frontiers in Virtual Reality

METHODS
published: 07 April 2021
doi: 10.3389/frvir.2021.641296

Using a Fully Expressive Avatar to Collaborate in Virtual Reality: Evaluation of Task Performance, Presence, and Attraction

Yuanjie Wu[1]*, *Yu Wang*[2], *Sungchul Jung*[1], *Simon Hoermann*[3] and *Robert W. Lindeman*[1]

[1]Human Interface Technology Lab New Zealand (HIT Lab NZ), University of Canterbury, Christchurch, New Zealand, [2]Beijing Institute of Technology, Beijing, China, [3]School of Product Design, College of Engineering, University of Canterbury, Christchurch, New Zealand

Avatar-mediated collaboration in virtual environments is becoming more and more prevalent. However, current consumer systems are not suited to fully replicate real-world nonverbal communication. We present a novel avatar system for collaboration in virtual reality, which supports high levels of nonverbal expression by tracking behavior such as body movement, hand gesture, and facial expression. The system was built using camera tracking technology only. Therefore, in contrast to many other high-level tracking systems, it does not require users to wear additional trackers on their bodies. We compared our highly expressive system with a consumer setup extended with two body-worn trackers in a dyadic study. We investigated users' performance, such as completion time and accuracy, as well as the presence and interpersonal attraction in a virtual charades game using an asymmetric control scheme. The results show that participants interacting with highly expressive avatars felt more social presence and attraction and exhibited better task performance than those interacting with partners represented using low-expressive avatars. Hence, we conclude that virtual reality avatar systems benefit from a higher level of nonverbal expressiveness, which can be achieved without additional body-worn trackers.

Keywords: CCS concepts: human-centered computing → virtual reality additional key words and phrases: avatar, virtual reality, shared virtual environment, communication, collaboration

OPEN ACCESS

Edited by:
Katja Zibrek,
The Inria Rennes-Bretagne Atlantique Research Centre, France

Reviewed by:
Aryabrata Basu,
Emory University, United States
Ana Serrano,
University of Zaragoza, Spain

***Correspondence:**
Yuanjie Wu
yuanjie.wu@pg.canterbury.ac.nz

Specialty section:
This article was submitted to Technologies for VR, a section of the journal Frontiers in Virtual Reality

Received: *14 December 2020*
Accepted: *26 January 2021*
Published: *07 April 2021*

Citation:
Wu Y, Wang Y, Jung S, Hoermann S and Lindeman RW (2021) Using a Fully Expressive Avatar to Collaborate in Virtual Reality: Evaluation of Task Performance, Presence, and Attraction. Front. Virtual Real. 2:641296. doi: 10.3389/frvir.2021.641296

1 INTRODUCTION

Current virtual reality (VR) technology can enable people to communicate and collaborate in shared virtual environments (SVEs) independently of their geographic locations. The quality and efficiency of communication and collaboration in VR, however, are often impacted by factors, such as virtual environment rendering (Gergle et al., 2013; McVeigh-Schultz et al., 2019), avatar representation (Bombari et al., 2015), latency (Friston and Steed 2014), and state synchronization (Pan and Steed 2017). In particular, avatars play an essential role in social VR, and avatar realism is one of the main factors affecting the sense of presence, interpersonal interactions, and copresence (Steed and Schroeder 2015; Jung and Hughes, 2016; Jung et al., 2017; Jung et al., 2018).

Avatar realism is often used to measure avatar quality, which can be divided into appearance and behavioral realism. Most previous work has been done on visual fidelity (Latoschik et al., 2016; Latoschik et al., 2017), and avatar appearance influences interaction in all shared VEs (Nilsson et al., 2002; Schroeder, 2012). The virtual character represents the user and presents all the verbal and

nonverbal behavior from the real-world player. For communication, humans actively use both verbal and nonverbal behavior for the best representation of their intentions. However, people tend to communicate more through nonverbal behavior (Matsumoto et al., 2012) during social interaction compared to the verbal channel. Therefore, it is essential to study the impact of nonverbal behavior on communication in VR.

Previous research studied some aspects of nonverbal behavior, such as eye gaze (Garau et al., 2003) and facial expressions (Bailenson et al., 2006), which have proven to be important factors in SVEs. Expressive avatar systems (integrating nonverbal behavior, such as body movement, hand gesture, facial expressions, and eye gaze) are limited in current immersive systems due to sensory technologies. Although we still have demand to improve the appearance realism (Bombari et al., 2015), the impacts of expressiveness of avatars in terms of nonverbal behavior have not yet been systematically investigated in communicative and collaborative virtual environments with fully embodied avatars.

In this article, we present a collaborative VR platform that supports asymmetric avatar mediated communication at different levels of avatar expressiveness in terms of nonverbal behavior. We implemented a charades game in the SVE with different expressive avatar conditions to measure copresence, social presence, and interpersonal attraction. "Charades is a game of pantomimes: you have to "act out" a phrase without speaking, while the other members of your team try to guess what the phrase is. The objective is for your team to guess the phrase as quickly as possible" (Dana, 2000). The reason we chose this game is to encourage participants to perform nonverbal behavior to complete an engaging collaborative task. We evaluated the avatar control systems with a dyadic user study, investigating performance in terms of accuracy and completion time.

This research makes the following main contributions: 1) we built a fully expressive avatar control system that supports eye-gaze and mouth rendering combined with tracking natural nonverbal behavior. The system works without the requirement of additional body-worn sensors and tracks hand-gestures in a large area by combining multiple Leap Motion tracking cameras. 2) We evaluated the effect of different levels of avatar nonverbal expressiveness on communication and collaboration in a shared virtual environment.

2 RELATED WORK

2.1 Collaboration in Shared Virtual Environments

Remote communication and collaboration between multiple users in different physical locations are increasingly taking place in SVE. Previous research on communication in SVEs explored questions about performance, social interactions, and leadership (Steed et al. 2019; Bailenson et al., 2002; Becker and Mark, 2002; Slater and Steed 2002; Schroeder, 2012). The SVE quality can impact the synchronous multiuser virtual experience if all the users do not perceive the same state of the VE. The VR system that supports social interaction requires replicating the user's appearance and behavior. The nonverbal cues delivered by the virtual characters in the collaborative virtual environment influence the efficiency of task performance (Roth et al., 2018), and the user's embodiment can lead to higher social presence ratings compare to face-to-face interactions (Smith and Neff 2018).

Pan and Steed (Pan and Steed 2017) developed an SVE to explore the impact of self-avatars on trust and collaboration using virtual puzzles with the HTC Vive and Unity UNET system, which is widely used for supporting multiuser networking. They compared self-avatar, no avatar, and face-to-face conditions, but the avatar was only a visual representation, and only the movement of the controller and not of the actual hand was tracked.

Smith and Neff (Smith and Neff 2018) implemented an SVE for negotiating an apartment layout and placing model furniture on an apartment floor to explore the communication behavior in embodied avatars. Participants could only use limited hand gesture driven by the controllers for communication. Roth et al. (Roth et al., 2019) proposed a software architecture using four data layers to augment social interactions by integrating behavior tracking such as body, eye gaze, and facial expressions into the SVE. Their system was able to support social communication, but participants missed hand-gesture cues. In summary, previous research either omitted tracking of nonverbal behavior or relied on controllers tracking for some limited gestures. This motivated us to explore in this study if increasing the level of expressiveness of an avatar in terms of nonverbal behavior can impact communication and collaboration behavior.

2.2 Avatar Control Systems and Representation

An avatar is a virtual representation of a user and is driven by the user's movements in the virtual world (Bailenson et al., 2004). An avatar system can provide an embodied experience (Slater et al., 2010), and the user can interact with the virtual world through the eyes of the virtual avatar from a first person point of view. Early avatar control systems could not provide a complete embodied experience due to limited tracking technology (tracking area and accuracy), which led to reduced interactivity such as limited or no possibilities of virtual body movement, hand gestures, and facial expressions. Currently, no single system exists that can capture and represent all nonverbal behavior. Hence, to create highly expressive avatars, integrating multiple sensors and systems is required but technically challenging.

