Jeremy N. Bailenson
bailenson@stanford.edu
Virtual Human Interaction Lab
Department of Communication
Stanford University
Stanford, CA 94305-2050

Andrew C. Beall
Jack Loomis
Jim Blascovich
Department of Psychology
University of California at
Santa Barbara
Santa Barbara, CA 93106

Matthew Turk
Department of Computer Science
University of California at
Santa Barbara
Santa Barbara, CA 93106

Transformed Social Interaction: Decoupling Representation from Behavior and Form in Collaborative Virtual Environments

Abstract

Computer-mediated communication systems known as collaborative virtual environments (CVEs) allow geographically separated individuals to interact verbally and nonverbally in a shared virtual space in real time. We discuss a CVE-based research paradigm that *transforms* (i.e., filters and modifies) nonverbal behaviors during social interaction. Because the technology underlying CVEs allows a strategic decoupling of rendered behavior from the actual behavior of the interactants, conceptual and perceptual constraints inherent in face-to-face interaction need not apply. Decoupling algorithms can enhance or degrade facets of nonverbal behavior within CVEs, such that interactants can reap the benefits of nonverbal enhancement or suffer nonverbal degradation. Concepts underlying transformed social interaction (TSI), the ethics and implications of such a research paradigm, and data from a pilot study examining TSI are discussed.

1 Introduction

> While conversing, you could look around the room, doodle, fine-groom, peel tiny bits of dead skin away from your cuticles, compose phone-pad haiku, stir things on the stove; you could even carry on a whole separate additional sign-language-and-exaggerated-facial-expression type of conversation with people right there in the room with you, all while seeming to be right there attending closely to the voice on the phone. And yet—and this was the retrospectively marvelous part—even as you were dividing your attention between the phone call and all sorts of other idle little fugue-like activities, you were somehow never haunted by the suspicion that the person on the other end's attention might be similarly divided (Wallace, 1996, p. 146).

In his hypothetical depiction of future, video-based remote interaction, Wallace looks back fondly on traditional phone conversations and notes a distinct advantage that telephone conversations hold over videoconferencing. While remote conferences mediated by telephony limit interactants to a single communication channel, the second channel (i.e., visual information) offered in conferences mediated by video may prove superfluous or even counterproductive to the quality of the interaction.

Collaborative virtual environments (CVEs) that employ 3D computer-

Presence, *Vol. 13, No. 4, August 2004,* 428–441

generated avatars to represent human interactants (as opposed to direct video feeds) may provide an ideal balance between the limited information offered via audio communication and the problems that seem inherent to videoconferences. In most current CVE implementations, interactants have the opportunity to utilize two perceptual channels: audition and vision. However, unlike a videoconference, a CVE operating system can be designed to render a carefully chosen subset of interactants' nonverbal behaviors, filter or amplify that subset of behaviors, or even render nonverbal behaviors that interactants may not have performed.

Transformed social interaction (TSI) involves novel techniques that permit changing the nature of social interaction (either positively or negatively) by providing system designers with methods to enhance or degrade interpersonal communication. Tracking nonverbal signals (e.g., eye gaze, facial gestures, body gestures) and rendering them via avatars allows for a strategic decoupling of transmitted nonverbal signals from one interactant from those received by another (i.e., rendered). For example, eye gaze directed from A to B can be transformed without A's knowledge, such that B experiences the opposite, gaze aversion. The idea of decoupling rendered behaviors from actual ones is not new (see discussion in Benford, Bowers, Fahlen, Greenhalgh, & Snowdon, 1995; on truthfulness, as well as Loomis, Blascovich, & Beall, 1999). Here, we explore this strategic decoupling. TSI can be applied to some, all, or no members of a CVE.

Distorting the veridicality of communication signals certainly raises ethical questions. We do not advocate the unconstrained use of TSI. However, we do believe that as CVEs become widespread, decoupling rendered behavior from actual behavior is inevitable. Indeed, current users of chat rooms and networked video games frequently represent themselves nonveridically (Yee, 2002). Consequently, the ethical implications of TSI warrant serious consideration by anyone who interacts via CVEs. At the very least, CVE system designers should anticipate and try to obviate misuse. Examining TSI now as a basic research question will increase the probability that we can ethically use and manage CVEs in the future.

The remainder of this paper is divided into three sections. First, we review some of the ideas and current implementations of CVEs, nonverbal behavior tracking technology, and the visual nonverbal behaviors in interaction. Second, we provide concrete examples of TSI and discuss possible implications for conversation. Finally, we conclude by discussing some of the ethical implications of TSI, provide pilot data from a study in which participants attempted to detect TSI, and point to future directions for research.

2 Nonverbal Behavior and CVEs

Social scientists have long understood that social interaction involves communication of both verbal and nonverbal signals. The former include spoken, written, and signed language; the latter include gaze, gestures and postures, facial expressions, touch, etc., as well as paralanguage cues such as variations in intonation and voice quality. If specific parallel signals were redundant in meaning across channels, little need would exist for multiple channels, and correspondingly little need would exist for sophisticated telecommunication technology beyond simple audio transmission.

However, signals often prove inconsistent across channels (e.g., "He's a winner" can communicate its literal meaning or the opposite, depending on tone). Furthermore, some channels appear less controllable by interactants, and hence are judged more veridical (e.g., nonverbal channels communicating feelings or emotions and motivation). Also, signals directed toward specific interactants convey messages to third parties. For example, if two interactants share mutual gaze to the exclusion of a third, the message to the third person can lead to feelings of ostracism (Williams, Cheung, & Choi, 2000).

Although much research on the role of nonverbal signals in social interaction has appeared (for reviews, see Argyle, 1988; Depaulo & Friedman, 1998; Patterson, 1982), for the most part, investigators have had to choose between ecological validity (i.e., a realistic setting or environment) and experimental control, forcing the sacrifice of one for the other. Ecologically realistic research has tended to involve qualitative observations. Experimental work, ideally, examines social behavior in

the lab via strict controls over most variables, sometimes even involving confederates or imagined scenarios, but without much in the way of external validity or generalizability. CVEs promise to produce major advances in the understanding of social interaction, both dyadic and group, by allowing much more ecological validity while maintaining a high level of experimental control (Blascovich et al., 2002; Loomis et al., 1999).

