# 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

Jeremy N. Bailenson[1], Nick Yee[1], Dan Merget[2] and Ralph Schroeder[3]
[1] Department of Communication, Stanford University, USA
[2] Department of Computer Science, Stanford University, USA
[3] Oxford Internet Institute, University of Oxford, UK

## Abstract

*The realism of avatars in terms of behavioral and form is critical to the development of collaborative virtual environments. In the study we utilized state of the art, real-time face tracking technology to track and render facial expressions unobtrusively in a desktop CVE. Participants in dyads interacted with each other via either a videoconference (high behavioral realism and high form realism), voice only (low behavioral realism and low form realism), or an "emotibox" that rendered the dimensions of facial expressions abstractly in terms of color, shape, and orientation on a rectangular polygon (high behavior realism and low form realism). Verbal and non-verbal self-disclosure were lowest in the videoconference condition while self-reported copresence and success of transmission and identification of emotions were lowest in the emotibox condition. Previous work demonstrates that avatar realism increases copresence while decreasing self-disclosure. We discuss the possibility of a hybrid realism solution that maintains high copresence without lowering self-disclosure, and the benefits of such an avatar on applications such as distance learning and therapy.*

## 1. Avatars

### 1.1. What is an avatar?

The study of virtual humans—from conceptual, design, and empirical perspectives—has progressed greatly over the past fifteen years. Traditionally, the field of research has delineated between *embodied agents* which are digital models driven by computer algorithms and *avatars* which are digital models driven by real-time humans. In terms of empirical behavioral research examining how people interact with virtual humans in social interaction, a majority of this work has utilized embodied agents (as opposed to avatars—see Bailenson & Blascovich [3] for a discussion of this disparity). One reason for this bias is because it is only over the past few years that readily available commercial technology has actually allowed people to make avatars that can look like and behave - via real-time tracking - like the user. In other words, up until now, producing real-time avatars that captured the user's visual features and subtle movements has been quite difficult to accomplish in a social science laboratory. Consequently, understanding the implications of the visual and behavioral veridicality of an avatar on the quality of interaction and on copresence is an important question that has received very little empirical attention. Schroeder [22] provides a review of the existing empirical work on avatars.

Avatars can be defined as digital models of people that either look or behave like the users they represent. In traditional immersive virtual environments, an avatar is the model that is rendered on the fly to reflect the user's behavior. However, the definition of an avatar certainly has blurry boundaries. For example, the definition including "looking like a user" would allow for a digital photograph of a person stored on a hard drive to be considered an avatar. Some would object that this archived image is not an avatar since it has no potential for behavior or for social interaction. On the other hand, some would include the photograph in the definition, arguing that people utilize static (i.e., non-animated) avatars with internet chat and emails. While people discuss the concept of avatars quite often in the literature on virtual humans and virtual environments, a standard definition of avatars has not emerged readily. But since avatars are playing an increasingly central role in virtual environments and other electronic media, it is important to investigate the suitability of different types of avatars for representing the user.

Figure 1 provides a preliminary attempt to provide a framework for considering representations of humans that is not limited just to digital avatars. The Y-axis denotes behavioral similarity—how much the behaviors of the representation correspond to the behaviors of a given person. The X axis indicates form similarity, how much the representation statically resembles features of a given person. On the left side are representations that correspond to a given person's form or behavior in real-time. On the right are representations that correspond to a person's form or behavior asynchronously. For example, a puppet is a representation of a person that has high behavioral similarity (the movements of the puppet are very closely tied to the person controlling it) but low form similarity (the puppet need not look at all like the person controlling it). Furthermore, the puppet's behaviors are expressed in real-time. On the other hand, an impressionist (i.e., someone who can very closely reproduce or mimic the behaviors of a person who is not physically present) has high behavioral similarity and low form similarity in that the impressionist need not look like the person being mimicked. Unlike the puppet, however, the impressionist is a non-real-time representation—the person being mimicked need not be present, aware of the impressionist's existence, or even still alive for that matter.

As Figure 1 demonstrates, there are lots of different types of representations of people utilized today. The shaded oval denotes the space in which we typically discuss avatars—digital representations of humans that are utilized in immersive virtual environments. Blascovich and colleagues [7] provide a theoretical framework to determine the interplay of behavioral and form realism for the avatars which fall into this shaded region.

### 1.2. Avatars and Copresence

A key reason why avatar form and behavior are so important is that they elicit an experience of being with another person; or copresence (also referred to as social presence). There are many definitions of copresence in the literature. Heeter defined copresence as the extent to which other beings, both living and synthetic, exist in a virtual world and appear to react to human interactants [15] . Slater and colleagues, in contrast, define copresence as the sense of being and acting with others in a virtual place [24] . Lee defines copresence as experiencing artificial social actors (agents) via objects that manifest humanness or actual social actors (avatars) connected via technology [18] . Finally, Blascovich and his colleagues have defined copresence as the extent to which individuals treat embodied agents as if they were other real human beings [7] .

Biocca, Harms and Burgoon [8] review the various definitions and measures of copresence. They discuss different media, including those in which the 'other' with whom one experiences presence can be an agent or other media-generated human-like appearance, and they include a broad range of phenomena within copresence partly so that they can compare different media (for example, para-social interaction with a film character). They also review several measures that have been proposed for copresence, including self-report, behavioural and psycho-physiological measures, but point out that little consensus has been reached on this issue. Their proposal to specify an extensive set of criteria and scope conditions for copresence is quite broad, including items such as "read[ing] minds" in both people and things' ([8] : 474). However, they also describe copresence as a more tightly defined subset of a larger phenomenon whereby people need to have a sensory experience of sharing the same space with someone else. This limits copresence to face-to-face experiences or experiences in which two (human) users *both* share the space and the sensory experience of each other (this also corresponds to Schroeder's strict definition of copresence [23] ).

