Comparing technologies for conveying emotions through realistic avatars in virtual reality-based metaverse experiences

Received <day> <Month>, <year>; | Revised <day> <Month>, <year>; | Accepted <day> <Month>, <year>

DOI: xxx/xxxx

SPECIAL ISSUE PAPER

Comparing Technologies for Conveying Emotions Through Realistic Avatars in Virtual Reality-based Metaverse Experiences

Alessandro Visconti*[1] | Davide Calandra[1] | Fabrizio Lamberti[1]

[1]DAUIN, Politecnico di Torino, Italy

Correspondence
*Alessandro Visconti Email:
{alessandro.visconti}@polito.it

Abstract

With the development of metaverse(s), industry and academia are searching for the best ways to represent users' avatars in shared Virtual Environments (VEs), where real-time communication between users is required. The expressiveness of avatars is crucial for transmitting emotions that are key for social presence and user experience, and are conveyed via verbal and non-verbal facial and body signals. In this paper, two real-time modalities for conveying expressions in Virtual Reality (VR) via realistic, full-body avatars are compared by means of a user study. The first modality uses dedicated hardware (i.e., eye and facial trackers) to allow a mapping between the user's facial expressions/eye movements and the avatar model. The second modality relies on an algorithm that, starting from an audio clip, approximates the facial motion by generating plausible lip and eye movements. The participants were requested to observe, for both the modalities, the avatar of an actor performing six scenes involving as many basic emotions. The evaluation considered mainly social presence and emotion conveyance. Results showed a clear superiority of facial tracking when compared to lip sync in conveying sadness and disgust. The same was less evident for happiness and fear. No differences were observed for anger and surprise.

KEYWORDS:
avatar representation; emotions; facial and eye tracking; lip sync approximation

1 | INTRODUCTION

In recent years, the development of immersive technologies like Virtual, Augmented, and Mixed Reality (VR, AR, and MR) or, in general, eXtended Reality (XR), led to an amount of studies investigating their impact in a variety of fields[1]. Among possible application areas, social VR and collaboration in XR are those which were characterized by the most rapid adoption, also due to the recent pandemic situation[2].

In this context, with the advent of the concept of metaverse(s), new research is being carried out to identify the best methods to represent the users' avatars in shared Virtual Environment (VEs). An avatar is one's interface to other users, through which the owner's movements and expressions are transferred into the VE[3]. Avatar representations can differ in various aspects. They can be realistic[4], or cartoon-looking[5]. They can also differ in terms of complexity[6], ranging from ergonomic and minimalist visualizations[7] in which only the user's head and hands are shown, up to configurations much closer to reality, with complete figures from head to toes[8]. The latter may be also provided with facial expressions[2]. Research on avatars and their representation, though, is still in an early stage. In particular, there are only a few studies aimed at measuring the impact of the representation of an avatar from a psycho-sociological and perceptual point of view, especially in XR applications[9]. Indeed, some studies

showed that users prefer avatars populating VEs to be realistic, full-body, and endowed with facial expression[8,2,10]. Capturing facial expressions, however, is still challenging when using, e.g., a Head-Mounted Display (HMD) for VR, although devices supporting this functionality are appearing on the market given the importance of these social cues for emotions conveyance[11]. An alternative way to reproduce facial expressions in these situations is to use lip synchronization (*lip sync*), i.e., synchronized animations on a speaking avatar[12].

Moving from these premises, this work operates a comparison between the above modalities for conveying emotions through realistic full-body avatars in a multi-user VR-based scenario. The two modalities, implemented with currently available consumer-level hardware (HTC Vive Facial Tracker for the first one) and software (SALSA Lip Sync Suite v2 for the second one), have been evaluated by means of a user study involving 28 participants. The participants were asked to wear the HMD and observe an avatar in a VE while playing six scenes, each representing one of the Ekman's basic emotions[13]. Each scene was played two times (one per modality) with a pseudo-randomized order of exposition. In this way, it was possible to evaluate the performance of the two modalities in a context as similar as possible to the conditions of a real social VR scenario. After spectating a given scene in both modalities, participants blindly evaluated them in terms of relevant social presence aspects like comfort, expressiveness, realism, naturalness, pleasantness, and emotion conveyance through direct comparisons. The experimental results showed the superiority of the facial tracking-based approach with respect to lip sync approximation for most of the emotions (precisely, sadness, fear, disgust, and happiness), whereas nothing could be said for two of them (surprise and anger).

2 | RELATED WORKS

To date, the impact of avatar appearance in multi-user XR scenarios has been investigated by many literature works[14,15,16,17,18,19]. In particular, researchers have examined how characteristics of an avatar reconstruction, among others, aesthetic traits, skills, movements, and behavior, actually influence relevant aspects related to interpersonal[20] and non-verbal[21] communication, advertising[22], etc.