Technology has long facilitated social interaction. For centuries, written correspondence has proven highly effective for communicating ideas and, to a lesser extent, feelings. The telegraph permitted more or less real-time interaction. However, the telephone constituted an enormous advance, both because it afforded real-time interaction and because it allowed communication via paralanguage cues so important for emotional exchange. More recently, videoconferencing has permitted the communication of some visual nonvisual (NV) cues, but with little opportunity for "side-channel" communication among nonconversing group members (e.g., meaningful glances), and typically without allowing for mutual gaze among group members (Gale & Monk, 2002; Lanier, 2001; Vertegaal, 1999). Now, CVEs promise to promote more effective dyadic (i.e., 2-person) and *n*-person interactions (Zhang & Furnas, 2002; Bailenson, Beall, & Blascovich, 2002; Slater, Sadagic, Usoh, & Schroeder, 2000; Normand et al., 1999; Leigh, DeFanti, Johnson, Brown, & Sandin, 1997; Mania & Chalmers, 1998; Schwartz et al., 1998) by sensing and rendering the visual NV signals of multiple interactants. Two approaches in this regard are: (1) capturing and interpolating 2D images from multiple video cameras and recovering the 3D models, and (2) tracking gestures using a variety of sensors, including video. The interpolated images or rendered 3D models can then be displayed to each of the interactants.

In addition to making remote human interaction possible, communication technology has important scientific value in terms of facilitating the assessment of the sufficiency or adequacy of transmitted verbal and nonverbal signals. For example, the fact that telephone conversants feel an intimate connection indicates that auditory information is often adequate for personally meaningful dyadic interaction. This sense of connectedness persists despite interactants' awareness that they are actually talking to devices, which indicates that the process of social interaction via telephone is to some extent "cognitively impenetrable" (Pylyshyn, 1980). Mirror talking provides another compelling example. If a room contains a large mirror, people often find themselves conversing with each other's mirror or "virtual" image. Interestingly, no discernible loss in effectiveness of the interaction appears to occur, even though each interactant knows that he or she is not engaging in face-to-face interaction with the actual person. This "transparency" of interaction is also observed in dyadic interaction over properly designed videoconferencing systems (i.e., ones that permit mutual eye gaze) and will be true of CVE systems in the near future, even though interactants know at some level that they see only digital models of other interactants. Transparency of interaction, reflected both in interactants' experience and in the effectiveness of group performance (e.g., in collaborative decision making), speaks to the sufficiency of the verbal and nonverbal signals and also indicates that social interaction is mediated by automatic processes that are quite separate from conscious cognition (Fodor, 1983; Pylyshyn, 1980). Thus, the creation of new communication media can provide insight into human social interaction.

3 Implementations of TSI

In this section, we outline three important TSI components. Each involves a number of theoretical ideas that warrant technical development as well as evaluation via behavioral research. The categories of TSI include: *self representations* (i.e., avatars), *sensory capabilities*, and *contextual situation*. Each category also provides researchers with powerful new tools to investigate and improve understanding of psychological processes underlying behavior (Blascovich et al., 2002; Loomis et al., 1999). Specifically, investigators can manipulate the underlying structure of social interaction using TSI by altering the operation of its individual components, and thereby "reverse engineering" social interaction. In this paper, however, the focus is to explore the theoretical nature of TSI as its own basic research question and to speculate on its potential implications for communication via CVEs. While we discuss these three categories

as separate entities, clearly, a system that employs TSI would be most effective as some combination of the three. We keep them separate for the purpose of clarity in this paper.

We realize that all of the necessary CVE technology may not yet be available (see Kraut, Fussell, Brennan, & Siegel, 2002). Furthermore, in order to adequately study and enable transformed social interaction in collaborative virtual environments, the technology used for tracking nonverbal signals must eventually be passive and unobtrusive. Sensors and markers that are worn on the body can limit the naturalness of interaction by causing participants to focus on the technology at the expense of the interaction. Computer vision technology offers the possibility of using passive, noncontact sensing to locate, track, and model human body motion. Subsequently, pattern recognition and classification techniques can be used to recognize meaningful movements and gestures.

In the past dozen or so years, there has been a significant and increasing interest in these problems within the computer vision research community (Turk & Kolsch, 2003; Black & Yacoob, 1997; Donato, Bartlett, Hager, Eckman, & Sejnowski, 1999; Feris, Hu, & Turk, 2003; Stiefelhagen, Yang, & Waibel, 1997; Viola & Jones, 2001). Motivated by various application areas, including biometrics, surveillance, multimedia indexing and retrieval, medical applications, and human-computer interaction, there has been significant progress in areas such as face detection, face recognition, facial expression analysis, articulated body tracking, and gesture recognition. The state of the art in these areas is not yet to the point of fully supporting CVEs, as many of these systems tend to be slow and lack robustness in real-world environments (with typical changes in lighting, clothing, etc.).

But the progress is promising, and we expect to see an increased utility of these technologies to track and model nonverbal behaviors in order to transmit and transform them within the context of CVEs. We believe that each of the TSI implementations discussed in the current work is foreseeable, perhaps even in the near future. For the purposes of the current discussion, details of the specific CVE implementation are not critical; TSI should be effective in projection-based CVEs, head-mounted display CVEs, CAVEs, or in certain types of augmented-reality CVEs.

A concrete example of a typical CVE interaction helps describe the specific types of transformations. Generally, TSI should enable interactants to communicate more effectively by providing them with more information as well as providing them (or systems designers) with more control in directing their nonverbal behaviors. The latter suggests, on a more cynical note, that the people who may profit most from TSI may be those who enter interactions with specific goals: for example, changing the attitudes of the other interactants (Slater, Pertaub, & Steed, 1999). In the subsequent sections, we describe an interaction with a leader and one or more community members evaluating a proposal in a CVE. However, one could just as easily substitute leader with politician, teacher, lawyer, leader, or missionary, and substitute community members with voters, students, jurors, members, or atheists. Hence, the theoretical parameters and implications of TSI have applications across many different contexts.

3.1 Transforming Self Representations

In CVEs, avatars representing interactants can bear varying degrees of photographic or anthropomorphic (Garau et al., 2003; Bailenson, Beall, Blascovich, Raimmundo, & Weisbuch, 2001; Sannier & Thalmann, 1998), behavioral (Bailenson et al., 2002; Cassell, 2000; Biocca, 1997), and even dispositional, resemblance to interactants they represent. Assuming that interactants (by their own design or through the actions of systems operators) have the freedom to vary both the photographic and behavioral similarity of their avatar to themselves, a number of subtle but potentially drastic (in terms of outcomes of CVE interactions) transformations can occur.