It is clear that different measures of copresence have drawbacks: self-report measures are subjective, but any objective (behavioural, cognitive, or psycho-physiological) measures will also be problematic since they will not directly reveal what people feel or how they interpret the

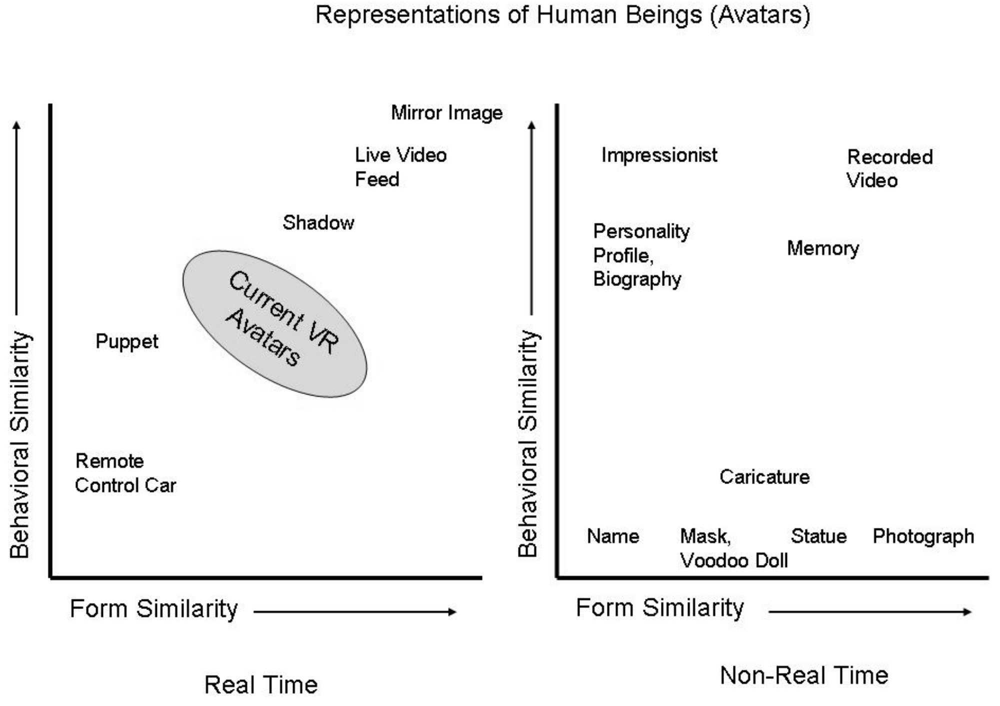

**Figure 1: A framework for classifying representations of humans in physical and digital space.**

presence of another. Obviously a combination of methods will provide the most well-rounded understanding and/or explanation of this phenomenon. Indeed, a recent empirical study by Bailenson, Swinth, Hoyt, Persky, Dimov, & Blascovich [4] directly compared the subjective, behavioural, and cognitive measures of copresence. Their results confirmed the hypotheses of Biocca et al.—by providing affective, behavioural and cognitive measures of copresence, they demonstrated that subjective reports alone were insufficient to highlight copresence differences between various types of agents. On the other hand linking the specific behavioural and cognitive results (which discovered differences not detected by self report measures) to the latent construct of copresence proved challenging.

A number of studies have explored avatar realism experimentally. Bailenson, Beall, & Blascovich [1] demonstrated that higher behavioural realism in terms of mutual gaze produced higher levels of copresence and produced changes in task performance. Garau [14] investigated photorealistic and behaviourally realistic avatars and showed that behavioural realism is more important than form realism in several different scenarios. Moreover, Bente [6] has shown that even avatars with very minimal levels of behavioural realism elicit responses from others.

There are also studies that have examined the interaction between avatars in 'naturalistic' settings. Becker and Mark [5] compared how social conventions are followed in three different online VEs: text-only, text-and-graphics, and voice plus talking head. They found, based on participant observation, that certain conventions from face-to-face interaction are followed in all three settings, but that certain of them are followed more in the more 'realistic' setting (i.e., interpersonal distance is kept more in the shared VE with voice). It has also been investigated what preferences people have for different avatar appearances. Cheng, Farnham and Stone [10] found that people in a text-and-graphics shared VE (V-chat) preferred representations of themselves that were neither too true-to-life to their own appearance nor too abstract. These studies demonstrate that people's habits and preferences will shape avatar appearance.

A related topic is the extent to which avatars are developed sufficiently enough to allow the transmission of 'social cues' of face-to-face communication, which includes all the information about one another (pitch of the voice, non-verbal gestures, etc.—see Whittaker [29] for a review). Walther [27] has argued against the widely held view that interaction with avatars lacks 'social richness' or 'media richness'. He has shown that it is not necessarily the case that less rich media prevent people from getting to know each other; it may just take more time. In fact, he argues that they may get to know each other better in certain respects in less rich media; he calls these 'hyperpersonal' relationships that are created among avatars and other representations in computer mediated communication in which people form extremely deep social ties with each other.

The literature on self-disclosure suggests that copresence mediates the effect of visual and behavioural realism on self-disclosure. For example, a meta-analysis of studies on self-disclosure in face-to-face interviews as compared with computer-administered interviews found that self-disclosure was higher in computer-administered interviews than in face-to-face interactions [28] . This suggests that less realistic avatars would elicit more self-disclosure from users. In a study where participants interacted with either a text-based or face-based agent, it was found that participants revealed more information about themselves when interacting with the text-based agent [25] . Previous researchers have also implemented and discussed self disclosure as a proxy for measures of copresence [20] .

### 1.3. Facial Expressions and Facial Tracking of Avatars

Research on transmitting as well as receiving facial expressions has received much attention from social scientists for the past fifty years. Some researchers argue that the face is a portal to the one's internal mental state (Ekman & Friesen [12] , Izard [16] ). These scholars argue that when an emotion occurs, a series of biological events follow that produce changes in a person—one of those manifestations is movement in facial muscles. Moreover, these changes in facial expressions are also correlated with other physiological changes such as heart rate changes or heightened blood pressure [11] .