Body movement is the primary source of data to control virtual avatars. To achieve high-quality embodied experiences, professional motion-capture systems and suits are often used in avatar-related research (Kilteni et al., 2013; Roth et al., 2016; Spanlang et al., 2014). However, these systems are expensive and motion-capture suites are cumbersome to wear, although providing high accuracy and potentially large tracking areas. In contrast, consumer VR devices such as the Oculus Rift or HTC Vive with spatial controllers are alternative solutions for tracking parts of the body. However, if more parts of the body, for example feet, need to be tracked, extra sensors are required. Most current VR systems are based on a three-tracking-point (one HMD plus two controllers)

solution, with only support "floating" avatars, such as Facebook spaces[1], VR Chat[2], and Mozilla Hubs[3]. Extra trackers are required along with sophisticated inverse kinematic algorithms if movements of arms or legs need to be virtually represented (Aristidou et al., 2017; Caserman et al., 2019). Compared to the HMD and trackers solution, RGB-D camera-based body tracking is a contactless way that can provide more information about body movement. The combination of an RGB-D sensor and VR device is another solution to support body tracking without wearing tracking sensors (Kwon et al., 2017). Users can experience improved articulation control of their avatar using these approaches.

Hand gestures are another essential data source, which can present important nonverbal information. Common VR controllers are possible to trigger specific gestures when certain buttons are pressed, but the remapping strategy is limited. To compensate for these constraints, camera-based tracking devices, such as the Leap Motion controller (LMC), can capture natural hand gestures without using any controller. For example, Wu et al. (2019a) developed a multisensor system that integrates multiple Kinects and an LMC to control an avatar. Other nonverbal cues for avatar control are eye gaze and facial expression. Roth et al. (2017) used an RGB-D sensor to track facial expressions and eye gaze and then mapped the data onto an avatar. In their work (Roth et al., 2019), they present a system architecture for the augmentation of social behaviors in multiuser environments. Their avatar framework can present nonverbal behavior such as facial expression and body posture, but it lacks hand gestures; hence, it is not fully suitable for communication and collaboration tasks that require hand gestures.

3 TECHNICAL SETUP

The experimental setup was implemented in a large room with two different physical systems. Two participants, one for each system, can simultaneously play the game with asymmetric avatar control connected through the local network. Both participants in the dyad can move freely within their 2 m circle, and the tracked movement and gestures are mapped on their avatars in the SVE. The details of the avatar system, network architecture, and software are provided in this section.

3.1 Avatar Control Systems

In this experiment, we adopted two avatar systems with different levels of expressiveness.

3.1.1 Highly Expressive Avatar Control System

Participants who used this avatar control system could control a highly expressive avatar representation with a contactless tracking system.

3.1.1.1 Body Tracking

The full-body movement data was collected by four Kinect v2 devices placed in the corners of the tracking area. This system was based on the work of Wu et al. (Wu et al., 2019a; Wu et al., 2019b) with both body (21 joints including the torso, arms, and legs) and hand-gesture tracking (19 joints with pointing, grasp, and pinch). However, we improved the avatar control algorithm compared to that of Wu et al. (2019a) in the following aspects: 1) we recalculated joint rotations based on joint positions and only used the joint rotations information provided by the Kinect cameras as a reference. 2) Unnatural joint twists were reduced by incorporating information of the skeletal tree and joint hierarchy relationship in the algorithm. The adding father-child nodes relationship can restrict the abnormal joints rotation. 3) Avatar movements were smoothed by calculating each joint's velocity and bone direction, which make the avatar control more natural and realistic.

3.1.1.2 Hand Tracking

The solution used in the work (Wu et al., 2019b) has a limited tracking range and area (single LMC), which requires the user to keep their hands in front of the eyes to avoid tracking loss. If the user moves the hands out of the tracking area, the data switches to the Kinect system, and the hands would only be tracked without finger movement. To address this issue, we built a multi-LMC system with five LMCs installed on a purpose-built mounting frame we attached to the HMD (**Figure 1A**). Each LMC sensor connects to a client machine sending the hand's frame data to a server machine. A shared-view calibration method was used based on the least-squares fitting (LSF) algorithm to process and integrate the multiple-hand data. To avoid incorrect tracking data from a single LMC interfering with the fusion result, we implemented a multi-LMC fusion algorithm based on a two-level evaluation method, namely, a prediction-based and a position-based method. After that, we combined the data from multiple LMCs using a Kalman Filter based on the evaluation results. Compared to the single LMC, our system can enlarge the hand tracking range to 202.16° horizontal and 164.43° vertical.

3.1.1.3 Eye and Mouth Movement

The avatar's eye-gaze direction was the same as the orientation of the head-mounted display (HMD), but small adaptations made it appear more natural. The avatar's eye-gaze direction shifted randomly every few seconds to simulate the eyeball's movement. For example, users look at some direction that is different from the facing direction. In addition, the virtual avatar performed random eye blinking (blink one time every 3 s). Fifteen visemes (Oculus Lipsync, 2019) were added to the virtual character as blend shapes. Each viseme depicted the mouth shape for a specific set of phonemes, which extended mouth-movement rendering possibilities compared to Wu et al. (2019a). The set of mouth shapes was driven by the Salsa LipSync v2 (Crazy Minnow Studio, 2019) Unity plugin, which approximates lip movement in real-time from the audio dialogue.

3.1.2 Further Information About the Multi-LMC System

System Setup

Five LMCs are used in our system. The central LMC is attached in the middle of HMD for capturing hand movement data in front of the user. The lateral LMCs at the four corners of the HMD

[1]https://www.facebook.com/spaces
[2]https://vrchat.com/
[3]https://hubs.mozilla.com/

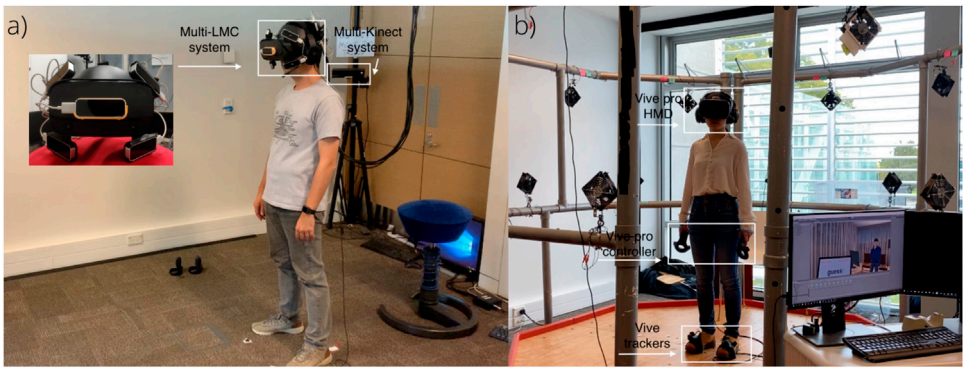

FIGURE 1 | The two avatar control systems used in the experiment: **(A)** Highly expressive avatar control system and **(B)** low expressive avatar control system (Note: extra items shown in these pictures, e.g., fans, were not used in this study).

TABLE 1 | Design parameter for multi-LMC system.

LMC position	Translation (mm)			Rotation (degree)		
	X	y	z	X	Y	z
Top-left	80	50	−60	−35	35	−30
Top-right	−80	50	−60	−35	−35	30
Bottom-left	80	−75	−80	30	35	30
Bottom-right	−80	−75	−80	30	−35	−30

provide supplementary tracking in the top-left, top-right, bottom-left, and bottom-right areas. The lateral LMCs are positioned relative to the observing coordinate system, whose origin is located at the center of the front surface of the HMD with the *x*-axis facing left, the *y*-axis facing up, and the *z*-axis facing forward. According to the maximum position that the human hand can reach (MacAtamney and Corlett 1993), the positioning parameters of the four lateral LMCs are presented in **Table 1**. The parameters ensure that the tracking area is large enough to cover the whole hand movement range, while keeping the overlapping areas to be sufficient for calibration. The error caused by infrared interference in our configuration is negligible (Placidi et al., 2017).

Calibration

We used the built-in recalibration function to calibrate the intrinsic parameters of the individual LMCs in our system. As for extrinsic calibration, we propose an efficient approach to calibrate the multi-LMC array with no dependence on external devices. Because the overlapping tracking range of the LMCs is sufficient for calibration, we devised a shared-view method based on the LSF algorithm to calibrate multiple LMCs. The data flow and process are shown in **Figure 2**.