In many instances, similarity breeds attraction (Byrne, 1971). We know that people treat avatars that look like themselves more intimately than avatars that look like others, as indicated by invasion of their personal space and willingness to perform embarrassing acts in front of them, and by how attractive and likable they believe the avatars to be (Bailenson, Blascovich, Beall, & Guadagno, 2004; Bailenson et al., 2001). Given this special rela-

tionship, a CVE interactant may use this principle to an advantage. Consider the situation in which a leader and a community member are negotiating via a CVE. A particularly devious leader can represent herself by incorporating characteristics of the member's representation. By making herself appear more similar to the member, the leader becomes substantially more persuasive (Chaiken, 1979; Simons, 1976). Indeed, a leader would be able to adjust the structural or textural similarity of her own avatar idiosyncratically to the members in her audience.

This similarity could be achieved in various manners, employing any of a number of techniques to parametrically vary the similarity of computer-generated models via 2D and 3D morphing techniques (Blanz & Vetter, 1999; Busey, 1998; Decarlo, Metaxas, & Stone, 1998). The leader could be represented as some kind of a hybrid, maintaining some percentage of her original facial structure and texture, but also incorporating percentages of the member's structure and texture. Alternatively, the leader could be represented completely veridically to her facial structure, but for a few frames per second could replace her own head with the head of the member. Priming familiarity with limited exposure to human faces has proven to be effective with 2D images (Zajonc, 1971). Finally, consider the situation in which the leader is interacting via CVE with two members. The leader can be differentially represented to both members simultaneously such that each member sees a different hybrid leader avatar incorporating aspects of each member. In other words, the leader does not need a consistent representation across interactants, because the CVE operator is free to render different leader avatars to each member.

Incorporating the self-identity of other interactants can also occur via behavioral characteristics. Psychological research has demonstrated that when an experimenter subtly mimics experimental participants (e.g., leans in the same direction as they do, crosses her legs when the participants do), participants subsequently report that they liked the experimenter more and smoother conversation flowed (Chartrand & Bargh, 1999). This "chameleon effect" could be extremely effective in CVEs. The leader (or the system operator) can use algorithms to detect motions of the other interactants at varying levels of detail and coordinate the animations of her avatar to be a blended combination of her own and those of the others.

Consider a CVE interaction consisting of a leader and two members. In the course of this interaction, patterns of nonverbal behaviors will emerge, and statistics based on a running tabulation can be automatically collected via CVE technology. In other words, if there is a certain rate of head nodding exhibited by person A and another rate exhibited by person B, the leader's head can be made to nod in a way consistent with the statistics (e.g., an average or median). Alternatively, the leader's avatar can just mimic each interactant individually and render those particular movements only to each corresponding interactant.

The leader could also morph her representation with that of an unrepresented party not present in the CVE, but who is previously known to possess qualities that inspire certain reactions. Depending on the context, for example, the leader can morph a percentage of famous politicians, historical figures, or even pop stars into her avatar. This feature blending can be explicit and blatant (e.g., the leader looks just like an expert or a religious figure) or more implicit and subterranean (e.g., the leader incorporates subtle features such as cheekbones and hairstyle). Alternatively, the leader can morph herself with a person who may not be famous but with whom the member maintains a deep trust (Gibson, 1984).

A second form of avatar transformation arises from the ability to selectively decouple and reconstruct rendered behavior in CVEs. In other words, not only can interactants render nonverbal behaviors different from the nonverbal behaviors that they actually perform, but, similarly to the discussion above, they can render those behaviors idiosyncratically for each of the other interactants.

Consider what we term Non-Zero-Sum-Mutual-Gaze (NZSMG). Ordinary mutual gaze occurs when individuals look at one another's eyes during discourse. In face-to-face conversation, mutual gaze is zero-sum. In other words, if interactant A maintains eye contact with interactant B for 70 percent of the time, it is not possible for A to maintain eye contact with interactant C for more than 30 percent of the time. However, interaction in CVEs is not bound by this constraint. With digital

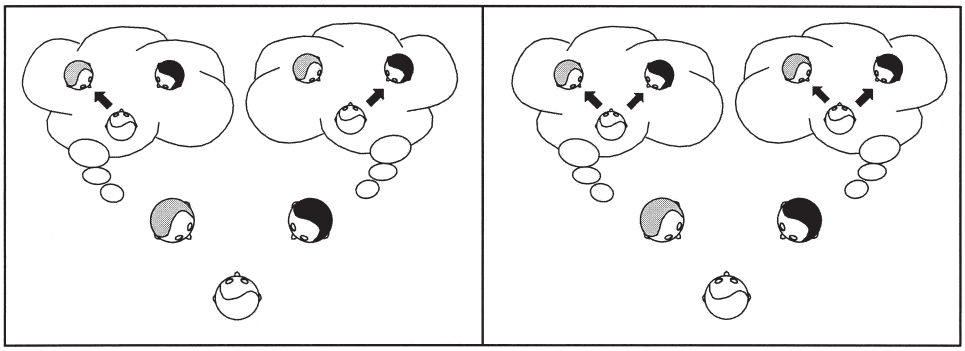

Figure 1. *Internal belief states from implicit NZMG (left) and explicit NZMG (right).*

avatars, A can be made to appear to maintain mutual gaze with both B and C for a majority of the conversation.

Gaze is one of the most thoroughly studied nonverbal gestures in research on social interaction (Rutter, 1984; Kleinke, 1986; Kendon, 1977). Direct eye gaze can provide cues for intimacy, agreement, and interest (Arygle, 1988). Furthermore, gaze can enhance learning during instruction as well as memory for information (Fry & Smith, 1975; Sherwood, 1987). The advantage of using CVEs is that normal nonverbal behaviors of interactants can be augmented via NZSMG. Furthermore, the interactants in a CVE can either be unaware of this transformation (i.e., implicit NZSMG) or aware of this transformation (i.e., explicit NZSMG), as Figure 1 demonstrates. Preliminary work studying implicit NZSMG has demonstrated that interactants are not aware of the decoupling from actual behavior. Furthermore, the interactants respond to the artificial gaze as if it were actual gaze (Beall, Bailenson, Loomis, Blascovich, & Rex, 2003). This method may prove to be most effective during distance learning in educational CVEs (Morgan, Kriz, Howard, das Neves, & Kelso, 2001) in which the instructor uses her augmented gaze as a tool to keep the students more engaged.

Decoupling can also be used to achieve the opposite effect. Consider the situation where the leader wants to scrutinize the nonverbal behaviors of member A, but does not want the member to feel uncomfortable from her unwavering gaze. The leader can render herself looking at her shoes, or perhaps at member B in the CVE, while in reality she is watching member A's every move.