The use of facial expressions to form attributions concerning others certainly changes during mediated communication. Telephone conversations clearly function quite well without any visual cues about another's face. As Whittaker [29] points out in a review of the literature examining visual cues in mediated communication, adding visual features is not always beneficial, and can sometimes be counterproductive. Specifically, Whittaker's survey of findings demonstrates that showing another person's face during interaction tends to be more effective when the goal of the interaction is social than when it is purely task oriented. However, a large part of the problems with previously studied visual mediated communication systems have been due to bandwidth delay in videoconferences or from the stark conditions offered by other virtual solutions [17] . However, as virtual reality systems and other visually mediated communications systems improve the accuracy of visual representations will become closer to that seen in face-to-face interaction. Consequently, facial expressions seen during human-computer interaction will be more diagnostic of actual facial movements.

There has recently been a great surge of work to develop automatic algorithms to identify emotional states from a video image of facial movements. Early work developed a system of facial action coding system in which coders manually identified anchor points on the face in static images [12] . Similarly, computer scientists have developed vision algorithms that automatically find similar anchor points with varying amounts of success (see Essa & Pentland [13] for an early example). As computer vision algorithms and perceptual interfaces become more elegant (see Turk & Kölsch [26] for a review), it is becoming possible to measure the emotional state of people in real-time, based on algorithms that automatically detect facial anchor points without using markers on the face and then and categorize those points into emotions that have been previously identified using some type of learning algorithm. These systems sometimes attempt to recognize specific emotions [19] or alternatively attempt to gauge binary states such as general affect [21] .

## 2. Study Overview

In the current study we empirically test two of the dimensions of avatars depicted in Figure 1—behavioural and form realism. We varied the extent to which an avatar's face resembled and gestured similarly to the users' faces. Dyads interacted via a desktop virtual display, and we tracked in real-time 22 anchor points on their faces as well as position of the faces and orientation of the faces. We are interested in how people behaved towards one another's avatars and whether or not they revealed more about themselves (in terms of how much information they revealed verbally as well as how much information they revealed through facial gestures) when they encountered avatars that were less realistic in form and behaviour. Furthermore, we measured the ability of subjects to transmit and receive specific emotional expressions at various levels of behavioural and form realism as both a cognitive measure of copresence as well as a test of our face-tracking system's effectiveness.

We had three conditions: 1) voice only, 2) videoconference, and 3) the emotibox—a polygon that changed shape, colour and orientation in response to the user's head position and facial expressions. Figure 2 shows screenshots of these three conditions.

The emotibox is reminiscent of the 'blockie' avatars of the avatars that were used in some of the earliest research on CVEs [9] . Here, we implement this type of avatar because it is a manner to represent high behavioural realism (via facial emotion) with low form realism. By high behavioural realism, we simply mean that the avatar behaves in ways that are contingent upon the behaviours of a human. In other words, our definition of behavioural realism in the current study requires a) a high number of behaviours to be tracked, and b) a high number of behaviours rendered on the avatar that are contingent upon those tracked behaviours. In some ways, this definition is counterintuitive, because the behaviours do not look like the actual behaviours of the user since they are abstracted. The hypothesis in the current study was that demonstrating behavioural contingency (though not behavioural similarity) was the best compromise between high behavioural realism and low form realism. Because it is not possible to have facial movements reflected realistically on an avatar without facial features, the emotibox maintained the best balance between high behavioural realism and low form realism.

If one of the main difficulties of shared VEs and other computer-mediated communication is going to be the live capture of people's facial appearance and expressions, then the amount of realism required for non-verbal facial communication becomes an important question. To our knowledge this is one of the first empirical studies of copresence that utilizes avatars capable of rendering real-time emotional expressions via face-tracking. By examining the unique contribution of facial expressions as an independent variable, as well as using the amount of emotions conveyed as a dependent variable, we can potentially examine a unique level of avatar realism.

## 3. Method

### 3.1. Participants

Thirty undergraduate students (12 men and 18 women) were paid ten dollars each for their participation in the study. The gender makeup of dyads was 3 male-male pairs, 6 mixed pairs, and 6 female pairs.

### 3.2. Design

There were three conditions in the study: 1) voice only, 2) videoconference, and 3) emotibox. In all 3 conditions, participants were seated in front of a computer terminal equipped with a Logitech QuickCam Messenger digital camera mounted on top of the monitor. A conferencing application (Microsoft Netmeeting) was used in all three

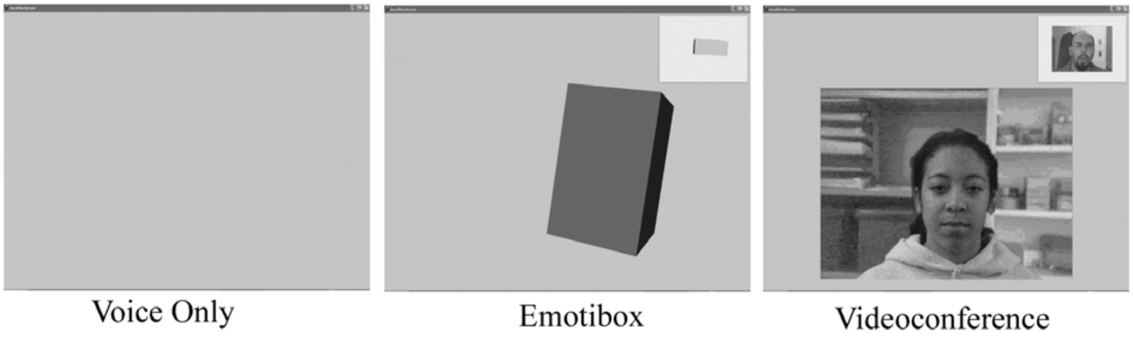

**Figure 2: A subject's eye-view of the three conditions. In the right two panels, the center of the screen shows the avatar of the other interactant while the top right corner shows the subject's own avatar**

conditions for voice.

