NONVERBAL COMMUNICATION IN VIRTUAL WORLDS:

UNDERSTANDING AND DESIGNING EXPRESSIVE CHARACTERS

Edited by

Joshua Tanenbaum, Magy Seif El-Nasr, & Michael Nixon

Nonverbal Communication in Virtual Worlds:
Understanding and Designing Expressive Characters

http://press.etc.cmu.edu/

Design Direction by Shirley Yee

ISBN: 978-1-304-81204-9
Library of Congress Control Number: 2014931252

15

AUTOMATION OF AVATAR BEHAVIOR

By Hannes Högni Vilhjálmsson

The graphical virtual worlds that sprang up in the mid-90s were truly inspiring. Finally the cyberspace that was promised to us by science-fiction was taking shape in front of our eyes. We could finally shed our real-life carcasses and enter a world of endless possibilities in shiny new digital bodies. The avatars we would embody would allow us to share this experience with others, just as if they were right there with us face-to-face.

1. CHAT VS. AVATARS

But something was not quite right. While we would now see all these avatars in the environment, they were suspiciously quiet and still. They hardly stirred, except for the occasional "slide" in or out of the room. It would have been easy to mistake the place for a wax museum, and yet, the place was literally bursting with excited conversations about this new frontier. The problem was that in order to actually notice the conversations, one had to open a chat window. The owners of the avatars essentially parked their bodies somewhere in the 3D environment and then started communicating with their fellow cybernauts using the traditional text interface (see Figure 15-1).

There were two modes of operation here. One where the user navigated their avatar around the virtual world to explore it, play in it or even construct it, and another where the user let go of the avatar and started communicating with other users through a separate modality. The avatar was hardly more than a graphical token that was associated with your location, rather than an embodiment that would deepen the social experience.

This split between the avatars and the conversations between their users even persisted when users were given buttons to play short animation sequences such as "waving" and "laughing". Of course users would play with these means of nonverbal communication, but once the typing of chat messages commenced, these buttons would hardly get used. One reason may be that using them would require letting go of the keyboard, therefore disrupting the flow of chatting. It required a certain level of dedication to explicitly control the avatar while at the same time composing meaningful contributions to an ongoing conversation. Therefore, it was common to see still avatars where people were in fact interacting. This sort of behavior has since been studied more closely, contributing further evidence that puppeteering of expressions and gestures simply takes too much effort to be frequently used (Seif El-Nasr et al. 2011; Shami, N.S. et al. 2010).

Figure 15-1: Worlds Chat was the first 3D chat environment on the Internet, launched in 1995 by Worlds Inc. It was a real breakthrough, but the avatars felt like chess pieces standing silently around the rooms, while conversations took place in the box below (picture from Bruce Damer)

2. IDLE ANIMATION

Recognizing that motionless avatars were lifeless and kind of creepy, it wasn't long until idle animation loops were added, which required no input from the users. These would essentially provide the illusion that the avatars were alive while users were doing things like chatting. We may take these idle movements for granted now, but it is a big deal to realize that users could not be trusted to keep their avatars alive by constantly animating them. It was not enough to give the users what we could call "button puppets" and expect a continuous performance that would bring them to life.

Idle loops really breathed life into the first avatar worlds, and the wax museum effect was gone. Instead, we had what seemed like crowds of people going about their business. Some of them would look around expectedly while others had a ponderous face, as if lost in thought. This was a first step, albeit a simple one, towards avatar automation. Still, things were not quite right. These animations were typically played cyclically or randomly, regardless of what was in fact going on. So, if one were to start a conversation with another user, through the chat window, the avatars would go on displaying the same nonverbal behavior as before the conversation started.

This could lead to mixed signals at best and downright inappropriate signals at worst. For example Alpha World, one of the more popular worlds (see Figure 15-2), used the action of looking at one's watch as an idle animation loop for some of its avatars. Few things are more socially demoralizing than watching your listeners repeatedly checking their watches while you are telling them something important. The knowledge that these were idle animations was not even enough to shrug off the strong nonverbal signal.