In order for such a system to be effective, there must be a convincing algorithm to drive the autonomous eye gaze. In other words, if the leader wants the freedom to employ NZSMG or to wander around the CVE scrutinizing different aspects of the conversation undetected, she (by her own device or assisted by the systems operator) must maintain the illusion that her avatar is exhibiting the typical and appropriate nonverbal gestures. There are a number of ways to achieve this. The first is some type of artificial intelligence algorithm that approximates appropriate gestures of the leader's avatar by monitoring the gestures and speech by the other interactants. While there have been significant advances in this regard (Cassell, 2000), the ability of an algorithm to process natural language, as well as generate believable responses, may still be many years off. A more likely method for achieving this goal would be to use actual humans instead of AI algorithms. In this scenario, the leader employs one or more nonverbal "cyranoids" (Milgram, 1992) to augment the nonverbal behaviors presented to each individual member. To do so, the leader solicits the help of several assistants each of whose job is to provide the nonverbal behaviors targeted toward a particular member. See Figure 2.

In this many-to-many, "Wizard of Oz" implementation, each member is presented a unified Leader who is rendered privately to her; this private representation would be a melding of the actual leader and one of her assistants so that when the leader's attention was di-

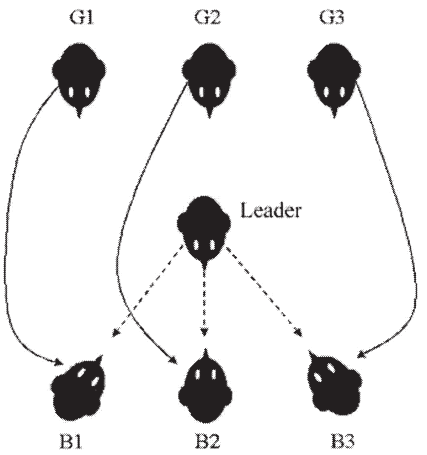

Figure 2. *A depiction of cyranoids. On the top row are three nonrendered gesturers. Each member on the bottom hears the leader's actual verbal behaviors (dashed lines). However, each member views the nonverbal behavior of her dedicated gesturer rendered onto the avatar of the leader (unbroken lines).*

verted away from that member for long periods of time, the assistant could step in and help maintain a believable interaction by seamlessly serving as the leader's proxy. The leader herself would then act as a conductor overseeing all the interactions yet being free to focus her attention on individual members when she so desires. In addition, the leader is free to wander about the digital space, consult her notes, take a rest, or conduct a sidebar meeting with another person. However, because her avatar is partially cyranic it can continue to exhibit the appropriate nonverbal behaviors all the while to each member. Furthermore, having a number of assistants whose sole focus is to respond with appropriate nonverbal gestures to each of the interactants in the CVE should maximize the members' involvement or sense of presence in the CVE. For important meetings, seminars, or presentations conducted via CVEs, individual interactants may want to utilize a number of assistants as a core presentation team.

3.2 Transforming Sensory Capabilities

Interactants can be assisted by technology that takes advantage of CVEs that can keep precise running tabs of certain types of behaviors, and then display summaries of those behaviors exclusively to individual interactants. For example, consider an educational CVE in which an instructor wants to ensure that she is directing her nonverbal behaviors in a desired fashion. Such as instructor may want to monitor her mutual gaze to ensure that she is not looking at any one student more than others during a presentation. The tracking equipment used to render the scene can keep an online total of the amount of time the instructor gazed at each individual student. The CVE can render a display of this gaze meter, as well as use visual or auditory alerts to inform the instructor of disproportionate applications of gaze.

Furthermore, interactants can use the tracking data summaries to learn more about the attitudes of the others. Nonverbal gestures are often correlates of specific mental states (Ekman, 1978; Zajonc, Murphy, & Inglehart, 1989). For example, in general we nod when we agree, smile when we are pleased, tilt our heads when we are confused, and look at something in which we are interested. Interactants will be able to tailor their CVE systems to keep track of nonverbal behaviors with the goal of aiding interactants to infer the mental states of the other interactants. For example, a teacher will be able to gauge the percentage of students exhibiting nonverbal behaviors that suggest confusion or not understanding a point in a lesson. Similarly, a leader could determine who in a room full of members is responding most positively to her behavior. Intuitively tabulating and assessing the nonverbal behaviors of others is certainly something that humans do constantly in face-to-face interactions. With CVEs, interactants will be able to tabulate these behaviors with greater precision. Interactants can use the objective tabulations from the tracking data to augment their normal intuitions about the gestures occurring in the interaction.

Another transformation involves filtering or degrading certain signals or nonverbal behaviors. There are some visual nonverbal behaviors that tend to distract interactants. Using filtering algorithms, interactants can prevent counterproductive distractions in a number of ways. For example, consider the situation in which a speaker in a CVE taps her pen rapidly as she speaks. In face-to-face meetings, this type of behavior can distract