In the *videoconference* condition, the conferencing application allowed participants to see each other via the digital cameras. The video feed was a gray-scale image with 256 levels, updated at 20 frames per second. The video was acquired at a resolution of 320x240, and then magnified to 760x570 so that it would fill most of a 1024x768 screen. While a videoconference image may not be traditionally categorized as an avatar, given that we were using digital video it does fit the definition discussed earlier on in this work. More importantly, for our purposes in this experiment, a videoconference worked most effectively as a high realism control condition.

In the *emotibox condition*, the Nevenvision Facial Feature Tracker, a real-time face-tracking solution, was integrated into Vizard 2.5, a platform for developing virtual environments, to capture key locations of the face. These anchor points, depicted in Figure 3, included 8 points around the contour of the mouth (three on each lip, and one at each corner), three points on each eye (including the pupil), two points on each eyebrow, and four points around the nose. The points were measured in a two-dimensional head-centred coordinate system normalized to the apparent size of the head on the screen; the coordinates were not affected by rigid head movements, and scaled well to different heads. The face-tracking software also tracked the pitch, yaw and roll of the face, the aspect ratio of the mouth and each eye, the coordinates of the face in the webcam image, and the scale of the face (which is inversely proportional to the distance from the face to the webcam). Our real-time face-tracking solution required no training, face-markers, or calibration for individual faces.

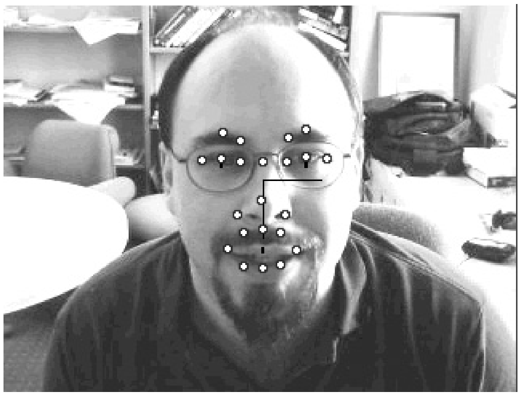

**Figure 3: The 22 anchor points automatically tracked without using facial markers by the Nevenvision Facial Feature Tracker at 30 Hz.**

The emotibox was based on the YUV colour scheme and had 11 degrees of freedom: 1) the eye aspect ratio controlled the Y-value (i.e., black-white spectrum) of the cube. In laboratory pilot studies, the aspect ratio of one eye was found to vary roughly between 0.10-0.35, so the aspect ratio of each eye was added together, truncated to the range 0.20-0.70, and linearly transformed to a Y-value from 0.5-1.0. Thus, the wider the person's eyes, the brighter the cube. A minimum Y of 0.5 kept the cube bright enough so the U and V could be seen. 2) The distance between the corners of the mouth and the eyes controlled the U-value (i.e., the blue-yellow spectrum) of the cube. The total distance was truncated to the range 85-100, and linearly transformed to a U of -0.4 to +0.4. Thus, the more a person smiled, the more yellow the cube became. And the more a person frowned, the more blue the cube became. These colours were chosen after extensive pre-testing indicated the most separability in terms of mapping discrete mental states. 3) The distance of the eyebrows from the pupils controlled the V-value (i.e., red-cyan spectrum) of the cube. Two different scales were used, since we found that relaxing the eyebrows brought them very close to their lowest extreme (at least according to our tracking software). Distances from 27-35 mapped to a V-value of 0.0 to +0.6, but distances from only 27 to 25 mapped to v-values of 0 to -0.6. The more you raised your eyebrows, the more cyan the cube would become. The 4) width and 5) height of the emotibox followed the width and height of the mouth: each dimension varied from 50% to 150% of the basic cube as the mouth width and height varied from 15-35 and 28-42, respectively. Finally, the emotibox followed the 6) pitch, 7) yaw, 8) roll, 9) x-coordinate, 10) y-coordinate, and 11) z-coordinate of the head.

The emotibox was also updated 20 times per second, even though the face-tracking software acquired images at 30 Hz. When the confidence of the face-tracking software fell below 40%, the data was discarded and the software was told to re-acquire the face from scratch. The other subject saw a frozen emotibox until this process was done. In the voice only condition, the sound system allowed participants to hear each other's voice.

In the *voice only* condition, subjects saw a blank screen and communicated through the audio software.

### 3.3. Materials

To generate two sets of questions (one for each interactant in the dyad) of a comparable degree of intimacy for the verbal self-disclosure task, 30 questions were pretested for their degree of intimacy. To pretest the materials, 15 undergraduates from a separate population from the experimental pool rated each of the questions on a 5-point, fully-labeled, construct-specific scale, ranging from "Not Personal At All" to "Extremely Personal". Six pairs of questions were chosen such that the questions in each pair did not differ significantly from each other in a t-test. In addition, we added a general self-disclosure question at the end of both sets - "Tell me a little more about yourself". These two sets of questions used in the main experiment are listed in the Appendix.

### 3.4. Procedure

Pairs of participants arrived at the laboratory for each session. Most participants did see one another in vivo before the experiment began. After signing informed consent, they were seated in front of the computer terminals in different rooms. Each pair of participants was assigned to the same condition using a predetermined randomization scheme. The study began with the verbal self-disclosure task. For all three conditions, the question sets were displayed textually on the monitor one at a time and alternated between the two participants. Participants were instructed to ask the other participant the question that was displayed (via text on the monitor) by speaking into a headset microphone. The participant that answered the question advanced to the next question by pressing the space bar when he or she was finished speaking. We randomized which participant would ask the first question. The audio from all interactions was recorded.