For instance, Roth et al.[23] investigated the effects of the reduction of social information and behavior channels in immersive VEs with full-body, humanoid avatar embodiment. Their results showed that the lack of realism of the avatar hindered and constrained social interactions. The absence of an implementation (and an analysis) of facial expressions and gaze, however, limited the generalizability of the work. The realism of humanoid avatars was also investigated in Dobre et al.[10]. In particular, the authors evaluated the effect of both realistic and cartoonish, full-body avatars on sense of presence in a MR-based telepresence scenario (i.e., an online meeting). The results of the investigation showed that the nonverbal behavior of the realistic avatar was perceived as more appropriate to the interaction and more useful for understanding others than that of the cartoonish one. Based on these studies, it appears that realistic avatars represent the optimal solution from various, fundamental perspectives in the context of social VR (i.e., presence and social interaction). Therefore, it was chosen to consider realistic avatars as the basis for the current investigation.

Once the overall appearance of the avatar was decided, the focus was shifted on possible alternatives for body reconstruction. For instance, Yoon et al.[4] analyzed the effect of avatar appearance on social presence with three levels of body parts visibility (head & hands, upper-body, and whole-body) in AR. The authors found that the realistic, whole-body avatar was perceived as being the best setup for remote collaboration. Similarly, Calandra et al.[8] studied two techniques for avatar representation in multi-user VR scenarios, i.e., a head & hands configuration consisting in displaying the 3D models of the VR equipment (HMD and controllers) worn by the user, and a full-body reconstruction obtained by blending Inverse Kinematics (IK) and animations. Voice communication was guaranteed by means of a VOIP channel, although none of the two techniques provided facial animations. The comparison was performed in the context of an emergency training scenario already presented in[24], in terms of embodiment, social presence, and usability. Results showed a preference for full-body reconstruction in critical aspects such as mutual awareness, mutual attention, mutual understanding, immersion, aesthetics and multi-user interaction. Taking into account these further works, it was determined that full-body avatars could represent a better solution in multi-user experiences compared to representations with lower visibility. For this reason, full-body avatars were considered in this study.

A fundamental feature of realistic humanoid avatars is the possibility to show facial expressions, which can be used as an additional layer of communication in addition to voice and body gestures. The inclusion of facial and eye tracking was explored, e.g., by Kasapakis and Dzardanova[25], who investigated the performance of a multi-user VR learning environment populated with a high-fidelity educator's avatar. The avatar featured facial cues, eye and body motion recorded in real-time (using a VIVE Pro HMD, five VIVE Trackers to track the motion of the pelvis, hands, and feet, ManusVR Gloves, Pupil Labs Eye Tracking

system, and BinaryVR to map facial expressions to blendshapes). According to the reported results, the body and eye movement, together with the facial cues of the educator's avatar helped 90% of the students to maintain their attention during the lecture, thereby increasing their understanding of the concepts presented.

The direct mapping of a user's facial expressions onto an avatar's face has indeed advantages over an expressionless avatar. This mapping, however, typically requires additional hardware or costly VR systems. Hence, more "approximate" software solutions are frequently preferred, which still offer higher functionality levels than a completely static face. One solution that falls into this category is the so-called lip sync approximation, which makes it possible to generate plausible animations of the lower part of the avatar's face based on real-time audio. The use of these techniques as an alternative to audio alone has been already studied. For instance, Hube et al.[12], extended a previous work[26] to examine the influence of facial visual parameters on virtual avatars in VR. Starting from a pre-determined set of audio files, the authors extracted the emotions represented into them and operated a real-time lip sync approximation (through SALSA Lip Sync Suite) with the aim to present this information on a virtual avatar in a VE. The experiment consisted in placing the participants in an immersive VE in front of a half-body, realistic-looking avatar, and making them observe the generated avatar behaviour while listening to the associated audio. This visualization was then compared with an audio-only variant, in which the avatar remained still. The obtained results indicated the superiority of using additional visual parameters on the avatars' face, as they could help to determine the emotions in the audio clip. This study however, did not consider the use of body language to enhance the conveyance of a state of mind. In particular, it only evaluated a small number of possible non-verbal cues, important features that users typically rely onto for better identifying a particular emotion. By considering the last two studies, it can be observed that available approaches to convey users' emotions through their avatars' facial expressions are characterized by different levels of functionality and deployability.

Summarizing the above review, it appears that full-body, realistic-looking avatars proved to be the most effective representation technique under various point of views and in a wide range of scenarios. Moreover, it also seems that the presence of facial cues, in the form of facial expressions, tended to provide an improved experience compared to expressionless avatars. These indications were used as a starting point for the realization of the work. According to the review, to implement facial expressions different approaches have been pursued. To the best of the authors' knowledge, however, a comparison of two of the most promising approaches, i.e., facial and eye tracking, and lip sync approximation, applied to full-body animated avatars in the context of multi-user social VR scenarios has not been performed yet. The current paper tackles this lack, formulating the hypothesis that the employment of a facial tracking technology could, in general, better convey the emotional content of user's behavior with respect to lip sync approximation, which only recreates the animation of the lower part of the face. The facial capture system is capable of reproducing a broad spectrum of facial expressions, whereas the automated lip-sync tool solely concentrates on synchronizing the lips and doesn't consider the rest of the face. As a result, the difference between the two modalities may appear obvious if one considers their contribution as isolated from everything else. In real-life social VR scenarios, however, the facial component cannot be separated from voice and body movements. Moreover, in some situations, facial expressions may not be the predominant way to convey emotions. Thus, for some emotions, a less pronounced difference between the two modalities may be expected.