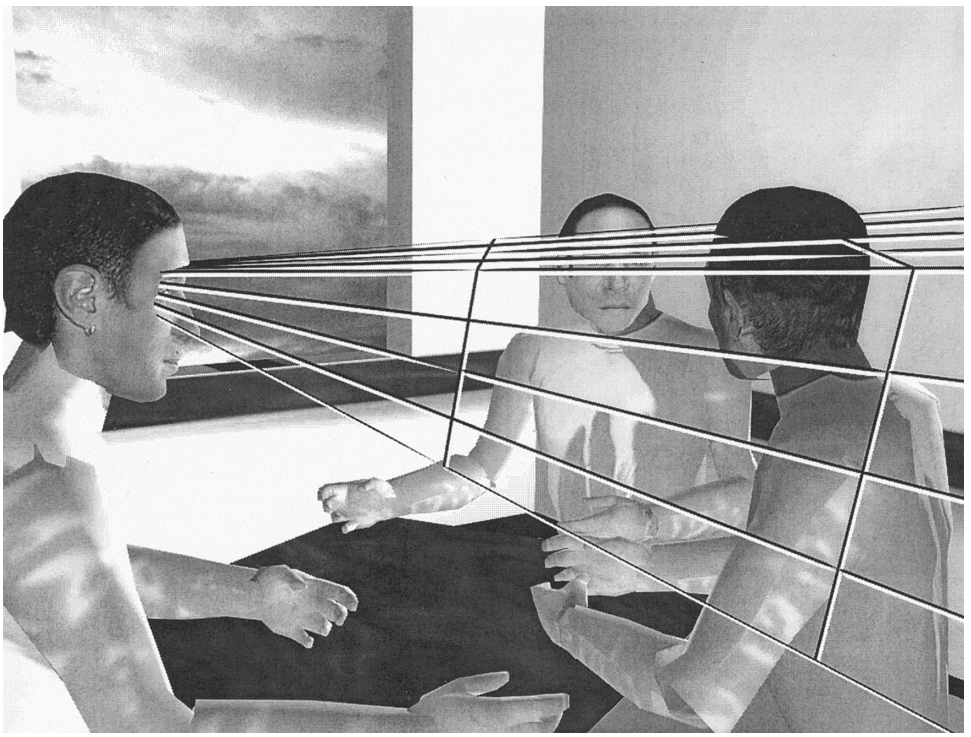

Figure 3. *View frustrums marking the field of view of interactants.*

interactants. Using a CVE, this type of behavior can be filtered in two ways. First, the speaker can filter the behavior on the transmitting end. If people know that they have difficulty suppressing certain nonverbal behaviors that tend to be perceived in a negative manner, such as a nervous tick, they can activate a filter that prevents the behavior from being rendered. Similarly, in certain situations, a CVE interactant may not want to render certain nonverbal behaviors. Consider the leader example. The potential member may benefit from rendering her "poker face," that is, not demonstrating any enthusiasm or disappointment via facial expressions. Consequently the member may accrue strategic advantage during a negotiation. Furthermore, interactants can filter behaviors on the receiving end. If a speaker's hand motions are distracting, then a listener can simply choose to not render that interactant's hand movements.

Another example of transforming sensory capabilities is producing a visual indicator regarding where each interactant's attention currently lies as revealed by their eye direction (Velichkovsky, 1995). We have explored a technique that involves rendering each person's view frustrum to indicate the field of view, as Figure 3 illustrates. In this example, the wire frame frustrums spotlight the 3D space visible to each person. This feature, color coded for each person, may be especially helpful to teachers in a distance learning CVE who could use such information to see where students are focusing their visual attention without having to look directly at the students' eyes.

There are a number of similar tools (i.e., specific objects rendered only to particular interactants) that can assist interactants in a CVE. For example, in our NZSMG studies an experimenter enters a CVE and attempts to persuade other interactants regarding a certain topic (Beall et al., 2003). In those interactions, we render the interactants' names over their heads on floating billboards for the experimenter to read. In this manner the experimenter can refer to people by name more easily. There are many other ways to use these floating billboards to assist interactants, for example, reminders

about the interactant's preferences or personality (e.g., "doesn't respond well to prolonged mutual gaze").

One of the most useful forms of transforming sensory capabilities may be to enlist one or more human consultants who are rendered to only one member in a CVE (i.e., virtual ghosts). Unlike a face-to-face interaction, a CVE will enable an interactant to have informed human consultants who are free to wander around the virtual meeting space, to scrutinize the actions of other interactants, to conduct online research and sidebar meetings in order to provide key interactants with additional information, and to generally provide support for the interactants. For example, the leader can have her research team actually rendered beside her in the CVE. Members of her team can point out actions by potential members, suggest new strategies, and even provide real-time criticism and feedback concerning the behavior of the leader without any of the other members having even a hint of awareness concerning the human consultants' presence. Alternatively, the leader herself can go into "ghost mode" and explore the virtual world with her team while her avatar remains seated, and is even controlled by yet another member of her team.

3.3 Transforming the Situation

In addition to transforming their representation and sensory capabilities, CVE interactants can also use algorithms to transform their general spatial or temporal situations. In a CVE, people generally adopt a spatially coherent situational context across all remote interactants that brings everyone together in the shared space. However, there is no reason that the details and arrangements of that virtual space need to be constant for all the interactants in the CVE. Consider the situation for three interactants. Interactant A may choose to form an isosceles triangle with the other two, while both interactants B and C may choose to form equilateral triangles. Interactant A may even choose to flip the locations of B and C. In this scenario, the CVE operating system can preserve the intended eye gaze direction by transforming the amplitudes or direction of head and eye movements in a prescribed manner. While this is a somewhat simple example, with as many as four interactants it is straightforward to design spatial transformations that allow the intended eye and head gaze cues to remain intact across all interactants. While eventually such discordance may cause the quality and smoothness of the interaction to suffer, there are a number of ways that transforming the situation can assist individual interactants.

One such transformation involves multilateral perspectives. In a normal conversation, each interactant has a unique and privileged perspective. That perspective is a combination of her sensory input (e.g., visual and acoustic fields of view) and internal beliefs about the interaction. In normal, face-to-face interactions, people continually use sensory input to update and adjust their internal beliefs (Kendon, 1977). Interactants in a CVE will possess a completely new mechanism to adjust and update internal beliefs. A person's viewpoint can be *multilateral,* as opposed to *unilateral* (normal). In other words, in a real-time conversation, interactant A can take the viewpoint of interactant B, and perceive herself as she performs various verbal and nonverbal gestures during the interaction. In this manner, she can acquire invaluable sensory information pertaining to the interaction, and update her internal beliefs concerning the interaction in ways not possible without the CVE.

Consequently, interactants in educational and persuasive interactions may be able to improve performance, because seeing oneself through the eyes of another may allow one to develop a more informed set of internal beliefs about others (Baumeister, 1998). Furthermore, it may be the case that being able to experience an interaction through someone else's eyes should reinforce the fact that one is indeed copresent in the CVE (e.g., Durlach & Slater, 2000). Finally, utilizing mulitlateral perspectives may assist students in distance learning CVEs in terms of training transfer effects (Rickel & Johnson, 2000) that might occur after an interactant who has been trained in multilateral perspective taking performs similar group tasks in nonmediated situations.

A second situational transformation involves partially recording the interaction and adjusting temporal properties or sequences in real time. Similar to commercial products sold for digitally recording and playing back broadcast television, interactants in a CVE should be able to accelerate and decelerate the perceived flow of time during the mediated interaction. Consider the fol-

lowing situation. The student in a distance learning CVE does not understand an example that the instructor provides. The student can "rewind" the recorded interaction, go back to the beginning of the confusing example, and then play back the example. Once the student has understood the confusing example, she can then turn up the rate of playback (e.g., watch the sequence at 2X speed), and eventually, she can catch up to the instructor again. By slowing down the rendered flow of time or speeding it up, the interactant can focus differentially on particular topics and can review the same scene from different points of view without missing the remainder of the interaction. Of course, doing so will result in costs to that interactant's contribution to the CVE in terms of interactivity (i.e., what does her avatar do while she rewinds?). Consequently the disruption of the temporal sequence will necessarily be coupled with some kind of an avatar autopilot.