The second task was an emoting task. Participants were given a list of seven emotions, one at a time in random order - disgusted, angry, sad, joyful, afraid, interested, and surprised. For each emotion, participants were asked to convey that emotion to the other participant for 10 seconds. The video-feed and emotibox subjects conveyed the emotion via facial expression, while the voice-only subjects used nonverbal sounds (i.e., no words allowed) to express themselves. While this condition is somewhat unnatural, this was the best way for us to not allow for the use of language or grammar to clue the specific emotion. After each emotion, the other participant would be asked which emotion was conveyed, and how sure they were of their answer. One participant would be instructed to emote through all seven emotions and then the other participant would be instructed to do the same. The last task was filling out the copresence questionnaire. Participants saw one question on the screen at a time in a random order and responded using the keyboard.

## 4. Measures and Hypotheses

### 4.1. Verbal Self-Disclosure

Two coders blind to experimental condition listened to the audio recordings of all interactants and rated each one's friendliness, honesty and how revealing their responses were on 5-point, fully-labeled, construct-specific scales. Thus, each participant had six ratings, three from each coder. The composite scale composed of these six items had a reliability of .85. We hypothesized that self-disclosure would be lowest in the videoconference condition and highest in the voice only condition, and that there would be more disclosure in front of the emotibox than the videoconference.

### 4.2. Non-Verbal Self-Disclosure

Previous research discussed above has indicated that people disclose more verbal information in a text interface than in an avatar-based interface. We were interested in testing for this effect in terms of non-verbal behaviors. We therefore predicted that participants in the voice only condition would disclose more non-verbal information than in the videoconference and emotibox conditions. The face tracking software was used to find 22 points on the face that varied with expression (see Figure 2), but were not affected by the position and/or orientation of the head as a whole. The standard deviation of each point (both x and y coordinates) measured how much activity occurred at that point, and the average of all 44 standard deviations served as a measure of how expressive the face was during the experiment. This metric is deliberately naïve, and some points, such as the corners of the mouth, were up to 6 times as mobile as others, and thus contributed more heavily to the face movement metric. Nonetheless, we used the simplest, least biased way of combining the measurements into a single score.[1] In future work, we plan on developing more elegant combinations of the facial feature points.

### 4.3. Copresence Ratings

Participants completed a 4-item copresence scale depicted in the Appendix, which was modeled after the scale developed by Biocca, Harms, & Burgoon [8]. The reliability of the composite scale was .62. We hypothesized that copresence would be highest in the videoconference condition and lowest in the voice only condition.

### 4.4. Emotion Detection

Participants were scored a 1 if they guessed the emotion correctly, a 0 if they were incorrect. The composite scale composed of the mean of the seven detection scores had a reliability of .62.

## 5. Results

### 5.1. Verbal Self Disclosure

We ran a between-subjects ANOVA with condition (voice only, emotibox, and videoconference) and subject gender as independent factors and self disclosure score as a dependent variable. There was a significant effect of condition, $F(2,24) = 5.80$, $p<.001$, partial Eta Squared = .33. As Figure 4 demonstrates, there was more disclosure in the voice only and the emotibox conditions than the videoconference conditions. The effect of participant gender was not significant, $F(1,24) = .02$, $p<.90$, partial Eta Squared = .00, and the interaction was not significant, $F(2,24) = 1.29$, $p<.29$, partial Eta Squared = .10.

[1] Participants were encouraged to always keep their heads in front of the camera, but we did not want to force artificial constraints into the interaction such as a chin-rest. Consequently, in the voice-only condition (in which subjects had no visual cue indicating their face was out of the camera tracking range), some participants kept their face out of the range of the tracking space for more than fifty percent of the time. When eliminating these subjects from the sample, the statistical significance of the results did not change at all. Consequently we leave all subjects in the analyses for the sake of simplicity.

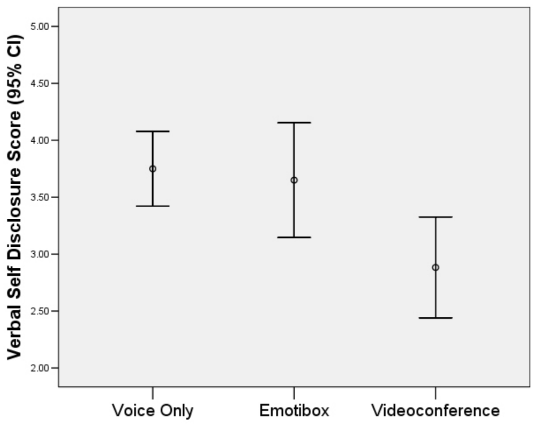

**Figure 4: Verbal self disclosure scores by condition.**

### 5.2. Nonverbal Disclosure

We ran a between-subjects ANOVA with condition (voice only, emotibox, and videoconference) and subject gender as independent factors and nonverbal disclosure score as a dependent variable. There was a significant effect of condition, *F*(2,24) = 6.45, *p*<.01, partial Eta Squared = .35. As Figure 5 demonstrates, there was more disclosure in the voice only condition than the emotibox or the videoconference conditions. The effect of gender was not significant, *F*(1,24) = .19, *p*<.67, partial Eta Squared = .01, and the interaction was not significant, *F*(2,24) = .65, *p*<.53, partial Eta Squared = .05.