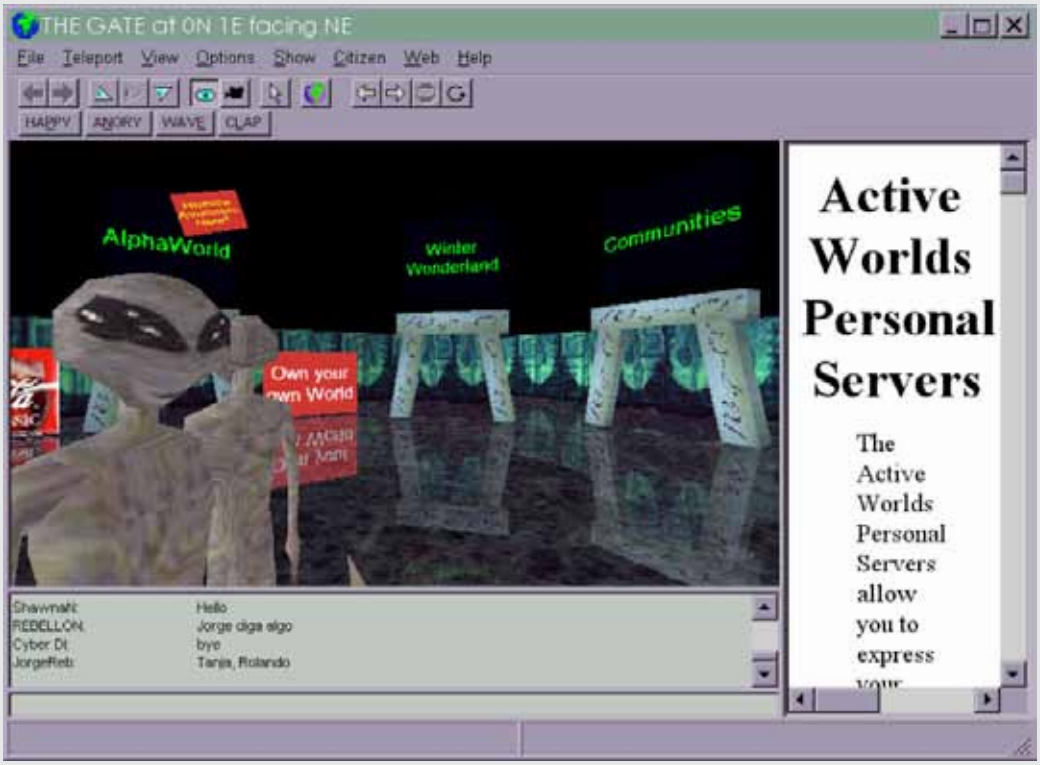

Figure 15-2: AlphaWorld (later ActiveWorlds) was one of the first 3D online environments, released in late 1995, to feature articulated button controlled avatars, and a set of idle animations. The "looking at watch" idle animation was a particularly memorable animation (picture from Bruce Damer)

3. LOCOMOTION ANIMATION

Bringing avatars to life through automating animation needs to be done carefully. Any nonverbal behavior exhibited by the avatar will be interpreted in the current context, both the social and environmental. Therefore, it makes sense to have the animation reflect the activity the avatar is currently engaged in. Perhaps the most readily accepted avatar animations that tie into the context are locomotion animations. Here the looping animations are chosen based on the user's intended mode of locomotion, for example walking, running or flying.

It is interesting to note that it is not the user that is choosing the animation to play during locomotion, but rather, the user may simply be pointing in the desired travel direction, while the avatar takes care of producing continuous body motion. This feels relatively natural to the user of the avatar, even though the automation may be adding various details to the behavior. For instance, when a user presses a button to go forward, the avatar may at first take short steps that gradually build up to a fast pace, slowing down again while crossing a river, and then transitioning smoothly over to a standing posture when the user lets go of the forward button.

Locomotion control for avatars, and corresponding continuous animation, has seen great advances in the last decade, mostly driven by the need to let game world avatars seamlessly traverse ever more complex environments made possible by more and more powerful graphics hardware. By using only one or two buttons players may have their avatars produce a wild dash through a crowded street, a daring jump from a pair of barrels up towards a window, where they barely manage to grab a hold of the ledge and pull themselves up into the safety of the room within.

4. ENGAGEMENT WITHOUT MICROMANAGEMENT

The fluid sequence of life-like motion and interaction with the environment that users see in their game world avatars arguably creates a stronger sense of being in the environment than if they were required to directly specify the animation to be played from moment to moment. Avatar micromanagement is something that games have tried to avoid when it distracts from core game play. Instead of micromanaging the avatar body, the level of control is placed at the level of player intent. For example, one might merely signal the desire to "jump" to produce all the corresponding preparation, execution and termination animation sequences.

What about social situations and communication? We were stuck with avatars that somehow were left out of conversations or displayed behavior that was in little relation to what was going on when people were attempting to communicate. Would it be possible to make the avatars provide social and communicative cues without requiring the users to execute specific animations? Similar to the locomotion scenario, might it be the case that the more interactivity with the social environment that the avatar can portray, the greater the sense of actually being there in the presence of others? It may seem counterintuitive that letting an avatar provide communicative signals on the behalf of the user would lead to a more integrated social experience. This became the subject of a research project called BodyChat at the MIT Media Lab in 1996 (Cassell and Vilhjalmsson, 1999) and later also became part of the so-called "Avatar Centric Communication" approach in There (Ventrella 2011).