3 | MATERIALS AND METHODS

In this section the technologies that were used to implement the two configurations considered in the study are discussed, along with the scenario against which they were evaluated.

3.1 | Configurations and Technologies

The VE was developed in Unity 2021.3 as a OpenXR application. The VR kit selected for the experiment was an HTC Vive Pro Eye[1]. The basis for both the considered representations was a realistic-looking, full-body avatar implementation developed in a previous work[8]. In particular, the VRIK module of FinalIK[2] was used to obtain a plausible full-body motion starting from the position and orientation of the HMD and the hand controllers. This module uses IK for the upper body, and a blending of animations to manage walking. The user's voice was captured through the HMD microphone.

[1]HTC Vive Pro Eye: www.vive.com/us/product/vive-pro-eye/overview/
[2]RootMotion FinalIK: assetstore.unity.com/packages/tools/animation/final-ik-14290

For the first modality being studied, in the following referred to as Facial+eye Tracking (FT), the selected HMD (already provided with eye tracking) was equipped with an HTC Vive Facial Tracker[3] (Figure 1 a). The reasons behind this choice are

(a) (b)

FIGURE 1 VR setup used for the experiment (HTC Vive Pro Eye + Facial Tracker) (a). Close-up of the user's avatar (b).

manifold. Firstly, the HTC Vive is one of the most appreciated ecosystems by social VR users due to the possibility to employ additional tracking devices to obtain a full-body reconstruction[27]. Secondly, the HTC Vive Pro Eye is already provided with eye tracking capabilities, and can be easily and seamlessly integrated with the HTC Vive Facial tracker to add the relative functionalities, thus representing a ready-to-use consumer solution for this purpose. This configuration allowed to perform a real-time mapping between the user's facial expressions and eye movements and a given mesh provided with eye and facial rig (in terms of blendshapes). In this case, the mesh was the full-body avatar, which was created with Autodesk Character Generator[4]. The character models created with this tool are automatically provided with a full rig (body and face), can be seamlessly integrated with FinalIK, and are almost compatible with the Vive Tracker facial rig (except for the "frown" blendshape, that had to be manually added in Autodesk Maya by merging other ones).

The second modality, later referred to as Lip-Sync approximation (LS), did not rely on additional hardware modules, as it used real-time algorithms to generate a plausible facial motion starting from an audio clip. To this purpose, similarly to what was done in[12], the SALSA LipSync Suite[5] asset was employed. Along with the main asset, the One-Click automatic configurator for Adobe Character Generator was used to guarantee a smooth integration with the considered humanoid mesh, along with the Eyes module to generate a plausible random eye motion. The choice of SALSA as lip sync approximation was due to the fact that it is a very common software solution to provide real-time, plausible facial animation to talking avatars in Unity-based VR experiences. In fact, it represents a good trade-off between the cost for the end-user (no hardware required) and animation quality (e.g., the upper face can be either animated randomly, or kept non-animated). Moreover, as seen in the above review, it has also been the subject of investigations in previous works on the topic.

[3] HTC Vive Facial Tracker: www.vive.com/us/accessory/facial-tracker/
[4] Autodesk Character Generator: charactergenerator.autodesk.com/
[5] SALSA LipSync Suite: assetstore.unity.com/packages/tools/animation/salsa-lipsync-suite-148442

3.2 | Scenario

In order to better evaluate the contribution of the studied modalities in terms of emotion conveyance, it was decided to perform an analysis per single emotion, similarly to what done in [12]. The authors of that work focused on three main emotions (happiness, sadness, and anger). In the present study, it was decided to widen this set to cover all the six Ekman's basic emotions [13] (anger, happiness, sadness, fear, disgust and surprise). For each emotion, the participants were shown an actor starring a short script. Each script was written with the aim to maintain the feeling and the overall perception within the field of the particular emotion represented, based on the scales of positivity and value that are generally used to define perceived emotions in studies on the subject [28]. Furthermore, an attempt was made to make the scripts as "social" as possible, and not to make them longer than a minute each. Figure 1 b shows some excerpts from the scenes that the participants spectated and rated. The original full scripts in Italian, their English translations, and the recordings of the six scenes are available for download[6].

In order to minimize the possible sources of bias associated with the need to watch the same scenes multiple times, it was decided to avoid relying on an a live performance of the actor in a real multi-user scenario, but rather to opt for a simulated one. Thus, the actor's performance was pre-recorded; an avatar driven by these recordings was then presented to the participants, effectively uniforming the experience. A single actor was involved in the acting of all the scenes. The actor was a 26-year-old man practicing acting as an amateur. The scenes were replayed and recorded several times. Finally, the best six recordings were selected, one per emotion. It is important to note that voice intonation can also influence emotion conveyance [29,30,31]. For this reason, the actor was requested to adopt the intonation that would best match the emotion to be conveyed in the scene. Then, in selecting the recording, the appropriateness of the intonation was also taken into account.

