments*, vol. 27, no. 2, pp. 163–182, Feb. 1, 2018, ISSN: 1531-3263. DOI: 10.1162/pres_a_00324.

[7] J. J. Cummings and J. N. Bailenson, “How immersive is enough? a meta-analysis of the effect of immersive technology on user presence,” *Media Psychology*, vol. 19, no. 2, pp. 272–309, Apr. 2, 2016, ISSN: 1521-3269, 1532-785X. DOI: 10.1080/15213269.2015.1015740.

[8] S. Y. Oh, J. Bailenson, N. Krämer, and B. Li, “Let the avatar brighten your smile: Effects of enhancing facial expressions in virtual environments,” *PLOS ONE*, vol. 11, no. 9, J. Najbauer, Ed., e0161794, Sep. 7, 2016, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0161794.

[15] J. N. Bailenson, A. C. Beall, and J. Blascovich, “Gaze and task performance in shared virtual environments,” *The Journal of Visualization and Computer Animation*, vol. 13, no. 5, pp. 313–320, Dec. 2002, ISSN: 1049-8907, 1099-1778. DOI: 10.1002/vis.297.

[16] J. N. Bailenson and N. Yee, “A longitudinal study of task performance, head movements, subjective report, simulator sickness, and transformed social interaction in collaborative virtual environments,” *Presence: Teleoperators and Virtual Environments*, vol. 15, no. 6, pp. 699–716, Dec. 1, 2006, ISSN: 1054-7460. DOI: 10.1162/pres.15.6.699.

[17] G. Makransky, L. Lilleholt, and A. Aaby, “Development and validation of the multimodal presence scale for virtual reality environments: A confirmatory factor analysis and item response theory approach,” *Computers in Human Behavior*, vol. 72, pp. 276–285, Jul. 2017, ISSN: 07475632. DOI: 10.1016/j.chb.2017.02.066.

[18] S. Alexdottir and X. Yang, “Phantom touch phenomenon as a manifestation of the visual-auditory-tactile synaesthesia and its impact on the users in virtual reality,” in *2022 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, Singapore, Singapore: IEEE, Oct. 2022, pp. 727–732, ISBN: 978-1-66545-365-3. DOI: 10.1109/ISMAR-Adjunct57072.2022.00218.

[19] A. Pilacinski, M. Metzler, and C. Klaes, “Phantom touch illusion, an unexpected phenomenological effect of tactile gating in the absence of tactile stimulation,” *Scientific Reports*, vol. 13, no. 1, p. 15 453, Sep. 18, 2023, ISSN: 2045-2322. DOI: 10.1038/s41598-023-42683-0.

[20] Q. Chen, M. M. Spapé, and G. Jacucci, “Understanding phantom tactile sensation on commercially available social virtual reality platforms,” *Proceedings of the ACM on Human-Computer Interaction*, vol. 8, pp. 1–22, CSCW1 Apr. 17, 2024, ISSN: 2573-0142. DOI: 10.1145/3637418.

Other Sources

[1] Valve Corporation. “Steam store.” (), [Online]. Available: https://store.steampowered.com/ (visited on 06/11/2024).

[2] Steam Charts. “VRChat player count over time,” VRChat - Steam Charts. (2024), [Online]. Available: https://steamcharts.com/app/438100 (visited on 04/28/2024).

[3] VRChat API Metrics. “VRChat player count over time as reported by VRChat API,” Grafana. (2024), [Online]. Available: https://metrics.vrchat.community/?orgId=1&refresh=30s&from=now-1y&to=now (visited on 04/28/2024).

[9] B. Thomas, *VRCFaceTracking*, version 5.2.3.0, 2024. [Online]. Available: https://github.com/benaclejames/VRCFaceTracking (visited on 06/13/2024).

[10] 200Tigersbloxed, *VRCFTOmniceptModule*, version 1.6.0, 2024. [Online]. Available: https://github.com/200Tigersbloxed/VRCFTOmniceptModule (visited on 06/13/2024).

[11] dfgHiatus, *VRCFT-babble*, version 2.0.7, 2024. [Online]. Available: https://github.com/dfgHiatus/VRCFT-Babble (visited on 06/13/2024).

[12] S. Suri, *ProjectBabble*, version 2.0.6 Alpha, 2024. [Online]. Available: https://github.com/Project-Babble/ProjectBabble (visited on 06/13/2024).

[13] Reallusion. “CC Character Base.” (2024), [Online]. Available: https://www.reallusion.com/character-creator/free-3d-character-base.html (visited on 06/13/2024).

[14] VRCFaceTracking. “Unified expressions | VRCFaceTracking.” version 5.0. (2024), [Online]. Available: https://docs.vrcft.io/docs/tutorial-avatars/tutorial-avatars-extras/unified-blendshapes (visited on 06/13/2024).