5. SOCIAL ANIMATION WITH BODYCHAT

BodyChat was a graphical chat system that automated some of the communicative behavior one would expect to see in a normal face-to-face setting. The automation was not random, but instead it was triggered by a combination of two things: (1) The social goal of the user; and (2) what was going on in the environment. For example, the user could click on another user's avatar to indicate that their goal was to start a conversation with them. This would result in a prolonged glance and a smile towards the other user, and if the other user was available for a chat (as indicated by an availability toggle), the other user's avatar would reciprocate the glance and smile (see Figure 15-3). This would then trigger a salutation sequence that would conclude with another smile and a head nod when the users would bring their avatars together.

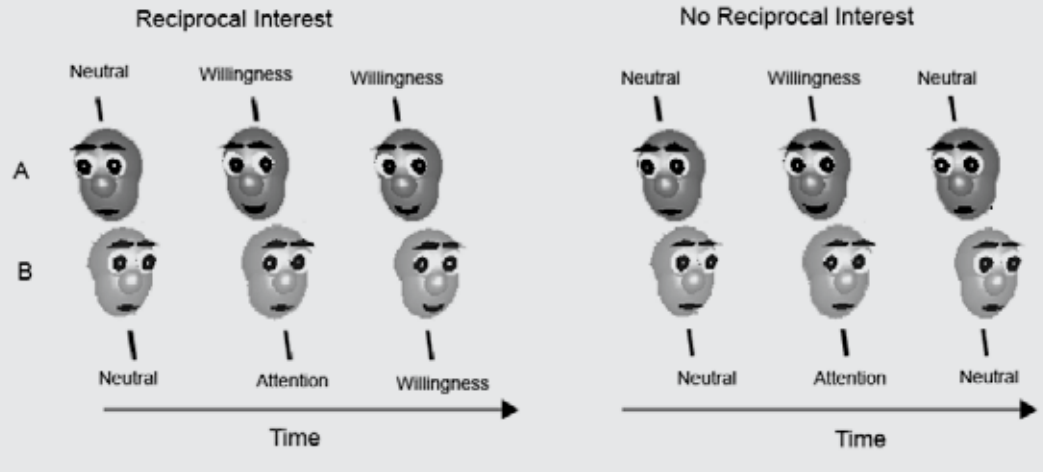

Figure 15-3: Avatars of two users, A and B, demonstrate a sequence of expressions that result from A wanting to start a conversation with B, first when B is available (left) and then when B is not available (right).

This whole process was modeled after documented human behavior, in particular studies on human greetings (Kendon, 1990). During chat, the avatars in BodyChat would animate eyes, eyebrows and head based on keywords and punctuation in the chat message, based for example on research on the relationship between facial displays and syntax during speech (Chovil, 1992). In Figure 15-4 we see from a first person perspective how another user's avatar is indicating that user's goal of talking to us. However, we have chosen to be unavailable (the toggle on the lower left), so our avatar will respond only with a quick glance, dismissing further interaction.

Figure 15-4: Another user's avatar displays willingness to chat with us in BodyChat. Notice that since we are not available for conversation (toggle on lower left), our avatar will dismiss the request with the appropriate expression.

6. MORE CONTROL WITH AUTOMATION?

To understand the effect of automated behavior on the social experience, a controlled user study was conducted. Four different versions of BodyChat were compared between subjects: *automatic* - Autonomy was responsible for all animated behavior (other than just moving the avatar around the environment); *hybrid* - In addition to the automation, all animations were also available to the user through a menu; *manual* - Automated animation turned off, but the animation menu was available; and *none* – Where avatars could not animate. For the sake of our discussion here, we will concentrate on the comparison between the automatic, hybrid and manual conditions. A complete discussion can be found in the original paper (Cassell and Vilhjalmsson 1999).

Subjects were tasked with entering a virtual environment with their avatar, meet other people and learn as much about them as they could. Unbeknownst to the subjects, all other avatars were under the control of a confederate that followed a strict behavior and conversation protocol. For example, the protocol dictated that the confederate produce similar nonverbal behavior as the subjects where manual control was possible. The confederate also provided scripted personal facts in response to questions from the subjects.

The results showed that the automation contributed to a better user experience. First some behavioral measures: conversations in the autonomous condition were significantly longer (a mean of 1111 seconds) than those in the manual (mean of 671 seconds) or hybrid (mean of 879 seconds) conditions. This can be taken as an index of the interest that subjects had in pursuing conversational interaction with people, when they were using the autonomous system. Moreover, subjects in the autonomous condition remembered more facts about the people they interacted with (a mean of 5.2) than did subjects in the manual condition (mean of 3.8) or the hybrid condition (mean of 4.5 facts). This can be taken as an index of how engaged subjects were in the conversation, perhaps because their attention was not divided between participating in the conversation and controlling the avatar.