Recorded inputs included voice (microphone), movements (HMD and controllers), and facial expressions (blendshapes), which would be the data to be transmitted over the network for driving the representation of each user's avatar in a real multi-user scenario. For the body, FinalIK was in charge of driving the body motion based on the head and hands movements for both the evaluated modalities. For the FT modality, all the recorded data were synchronized and reproduced at the same time, obtaining a result similar to having an actor playing in real-time in a real multi-user scenario. The same recordings were used to drive the representation for the LS modality, simply discarding the data related to the facial blendshapes and eyes motion, keeping the same body and voice input, and using SALSA to approximate the lip sync. This choice made it possible to keep the body movements and voice intonation perfectly identical in the two conditions investigated. Given the focus on facial expressions, this was essential to remove possible sources of bias.

3.3 | Experimental Procedure

The experimental activity was designed as a within-subjects user study, with a sample of 28 participants (16 males, 12 females) aged between 21 and 68 years, recruited among the students and staff at the authors' university. The hardware used for the experiment was the same used for the recording of the six scenes.

Initially, the participants were requested to fill in a brief demographic questionnaire regarding age, gender, general experience with VR, and experience with multi-user social VR/metaverse applications. A large part of the sample answered that they had limited experience with VR technology, whereas the vast majority indicated they did not have previous experience with social VR or metaverse(s). In particular, ∼61% of the participants rated their experience with VR technology with a score equal to or below 2 on a scale from 1 to 5, where 1 indicated almost no experience and 5 indicated a daily use. Regarding the fruition of social VEs and metaverse(s), the average scores were even lower, with ∼78% of the participants who stated they almost never experienced this kind of applications. The participants were told that the study was focusing on the expressiveness and communication ability of the avatar, in particular, of the upper body, where the three main communication factors that intervene in any conversation in the real world (tone of the voice, arm and hand gestures, and facial expressions) are more evident.

After this introduction, the participants were provided with the HMD and invited to enter a VE depicting a theater [7], where they could see the actor's recorded avatar. In each scene, the participants were free to move on the stage via natural walking around the avatar to observe gestures and expressions from the preferred position (Figure 2). This freedom allowed to recreate conditions similar to those users would find in a social VR environment. Latin square order was used to balance the exposition to the six scenes (emotions). Each scene was played one time per modality, in a randomized order. Thus, the participants were not aware of which modality they were actually experiencing (modality 1 or modality 2, based on the order of exposition). After

[6] Scene scripts and recordings: http://bit.ly/3xNnH52
[7] Madame Walker Theatre: https://sketchfab.com/3d-models/madame-walker-theatre-98ba4154bbb644bb9cb4d9c68d7dd87b

FIGURE 2 Area of the stage scenario in which the participant (represented by the white HMD) can move around to spectate the avatar during the acting.