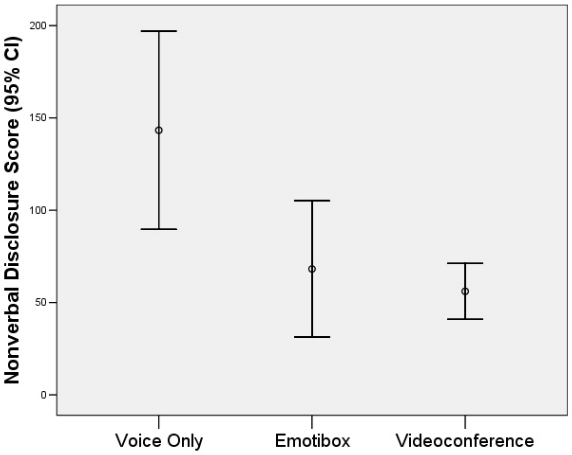

**Figure 5: Average nonverbal disclosure score by condition. The scale of the Y-axis is normalized to the size of the head within the screen image and does not map onto a standard metric such as centimeters.**

### 5.3. Copresence Ratings

We ran a between-subjects ANOVA with condition (voice only, emotibox, and videoconference) and subject gender as independent factors and self-report copresence score as a dependent variable. There was a significant effect of condition, *F*(2,24) = 3.55, *p*<.05, partial Eta Squared = .23. As Figure 6 demonstrates, there was less copresence in the emotibox condition than the voice only condition. The effect of gender was marginally significant, *F*(1,24) = 3.24, *p*<.08, partial Eta Squared = .12, and the interaction was not significant, *F*(2,24) = 1.36, *p*<.28, partial Eta Squared = .10.

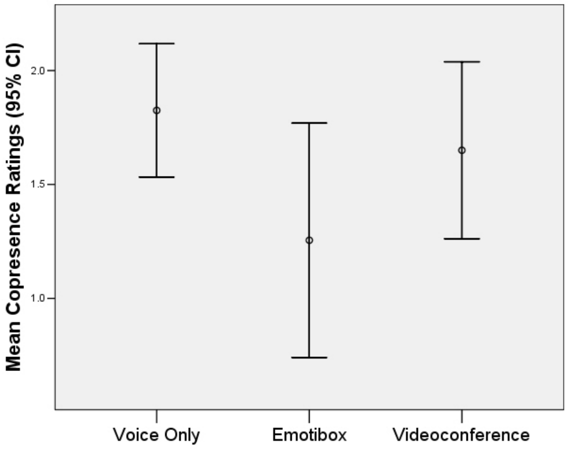

**Figure 6: Mean copresence ratings by condition.**

### 5.4. Emotion Detection

We ran a between-subjects ANOVA with condition (voice only, emotibox, and videoconference) and subject gender as independent factors and emotion detection score as a dependent variable. There was a significant effect of condition, *F*(2,24) = 18.05, *p*<.001, partial Eta Squared = .60. As Figure 7 demonstrates, there was worse performance in the emotibox condition than the voice only or the videoconference conditions. The effect of gender was not significant, *F*(1,24) = .12, *p*<.73, partial Eta Squared = .01, and the interaction was not significant, *F*(2,24) = .18, *p*<.83, partial Eta Squared = .02. In all three conditions, subjects were significantly above chance (depicted by the dotted line in Figure 7) at emotion detection.

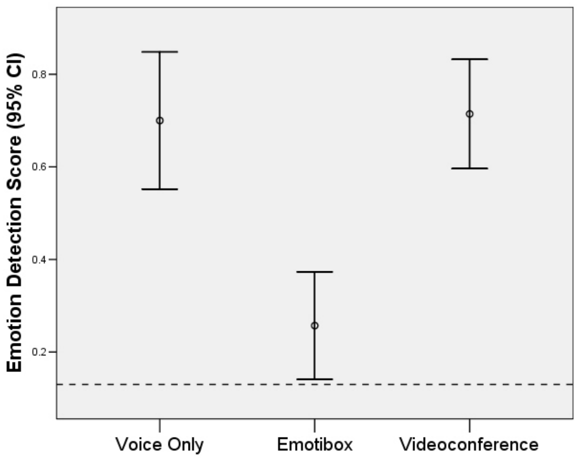

**Figure 7: Mean percent correct on emotion detection task by condition. The dotted line indicates chance performance.**

# 6. Discussion

## 6.1. Summary of Results

In this study, we compared the behavioral similarity and form similarity of avatar faces during real-time dyadic interaction. Our results demonstrated that, both verbally and nonverbally, people disclosed more information to avatars that were low in realism. In terms of verbal disclosure, subjects were perceived as less revealing, honest, and friendly in a videoconference then they were when interacting with either a text-only display or an avatar high in behavioral similarity but low in form similarity (the emotibox). In terms of nonverbal disclosure, subjects utilized more facial gestures and movements in a voice only interaction than in an interaction with either high behavioral realism (the emotibox) or high behavior and form realism (the videoconference). In other words, people emote more freely when their avatar does not express those emotions.

Overall, the emotibox proved to be a less effective interface than either of the two other alternatives in terms of copresence ratings and effectiveness in transmitting emotions. Nonetheless, without any training at all, on average subjects were above chance when attempting to identify the seven emotions with the emotibox, and on certain emotions were much higher than chance (e.g., 42% correct with "joyful"), which is encouraging considering that these emotions were expressed in a completely abstract fashion. With more elegant algorithms it should be quite possible to make more effective avatars that are high in behavioral similarity and low in form similarity.

## 6.2. Implications, Limitations and Future Directions

Earlier we discussed the defining characteristics of an avatar, and argued that a representation needs to have either high behavioral or form similarity in order to be utilized as an effective avatar in an interaction. In the current study, the emotibox was designed to elicit high behavioral similarity with low form similarity. However, by abstracting emotional expressions (as opposed to rendering the movements on a face-like digital model) we may have fallen short of our goal of producing high behavioral similarity. Participants may have been distracted by the foreign parameters of the box. In future work we plan on developing algorithms that are more stable (the same patterns emerge more readily across participants) and more intuitive (the mapping of color, shape, and orientation of the box is naturally tied to what we see on actual facial expressions).