The results of a subjective questionnaire indicated that the avatars in the purely automated condition were judged significantly more natural than both the manual and the hybrid versions. Furthermore, the system in the purely automated condition was judged to support significantly greater expressivity than the manual version, and somewhat greater than the hybrid version although not to a significant degree (see Figure 15-5). The measures of naturalness and expressivity are aggregates of several questions that probe these factors as detailed in (Cassell and Vilhjalmsson, 1999). It was clear that the automation was adding something to the experience.

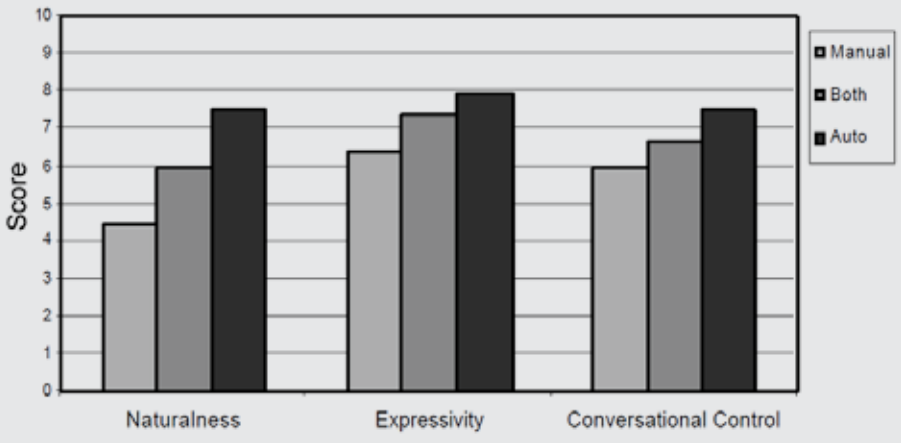

Figure 15-5: Subjective between-subject scores of avatar naturalness, expressivity and conversation control across manual, hybrid and fully automated animation conditions

The most controversial result however, regarded the issue of control of the situation. Subjects were asked (a) how much control did you have over the conversation? and (b) how much control do you think the other users had over the conversation? An ANOVA, and subsequent post-hoc t-tests, analyzing these two questions together revealed that users of the autonomous system considered the conversation significantly more under users' control than did the users of either the manual or the hybrid systems. This may sound counterintuitive because the nonverbal conversational behaviors were not under their direct control at all. However, one could argue that since the users were freed from the overhead of managing nonverbal behavior, they could concentrate on steering the course of the conversation itself. The fact that the users felt more in control of the conversation in the purely automatic version versus the hybrid version, where manual control is added as an additional feature, may indicate that the mere knowledge of a menu introduced some kind of a distraction.

Thinking too much about what you are doing with your face and hands while having a conversation with someone will make it harder for you to conduct yourself naturally. Try it the next time you have a face-to-face conversation. Decide ahead of time that you will not show any gesture or facial expression except a thumbs-up gesture and a smile, but you have to pick the right moment to use them. Pay attention to what happens to the rest of the conversation, especially with regards to timing.

We sometimes see odd synchrony between the spoken channel and the gestural channel in amateur actors or new politicians that have just been coached in the use of gesture. They are conscious of what they consider proper gesturing, but they may not always get the timing right, for example when making a point, they might throw a clenched fist into the air just a moment too late, essentially losing the momentum and breaking the flow. Good actors and good public speakers no longer need to micro manage their behavior, as they have already internalized their chosen speaking style. Their gestures and facial expressions arise from the same place as their spoken words – a place that represents their communicative intent. It is almost as if the realization of that intent simply grabs a hold of the body, and animates it in the manner that is most effective, without conscious effort.

7. FACE-TO-FACE COMMUNICATION

It doesn't take professional speakers to coordinate multiple communication modalities. When we engage in casual conversation with friends or when we interact with a cashier at the grocery store, we spontaneously produce an elaborate multimodal performance. Behaviors that range from posture and facial expressions to our tone of voice and words spoken are woven together into coherent and seamless communication. Are these behaviors serving any particular purpose? Do they somehow contribute to the success of our everyday conversations and transactions? To answer that question, we must first understand the process of face-to-face communication and what kind of activity underlies its success.