We set the front LMC as the reference camera. The input is the hand trajectories of the specified hand-joint sampled by the reference LMC and the lateral LMCs in the overlapping tracking area. During sampling, the user needs to flatten their hands and move the hands randomly in the overlapping tracking area (left hand in the overlapping area of central, top-left and bottom-left LMCs, right hand in the overlapping area of

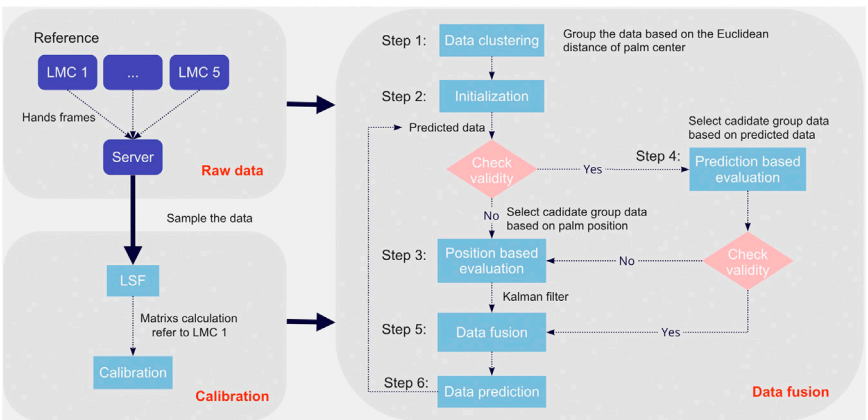

FIGURE 2 | Multi-LMC data flow chart.

central, top-right and bottom-right LMCs). In order to eliminate the error caused by the sampling latency between each LMC, the moving speed of hands should be slow (less than 10 mm per second according to our experience). We use C_r to represent the trajectory from the reference LMC, and use $C_s, s = 1, 2, 3, 4$ to represent the trajectories from the lateral LMCs. After sampling, two point sets $P_r = \{p_r | p_r^{(i)} \in C_r, i = 1 \ldots n\}$ and $P_s = \{p_s | p_s^{(j)} \in C_s, s = 1 \cdots 4, j = 1 \ldots n\}$ are generated from C_r and C_s, respectively.

The next step is to calculate the calibration matrix using the LSF algorithm. The theory of LSF is to find the optimal transformation (consisting of rotation R and translation t), which minimizes the sum of the distance between the coordinates of the matching pairs (Zhang, 1994). We use R_s and t_s to represent the transformation parameters of the sth lateral camera. The objective function of the LSF algorithm can be represented using

$$f(R_s, t_s) = \sum_{i=1}^{n} \|R_s x_i + t_s - y_i\|^2, x_i \in P_s, y_i \in P_r, \quad (1)$$

in which x_i and y_i are a pair of corresponding points between P_s and P_r. Because our method samples the data from the reference camera and the calibrating camera simultaneously, the corresponding-point pair in our method is given by

$$(x_i, y_i) = (p_s^{(i)}, p_r^{(i)}), i = 1 \ldots n. \quad (2)$$

Substituting **Eq. 2** into **Eq. 1**, the objective function then becomes

$$f(R_s, t_s) = \sum_{i=1}^{n} \|R_s p_s^{(i)} + t_s - p_r^{(i)}\|^2. \quad (3)$$

Equation 3 can be solved using the singular value decomposition (SVD) method (Arun, 1987). In our system, the software automatically ran the SVD solver from the PCL library to calculate the calibration matrix after the sampling. According to our pilot test, an experienced user can complete the entire calibration process in 2 min. Once calibration is complete, there is no need to recalibrate unless the LMCs are moved.

Multi-LMC Data Fusion

The main task of our fusion algorithm is to find the most reliable data set, named candidate group g_c, from the raw skeleton data provided by the LMC SDK. Then, the algorithm combines the data based on the data confidence μ. A vector $\mu_c = \{\mu_c^l, \mu_c^r\}^T$ is introduced to ensure the chirality correctness (handedness) of the fused hands. The values of μ_c^l and μ_c^r indicate the confidence of the left or right chirality expressed by the group of data.

Algorithm Overview

An overview of the steps of our algorithm is shown below.

- STEP 1: *Data Clustering*. This step firstly clusters the data from the raw data set, which contains all the detected hand data from multiple LMCs, into a group set $G = \{g_1, \cdots g_j\}$ according to the palm center position. The group $g_j = \{h_1, \ldots, h_k\}$ is a collection of tracking data of hand h from k LMCs. The Euclidean distance between the palm center of h_k in the group g_j is within a threshold $\in$. The calibration error determines the value of $\in$.
- STEP 2: *Initialization*. The validity of the predicted data h_p made in the last frame is checked. If h_p is valid, the algorithm will go to STEP 4. Otherwise, the algorithm will go to STEP 3.
- STEP 3: *Position-Based Evaluation*. The data confidence μ of all detected hand data is calculated based on the palm center position. Then, the μ_c of each group in G is calculated using the evaluation result. After that, the candidate groups are selected out by comparing μ_c among all groups and sent to STEP 5 for data fusion.
- STEP 4: *Prediction-Based Evaluation*. First, the groups closest to h_p are chosen as the candidate groups. Then, the μ of hands in the candidate group is calculated based on the skeleton data difference between the tracking data and the predicted data. Finally, the evaluation results are verified using a chirality verification method. If the result is valid, the algorithm will go to STEP 5. Otherwise, the algorithm will go to STEP 3.
- STEP 5: *Data Fusion*. The fused results are obtained by fusing the hand data in the candidate groups according to the confidence μ_c. The chirality of the fused results is decided according to the u_c. The data of the candidate group will be fused with h_p using a Kalman filter if h_p is valid.
- STEP 6: *Prediction*. If the last frame data is valid, the hand motion of the next frame will be predicted based on kinematic theory (described in detail below). Then, the fusion data of the current frame is stored for the prediction process of the next frame.

More details of our algorithm are given in the following parts.

Position-Based Evaluation

The theory of the position-based method is based on the inconsistency of LMC tracking quality (Guna et al., 2014), which considers that the hand-tracking quality will be good if the hand is close to the center of its observing LMC's tracking range. The function of the confidence calculation is given in

$$\mu^{position} = \frac{\epsilon_a}{\epsilon_a + d_c}, \quad (4)$$

in which d_c (mm) is the distance between the palm center of the detected hand and the y-axis of the observing LMC's coordinate system and ϵ_a is an empirical parameter which represents the range of good tracking quality. In this system, we set ϵ_a to 250 mm according to Joze's work (Guna et al., 2014). The μ_c of each group is calculated using

$$\mu_c = \sum_{i=1}^{k} \mu^{(i)} \mu_c^{(i)}, \quad (5)$$

where $\mu_c^{(i)}$ is the chirality confidence of each hand in the group, acquired from the estimation result of the LMC SDK. $\mu_c^{(i)}$ equals $(1, 0)^T$ if the hand is estimated as a left hand or equals $(0, 1)^T$ if the hand is estimated as a right hand. The group with the highest

value of μ_c^l is selected as the candidate group for left-hand fusion. For μ_c^r, the rule is the same for the right-hand fusion.

It should be noted that the position-based evaluation method is a rule-of-thumb method. The result of the confidence evaluation will be unreliable sometimes. However, this method does not require the data of previous frames. Thus, it is used to calculate the initial value for the prediction-based method and as a supplementary method when the prediction-based method generates invalid results.

Prediction-Based Evaluation

The theory of the prediction-based method is based on the spatiotemporal continuity of hand motion (Anjum and Cavallaro, 2009), which considers the data difference between the current frame and the prediction from the last frame to be smaller for the correct tracking data compared to poor tracking data. Because the four metacarpal bones of the palm can be regarded as a rigid body, it is reasonable to predict the motion of these bones using the palm center velocity. Thus, we choose the position of the Prev-Joint and Next-Joint of the four palm metacarpal bones to calculate the data difference. The function of the prediction-based evaluation is given as

$$\mu^{prediction} = \frac{\epsilon_p^3}{\epsilon_p^3 + d_m^3}, \tag{6}$$

in which ϵ_p is the indicator of 50% confidence and d_m is the sum of the distance of the metacarpal joint between the tracking data and the predicted data, which is calculated as

$$d_m = \sum_{i=1}^{8} \sqrt{\left(p_t^{(i)} - p_p^{(i)}\right)^2}, \tag{7}$$

in which $p_t^{(i)}$ and $p_p^{(i)}$ are the position vector of the metacarpal joints of the tracking data and the predicted data, respectively.

The data confidence indicator ϵ_p is related to the distribution of d_m under the normal and poor tracking conditions. To ensure a safe classification, we choose the mid-value between the upper-bound and the lower-bound of the 99.7% confidence interval of the normal and poor tracking distribution respectively as ϵ_p.

We use a verification process to ensure the correctness of the evaluation result because the prediction results are not reliable when the hand moves quickly. In the process, the algorithm compares the μ_c^l with μ_c^r for each candidate group and chooses the larger one as the chirality according to the evaluation result. If the chirality of the evaluation result is coincident with the prediction, we consider the evaluation of this group as reliable. Otherwise, we discard the result and use the position-based method to evaluate the current frame.