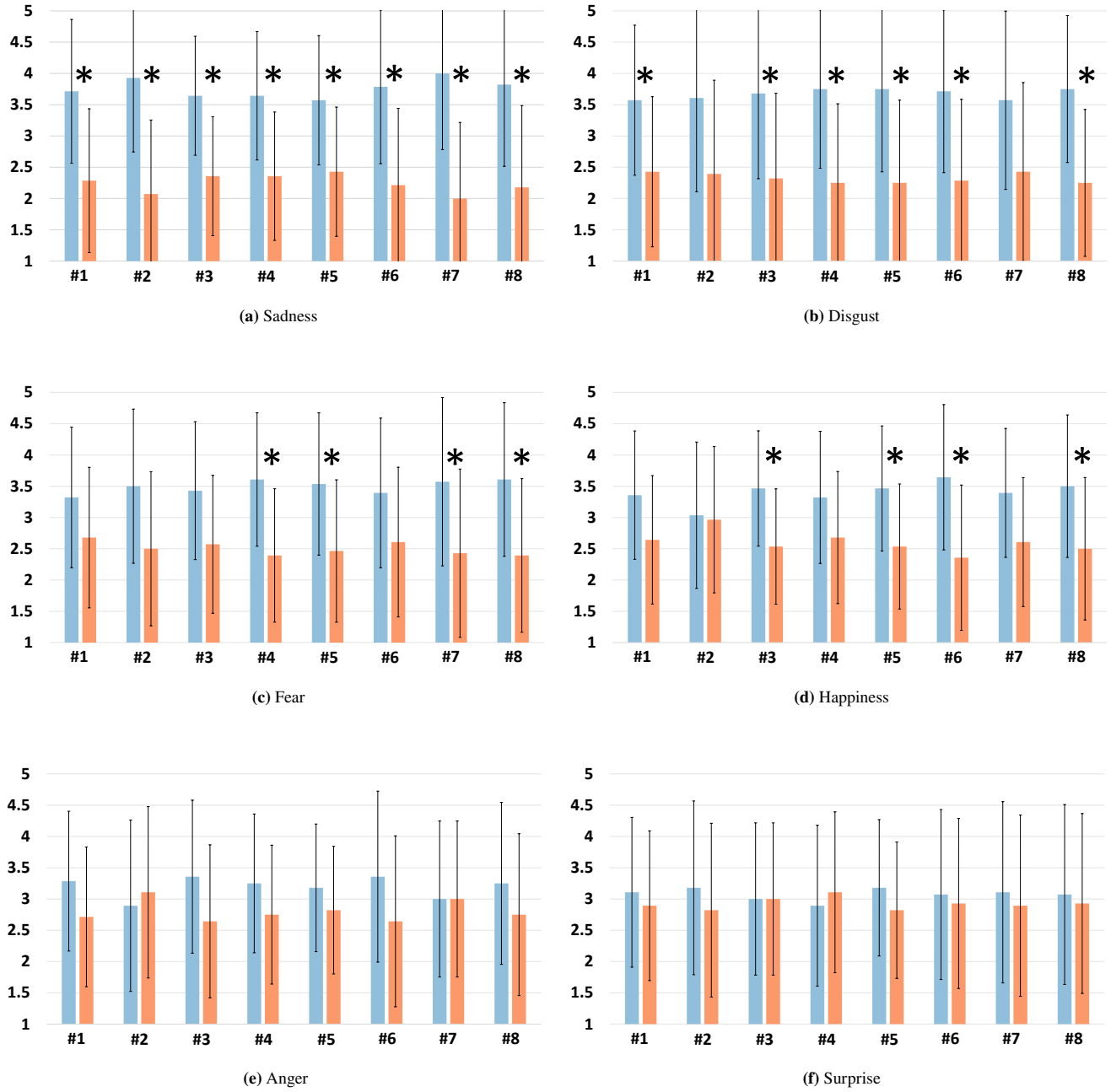
having watched a pair of scenes referring to the same emotion, the participants were asked to answer a comparative questionnaire. Previous works in this field did not provide a standard methodology suitable for this kind of investigation. Therefore, it was necessary to create an ad-hoc set of questions aimed to measure relevant aspects, such as the comfort with the representation, expressiveness, realism, naturalness, likelihood, pleasantness, emotional conveyance, and overall preference. This process was repeated for the six emotions. In order to limit the possible ambiguity of questions related to comfort and pleasantness when applied to negative emotions (i.e., disgust, anger, and sadness), the participants were told to approach the scenes with a neutral and detached point of view with respect to the emotion expressed. Finally, the participants were asked to provide comments in the form of open feedback. The full questionnaire is available for download [8].

4 | RESULTS

In this section, the experimental results (summarized in Figure 3) are presented and discussed. The Shapiro-Wilk test was used to evaluate the normality of data. Since data resulted as non-normally distributed, the non-parametric Wilcoxon signed-rank test was performed, with a 5% significance threshold (p-value $< .05$). After running the experiments, a power analysis was performed, and a value of ∼0.99% with a minimum effect size equal to ∼0.85% was obtained. This outcome confirmed the significance of results, which in the following are discussed on a per-scene basis.

In particular, the results (in terms of average scores standard deviations) for the "Sadness" scene are provided in Figure 3 a. In this case, the FT modality was judged as significantly superior to the LS modality across all the eight dimensions investigated through the questions asked to the participants. In particular, FT was more convincing than LS with regard to the general expressiveness of the avatar (3.93 vs 2.07, $p = .002$) and the conveyance of emotions (4.00 vs 2.00, $p = .001$). Similar trends can be observed for overall pleasantness (3.78 vs 2.22, $p = .006$), feeling of being at ease (3.71 vs 2.29; $p = .01$), realism (3.64 vs 2.36, $p = .004$) and overall preference (3.82 vs 2.18, $p = .009$). Less different, but still significant, are the results regarding

[8] Questionnaire: http://bit.ly/41nn13G



(a) Sadness **(b)** Disgust

(c) Fear **(d)** Happiness

(e) Anger **(f)** Surprise

FIGURE 3 FT LS

Results of the questionnaires for each scene/emotion (a–f). Statistically significant results ($p < .05$) are marked with a star (*) symbol. Standard deviation is expressed through error bars. Questions numbered from #1 to #8. Which of the two modalities: #1. Let you feel more comfortable? #2. Was more expressive? #3. Was more realistic? #4. Was more natural? #5. Was more plausible? #6. Was more pleasant? #7. Better conveyed the emotional content? #8. Did you prefer (overall)?

naturalness of expressions (3.64 vs 2.36, $p = .007$) and likelihood (3.57 vs 2.43, $p = .014$). This outcome is in line with the expectations, since the emotion of sadness is mostly conveyed by eye and mouth movements [13] and, based on observations made in the development steps, the output of the FT modality was found to be more accurate than the approximation provided by the LS.