Changing the rate of time in a CVE brings up another interesting transformation. Traditionally, CVEs are roughly defined as "geographically separated interactants" interacting over some kind of a computer-mediated network in a shared environment. However, by combining some of the concepts discussed in previous sections, it may be possible to include in the definition of a CVE "temporally separated interactants" in a shared environment. Consider a videoconference of a business meeting. Oftentimes, interested parties who cannot attend the meetings will later review a videotape of the meeting. In a CVE, the temporally absent member has an option to more deeply involve herself in the interaction. Specifically, she can situate her avatar in a specific place in the CVEs seating arrangement and use an autopilot to give her representation rudimentary nonverbal behaviors. Furthermore, the absent member can program her avatar to perform simple interactive tasks—prerecorded introductions, answers to certain questions about the CVE topic, or perhaps more realistically for the near-term, direct the avatar to play back a recorded performance. Then, the CVE interaction can proceed in real time with the temporally absent member's avatar approximating the types of behaviors that she would do and say. As a result, temporally present members would actually direct pieces of the conversation towards the absent member as well as transmit nonverbal gestures towards her. Later on, instead of just reviewing the recording, the temporally absent member can take her place in the CVE and actually feel present in the dialogue, receiving appropriate nonverbal behaviors and maximizing the degree of copresence. Moreover, the members of the CVE who were present at the scheduled time can program their avatars, during the replay of the interaction, to respond to any post hoc questions that the absent member might have. In this way the degree of interactivity during the replay can be increased, and perhaps at some point in the not-too-distant future the line between real-time and non-real-time interactions will become interestingly blurred.

4 Implications of TSI and Research Directions

For better or for worse, TSI implemented through CVEs has great potential to change the nature of mediated interaction. The strategic decoupling of rendered behavior from actual behavior allows interactants to break many constraints that are inherent in face-to-face interaction as well as other forms of mediated interactions, such as telephone and videophone conferences. The effects of TSI remain to be seen. Assuming that implementation of the TSI techniques are technically feasible, and that using TSI implementations is conceptually workable for the interactants (both of which are substantial assumptions), one could predict a number of consequences. First, TSI may develop into a worthwhile tool that assists interactants in overcoming the inadequacies of communicating from remote locations. By augmenting their representational, sensory, and situational characteristics, interactants of CVEs may be able to achieve levels of interaction that actually surpass face-to-face interaction.

On the other hand, people in fact may find the use of these transformations extremely unsettling. There is the potential for the difference between TSI and current CVE implementations to be as drastic as differences between email and the written letter. As this technology is developed, it is essential to examine people's responses to this new medium (i.e., Reeves & Nass, 1996). It is essential to examine these impor-

tant potential implications of TSI before the technology becomes widespread.

Along the same lines, the threat of TSI may be the very downfall of CVE interaction. In face-to-face interaction, there tends to be some degree of deception, for example people using facial expressions to mask their emotions. Clearly, this deception has the potential to be much greater with TSI. If interactants have no faith that their perceptual experience is genuine, they may have little reason to ever enter a CVE. A complete lack of trust in the truthfulness of gestures, one-to-one correspondence of avatars, and temporal presence of interactants has the potential to rob the CVE of one of its greatest strengths, namely interactivity, since the interactants may not know who, what, or when they are interacting with others. Similarly, given an expectation of TSI, interactants may be constantly suspicious during interactions; this lack of trust of fellow interactants may lead to unproductive collaborations.

A solution to this breakdown may require the development of TSI detectors for interactants, either based on computer algorithms that analyze nonverbal behaviors or based on actual humans that scrutinize the interaction. To examine the possibility of using human TSI detectors, we now discuss what we call the non-verbal Turing Test (NVTT).

In the popular reinterpretation of the Turing Test (Turing, 1950), a judge attempts to determine which of two players (one human, one machine) is a fellow human. In our NVTT pilot studies, experimental participants acting as judges enter a CVE with two virtual humans: one avatar whose head movements are veridical and playing back the movements of another human in real time (i.e., without TSI), and one avatar whose head movements are actually a transformation of the judge's own head movements (i.e., time-delayed and reduced motion range). The judge sees the head movements from a real person on one avatar and some sort of self-mimicked movements on the other. During the interaction, only head movements are permitted (i.e., no verbal communication allowed) and participants must devise ways to engage and test the two virtual humans through nonverbal means in order to ascertain which is human and which is a mimicker. Figure 4 illustrates.

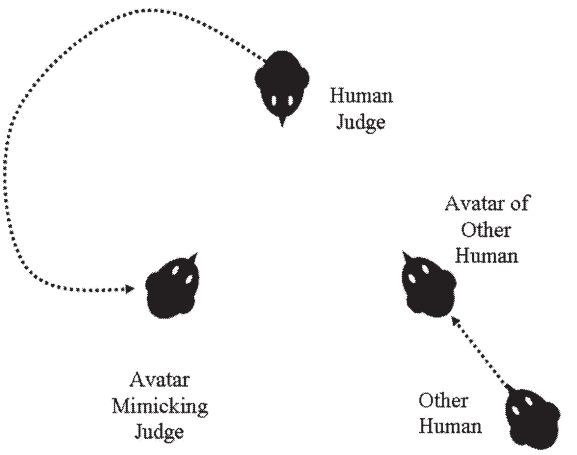

Figure 4. *A schematic of the NVTT. The human judge is forced to determine which of the two avatars exhibiting head movements is the real human and which is the computer-generated human mimicking the judge's own movements.*

In the current initial pilot study, we manipulated three independent variables: test trial length (either 16 or 32 seconds), mimic delay (i.e., the computer-agent mimics either 1, 2, 4, or 8 seconds after the judge's movements) and range of motion (high: pitch, yaw, and roll, or low: yaw only).