The study of face-to-face communication has been undertaken by many different fields, and typically fields that straddle boundaries between more traditional disciplines. These include discourse analysis, context analysis, sociolinguistics and social psychology. Scientific enquiry has revealed patterns of behavior, coordinated between all participants of social encounters. While the patterns are many and varied, the processes they represent generally fall into two main categories: *interactional* and *propositional*. In essence the former deals with establishing and maintaining a channel of communication, while the latter deals with providing effective communication across that channel (Cassell et al. 1999; Cassell et al. 2001). There is more going on in face-to-face interaction, but these processes provide fundamental building blocks.

On the interactional side, two important functions are *turn management* and *feedback*. Properly managing turns is necessary to ensure that everyone is not speaking at the same time, and can therefore be clearly

heard. Turns are requested, taken, held and given using various signals, often exchanged in parallel with speech over nonverbal channels such as gaze, intonation, and gesture (Duncan 1974; Goodwin 1981). Taking or requesting turns most often coincides with breaking eye contact (Argyle and Cook 1976) while raising hands into gesture space (Kendon 1990). A speaker gives the turn by looking at the listener, or whoever is meant to speak next, and resting the hands (Duncan 1974, Goffman 1983; Rosenfeld 1987).

Speakers often request feedback while speaking and expect at the very least some sign of attention from their listeners. Feedback requests typically involve looking at the listeners and raising eyebrows (Chovil 1991). To request a more involved feedback, this behavior can be supplemented with pointing the head towards the listener or conducting a series of low amplitude head nods ending with a head raise (Rosenfeld 1987). Listener response to feedback requests can take on a variety of forms. Brief assertion of attention can be given by the dropping of the eyelids and/ or a slight head nod towards the speaker. A stronger cue of attention may involve a slight leaning and a look towards the speaker along with a short verbal response or a laugh. A speaker's ability to formulate messages is critically dependent on these attentive cues, and therefore, even if only one person is speaking, everyone present is engaged in some kind of communicative behavior.

The propositional side deals with what we say and how we make sure those who are listening pick it up correctly. In addition to the content itself, there are three types of communicative functions that play an important role: *emphasis*, *reference* and *illustration*. Emphasis signals to listeners what the speaker considers to be the most important contribution of each utterance. It commonly involves raising or lowering of the eyebrows and sometimes vertical head movement as well (Argyle et al. 1973; Chovil 1991). A short formless beat with either hand, striking on the stressed syllable, is also common (McNeill 1992). Reference is most commonly carried out by a pointing hand. The reference can be made to the physical surroundings such as towards an object in the room, or to imaginary spots in space that for example represent something previously discussed (McNeill 1992). References through pointing are also made towards the other participants of a conversation when the speaker wishes to acknowledge their previous contributions or remind them of a prior conversation (Bavelas et al. 1995). Illustration is the spontaneous painting with the hands of some semantic feature of the current proposition. The particular features may lend themselves well to be portrayed by a visual modality, such as the configuration or size of objects, or the manner of motion (Kendon 1987; McNeill 1992).

From this overview we can see that the processes of interactional and propositional management are each carried out by a series of communicative functions. Each function in turn is realized through one or more nonverbal behaviors. The behaviors are not arbitrary, their function is often clear. It is good to keep in mind that human communication evolved in a face-to-face setting, and therefore the body has had an integral communicative role for most of human history.

8. ARE COMMUNICATIVE FUNCTIONS SUPPORTED IN MEDIATED ENVIRONMENTS?

It is a relatively recent development that humans can have conversations without their own bodies being present. So, what happens when the body is removed? The nonverbal behaviors are no longer available to support the crucial functions we just described. This sometimes leads to difficulties in communication. For example, think about a meeting where a couple of participants are only present through a voice conference system. What invariably happens is that the physically present participants tend to dominate the discussion while those on the phone have a hard time synchronizing their turns with the rest of the team. The body proves to be an advantage.

In the 1960s, AT&T introduced the Picturephone, a combination of a telephone and television that was meant to make remote conversations feel like they were truly face-to-face. The technology did not catch on, and even though available bandwidth and cheap hardware has made video mediated communication (VMC) very accessible, this mode of communication has still not become as commonplace as expected. A number of studies attempting to explain the slow adoption have shown that while VMC provides many important benefits over audio-only, it is also hampered by important limitations and in some cases may introduce negative artifacts that compromise the interaction.

Some of the benefits provided by VMC include the availability of nonverbal feedback and attitude cues, and access to a gestural modality for emphasis and elaboration (Isaacs and Tang 1994; Doherty-Sneddon, Anderson et al. 1997; Isaacs and Tang 1997). Seeing evidence of attention and attitude may be the reason why VMC has been shown to particularly benefit social tasks, involving negotiation or conflict resolution. In fact, groups that communicate with video tend to like each other better than those using audio only (Whittaker and O'Conaill, 1997). However, benefits for problem-solving tasks have been more evasive (Doherty-Sneddon, Anderson et al. 1997), and there one of the greatest limitations of classic VMC may play a role: The participants are not sharing the same space, they are each trapped in their own window.