Developing avatars that have high behavioral similarity and low form similarity is a worthy goal. The current study demonstrates that people are willing to disclose more personal information with an emotibox than with the avatar which is more realistic in form used in a videoconference. Unfortunately, the current instantiation of the emotibox elicited low copresence according to self report ratings and emotion recognition performance. If we can improve the quality of emotional transmittance of the emotibox, we can then create avatars in which people feel more comfortable using than ones highly realistic in form. Such avatars may be extremely useful for introverted students talking in front of a class in a distance learning scenario, patients interacting with a virtual therapist, and many other applications in which people interact with avatars in highly self-relevant and personal situations.

The current study is one of the first to use facial expressiveness as a dependent variable of copresence. Measuring people's nonverbal facial disclosure can be an extremely powerful tool to uncover the elusively latent construct of copresence. Indeed, the finding that people utilize more facial expressions when the other interactants cannot see their avatars is quite counterintuitive, as one might predict more facial expressions to be used when another person can actually see those facial expressions. This counterintuitive finding supports the notion raised in the introduction that facial expressions are direct correlates of emotions, as opposed to a social tool that can be turned on and off strategically. Future work examining people interacting via avatars and embodied agents should build upon this methodology.

For example, research should explore the interplay between avatar realism and context. Even if the emotibox elicited low copresence and emotion recognition, this may not be important for some tasks or settings - and may in fact be an advantage. For certain object-focused tasks in CVEs, for example, participants may be completely focused on the task and not focus on each other's faces. In this case an emotibox-type avatar could transmit only certain basic emotions that are designed to support the task (e.g., raising eyebrows translated into cyan cube color could transmit 'I am concentrating') which the collaborator could glance at occasionally without losing his or her concentration. Another type of avatar face might be developed for particular types of interpersonal interactions. The emotibox might, for example, transmit or signal only certain personal states, such as a smile translated into a yellow cube to signal 'I am happy to continue our conversation'.

In turn, exploring different types of contexts will allow us to converge upon an optimal avatar design. In the current work, the emotibox avatar is at the most basic end of the continuum of form realism of avatar representations in CVEs. But it will be possible to 'ramp up' avatar realism by degrees. Further towards the realistic end of the continuum, there could, for example, be a cube with a human-like appearance (such as a cartoon face, not necessarily resembling the real user) and this could be given a more subtle range of emotions that are conveyable (for example, colors on the cheeks to convey degrees of shyness).

The current work also suggests new direction for measurement criterion in CVEs. Although presence and copresence are largely regarded as the 'holy grail' of virtual environments research, as CVE (and other new media) use increases, avatars will require different levels of self-disclosure and expressiveness, with the traditional notion of copresence weighed only as an additional factor in the mix. Findings such as those presented in the current paper will provide a useful tool for gauging the kinds of representations required for different forms of mediated communication, as well as providing insights into the nuances of face-to-face behavior that may be easier to measure and manipulate within CVE environments.

Furthermore with face-tracking and other technologies, users will be able to use self presentation as a mechanism to transform their avatar's expressiveness. The possibilities for different forms of *transformed social interaction*– wearing different faces with capabilities for self-disclosure and emotional expressivity which can be changed 'on the fly' – offers potential for a number of training and related areas (see Bailenson & Beall [1] for other examples).

One of the most useful implications of the design of the emotibox is the idea of creating a framework within the notion of behavioral realism. Currently, behavioral realism is rarely discussed in a series of sub-dimensions. The emotibox raises issues in this regard. One dimension of behavioral similarity is the idea of *contingency*, the idea that for every performed behavior by the user, that behavior is tracked and then rendered on an avatar. Another one is *veridicality*, how much rendered behaviors resemble in terms of the actual animation. In other words, the emotibox from the current study was high in contingency but low in veridicality. A third type of realism is *correlation realism*. If not all behaviors of the human can be tracked, are there any behaviors that should be rendered? In other words, if it is not possible to track pupil dilation, but we know that pupil dilation correlates quite highly with heart-rate (which we can track), should we use probabilistic rendering of pupil dilation based on heart data? This is extremely important, given that tracking of human behaviors in real-time is currently quite difficult.

These areas of research and development will overlap, and there will be requirements for a variety of *degrees* of form and behavior realism in emerging media. Thus it is possible to envisage a range of avatar faces that could be combined in a pick-and-mix fashion to suit different types of interaction in CVEs, depending on the requirements for expressiveness and the task.

### 6.3. Conclusion

It is clear that avatar realism is critical to the future of collaborative virtual environment development. Highly realistic avatars with real-time facial form and tracking require more resources – both computationally and in terms of person-hours required to implement them. Moreover, the issue of the realism of digital human representations is a key question for a range of new media other than immersive virtual environments, such as videoconferencing, mobile telephony, online gaming, instant messaging and any other media that includes online representations of users. Understanding the relationship between form and behavioural realism is critical to begin examining the use of these new forms of media.

## Appendix

Verbal Disclosure Question Set A:

1. Where do you live on campus?
2. Where did you grow up?
3. What do your parents do?
4. What has been the most stressful event of the last six months for you?
5. Of all the people you know, whose death would bring you the most sadness?
6. What's the longest relationship you've ever been in?
7. Tell me a little more about yourself.

Verbal Disclosure Question Set B:

1. What are you majoring in?
2. Do you have any siblings?
3. What's the scariest thing that's ever happened to you?
4. Do you think you're ready for a long-term romantic relationship? Why do you feel that way?
5. Which part of your body are you most uncomfortable with?
6. How much money do your parents make?
7. Tell me a little more about yourself.

Copresence scale:

1. How easily distracted were you during the interaction?
2. How easy was it for you to tell how your partner felt?
3. How responsive was your partner?
4. How often were your partner's behaviors clearly a reaction to your own behaviors?

## Acknowledgements

The authors would like to thank Megan Miller for managing this project and Robby Ratan and Paola Worsley for assistance in data entry and analysis.

## References

[1] Bailenson, J. and Beall, A. Transformed Social Interaction: Exploring the Digital Plasticity of Avatars. In R. Schroeder and A.-S. Axelsson (Eds.). *Avatars at Work and Play: Collaboration and Interaction in Shared Virtual Environments.* London: Springer, 2005.