Participants in this study wore head-mounted displays while a render computer tracked their head orientation; tracking, rendering, and networking latencies were all low enough to impart a compelling sense of copresence (see Bailenson et al., 2002 for detailed descriptions of the hardware and software used). Participants were instructed to sit in a virtual room with two virtual people: a human agent (i.e., a representation whose movements are controlled by a real person in another room) and a computer agent (i.e., a computer program that is designed to mimic the user's movements in some way). Participants were instructed to interact with the two other virtual people using head movements in order to determine which one is the human agent. Participants were run in groups of two, with each one acting as the human agent for the other. Each participant sat at a virtual conference table with two virtual humans (similar to that shown in Figure 3). Each participant received a random order of 32 test trials (two instances of the 16 conditions resulting from the crossing of the three inde-

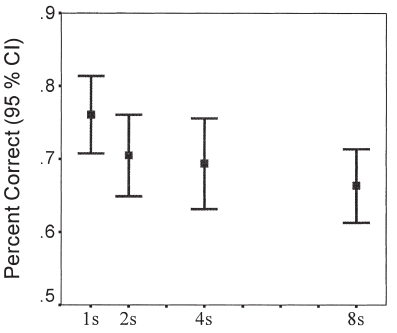

Figure 5. *Percent correct by mimic delay in seconds. This data excludes subjects at chance performance.*

pendent variables). Forty-one undergraduates participated in this study.

For the purposes of brevity, we focus on two results in particular. First, despite the fact that we explicitly told participants that the computer agent was directly mimicking them, they performed surprisingly poorly when attempting to identify the human avatar. The overall average score was only 66% correct (SD = 10%, chance = 50%, maximum score = 100%). Moreover, of the 41 participants in the study, more than one fourth was not reliably different from chance (i.e., less than 3 SEM from 50%: between 44% and 56%). Second, as Figure 5 demonstrates, participants' scores diminished inversely with the magnitude of the mimic delay, in that there was a linear trend in the logarithm of the delay variable, $F(1,32) = 8.85$, $p < .01$. When the delay was greater than 1 second, participants had more difficulty identifying a mimicker.

These data are particularly striking in that we had initially predicted that participants would be able to recognize their own head movements much more easily when aware that one avatar was designed to mimic them. In this study, we explicitly told participants about the use of TSI and they still had great difficulty in detecting it. The effects of implicit TSI (that is, not disclosed) can only have a higher impact. While this pilot study is extremely simple, and only scratches the surface of a paradigm that examines TSI, it is still noteworthy that participants did not detect the mimicker across the board.

We are currently exploring other factors underlying the discrimination of human nonverbal behavior from computer-generated behaviors. In future studies we will use NVTTs to study other nonverbal behaviors such as facial gestures, eye-head gestures (pointing indications by either system), hand gestures, and interpersonal distance. We have shown that in albeit simple scenarios it is possible to pass the NVTT for a percentage of our test population using TSI. We are confident that as this percentage grows in the near future, important scientific and sociological discoveries will surface along the way.

In conclusion, there are many reasons one might want to avoid TSI; these reasons range from Orwellian concerns to the fear of rendering CVEs (perhaps even the telephone) functionally useless. We are not advocates of TSI as a means to replace normal communication, nor are we staunch believers in avoiding TSI in order to preserve the natural order of communication and conversation. However, we do acknowledge the fact that, as CVEs become more prevalent, the strategic decoupling of representation from behavior is inevitable. For that reason alone, the notion of TSI warrants considerable attention.

Acknowledgments

The authors would like to thank Robin Gilmour and Christopher Rex for helpful suggestions. Furthermore, we thank Christopher Rex and Ryan Jaeger for assistance in collecting data. This research was sponsored in part by NSF Award SBE-9873432 and in part by NSF ITR Award IIS 0205740.

References

Argyle, M. (1988). *Bodily communication* (2nd ed.). London: Methuen.

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.

Bailenson, J. N., Beall, A. C., Blascovich, J., Raimmundo, R., & Weisbuch, M. (2001). Intelligent agents who wear your face: Users' reactions to the virtual self. *Lecture Notes in Artificial Intelligence, 2190,* 86–99.

Bailenson, J. N., Blascovich, J., Beall, A. C., & Guadagno,

R. E. (submitted). Self representations in immersive virtual environments.

Baumeister, R. F. (1998). The self. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *Handbook of social psychology* (4th ed.; pp. 680–740). New York: McGraw-Hill.

Beall, A. C., Bailenson, J. N., Loomis, J., Blascovich, J., & Rex, C. (2003). Non-zero-sum mutual gaze in immersive virtual environments. *Proceedings of HCI International 2003.*

Benford, S., Bowers, J., Fahlen, L., Greenhalgh, C., & Snowdon, D. (1995). User embodiment in collaborative virtual environments. *Proceedings of CHI'95* (pp. 242–249). ACM Press.

Biocca, F. (1997). The cyborg's dilemma: Progressive embodiment in virtual environments. *Journal of Computer-Mediated Communication [online], 3.* Retrieved from http://www.ascusc.org/jcmc/vol3/issue2/biocca2.html

Black, M., & Yacoob, Y. (1997). Recognizing facial expressions in image sequences using local parameterized models of image motion. *International Journal of Computer Vision, 25*(1): 23–48.

Blanz, V., & Vetter, T. (1999). A morphable model for the synthesis of 3D faces. *SIGGRAPH '99 Conference Proceedings,* 187–194.

Blascovich, J., Loomis, J., Beall, A. C., Swinth, K. R., Hoyt, C., & Bailenson, J. N. (2002). Immersive virtual environment technology as a methodological tool for social psychology. *Psychological Inquiry, 13,* 103–124.

Busey, T. A. (1988). Physical and psychological representations of faces: Evidence from morphing. *Psychological Science, 9,* 476–483.

Byrne, D. (1971). *The attraction paradigm.* New York: Academic Press.

Cassell, J. (2000). Nudge nudge wink wink: Elements of face-to-face conversation for embodied conversational agents. In J. Cassell et al. (Eds.), *Embodied conversational agents.* Cambridge, MA: MIT Press.

Chaiken, S. (1979). Communicator physical attractiveness and persuasion. *Journal of Personality and Social Psychology, 37,* 1387–1397.

Chartrand, T. L., & Bargh, J. (1999). The chameleon effect: The perception-behavior link and social interaction. *Journal of Personality & Social Psychology, 76*(6), 893–910.

Decarlo, D., Metaxas, D., & Stone, M. (1998). An anthropometric face model using variational techniques. *Proceedings of SIGGRAPH '98,* 67–74.

Depaulo, B. M., & Friedman, H. S. (1998). Nonverbal communication. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed., Vol. 2, pp. 3–40). Boston: McGraw-Hill.

Donato, G., Bartlett, M. S., Hager, J. C., Ekman, P., & Sejnowski, T. J. (1999). Classifying facial actions. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 21*(10), 974–989.

Durlach, N., & Slater, M. (2000). Presence in shared virtual environments and virtual togetherness. *Presence: Teleoperators and Virtual Environments, 9,* 214–217.