Data Fusion

After acquiring the data confidence of each hand from the above evaluation method, we use a weighted-sum method to obtain the fused result of the candidate group g_c. The function is presented as

$$h_f = \sum_{i=1}^{k} \omega_i h_i, \tag{8}$$

$$\omega_i = \frac{\mu_i}{\sum_{j=1}^{k} \mu_j}, \tag{9}$$

where h_f and h_i represent the skeleton joint pose of the fused hand and original hand, respectively, and ω_i is the weighting value of h_i. Because the difference of the rotation data between the hand in g_c is small after calibration, we use a linear method to calculate quaternion interpolation approximately.

We use a Kalman filter to improve the fusion quality if the prediction data is valid. Assuming that the tracking error of all hand joints follows the same distribution, the update function of the Kalman filter (Bishop et al., 2001) can be given as

$$h_f' = h_p + K\left(h_f - h_p\right), \tag{10}$$

$$P' = (1 - K)P, \tag{11}$$

$$K = \frac{P}{P + R} \tag{12}$$

In the above equations, h_f' is the final fusion result of the current frame and P' and p are the variance of the final fusion results and the prediction results, respectively. k represents the Kalman gain. R represents the variance of the fused tracking data of the current frame. Because the calibration accuracy affects the absolute position measurement of each LMC, R can be represented by

$$R = \sum_{i=1}^{n} \mu_i^2 R_i, \tag{13}$$

in which R_i is the calibration error of each LMC.

Prediction

The prediction of the hand motion in the next frame is based on kinematic rules. We first calculate the velocity of the hand v_t using

$$v_t = \frac{p_t - p_{t-1}}{\Delta_{t-1}}, \tag{14}$$

in which p_t and p_{t-1} are the palm center position of the current and previous frames, respectively, and Δ_{t-1} represents the time interval between the current and previous frames. Then, the predicted result of next frame is obtained by

$$h_{p,t} = h_{f,t-1}' + v_t \Delta_t, \tag{15}$$

in which, Δ_t is the time interval between the current frame and the next frame.

3.1.3 Low Expressive Avatar Control System

Participants who used this avatar control system needed to wear additional tracking sensors to control their virtual avatar. In addition to tracking the HMD for the head, two tracked controllers for the hands and two extra HTC Vive trackers for the feet were used.

Body Tracking
The Final IK (RootMotion, 2019) Unity plugin was used to calculate and estimate the positions and rotations of other joints of the body, excluding the head, hands, and feet.

Hand Tracking
Virtual hand position and rotation data were mapped based on the data from the two controllers. We customized and mapped specific hand gestures to button presses on the controllers to allow some interaction with hand gestures. For example, squeezing the trigger button made a pointing gesture, pressing the controller grip buttons made a "V"-sign, pushing on the touchpad made a fist, and doing nothing was an open hand gesture.

Eye and Mouth Movement
Eye and mouth movements were similar to Wu et al. (2019a), where the eye-gaze direction followed the HMD facing direction, along with random eye blinking. We approximated mouth movements using small, medium, and large mouth openings, triggered by the loudness captured by the microphone using the Salsa LipSync v1 (Crazy Minnow Studio, 2014) Unity plugin.

In summary, the HEA control system uses contactless camera-based tracking to track a large range of behaviors used in social interaction, especially nonverbal behavior. It does not require the user to wear any additional trackers other than the ones on with the HMD. The LEA control system, in contrast, can be built with off-the-shelf hardware only. It is relatively easy to set up but has limited capability to track user behavior and requires additional body-worn sensors. In this article, we compare the two systems as a whole and do not consider the impact of a single factor such as hands or facial expressions on communication and collaboration. The aim is to find out if and to what extent it is beneficial to use a highly expressive avatar system for collaboration in a shared virtual environment.

3.2 System Overview

3.2.1 Hardware and Software

The HEA control system is a network solution, and the Kinect and Leap Motion Control (LMC) are working in a client-server mode. An Intel NUC (Intel Core i5-8259U at 2.3 GHz, 8GB RAM, and Iris Plus Graphics 655) is used for the client machines to drive the connected Kinect and LMC sensors. To avoid infrared interference between the Kinects and VR devices, we used the inside-out tracking of the Oculus Rift S (Facebook. 2019) as the HMD, which was driven by the server machine (Windows 10 desktop computer, Intel Core i7-7700K at 4.2 GHz, 32GB RAM, and NVIDIA GeForce 1080Ti). The front LMC sensor was connected directly to the server machine and was used as a primary reference for the calibration of the other four external LMC attached to the HMD (**Figure 1A**). All four client machines and the server machine were connected to a Gigabit Switch (NETGEAR GS110MX) through Ethernet cables for network data transmission (**Figure 3A**). The software on each NUC serialized body frame and hand frame data retrieved from Kinect and Leap Motion SDKs, wrapped them in Open Sound Control (OSC), and transmitted it to the network. We configured a standard VR setup for the LEA control system (HTC Vive Pro HMD (HTC, 2018) with two handheld controllers). A Windows 10 desktop computer (Intel Core i7-8700 at 3.2 GHz, 32GB RAM, and NVIDIA GeForce 2080) drove the Vive with two second-generation Lighthouse stations.

We developed the SVE using the Unity game engine version 2019.2.0f1 (Unity Technologies, 2019) with SteamVR for Unity (Valve Corporation, 2019), and Leap Motion plugin for Unity (Core 4.4.0) (Ultrahaptics Ltd., 2017). The generic virtual avatars were created through Makehuman (MakeHuman, 2018) with customized mouth shapes from Blender 2.79b (blender, 2019). The Point Cloud Library (PCL) (Radu and Cousins, 2011) version 1.6.0 was used to perform LSF calculations during the multiple-LMC calibration process. The Rug.OSc library (Rugland Development Group, 2010) was used for wrapping the transmitted frames as OSC messages and for handling them in Unity.

3.2.2 Networking and Latency

Several optimizations were implemented to realize low latency, stable connections, and accurate status synchronization for synchronous multiuser virtual experiences in our system. Instead of using a general server-client mode to synchronize participants' states, we set up a peer-to-peer (P2P) network mode to directly update avatars and events in the SVE on both sides. **Figure 3B** shows the working mechanism. We set up the same virtual scene on both peers with either HEA or LEA configurations, and either peer could be launched first as the virtual server machine waiting for the connection. The participants who played on the Client A side (HEA control system) used the body and hand data from the multiple Kinects and LMC sensors to update the local scene first. After data fusion, Avatar A was rendered and the system waited to transmit. Once the Unity program launched on the Client B side (LEA control system), the network connection was established. The data from the controllers and trackers drove Avatar B, which sent its data to Client A. The event listener was running on both sides and prepared for commands from the outside.

Figure 3A shows that the four Kinects and four of the LMCs were directly connected to the four NUCs, which continuously streamed serialized body-frame data to the server machine at a rate of 1.5 Mbps. The bandwidth requirements for each LMC was 2.4 Mb/s for single-hand data and 14.4 Mb/s at peak. All data transmitted within the Client A system or between Client A and Client B used the UDP protocol. The OSC message handling in the HEA control system was running in the background independently of the shared virtual environment. Through the optimization of the message processing pipeline compared to the method in Wu et al. (2019a), the latency of the local avatar rendering in the HEA control system was reduced to about 10 ms. The synchronization of the other avatar's status in the shared virtual environment added 30 ms and included the transforms of each bone and mesh of the head (eye and mouth movement). Therefore, this multiuser VR system's total latency accumulated for data transmission and rendering is about 40 ms. In other words, users see their movement on their avatar in about 10 ms, and the other person sees it in about 40 ms.

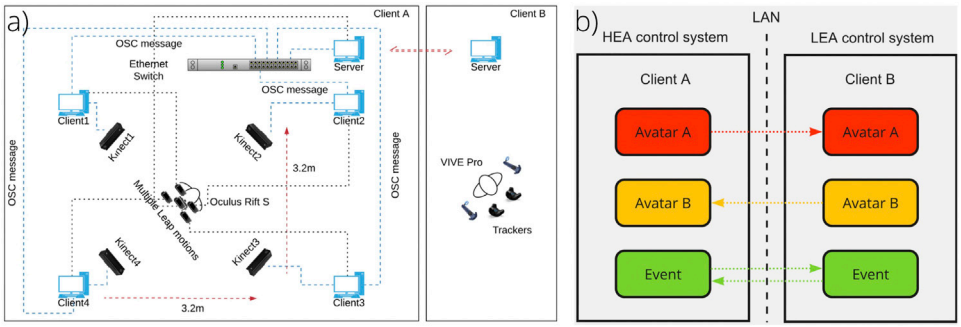

FIGURE 3 | Multiuser VR system. **(A)** System setup and **(B)** network.

The audio communication was set up using discord (Discord Inc., 2019). Participants wore similar sized over-ear gaming headphones in both systems, Logitech G433 headphones on the HEA side and Razer Nari Ultimate headphones on the LEA side.