[21] Adjerry91, *VRCFaceTracking-templates*, version 6.0.0, 2024. [Online]. Available: https://github.com/Adjerry91/VRCFaceTracking-Templates (visited on 06/13/2024).

[22] J. B. Slik. “Open Science publication of assets and code used for this paper.” (Jun. 24, 2024), [Online]. Available: https://github.com/JoshCode/thesis-vr-presence-2024.

[23] J. B. Slik, *Data underlying publication: Effect of facial realism on presence in collaborative virtual environments*, Jun. 23, 2024. DOI: 10.4121/5ef9e932-fb11-46f5-9698-60174bfac5ff.

A Open Science publishing

This paper is published publicly under a CC-BY license. An electronic version is freely available at https://repository.tudelft.nl/.

Avatar assets, the Unity avatar project, and data analysis code is freely available at [22]. Original assets owned by other parties are present in this repository, licensed under their respective original licenses. Modifications to those assets, and other original work in this repository are licensed CC-BY. Further documentation and configuration instructions are available.

Anonymous data obtained from the controlled experiment is available publicly under a CC-BY license [23].

B Presence Questionnaires

These three subscales were part of one questionnaire. Items on the questionnaire were asked in this order, and were not divided into subscales to the particpants.

How strongly do you agree or disagree with the following statements?
1 = Strongly Disagree — 7 = Strongly Agree

Physical presence

Q1 The virtual environment seemed real to me.

Q2 I had a sense of acting in the virtual environment, rather than operating something from outside.

Q3 My experience in the virtual environment seemed consistent with my experiences in the real world.

Q4 While I was in the virtual environment, I had a sense of “being there”.

Q5 I was completely captivated by the virtual world.

Social presence

Q6 I felt like I was face-to-face with my partner.

Q7 I felt like I was able to assess my partner’s reactions to what I said.

Q8 I felt like my partner was watching me.

Q9 I felt like my partner was aware of my presence.

Q10 I felt like my partner was present.

Q11 I felt like during the simulation there were times where the computer interface seemed to disappear, and I was working directly with another person.

Q12 I felt like I was interacting with other people in the virtual environment, rather than a computer simulation.

Self-presence

Q13 I felt like I was my avatar’s body.

Q14 During the simulation, I felt like my avatar and my real body became one and the same.

Q15 I felt like my avatar was an extension of me.

Q16 I felt like my avatar was me.

Q17 When something happened to my avatar, I felt like it was happening to me.

C Avatar Shape Keys

The following expression shape keys in Table 4 (next page) are all implemented shapes on the avatar. The names are references to defined shapes in the Unified Expressions standard v5.0 [14].

Table 4 Implemented Unified Expressions [14] Shape Keys

Base Shapes		
Basis		
Eyes		
EyeLookOutRight	EyeLookInRight	EyeLookUpRight
EyeLookDownRight	EyeLookOutLeft	EyeLookInLeft
EyeLookUpLeft	EyeLookDownLeft	EyeClosedRight
EyeClosedLeft	*EyeDilation*^{C}	*EyeConstrict*^{C}
Brows		
BrowPinchRight^U	BrowPinchLeft^U	BrowLowererRight^U
BrowLowererLeft^U	BrowInnerUpRight^U	BrowInnerUpLeft^U
BrowOuterUpRight^U	BrowOuterUpLeft^U	
Nose		
NoseSneer^C		
Cheeks		
CheekSquintRight	CheekSquintLeft	CheekPuffRight
CheekPuffLeft	CheekSuckRight	CheekSuckLeft
Jaw		
JawOpen	MouthClosed	JawRight
JawLeft	JawForward	JawBackward
Lips		
LipSuckCornerRight	LipSuckCornerLeft	*LipSuckUpper*^{C}
LipSuckLower^C	*LipFunnel*^{C}	*LipPucker*^{C}
Mouth		
MouthUpperUpRight	MouthUpperUpLeft	MouthUpperDeepenRight
MouthUpperDeepenLeft	MouthFrownRight	MouthFrownLeft
MouthStretchRight	MouthStretchLeft	MouthDimpleRight
MouthDimpleLeft	MouthRaiserUpper	MouthRaiserLower
MouthTightenerRight	MouthTightenerLeft	
MouthRight^C	*MouthLeft*^{C}	*MouthPress*^{C}
MouthSmileRight^C	*MouthSmileLeft*^{C}	*MouthLowerDown*^{C}
Tongue		
TongueOut	TongueUp	TongueDown
TongueRight	TongueLeft	TongueRoll
TongueTwistRight	TongueTwistLeft	

All shapes are base shapes unless noted otherwise
Shape^C: Combined shape, base shape keys of combination also exist.
^{U}: Unused shape, unmapped by tracking directly. Might still be operated by a different feature