Many important communicative functions break down when participants don't have a common frame of spatial reference, especially for group conversations (Isaacs and Tang 1994; Whittaker and O'Conaill 1997; Neale and McGee 1998; Nardi and Whittaker 2002). For example, turn-taking and judging the focus of attention becomes difficult when gaze direction is arbitrary. Pointing and manipulation of shared objects becomes troublesome and side conversations cannot be supported.

Variations on the classic video conferencing system have been developed to address some of the limitations. For example, "video surrogates" can be created by physically embedding two-way video units in a room for each remote participant. In some cases, such surrogates are even attached to robotic platforms. These systems can provide some level of gaze awareness and increased social presence (Inoue, Okada et al. 1997; Yankelovich et al. 2007; Adalgeirsson and Breazeal 2010), but often rely on static configurations or manual control, which misses the dynamic and spontaneous nature of fully embodied communication. Besides, specialized hardware solutions can become preventively expensive.

How about avatars in virtual worlds then? They certainly have one advantage over video: They all occupy a shared virtual place. But can they support the crucial communicative functions? As we have already mentioned, we cannot expect the users of the avatars to manually produce all the right nonverbal behaviors that normally are produced unconsciously face-to-face.

If we intend to capture the spontaneous movements of the users and transfer that movement over to the avatar (e.g. using computer vision or other real-time tracking technology), we run into a similar problem to that of the transmitted video: the user behind the screen doesn't occupy the same space as the user's avatar. Therefore, any spatially related behaviors may not read correctly on the avatar's body.

This could be addressed by fully immersing the user in the virtual environment through a head-mounted display, but the user would still be confined to the immediate physical space and facial expressions may be difficult to read when the eyes are completely covered. Furthermore, the user's performance may not be "large" enough for the virtual world, both literally in terms of movement ranges, and in terms of style befitting a world that might be larger than life.

9. AUTOMATING COMMUNICATIVE BEHAVIOR

That brings us back to avatar automation. Can communicative nonverbal behavior be automated in avatars to the extent that crucial communicative functions are being served? If the avatar could exhibit the needed spontaneous behavior and demonstrate full immersion in the social environment, we could finally talk about a virtual encounter that simulates meeting face-to-face.

The key to making this possible is to let go of the notion that the avatar is a mere puppet. Instead, we can think of it as an autonomous agent that serves as our primary interface to the virtual world and to its other inhabitants.

A couple of very early experiments with avatar automation include Comic Chat (Kurlander, Skelly et al. 1996) and Illustrated Conversation (Donath 1995). In Comic Chat the avatars were not animated models, but characters in a comic strip that got automatically generated, frame by frame, from an ongoing text chat (mainly based on keywords in the text) and a special emotion selection wheel. The real genius was that the generated comic strip gave a strong visual impression of a face-to-face group interaction. In Illustrated Conversation users were represented by their portraits on the screen, but each portrait was picked from a set of photos taken from different viewing angles. The system automatically picked photos that would result in gaze alignments that properly reflected who was attending to whom. While the avatars in these systems had very little articulation, they hinted at the power of machine augmented expression.

The BodyChat study showed that automation is useful, but how closely do we need to model actual human communicative behavior to get the benefits of automation? Two early studies demonstrated the importance of using principled approaches to automated animation, rather than resorting to the much cheaper randomized behavior.

The first study compared how two subjects interacted with each other in an audio only condition, random avatar gaze condition, algorithmic avatar gaze condition and through a video tunnel (Garau, Slater et al. 2001). The timings for the gaze algorithm were taken from research on face-to-face dyadic conversations and based on who was speaking and who was listening. A questionnaire assessing perceived naturalness of interaction, level of involvement, co-presence and attitude toward the other partner showed that the algorithmic gaze outperformed the random case consistently and significantly. This suggested that for avatars to meaningfully contribute to communication it is not sufficient for them to simply appear lively. In fact, the algorithmic gaze scored no differently than the video tunnel with regards to natural interaction and involvement, demonstrating at least subjectively that even crude and sparse (only gaze) but appropriate behavior in avatars brings the interaction closer to a face-to-face experience.

Another study showed the impact of principled automation on actual task outcomes in a collaborative scenario. The study compared a random gaze avatar with an algorithmic gaze avatar where a subject had to collaborate with double-blind actors on constructing syntactically correct permutations of sentence fragments (Vertegaal and Ding 2002). The subjects in the algorithmic gaze condition gave significantly more correct answers than in the random gaze condition, showing that appropriate communicative behavior helps us get things done.