4 METHODS

We conducted a controlled laboratory experiment to investigate the impact of avatar expressiveness on communication and collaboration. The experiment was approved by the Human Ethics Committee of our University.

4.1 Participants

We recruited 20 dyads, 40 participants (25 male, 15 female) from our university through advertisements posted on campus and on the university social media platforms. Participants were aged 18–46 (M = 29.3, *SD* = 6.7), and all were students or academic staff. We collected basic demographic data such as level of English (13 Native speakers and 27 fluent but nonnative English speakers) and dyad relationship (34 friends, six classmates or colleagues). Most participants (34/40) had previous VR experience with an average rating of (M = 2.4, *SD* = 0.93) on a 5-point Likert scale, from 1 (never), 3 (a few times a month), to 5 (daily use). The frequency of Social VR platform use was never (62.5%), a few times a year (32.5%), and a few times a month (5%). From the demographic information, most participants had VR experience, but only 37.5% of subjects had tried social VR applications before. Previous experience of participants with charades game was distributed never (37.5%), beginner (37.5%), intermediate (22.5%), and one person rating themselves as an expert (2.5%).

4.2 Study Design

The present study adopted a within-subjects design with one independent variable (expressiveness) with two levels: highly-expressive avatar (HEA) and low-expressive avatar (LEA) as described previously. To evaluate the user behavior and experience in different avatar control systems during mutual communication and collaboration, we set up a charades game playing scenario. The experiment had four game-play sessions per dyad. In each session, the dyad used both sides and embodied the relevant avatar, either the word performer or the guesser. The purpose was to make sure the dyad could try both avatar systems and take turns in the different roles. We decided to ask participants to rate their experience with the system after using one system in both roles (word performer, guesser). We randomized participants' orders using Research Randomizer (Geoffrey, 1997).

4.2.1 The Scene and Charades Game

The SVE was a virtual living room, with the two virtual avatars facing each other (see **Figure 4**). The distance between them was around 2 m. Virtual displays were placed on small tables in front of the avatars to show the words to mime and the number of words left. There was a countdown timer displayed on the wall once the game started. In the physical world, an experimenter sat at on the Client A side and used a keyboard to control the whole process. After the pair of participants put on the HMD and were ready for the study, the observer pressed a button to start the game, and the participants in the virtual world could see a text message about the game start from a first-person perspective. The experiment consisted of four sessions. For each session, a set of ten words was randomly selected from The Game Gal, 2020; HubPages Inc, 2020 with different difficulty (six easy words and four hard words). The sets were as follows:

- Set 1: pillow, tail, drum, mouth, finger, hungry, haircut, password, fast food, traffic jam.
- Set 2: swimming, love, hugs, itchy, grab, basketball, glue gun, sushi, cushion, police.
- Set 3: boxing, weightlifting, lobster, applaud, dancing, walking, lunch box, painting, elevator, earthquake.
- Set 4: scissor, crouching, hammer, piano, guitar, robot, thief, assemble, barber, pocket.

Once the game started, the participant at Client A saw a word shown on the virtual display, and he/she could only use nonverbal cues such as body posture and hand gestures to describe the word. The other participant could use verbal and nonverbal communication to guess or ask the performer for more. They

FIGURE 4 | The charade game scene.

FIGURE 5 | The experiment process. **(A)** Session 1, word performer using HEA control, **(B)** Session 2, word performer using LEA control, **(C)** Session 3, word performer using HEA control, and **(D)** Session 4, word performer using LEA control.

needed to collaborate to finish the ten words within 5 min. In the second session, participants stayed in their positions, but switched roles; the player at Client B mimed the next set of words for the player at Client A to guess. For the next two sessions, participants swapped avatar systems and repeated the process with different sets of words.

Figure 5 shows the four sessions, and the virtual view in each picture is from the partner. During the game, the researcher listened to the guesser and if he/she said the correct word, pressed the "Next" button, and, in the virtual world, the participants progressed to the next word. If a participant thought the current word was too hard to perform or his/her partner was taking too long to guess it, the guesser could ask the researcher to skip the word. The observer would then press the "Pass" button to skip the word, and the system would record which word was passed for later analysis.

4.2.2 Hypotheses

We expected that charades game performers using the expressive avatar control system perform better and make their counterparts

feel more socially connected. To test these expectations we formulated the following four hypotheses:

- H_1: Participants will feel greater copresence interacting with the highly expressive avatar.
- H_2: Participants will feel greater social presence interacting with the highly expressive avatar.
- H_3: Participants will feel greater attraction interacting with the highly expressive avatar.
- H_4: Participants using the highly expressive avatar control system will be able to explain more words successfully.

In addition, we hypothesized that participants prefer the HEA system over the LEA system and find it more helpful as a performer.

4.3 Measurements

Objective and subjective data were collected. Participants completed questionnaires after every two sessions, i.e., after they had used a system as performer and guesser, as well as at the end of the intervention. Additionally, the system automatically recorded completion time and the number of passed words.

4.3.1 Copresence

Copresence is the feeling that the user is with other entities (Schroeder, 2002). We measured copresence with two separate scales, their involvement in the interaction (self-reported copresence), and perception of their partner's involvement in the interaction (perceived other's copresence). The questionnaires for copresence were from Nowak et al. (Nowak and Biocca 2003), which were also used in the previous research from Roth et al. (Roth et al., 2018). The self-reported copresence scale included six items asking the participants to self-report their level of involvement in the interaction. The perceived other's presence scale included twelve indicators for intimacy, involvement, and immediacy. Participants rate their level of agreement with statements like, "I was interested in talking to my interaction partner" and "The interaction partner communicated coldness rather than warmth," on a 7-point Likert scale (1 = strongly agree, 7 = strongly disagree). We tested the reliability of the scales using the data collected in our experiment and found the copresence scales had good internal consistency: self-reported presence (Cronbach's α = 0.726), perceived other's presence (Cronbach's α = 0.810).

4.3.2 Social Presence

Social presence is the feeling of the user, which makes people feel connected with others through the telecommunication system, according to Rice (Rice, 1993), Short et al. (Short et al., 1976), and Walther (Walther, 1996). The questionnaire for social presence was from Nowak et al. (Nowak and Biocca 2003). The scale consisted of six items, and participants used a sliding scale (0–100) to answer questions like "To what extent did was this like you were in the same room with your partner?" The reliability of the scale was good (Cronbach's α = 0.768).

4.3.3 Interpersonal Attraction

The measure for liking and attraction was adapted from Oh et al. (Oh et al. 2016), which consisted of six items. Sample items include "I would enjoy a casual conversation with my partner" and "I would get along well with my partner." It was using a 7-point Likert scale (1 = strongly disagree, 7 = strongly agree). The reliability of the scale was good (Cronbach's α = 0.921).

Finally, we asked the participants to fill out the postquestionnaire about system preference and comments. Sample items include "Which VR system was most helpful when you were describing words to your partner?" and "Which VR system do you prefer?"

4.4 Procedure

The participants were asked to fill out the demographic survey and consent form at the beginning of the experiment. The experimenter introduced Charades game rules and the experiment process and explained how to use the devices involved in this user study. Charades is a communicative and collaborative game that requires players to use specific body postures or hand gestures. The rules for describing the words, and the level of expertise, can vary from person to person, so both participants were asked to familiarize themselves with the rules and discuss strategies in a face-to-face discussion before the VR game. This was to reduce the risk of a bad game experience or conflicts due to disagreements about the rules, etc.

After these preparation steps, both participants were guided to their respective avatar control systems. The experimenter helped them put on the HMD, gave them the relevant devices, and let them get familiarized with the system and interaction devices. Once the connection was established, the participants on both sides were asked to practice communication only using nonverbal behavior. Then the Discord program was launched for an audio communication test.

When they were ready, the experimenter started the game for the first two sessions. After that, participants were required to fill out the first set of questionnaires about their experience with the system. The experimenter then cleaned all of the devices and changed the configuration so that participants could swap avatar control systems for the remaining two sessions. Finally, participants were given one additional survey to gather information about their preference and ease of use of the avatar control schemes. The researcher then performed an experimental debrief with the participants, encouraged them to write comments about the two systems, discuss their survey answers, and talk about their general impressions of the two systems.

4.5 Statistical Analysis

For the analysis, we used the collected data sets of our 40 participants (20 dyads). A paired-samples t-test was used to compare participants' ratings of copresence, social presence, and interpersonal attraction for the two system. In **Table 2**, we present t-test values, means, and standard deviations for the questionnaires. We used α = 0.05 as level for statistical significance. We ran Shapiro-Wilk test before the t-test to check if our collected data are normally distributed and found that self-reported copresence [HEA (p = 0.203), LEA (p = 0.054)], perceived

TABLE 2 | Statistical results for copresence, social presence, and interpersonal attraction.