Ekman, P. (1978). Facial signs: Facts, fantasies, and possibilities. In T. Sebeok (Ed.), *Sight, sound and sense.* Bloomington, IN: Indiana University Press.

Fodor, J. A. (1983). *The modularity of mind: An essay on faculty psychology.* Cambridge, MA: MIT Press.

Fry, R., & Smith, G. F. (1975). The effects of feedback and eye contact on performance of a digit-encoding task. *Journal of Social Psychology, 96,* 145–146.

Gale, C., & Monk, A. F. (2002). A look is worth a thousand words: Full gaze awareness in video-mediated conversation. *Discourse Processes, 33.*

Garau, M., Slater, M., Vinayagamoorhty, V., Brogni, A., Steed, A., & Sasse, M. A. (2003). The impact of avatar realism and eye gaze control on perceived quality of communication in a shared immersive virtual environment. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.*

Gibson, W. (1984). *Neuromancer.* New York: Ace Books.

Hu, C., Ferris, R., & Turk, M. (2003). Active wavelet networks for face alignment. *Proceedings of the British Machine Vision Conference,* Norwich, UK.

Kendon, A. (1977). *Studies in the behavior of social interaction.* Bloomington, IN: Indiana University.

Kleinke, C. L. (1986). Gaze and eye contact: A research review. *Psychological Bulletin, 100,* 78–100.

Kraut, R. E., Fussell, S. R., Brennan, S. E., & Siegel, J. (2002). Understanding effects of proximity on collaboration: Implications for technologies to support remote collaborative work. In P. Hinds & S. Kiesler (Eds.), *Distributed work.* Cambridge, MA: MIT Press.

Lanier, J. (2001). Virtually there. *Scientific American,* April, 2001.

Leigh, J., DeFanti, T., Johnson, A., Brown, M., Sandin, D. (1997). Global telemersion: Better than being there. *Proceedings of ICAT '97.*

Loomis, J. M., Blascovich, J. J., & Beall, A. C. (1999). Immersive virtual environments as a basic research tool in psychology. *Behavior Research Methods, Instruments, and Computers, 31*(4), 557–564.

Mania, K., & Chalmers, A. (1998). *Proceedings of the Fourth International Conference on Virtual Systems and Multimedia* (pp. 177–182). Amsterdam: IOS Press-Ohmsha.

Milgram, S. (1992). *The individual in a social world: Essays and experiments* (2nd ed.). New York: McGraw-Hill.

Morgan, T; Kriz, R., Howard, T., Dias Neves, F., & Kelso, J. (2001). Extending the use of collaborative virtual environments for instruction to K–12 schools. *Insight 1*(1).

Normand, V., Babski, C., Benford, S., Bullock, A., Carion, S., Chrysanthou, Y., et al. (1999). The COVEN project: Exploring applicative, technical and usage dimensions of collaborative virtual environments. *Presence: Teleoperators and Virtual Environments, 8*(2), 218–236.

Patterson, M. L. (1982). An arousal model of interpersonal intimacy. *Psychological Review, 89,* 231–249.

Pylyshyn, Z. W. (1980). Computation and cognition: Issues in the foundations of cognitive science. *Behavioral & Brain Sciences, 3,* 111–169.

Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people and places.* New York: Cambridge University Press.

Rickel, J., & Johnson, W. L. (2000). Task-oriented collaboration with embodied agents in virtual worlds. In J. Cassell, J. Sullivan, S. Prevost, & E. Churchill (Eds.), *Embodied conversational agents.* Cambridge, MA: MIT Press.

Rutter, D. R. (1984). *Looking and seeing: The role of visual communication in social interaction.* Suffolk, UK: John Wiley & Sons.

Sannier, G., & Thalmann, M. N. (1998). A user friendly texture-fitting methodology for virtual humans. *Computer Graphics International '97.*

Schwartz, P., Bricker, L., Campbell, B., Furness, T., Inkpen, K., Matheson, L., et al. (1998). Virtual playground: Architectures for a shared virtual world. *Proceedings of the ACM Symposium on Virtual Reality Software and Technology 1998,* 43–50.

Sherwood, J. V. (1987). Facilitative effects of gaze upon learning. *Perceptual and Motor Skills, 64,* 1275–1278.

Simons, H. (1976). *Persuasion: Understanding, practice, and analysis.* Reading, MA: Heath.

Slater, M., Pertaub, D., & Steed, A. (1999). Public speaking in virtual reality: Facing an audience of avatars. *IEEE Computer Graphics and Applications, 19*(2), 6–9.

Slater, A., Sadagic, M., Usoh, R., & Schroeder R. (2000). Small group behavior in a virtual and real environment: A comparative study. *Presence: Teleoperators and Virtual Environments 9*(1), 37–51.

Stiefelhagen, R., Yang, J., & Waibel, A. (1997). Tracking eyes and monitoring eye gaze. In M. Turk & Y. Takabayashi (Eds.), *Proceedings of the Workshop on Perceptual User Interfaces.*

Turing, A. (1950). Computing machinery and intelligence. *Mind, 59* (236).

Turk, M., & Kolsch, M. (in press). Perceptual interfaces. In Medioni, G. & Kang, S. B. (Eds.), *Emerging topics in computer vision.* Boston: Prentice Hall.

Velichkovsky, B. M. (1995). Communicating attention: Gaze position transfer in cooperative problem solving. *Pragmatics and Cognition, 3*(2), 199–222.

Vertegaal, R. (1999). The GAZE groupware system: Mediating joint attention in multiparty communication and collaboration. *Proceedings of the CHI '99 Conference on Human Factors in Computing Systems: The CHI is the Limit,* 294–301.

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

Wallace, D. F. (1996). *Infinite jest.* Boston: Little Brown.

Williams, K., Cheung, K. T., & Choi, W. (2000). Cyberostracisms: Effects of being ignored over the internet. *Journal of Personality and Social Psychology, 79,* 748–762.

Yee, N. (2002). Befriending ogres and wood elves—Understanding relationship formation in MMORPGs. Retrieved from http://www.nickyee.com/hub/relationships/home.html

Zajonc, R. B. (1971). Brainwash: Familiarity breeds comfort. *Psychology Today, 3*(9): 60–64.

Zajonc, R. B., Murphy, S. T., & Inglehart, M. (1989). Feeling and facial efference: Implication of the vascular theory of emotion. *Psychological Review, 96,* 395–416.

Zhang, X., & Furnas, G. (2002). Social interactions in multiscale CVEs. *Proceedings of the ACM Conference on Collaborative Virtual Environments 2002 (CVE 2002).*