If we now think about our avatar as an autonomous agent, what capability does it have to possess to mediate communication usefully? Minimally it requires two things to do its job: a *model* and a *context*. The model describes the important communication processes and the rules that underlie appropriate behavior, for example a face-to-face model would need an algorithmic representation of turn-management and a rule for mapping the function of taking the turn into a coordinated behavior of averted gaze and increased gestural activity. The context represents the information that is needed to correctly interpret what is going on in the social interaction, and therefore choose what parts of the model are relevant at any given time. For example,

a part of the context needs to keep track of whether your avatar is already engaged in conversation with someone, and what kind of a social situation you are in.

10. THE SPARK SYSTEM

BodyChat took the first steps towards turning avatars into agents with communication skills, but the focus there was on automating only a few interactional functions, so crucial propositional functions were left out. The full range of interactional and propositional function support was introduced in a system called Spark (see Figure 15-6), developed in 2003 (Vilhjalmsson 2004; Vilhjalmsson 2005). Spark consisted of virtual world clients and a central server that not only synchronized the clients, but also analyzed all chat messages using real-time natural language processing. The idea was to give the avatar agents plenty to build their communicative behaviors on.

The bulk of the communication *model* in Spark was represented by a series of language and event processing modules inside the server that each kept track of a critical communication process for every ongoing conversation. For example, there were separate modules for turn-taking, grounding (e.g. feedback), visual and textual reference (e.g. for pointing), emphasis, illustration (e.g. for elaborate hand gestures) and topic shifts.

The *context* got represented by three important structures: discourse model, domain knowledge and participation framework. The *discourse model* was a dynamic structure that reflected the state of the ongoing conversation. A key component was the discourse history, essentially a list of objects referenced so far in a conversation. The discourse model also included a list of objects visible in the immediate environment, since those are considered part of what is shared information. The *domain knowledge* was a static structure that described the ontology of a particular domain that related to a conversation. While helpful for resolving ambiguities and for suggesting richer semantics for gestures, conversations about topics not covered by the domain knowledge base would still generate behavior. Finally the *participation framework* kept track of the status of every person in a particular gathering, such as whether they are speaking, being addressed, general listeners or not attending to the conversation at all.

Figure 15-6: Three users discuss how to solve a route planning puzzle, while their automated avatars express relevant communicative behavior, both interactional and propositional. The old tree gets a pointing gesture (the arm has already retracted a bit in the screen shot) when it is first mentioned, since it is clearly visible to everyone.

The users of Spark communicated only via text messages, but everything that got written, was processed by the system. The processing progressed through a couple of major steps: functional annotation and behavioral annotation. The first step identified the communicative functions, interactional and propositional, that would need to accompany the delivery of the message. For example, if this was the first time the user "spoke", a turn taking function would need to be added, and if the message contained new information, that information would get associated with an emphasis function. In the second step, the functionally annotated message would be given to the user's avatar for delivery. The avatar would then transform the functional annotation into behavioral annotation, according to the model mapping rules (see Table 15-1). This would ensure that every communicative function would produce a corresponding supporting nonverbal behavior.

The avatar would finally deliver the message after a typical processing lag of about 2-5 seconds (either as scrolling text or synthesized audio) along with a fully synchronized embodied performance. The textual chat and the animated body would become completely integrated.

Table 15-1: An example of functional annotations automatically added to chat messages by the Spark server and the corresponding behavior generated by the avatar agents upon delivery

FUNCTION ANNOTATION	BEHAVIOR ANNOTATION
EMPHASIS_WORD	HEADNOD
EMPHASIS_PHRASE	EYEBROWS_RAISE
GROUND_REQUEST	GLANCE (ADDRESSEE)
TURN_GIVE	LOOK (ADDRESSEE)
TURN_TAKE	GLANCE_AWAY
TOPICSHIFT	POSTURESHIFT
REFERENCE_TEXTUAL	GLANCE (LAST REF. SPEAKER)
REFERENCE_VISUAL	GLANCE (OBJECT)
CONTRAST	GESTURE_CONTRAST
ILLUSTRATE	GESTURE_FEATURE

11. IMPACT ON COMMUNICATION

The goal with Spark was to turn the relatively narrow channel of text into a fully embodied face-to-face experience, where the benefits of having a body during communication would become clear. The only way to find out whether this goal was achieved was to carefully study the impact of the Spark avatars on a communicative situation. A study was devised where groups of three users would use Spark to solve a visual puzzle. The puzzle was in the form of a map and the group's goal was to decide on the shortest path between two locations, given information about various hazards on the way.