	Copresence		Social presence	Interpersonal attraction
	Self-reported	Perceived other's		
t-test	$p = 0.661$	$p = 0.819$	$p = 0.0008$	$p = 0.0007$
HEA M (*SD*)	4.2 (0.60)	4.0 (0.47)	63.0 (18.87)	5.4 (1.23)
LEA M (*SD*)	4.2 (0.59)	4.0 (0.46)	72.8 (7.99)	6.1 (0.46)

TABLE 3 | Summary of objective measurement results.

Group	Session	Completion time	Number of passed words	Session role	
				HEA	LEA
A	1 and 3 M (*SD*)	290.3 (8.5)	1.8 (0.8)	Performer	Guesser
B	2 and 4 M (*SD*)	291.5 (7.4)	3.1 (1.4)	Guesser	Performer

partner's copresence [HEA ($p = 0.875$), LEA ($p = 0.069$)], social presence [HEA ($p = 0.064$), LEA ($p = 0.395$)], and interpersonal attraction [HEA ($p = 0.432$), LEA ($p = 0.056$)] did not significantly deviate from it. The Shapiro-Wilk test for users performance data on completion time [Group A ($p = 0.916$), Group B ($p = 0.119$)] and Number of passed words [Group A ($p = 0.977$), Group B ($p = 0.817$)] was also not significant.

5 RESULTS

In this section, we show the summarized data and results of the statistical analyses. **Table 2** and **Table 3** as well as **Figure 6**, **7** provide overview of the collected data. The questionnaires to measure the social presence and interpersonal attraction we used from Nowak and Biocca 2003, and Oh et al. 2016, focus on the experience by reviewing a partner's performance. Hence the scores in the table are based on the system that their counterpart used.

5.1 Social Presence

There was a significant difference ($t(39) = 3.632, p < 0.001$) on how participants rated social presence for the two systems. Participants interacting with a HEA counterpart rated social presence significantly higher ($M = 72.8, SD = 7.99$) than when they interacted with a LEA counterpart ($M = 63.0, SD = 18.87$).

5.2 Interpersonal Attraction

Similarly, participants ratings showed that there was a significant difference for interpersonal attraction ($t(39) = 3.685, p < 0.001$) again showing higher results for participants interacting with a HEA counterpart ($M = 6.1, SD = 0.46$) compared to the LEA condition ($M = 5.4, SD = 1.23$).

5.3 Copresence

The collected data did not show any significant differences between the HEA and LEA systems for copresence, neither the subcomponent self-reported copresence ($t(39) = 0.442, p = 0.661$) nor the perceived partner's copresence ($t(39) = 0.231, p = 0.819$).

5.4 Performance

We recorded the completion time and the number of passed words. For each session, participants saw a timer of 5 min to finish display in the virtual world, but they were allowed to continue if they did not manage to go through all ten words within that time. We split the collected data into two groups. Group A for conditions in which participants were using HEA as the performer and partners using LEA as the word guesser and Group B in which participants were using LEA as the performer and partners using HEA as the word guesser. There was no significant difference between the amount of time participants took to finish each session when the performer used either HEA ($M = 290.3, SD = 8.5$) or LEA ($M = 291.5, SD = 7.4$) to describe the words ($t(39) = 0.698, p = 0.489$). The results, however, show that there is a significant difference ($t(39) = 5.551, p < 0.001$)

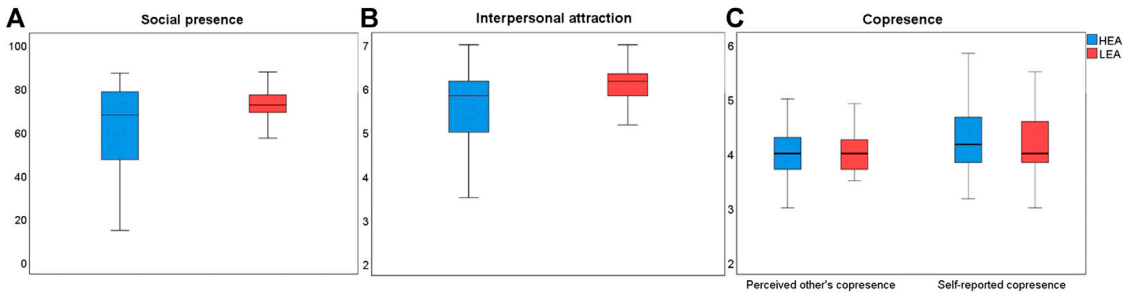

FIGURE 6 | Statistical results. **(A)** Social presence, **(B)** interpersonal attraction, and **(C)** copresence.

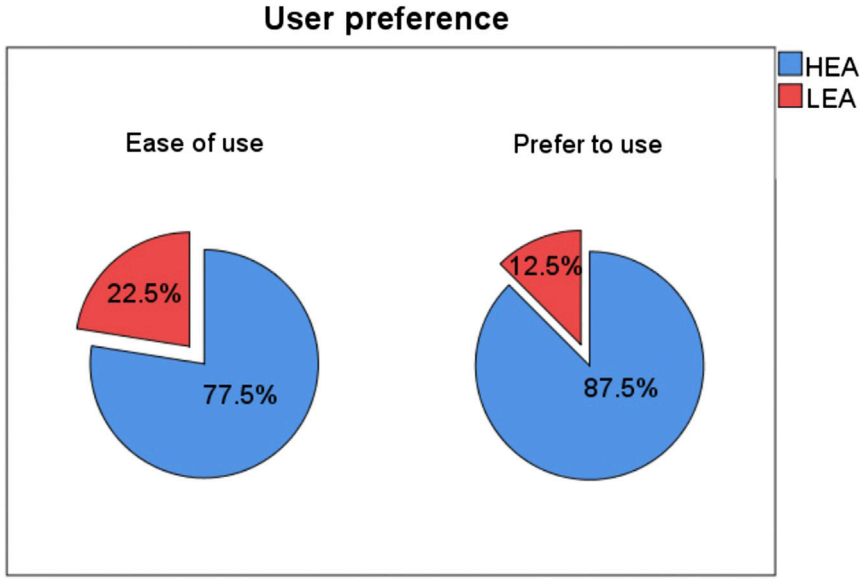

FIGURE 7 | Preference.

between the two groups for the number of passed words. When participants used the HEA control system to describe the words, they passed fewer words ($M = 1.8, SD = 0.8$) compared to performers using the LEA ($M = 3.1, SD = 1.4$).

5.5 Preference

The results show that 31 (77.5%) of participants thought it was easier to use HEA as a word performer to describe words to their partner. Furthermore, about 35 (87.5%) of participants preferred the HEA overall.

Participants also provided comments about the experiment and their embodied avatar control experience during mutual collaboration. Many comments reflect the importance of natural and accurate nonverbal behavior for high-quality communication, which can let user immerse themselves in the SVE and experience communication more like a face-to-face meeting.

- "High quality of experience about the person-to-person meeting, easy to understand what my partner wants to show/say. To compare, the HEA control system brings more real experience. It shows a clear movement of my partner's whole body."
- "In the LEA control system, I sometimes felt despair because I knew a simple gesture that would have explained the word immediately, but I could not do it and could not come up with something to replace or mimic it with the limited capabilities."
- "I tried both systems, I prefer the HEA rather than LEA. HEA control system is like the real world much more than the LEA control system, more activities, more details. It feels we have more communication between us. Besides, when I used the controller, I only can use my arms, legs and two fingers."
- "I like the HEA control system because it is more flexible. It was still quite different from real-life face-to-face experience, but it acts as a benefit to me like I don't feel shy to perform something that I might not perform in real life."

6 DISCUSSION

In this study, we found that with participants who interacted with people using avatars that had high expressiveness, nonverbal behavior felt greater social presence, which supports hypothesis H_2. Furthermore, the participants felt more attracted when they communicated and collaborated with the users who used the HEA control system, which supports hypothesis H_3. Another important aspect to note is that the majority of the participants preferred the HEA control system and felt it was easier to use. As for hypothesis H_4, the statistical results partly support it. The average number of successfully explained words for a user using HEA as a performer was 8.2 (82%), which is higher than the condition when participants used the LEA control system as a performer 6.9 (69%). However, there was no statistical difference between completion times. This could be partially because participants were presenting a timer of 5 min, which led to a ceiling effect that most participants took close to 5 min. As other factors, such as the amount of skipped word, can also impact the completion time for hypothesis H_4. Hence, the number of completed words can be seen as the only suitable measure of users' performance. We did not find evidence to support hypothesis H_1. The embodied experience can provide a

similar sense of presence when the participants use a simple avatar control system, as was found in (Wu et al., 2019a). Therefore, if the SVE system is stable with low network latency and the participants can both communicate with each other based on their real behavior, it is not hard to understand that there is no significant difference between the high and low expressive avatar control system in the either self-reported and perceived copresence. However, from the user comments and the pie chart in **Figure 7**, we can conclude that participants preferred using the HEA control system to communicate and collaborate in VR because it was flexible and more natural.