[2] Bailenson, J.N., Beall. A.C., & Blascovich, J. (2002). Mutual gaze and task performance in shared virtual environments. *Journal of Visualization and Computer Animation, 13*, 1-8.

[3] Bailenson, J.N., & Blascovich, J. (2004) Avatars. *Encyclopedia of Human-Computer Interaction*, Berkshire Publishing Group, 64-68.

[4] Bailenson, J. N., Swinth, K. R., Hoyt, C. L., Persky, S., Dimov, A., & Blascovich, J. (in press). The independent and interactive effects of embodied agent appearance and behavior on self-report, cognitive, and behavioral markers of copresence in immersive virtual environments. *PRESENCE: Teleoperators and Virtual Environments*.

[5] Becker, B. and Mark, G. (2002), Social Conventions in Computer-mediated Communication: A Comparison of Three Online Shared Virtual Environments, in R. Schroeder (ed.), *The Social Life of Avatars*: *Presence and Interaction in Shared Virtual Environments*. London: Springer, 19-39.

[6] Bente (2004). Measuring Behavioral Correlates of Social Presence in Virtual Encounters. Paper presented in the *54th Annual Conference of the International Communication Association*, New Orleans, LA

[7] Blascovich, J., Loomis, J., Beall, A. C., Swinth, K. R., Hoyt, C. L., & Bailenson, J. N. (2002). Immersive virtual environment technology as a methodological tool for social psychology. *Psychological Inquiry, 13*, 146-149.

[8] Biocca, F.; Harms, C. and Burgoon, J.K. (2003). Toward a More Robust Theory and Measure of Social Presence: Review and Suggested Criteria', *Presence: Journal of Teleoperators and Virtual Environments*, vol.12, no.5, pp. 456-80.

[9] Bowers, J., Pycock, J. & O'Brien, J. (1996). Talk and embodiment in collaborative virtual environments. *Conference on Human Factors in Computing Systems (CHI'96)*, 58-65.

[10] Cheng, L., Farnham, S. and Stone, L. (2002), Lessons Learned: Building and Deploying Shared Virtual Environments, in R. Schroeder (ed.), *The Social Life of Avatars: Presence and Interaction in Shared Virtual Environments*. London: Springer, 90-111.

[11] Ekman P, Friesen WV (1976): Measuring facial movement. *Journal of Environmental Psychology and Nonverbal Behavior* 1:56-75.

[12] Ekman P, Friesen WV (1978): *Facial action coding system: A technique for the measurement of facial movement*. Palo Alto, CA: Consulting Psychologists Press.

[13] Essa, I. & A. Pentland. ``A Vision System for Observing and Extracting Facial Action Parameters'', In *Proceedings of IEEE Computer Vision Pattern Recognition Conference 1994*, Seattle, WA. June 1994.

[14] Garau, M. (2003). *The Impact of Avatar Fidelity on Social Interaction in Virtual Environments*. Ph.D. thesis, Department of Computer Science, University College London.

[15] Heeter, C. (1992). Being there: The subjective experience of presence. *PRESENCE: Teleoperators and Virtual Environments, 1(2)*, 262-271.

[16] Izard CE. 1971. *The Face of Emotion*. New York: Appleton Century Crofts.

[17] Lanier, J. (2001). Virtually there. *Scientific American,* April, 66-75.

[18] Lee, K., M. (2004). Presence, Explicated. *Communication Theory, 14*, 27-50.

[19] Michel, P. and El Kaliouby, R. (2003). Real time facial expression recognition in video using support vector machines. *Proc. 5th International conference on multimodal interfaces (ICMI)*, November 2003.

[20] Moon, Y. (200). Intimate Exchanges: Using Computers to Elicit Self-Disclosure from Consumers. *Journal of Consumer Research, 26*, 323-339

[21] Picard, R. W. & S. Bryant Daily (2005), "Evaluating affective interactions: Alternatives to asking what users feel," *CHI Workshop on Evaluating Affective Interfaces: Innovative Approaches*, Portland Oregon, April 2005.

[22] Schroeder, R. (2002). Social Interaction in Virtual Environments: Key Issues, Common Themes, and a Framework for Research, in R. Schroeder (ed.), *the Social Life of Avatars: Presence and Interaction in Shared Virtual Environments*. London: Springer, pp.1-18.

[23] Schroeder, R. (2002). Copresence and Interaction in Virtual Environments: An Overview of the Range of Issues. In *Presence 2002: Fifth International Workshop. Porto*, Portugal, Oct.9-11, 274-295.

[24] Slater, M., Sadagic, A., Usoh, M., Schroeder, R. (2000), Small Group Behaviour in a Virtual and Real Environment: A Comparative Study. *Presence: Journal of Teleoperators and Virtual Environments, 9(1)*: 37-51.

[25] Sproull, L., Subramani, M., Kiesler, S., Walker, J., Waters, K. (1996). When the interface is a face. *Human-Computer Interaction, 11*, 97-124.

[26] Turk, M. and Kölsch, M., (2004), "Perceptual Interfaces," G. Medioni and S.B. Kang (eds.), *Emerging Topics in Computer Vision*, Prentice Hall.

[27] Walther, J. B. (1996). Computer-mediated communication: Impersonal, interpersonal and hyperpersonal interaction. *Communication Research, 23*, 3-43.

[28] Weisband, S. & Kiesler, S. (1996). Self disclosure on computer forms: Meta-analysis and implications. Presented at p*roceedings of the SIGCHI conference on human factors in computing systems*.

[29] Whittaker, S. (2002). Theories and Methods in Mediated Communication. In Graesser, A., Gernsbacher, M., and Goldman, S. (Ed.) *the Handbook of Discourse Processes*, 243-286, Erlbaum, NJ.