Moving to the "Disgust" scene (Figure 3 b), the results again highlighted a significant preference of the participants in favour of the FT modality over the LS one, albeit not completely as for the previous scene. For this emotion, the participants found the FT as more realistic (3.67 vs 2.33, $p = .015$), more natural (3.75 vs 2.25, $p = .008$), more likely overall (3.75 vs 2.25, $p = .009$), more pleasant (3.71 vs 2.29, $p = .02$), more comfortable (3.57 vs 2.43 , $p = .04$) and overall preferable (3.75 vs 2.25, $p = .004$) than the counterpart. Concerning the expressiveness and the emotion conveyance, instead, results did not significantly differ. This outcome may be related to the fact that some emotions (in particular, disgust and surprise) are often expressed impulsively and, thus, on a shorter time frame compared to others that may be more persistent over time (anger, sadness, and happiness). Thus, it may be harder to notice differences between the two modalities, especially if the participant happens to be not particularly focused on the other peer's face at the very moment in which the emotion is impulsively expressed. This aspect was also reported by some participants in the open feedback section.

For what it concerns the "Fear" scene, (Figure 3 c), the FT modality was judged as superior to the LS one for four dimensions. In particular, the participants found FT as more natural (3.60 vs 2.40, $p = .01$), more plausible (3.53 vs 2.47, $p = .02$), more capable of conveying the emotional content of the scene (3.57 vs 2, 43, $p = .04$) and preferred it, overall (3.60 vs 2.40, $p = .02$). For the other dimensions, no significant differences were found. This outcome may be explained by the important role of the upper face for expressing this emotion, e.g., in terms of cheekbones eyebrows motion, which is not managed by LS. Furthermore, based on the open feedback, some participants noticed that the LS did not generate any mouth movement when the actor was gasping, which is a common action in this particular scene. This behavior is likely due to a limitation of the software, which does not trigger animations for sounds different than speech.

Regarding the "Happiness" scene (Figure 3 d), no clear winner was observed again. FT was perceived as significant better than LS in terms of realism (3.46 vs 2.54, $p = .03$), general likelihood (3.46 vs 2.54, $p = .04$), pleasantness (3.64 vs 2.36, $p = .016$) and general preference (3.50 vs 2.50, $p = .05$). It should be noted that, for what it concerns expressiveness, average scores for the two modalities were extremely similar. This result may be related to the fact that the HTC Vive Facial Tracker showed some issues in detecting the user's smile or in properly triggering the blendshape associated with the smiling action. This explanation is in line with some of the comments by the participants, who stated that they had not seen the avatar smiling despite the context, and therefore both the modalities appeared as not particularly expressive. For what it concerns emotion conveyance, results were close to the significance threshold (3.39 vs 2.60, $p = .073$), suggesting that with a larger sample size the difference could possibly become significant.

Nothing can be said for the "Anger" scene (Figure 3 e), probably again due to the impulsive nature of the given emotion. Moreover, some participants commented in the open feedback section that the scene heavily relied on speech, for which LS is probably on par with FT.

Similarly, also for the "Surprise" scene (Figure 3 f) no significant differences were found. This outcome may be related to the fact that surprise is a very ephemeral emotion, very expressive for short periods but then difficult to be maintained for a long time, similarly to anger or sadness. Thus, it may be possible that the presence of other elements (such as body movements) and the audio clip itself provided already enough cues that other differences in the face could not be spotted, despite the theoretical higher fidelity of FT with respect to LS for facial expressions not related with speech.

5 | CONCLUSION AND FUTURE WORKS

In this work, two avatar representations were compared by means of a user study in terms of emotion conveyance in the context a social VR/metaverse-oriented scenario. Both the modalities rely on a realistic full-body avatar representation obtained by blending the contribution of IK and animations. The first modality, labeled Facial+eye Tracking (FT), achieves the described representation by employing additional hardware elements (HTC Vive Pro Eye + Facial Tracker) to track the user's eyes and facial expressions. The second modality, labeled Lip-Sync approximation (LS), takes advantage of a commercial software solution (SALSA Lip Sync) to drive a visually-plausible approximation of the facial movements related to speech based on the audio captured by the HMD microphone.

The study involved 28 participants who blindly evaluated the two configurations by separately spectating an avatar performing six pre-determined scenes related to the six Ekman's basic emotions[13]. The evaluation covered important social presence aspects such as comfort, expressiveness, realism, naturalness of the expressions, likelihood, pleasantness, emotion conveyance, and overall preference. The experimental results showed a clear superiority of FT with respect to LS for what it concerns the

conveyance of sadness and disgust for most of the analyzed dimensions, highlighting the importance of having a faithful representation of the user's eyes and facial cues in the selected use case. The same trend was observed for happiness and fear, but to a lesser extent, probably because of the mentioned issue with the hardware regarding the detection of the smiling action, as well as of the more ephemeral and impulsive nature of the emotion for the latter emotion. Finally, nothing can be said for the remaining emotions. For anger, this outcome may be related to the fact that the scene depiting it was mostly based on speech, for which LS (via SALSA Lip-Sync) showed solid performance. Regarding surprise, the ephemerality of the emotion may have played again a role in making it not so noticeable with both FT and LS.