6.1 Implications

Our findings have practical implications for designers and developers of shared virtual environments. A highly expressive avatar control system that can support natural nonverbal behavior can lead to a more positive and realistic experience between players. It is intuitive and straightforward to express themselves with body posture or hand gestures when they communicate and collaborate in the SVE. The positive effects on social presence and interpersonal attraction from our highly expressive avatar control system can make virtual communication more like a face-to-face experience.

6.2 Limitations

Some limitations of the study need to be addressed. First, some participants reported that the HMD was a little bit heavy for the HEA control system due to the presence of five LMCs mounted on the HMD, along with the necessary extension cables. Although we tried to manage the cables by hanging them from the ceiling, they still may have bothered participants during gameplay. Second, some participant actions went beyond the hand tracking area, even though our system greatly enlarges the area compared to normal tracking. For example, sometimes they moved their hands over their heads. Also, participants sometimes touched the mounting frame of the multi-LMC system on the HMD, which in some cases resulted in the need to recalibrate the system to guarantee quality hand tracking. Therefore, we asked the participants to avoid touching the sensors on the HMD and reduce arm movement amplitude when they moved their hands over their heads. This could have affected the participant's perceptual and cognitive load. Third, in this study, we paired participants regardless of gender. The performance may be different when females collaborate with males compared to other gender combinations. We need to think about gender as a factor when we design collaborative studies.

7 CONCLUSIONS AND FUTURE WORK

We implemented a shared virtual environment using an asymmetric avatar control system and investigated the impact of different levels of nonverbal expressiveness on communication and collaboration behavior through a virtual charades game. We found a significantly higher social presence and interpersonal attraction when participants interacted with a user using the HEA control system. Participants prefer using the highly expressive avatar control system, which improves the task performance with a higher number of successful explained words.

In future work, we plan to improve the multi-LMC system by replacing the five extension cables with wireless transmitters and receivers. We also plan to refine the calibration algorithm for the multi-LMC system to a self-adaptive version, so the player does not need to recalibrate the system if the frame mount is moved. Furthermore, we consider to add tactile feedback into this multiuser VR system to explore the effect of haptic cues on communication and collaboration behavior.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Human Ethics Committee of the University of Canterbury. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

RL and SJ helped with the study's methodology and provided good feedback on the user study design. YW helped design and test part of the system. RL and SH helped review and edit the article and gave good feedback after the draft was finished.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frvir.2021.641296/full#supplementary-material.

REFERENCES

Anjum, N., and Cavallaro, A. (2009). "Trajectory association and fusion across partially overlapping cameras," in Sixth IEEE International conference on advanced video and signal based surveillance, AVSS 2009. Genova, Italy, September 2–4, 2009 (IEEE), 201–206. doi:10.1109/AVSS

Anthony, S., and Schroeder, R. (2015). Collaboration in immersive and non-immersive virtual environments, *Immersed in media*, 263–282. doi:10.1007/978-3-319-10190-3_11

Anthony, S., Slater, M., Sadagic, A., Bullock, A., and Tromp, J. (1999). "Leadership and collaboration in shared virtual environments," in Proceedings IEEE virtual reality (cat. No. 99CB36316), Houston,TX, March 13–17, 1999 (IEEE), 112–115.

Aristidou, A., Lasenby, J., Chrysanthou, Y., and Shamir, A. (2017). Inverse kinematics techniques in computer graphics: a survey. *Comput. Graphics Forum,* 37 (6), 35–58. doi:10.1111/cgf.13310

Arun, K. S., (1987). Least-squares fitting of two 3-D point sets. *IEEE Trans. Pattern. Anal. Mach. Intell.* 9, 698–700. doi:10.1109/tpami.1987.4767965

Bailenson, J. N., Beall, A. C., and Jim, B (2002). Gaze and task performance in shared virtual environments. *The J. Visual. Comput. Anima.* 13, 313–320. doi:10.1002/vis.297

Bailenson, J. N., Beall, A. C., Loomis, J., Jim, B, and Turk, M. (2004). Transformed social interaction: decoupling representation from behavior and form in collaborative virtual environments. *Presence: Teleoperators Virtual Environ.* 13, 428–441. doi:10.1162/1054746041944803

Bailenson, J. N., Yee, N., Merget, D., and Schroeder, R. (2006). 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. *Pres: Tele. Vir. Environ.* 15, 359–372. doi:10.1162/pres.15.4.359

Becker, B., and Mark, G. (2002). *Social Conventions in Computermediated communication: a Comparison of three online shared virtual environments*. London, United Kingdom: Springer London, 19–39. doi:10.1007/978-1-4471-0277-9_2

Bishop, G., and Welch, G. (2001). An introduction to the Kalman filter. Proceeding SIGGRAPH, Course 8. Chapel Hill, NC: University of North Carolina at Chapel Hill.

Blender (2019). blender.org - home of the blender project - free and open 3D creation software. Available at: https://www.blender.org/ (Accessed March 19, 2021).

Bombari, D., Schmid Mast, M., Canadas, E., and Bachmann, M. (2015). Studying social interactions through immersive virtual environment technology: virtues, pitfalls, and future challenges. *Front. Psychol.* 6, 869. doi:10.3389/fpsyg.2015.00869

Caserman, P., Garcia-Agundez, A., Konrad, R., Göbel, S., and Steinmetz, R. (2019). Real-time body tracking in virtual reality using a vive tracker. *Vir. Real.* 23, 155–168. doi:10.1007/s10055-018-0374-z

Crazy Minnow Studio (2014). SALSA LipSync V1 — animation — unity asset store. Avaliableb at: https://assetstore.unity.com/packages/tools/animation/salsa-lipsync-suite-148442 (Accessed March 19, 2021).

Crazy Minnow Studio (2019). SALSA LipSync V2 — animation — unity asset store. Available at: https://assetstore.unity.com/packages/tools/animation/salsa-lipsync-suite-148442.

Dana, S. (2000). Nau. 2000Rules for charades. Available at: https://www.cs.umd.edu/users/nau/misc/charades.html (Accessed March 19, 2021).

Discord Inc (2019). Discord – chat for communities and friends. Available at: https://discordapp.com/ (Accessed March 19, 2021).

Facebook (2019). Oculus rift. Available at: https://www.oculus.com/rift-s/ (Accessed March 19, 2021).

RootMotion. (2019). Final IK - RootMotion. Available at: http://www.root-motion.com/final-ik.html (Accessed March 19, 2021).

Friston, S., and Anthony, S. (2014). Measuring latency in virtual environments. *IEEE. Trans. Vis. Com. Graph.* 20, 616–625. doi:10.1109/tvcg.2014.30

Garau, M., Slater, M., Vinayagamoorthy, V., Brogni, A., Anthony, S., and Sasse, M. A. (2003). "The impact of avatar realism and eye gaze control on perceived quality of communication in a shared immersive virtual environment," in Proceedings of the SIGCHI conference on Human factors in computing systems, Florida, United States, April 5–10, 2003 ACM, 529–536.

Geoffrey, C. (1997). Urbaniak and scott. Plous Research Randomizer. Available at: https://www.randomizer.org/ (Accessed March 19, 2021).

Gergle, D., Kraut, R. E., and Fussell, S. R. (2013). Using visual information for grounding and awareness in collaborative tasks. *Human. Com. Inter.* 28, 1–39. doi:10.1080/07370024.2012.678246

Guna, J., Jakus, G., Pogačnik, M., Tomažič, S., and Jaka, S. (2014). An analysis of the precision and reliability of the leap motion sensor and its suitability for static and dynamic tracking. *Sensors* 14 (2), 3702–3720. doi:10.3390/s140203702

Harrison, J. S, and Michael, N. (2018). "Communication behavior in embodied virtual reality," in Proceedings of the 2018 CHI conference on human factors in computing systems, Montreal, Canada, April 21, 2018, 1–12. doi:10.1145/3173574.3173863

HTC (2018). VIVE pro— the professional-grade VR headset, Available at: https://www.vive.com/nz/product/vive-pro/.

HubPages Inc (2020). Charades: topic ideas, word lists, and how to play — HobbyLark. Available at: https://hobbylark.com/party-games/Charades-Ideas (Accessed March 19, 2021).

Jung, S., and Hughes, C. E. (2016). "The effects of indirect real body cues of irrelevant parts on virtual body ownership and presence," in ICAT-EGVE 2016 - international conference on artificial reality and telexistence and eurographics symposium on virtual environments, Goslar, Germany, October, 2016, (The Eurographics Association) 107–114.

Jung, S., Sandor, C., Pamela, J. W., and Charles, E. H. (2017). "RealME: the influence of body and hand representations on body ownership and presence," in Proceedings of the 5th symposium on spatial user interaction, Brighton, UK, October 16–17, 2017, (ACM), 3–11.

Jung, S., Pamela, J. W., and Charles, E. H. (2018). "In limbo: the effect of gradual visual transition between real and virtual on virtual body ownership illusion and presence," in Proceedings of the 25th IEEE conference on virtual reality and 3D user interfaces. Reutlingen, Germany, March 18–22, 2018, (IEEE).

Kilteni, K., Bergstrom, I., and Slater, M. (2013). Drumming in immersive virtual reality: the body shapes the way we play. *IEEE. Trans. Vis. Com. Graph.* 19, 4597–4605. doi:10.1109/tvcg.2013.29

Kwon, B., Kim, J., Lee, K., Yang, K., Park, S., and Lee, S. (2017). Implementation of a virtual training simulator based on 360° multi-view human action recognition. *IEEE. Access* 5, 12496–12511. doi:10.1109/access.2017.2723039

Latoschik, M. E., Lugrin, J. L., and Roth, D. (2016). "FakeMi: a fake mirror system for avatar embodiment studies," in Proceedings of the 22nd ACM conference on virtual reality software and technology, New York, United States, November 22, 2016, (ACM), 73–76.

Latoschik, M. E., Roth, D., Gall, D., Achenbach, J., Waltemate, T., and Botsch, M. (2017). "The effect of avatar realism in immersive social virtual realities," in Proceedings of the 23rd ACM symposium on virtual reality software and technology, New York, United States, November, 2017, (ACM), 1–10.

Lynn, M. A., and Corlett, E. N. (1993). RULA: a survey method for the investigation of work-related upper limb disorders. *Appl. Ergon.* 24 (2), 91–99. doi:10.1016/0003-6870(93)90080-s

MakeHuman. (2018). Makehuman community. Available at: http://www.makehumancommunity.org/ (Accessed March 19, 2021).

Matsumoto, D., Frank, M. G., and Hwang, H. S. (2012). *Nonverbal communication: science and applications*. San Francisco: Sage Publications.

McVeigh-Schultz, J., Kolesnichenko, A., and Isbister, K. (2019). "Shaping pro-social interaction in VR: an emerging design framework," in Proceedings of the 2019 CHI Conference on human factors in computing systems, Newyork, United States, May 8–13, 2021, 1–12.

Nilsson, A., Ann-Sofie, A., Heldal, I., and Schroeder, R. (2002). "The long-term uses of shared virtual environments: an exploratory study," in *The social life of avatars*, London, United Kingdom. Springer, 112–126.

Nowak, . L., and Frank, B. (2003). The effect of the agency and anthropomorphism of users' sense of telepresence, copresence, and social presence in virtual environments. *Presence: Teleopertors Virtual Environ.* 12, 5481–5494. doi:10.1162/105474603322761289

Oculus Lipsync. (2019). Viseme reference. Available at: https://developer.oculus.com/documentation/unity/audio-ovrlipsync-viseme-reference/.

Pan, Y., and Anthony, S. (2017). The impact of self-avatars on trust and collaboration in shared virtual environments. *PLoS One* doi:10.1371/journal.pone.0189078

Placidi, G., Cinque, L., Petracca, A., Polsinelli, M., and Spezialetti, M. (2017). "A virtual glove system for the hand rehabilitation based on two orthogonal LEAP motion controllers," in Proceedings of the 6th international conference on pattern recognition applications and methods, Porto, Portugal, February 24–26, (SciTePress)184–192.

Radu, B., and Cousins, S. (2011). "Rusu3D is here: point Cloud Library (PCL)," in IEEE international conference on robotics and automation (ICRA). Shanghai, China, May 9–13, 2011, (IEEE).

Rice, R. E. (1993). Media appropriateness: using social presence theory to compare traditional and new organizational media. *Human. Comm. Res.* 19 (4), 451–484. doi:10.1111/j.1468-2958.1993.tb00309.x

Roth, D., Lugrin, J., Julia, B, Gary, B., Fuhrmann, A., and Latoschik, M. E. (2016). "A simplified inverse kinematic approach for embodied VR applications. In

2016 IEEE Virtual Reality (VR)," in IEEE virtual reality (VR), Greenville, SC, July 7, 2016, (IEEE), 275–276.

Roth, D., Waldow, K., Latoschik, M. E., Fuhrmann, A., and Gary, B. (2017). "Socially immersive avatar-based communication2017IEEE virtual reality (VR)." in IEEE virtual reality (VR), Los Angeles, CA, March 18–22, 2012, IEEE, 259–260.

Roth, D., Mal, D., Purps, C. F., Kullmann, P., and Latoschik, M. E. (2018). "Injecting nonverbal mimicry with hybrid avatar-agent technologies: a Naïve approach," in Proceedings of the symposium on spatial user interaction, Berlin, Germany, 13-14 October 2018, 69–73. doi:10.1145/3267782.3267791

Roth, D., Gary, B., Kullmann, P., Mal, D., Purps, C. F., Vogeley, K., et al. (2019). "Technologies for social augmentations in user-embodied virtual reality," in 25th ACM symposium on virtual reality software and technology, (Parramatta, NSW:Social Virtual Reality), 1–12.

Rugland Development Group (2010). Rug OSC library, Available at: https://bitbucket.org/rugcode/rug.osc/src/master/.

Schroeder, R. (2002). "Social interaction in virtual environments: key issues, common themes, and a framework for research," in *The social life of avatars*. Springer, 1–18.

Schroeder, R. (2012). *The social life of avatars: presence and interaction in shared virtual environments*. London, United Kingdom, Springer Science & Business Media.

Short, J., Williams, E., and Christie, B. (1976). *The social psychology of telecommunications*. New York, NY: John Wiley & Sons.

Slater, M., and Anthony, S. (2002). *Meeting people virtually: experiments in shared virtual environments*, London: Springer London, 146–171. doi:10.1007/978-1-4471-0277-9_9

Slater, Mel., Spanlang, B., Sanchez-Vives, M. V., and Olaf Blanke (2010). First person experience of body transfer in virtual reality. *PLoS One* 5 (5). e10564. doi:10.1371/journal.pone.0010564

Soo Youn., O, Bailenson, J., Krämer, N., and Li, B. (2016). Let the avatar brighten your smile: effects of enhancing facial expressions in virtual environments. *PLoS One* 11, 91–18. doi:10.1371/journal.pone.0161794

Spanlang, B., Normand, J. M., Borland, D., Kilteni, K., Giannopoulos, E., Pomés, ., et al. (2014). How to build an embodiment lab: achieving body representation illusions in virtual reality. *Front. Robot. AI* 1, 9. doi:10.3389/frobt.2014.00009

The Game Gal (2020). Game word generator - the game gal. Available at: https://www.thegamegal.com/word-generator/.

Ultrahaptics Ltd (2017). Unity, leap motion developer. Available at: https://developer.leapmotion.com/unity.

Unity Technologies (2019). Unity real-time development platform — 3D. 2D VR & AR visualizations. Available at: https://unity.com/ (Accessed March 19, 2021).

Valve Corporation (2019). Steam VR, Available at: https://www.steamvr.com/en/ (Accessed March 19, 2021).

Walther, J. B. (1996). Computer-mediated communication: impersonal, interpersonal, and hyperpersonal interaction. *Commun. Res.* 23, 3–43. doi:10.1177/009365096023001001

Wu, Y., Wang, Y., Jung, S., Simon, H., and Lindeman, R. W. (2019a). "Exploring the use of a robust depth-sensor-based avatar control system and its effects on communication behaviors," in 25th ACM symposium on virtual reality software and technology (VRST' 19). (New York, NY: Association for Computing Machinery). doi:10.1145/3359996.3364267

Wu, Y., Wang, Y., Jung, S., Simon, H., and Lindeman, R. W. (2019b). Towards an articulated avatar in VR: improving body and hand tracking using only depth cameras. *Entert. Comput.* 31, 100303. doi:10.1016/j.entcom.2019.100303

Zhang, Z. (1994). Iterative point matching for registration of free-form curves and surfaces. *Int. J. Com. Vis.* 13, 119–152. doi:10.1007/bf01427149

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.