Future developments will be devoted to extend the analysis reported in this work, for instance by considering in the evaluation further approaches and technologies that can be used to convey emotions in social VR scenarios, either hardware-based (e.g., the facial tracking of the Meta Quest Pro) or software-based (e.g., emotion detection algorithms, in combination with SALSA to better approximate the whole face). The scenes representing the emotions may be refined too, in order to address the limitations related to the reduced "screen time" of some emotions (e.g., disgust and surprise), with respect to other more persistent emotions. Similarly, the anger scene may be modified to be less speech-based, and to include more anger-related facial expressions (e.g., shouts). The incorporation of multiple scenes to represent each emotion could be also considered. In fact, since emotions may be expressed in different ways depending on the situation, the use of different scripts for a given emotion would allow to achieve more general results. Furthermore, other factors influencing non-verbal communication may be included in the evaluation, like body gestures and voice intonation. To this purpose, full body tracking solutions may be considered as alternative to IK. Finally, further avatar representations may be considered (e.g., cartoonish, non-human, abstract), possibly integrated with other techniques for conveying emotions (e.g., with faceless avatars, additional user interface elements to communicate the user's emotional status), to study their suitability for social VR scenarios and metaverse(s).

5.1 | Acknowledgements

This work has been carried out in the frame of the VR@POLITO initiative. The authors want to to acknowledge the help of Gianmario Lupini, who contributed to the development of the system and to the experimental activity.

References

1. Xi N, Chen J, Gama F, Riar M, Hamari J. The challenges of entering the metaverse: An experiment on the effect of extended reality on workload. *Information Systems Frontiers* 2023; 25(2): 659–680. doi: 10.1007/s10796-022-10244-x

2. Hart JD, Piumsomboon T, Lee GA, Smith RT, Billinghurst M. Manipulating Avatars for Enhanced Communication in Extended Reality. In: *Proc. of 2021 IEEE International Conference on Intelligent Reality (ICIR)*; 2021: 9-16. doi: 10.1109/ICIR51845.2021.00011

3. Freeman G, Zamanifard S, Maloney D, Adkins A. My Body, My Avatar: How People Perceive Their Avatars in Social Virtual Reality. In: *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*; 2020: 1–8. doi: 10.1145/3334480.3382923

4. Yoon B, Kim Hi, Lee GA, Billinghurst M, Woo W. The Effect of Avatar Appearance on Social Presence in an Augmented Reality Remote Collaboration. In: *Proc. of 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*; 2019: 547-556. doi: 10.1109/VR.2019.8797719

5. Rhee CH, Lee CH. Cartoon-like Avatar Generation Using Facial Component Matching. *International Journal of Multimedia and Ubiquitous Engineering* 2013; 8(4): 69–78.

6. Schäfer A, Reis G, Stricker D. A Survey on Synchronous Augmented, Virtual and Mixed Reality Remote Collaboration Systems. *arXiv preprint arXiv:2102.05998* 2021: 1–26. doi: 10.48550/arXiv.2102.05998

7. De Lorenzis F, Pratticò FG, Lamberti F. HCP-VR: Training First Responders through a Virtual Reality Application for Hydrogeological Risk Management.. In: *Proc. of 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 2: HUCAPP*; 2022: 273-280. doi: 10.5220/0011007800003124

8. Calandra D, Pratticò FG, Lupini G, Lamberti F. Impact of Avatar Representation in a Virtual Reality-Based Multi-user Tunnel Fire Simulator for Training Purposes. In: *Computer Vision, Imaging and Computer Graphics Theory and Applications*. 1691. ; 2023: 3–20. doi: 10.1007/978-3-031-25477-2_1

9. Nowak KL, Fox J. Avatars and computer-mediated communication: A review of the definitions, uses, and effects of digital representations. *Review of Communication Research* 2018; 6: 30-53. doi: 10.12840/issn.2255-4165.2018.06.01.015

10. Dobre GC, Wilczkowiak M, Gillies M, Pan X, Rintel S. Nice is Different than Good: Longitudinal Communicative Effects of Realistic and Cartoon Avatars in Real Mixed Reality Work Meetings. In: *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*; 2022: 1–7. doi: 10.1145/3491101.3519628

11. Hart JD, Piumsomboon T, Lawrence L, Lee GA, Smith RT, Billinghurst M. Emotion Sharing and Augmentation in Cooperative Virtual Reality Games. In: *Proc. of Annual Symposium on Computer-Human Interaction in Play Companion (Extended Abstracts)*; 2018: 453–460. doi: 10.1145/3270316.3271543

12. Hube N, Vidackovic K, Sedlmair M. Using Expressive Avatars to Increase Emotion Recognition: A Pilot Study. In: *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*; 2022: 1–7. doi: 10.1145/3491101.3519822

13. Ekman P. Are there basic emotions?. *Psychological Review* 1992; 99(3): 550–553. doi: 10.1037/0033-295X.99.3.550

14. Heidicker P, Langbehn E, Steinicke F. Influence of avatar appearance on presence in social VR. In: *Proc. of 2017 IEEE Symposium on 3D User Interfaces (3DUI)*; 2017: 233-234. doi: 10.1109/3DUI.2017.7893357

15. Otto O, Roberts D, Wolff R. A review on effective closely-coupled collaboration using immersive CVE's. In: *Proc. of Proceedings of the 2006 ACM international conference on Virtual reality continuum and its applications*; 2006: 145–154. doi: 10.1145/1128923.1128947

16. Kim K, Boelling L, Haesler S, Bailenson J, Bruder G, Welch GF. Does a Digital Assistant Need a Body? The Influence of Visual Embodiment and Social Behavior on the Perception of Intelligent Virtual Agents in AR. In: *Proc. of 2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*; 2018: 105-114. doi: 10.1109/ISMAR.2018.00039

17. Krekhov A, Cmentowski S, Krüger J. The illusion of animal body ownership and its potential for virtual reality games. In: *Proc. of 2019 IEEE Conference on Games (CoG)*; 2019: 1–8. doi: 10.1109/CIG.2019.8848005

18. Salem B, Earle N. Designing a Non-Verbal Language for Expressive Avatars. In: *Proc. of 3rd International Conference on Collaborative Virtual Environments*; 2000: 93-–101. doi: 10.1145/351006.351019

19. Bailenson JN, Yee N, Merget D, Schroeder R. The effect of behavioral realism and form Realism of Real-Time Avatar Faces on Verbal Disclosure, Nonverbal Disclosure, Emotion Recognition, and Copresence in Dyadic Interaction. *Presence: Teleoperators and Virtual Environments* 2006; 15(4): 359-372. doi: 10.1162/pres.15.4.359

20. Kotlyar I, Ariely D. The effect of nonverbal cues on relationship formation. *Computers in Human Behavior* 2013; 29(3): 544–551. doi: 10.1016/j.chb.2012.11.020

21. Bente G, Krämer NC. Virtual gestures: embodiment and nonverbal behavior in computer-mediated communication. *Face-to-face communication over the internet: Issues, research, challenges* 2011: 176–209. doi: 10.1017/CBO9780511977589.010

22. Ahn SJ, Bailenson JN. Self-Endorsing Versus Other-Endorsing in Virtual Environments. *Journal of Advertising* 2011; 40(2): 93–106. doi: 10.2753/JOA0091-3367400207

23. Roth D, Lugrin JL, Galakhov D, et al. Avatar realism and social interaction quality in virtual reality. In: *Proc. of IEEE Virtual Reality*; 2016: 277-278. doi: 10.1109/VR.2016.7504761

24. Calandra D, Pratticò FG, Migliorini M, Verda V, Lamberti F. A Multi-role, Multi-user, Multi-technology Virtual Reality-based Road Tunnel Fire Simulator for Training Purposes. In: *Proc. of 16th International Conference on Computer Graphics Theory and Applications*; 2021: 96–105. doi: 10.5220/0010319400960105

25. Kasapakis V, Dzardanova E. Using High Fidelity Avatars to Enhance Learning Experience in Virtual Learning Environments. In: *Proc. of IEEE Conference on Virtual Reality and 3D User Interfaces (Abstracts and Workshops)*; 2021: 645-646. doi: 10.1109/VRW52623.2021.00205

26. Hube N, Lenz O, Engeln L, Groh R, Sedlmair M. Comparing Methods for Mapping Facial Expressions to Enhance Immersive Collaboration with Signs of Emotion. In: *Proc. of 2020 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*; 2020: 30–35. doi: 10.1109/ISMAR-Adjunct51615.2020.00023

27. Campbell IC. This VR doc shows how vibrant virtual life already is without Meta's meddling. https://www.inverse.com/input/reviews/we-met-in-virtual-reality-documentary-shows-metaverse-vibrant-without-facebook; 2022. [accessed 13 April 2023].

28. Russell JA, Weiss A, Mendelsohn GA. Affect grid: A single-item scale of pleasure and arousal. *Journal of personality and social psychology* 1989; 57(3): 493. doi: 10.1037/0022-3514.57.3.493

29. Pierre-Yves O. The production and recognition of emotions in speech: Features and algorithms. *International Journal of Human-Computer Studies* 2003; 59(1): 157-183. doi: 10.1016/S1071-5819(02)00141-6

30. Bänziger T, Scherer KR. The role of intonation in emotional expressions. *Speech Communication* 2005; 46(3): 252-267. Quantitative Prosody Modelling for Natural Speech Description and Generationdoi: 10.1016/j.specom.2005.02.016

31. Rodero E. Intonation and Emotion: Influence of Pitch Levels and Contour Type on Creating Emotions. *Journal of Voice* 2011; 25(1): e25-e34. doi: 10.1016/j.jvoice.2010.02.002

How to cite this article: Visconti A., Calandra, D., and Lamberti F. (202X), Comparing Technologies for Conveying Emotions Through Realistic Avatars in Virtual Reality-based Metaverse Experiences, *Comput Anim Virtual Worlds*, *202X;XX:1–X*.