In half of the sessions, users would be represented by Spark avatars; in the other half, no avatars were visible. In both types of sessions, an interactive puzzle map and chat window were available. Since it had already been established that randomly animated avatars perform very poorly compared to principled animation (Colburn et al. 2001; Garau 2001; Vertegaal and Ding 2001), the avatars were compared to having no avatars, instead of "dumber" random avatars. Comparing this new breed of avatars to what people can achieve with text chat, would instead address a widely known and used communication medium without any distraction.

Two other conditions, crossed with the avatar versus no-avatar conditions, were the use of synthesized speech versus scrolling text. Apart from noting that people typically didn't like the synthesized voices, this part of the study won't be discussed further here. However, the fact that each subject could only be assigned to 2 instead of all 4 conditions in this 2x2 design (although balanced for order effects and only assigned to

adjacent cells) due to practical constraints, , made the analysis of the data more difficult and contributed to lower power than with standard within-subject experiments. Nevertheless, some clear results emerged.

The 14 subjects that tried both an avatar system and a system without avatars were first asked to compare the systems on a 9 point Likert scale from a high preference for no avatars to a high preference for avatars along 6 dimensions including which system was "more useful", "more fun", "more personal", "easier to use", "more efficient" and "allowed easier communication". One tailed t-tests showed that the preference for avatars was significant ($p<0.05$) for all but the "easier to use" question where no significant preference either way was found. These subjective results clearly indicated that people found the avatars compelling and helpful. Of particular interest is that the avatars delivered this greater subjective impact at no extra cost to usability.

To test whether the Spark avatars improved the overall process of conversation, compared to text-only messaging, 11 different measures of quality of conversation process were taken. Seven were objective behavioral measures from the chat logs, including the portion of utterances without explicit grounding (i.e. verbal verification of reception), portion of questions that got replies, portion of non-overlapping utterances and portion of on-task utterances. Four were subjective Likert scale questionnaire measures, including sense of ability to communicate and sense of control over conversation. All but one measure was found higher in the avatar condition and a t-test of the grand mean (across all 11 normalized measures) showed that indeed it was significantly higher ($p<0.02$) in the avatar condition than in the non-avatar condition, supporting the hypothesis that the Spark avatars were improving communication.

To test whether the Spark avatars would improve the outcome of the collaboration, compared to text-only messaging, 8 different measures of the quality of task outcome were taken. Two were objective measures, one being the quality of the map route that the subjects chose together and the other being the completion time (which ranged from 5 to 40 minutes). Six were subjective Likert scale questionnaire measures including, "How well did you think the group performed on the task?", "How strong do you think the group's consensus is about the final solution?" and "How much do you think you contributed to the final solution?" Again, all but one measure was higher in the avatar condition, and again, a t-test of the grand mean (across all 8 normalized measures) showed that it was significantly higher ($p<0.02$) in the avatar condition than in the non-avatar condition, supporting the hypothesis that the Spark avatars were improving the collaboration in various ways, although interestingly the resulting solution to the puzzle was not significantly different.

12. BALANCE OF CONTROL

How do you know how much of the avatar behavior in general should be left up to automation? The short answer is that it depends entirely on the context of use. But for each context there are several factors that need to be considered. Perhaps the most important thing to have in mind is that ultimately the users should feel in absolute control of the situation they are dealing with, which possibly may be achieved through greater automation at the behavioral level. For example, being able to tell your avatar that you wish to avoid certain people may free you from having to worry about accidentally inviting them to chat by making an unexpected eye contact.

There are other factors to consider as well. First of all, the avatar may have access to more resources than the user to base its behavior on. These resources include the virtual world in which the avatar resides. Beyond what is immediately visible, the avatar may even be able to use senses not available to the human user. In the example above, the avatar would be able to know whether the person you are trying to avoid is standing

behind you and therefore would not make the mistake of turning around to face them. Time is also a resource, and sometimes it is crucial that an avatar reacts quickly to a situation. A time delay from the user to the avatar could force control over the situation out of the user's hands.

Related to the resource of time, the avatar can maintain consistent continuous presence in the virtual world even if the link from the user is a discrete one. The discreteness may be the result of a physical link that can only support control commands in short bursts, or it could be that high cognitive load requires the user to multi-task. In either case, delegating control to the avatar may ensure that the remote operation is not interspersed with abrupt standstills.

Although an avatar is meant to be a representation of a user, it does not necessarily mean that the avatar can only mimic what the user would be able to do. In fact, the avatar is an opportunity to extend the capabilities of the user, even beyond the capability of being in a remote place. For example tele-operated robots, which in a sense are physical avatars, may be able to perform operations such as changing a valve at super-human speeds. The user, or operator, may therefore want to leave the execution up to the robot after making sure it has been maneuvered into the right spot. Similarly, in a social setting, an avatar could have certain nonverbal behavior coordination skills programmed that are beyond what the user would be able to orchestrate. A user could for example choose an avatar that knew how to produce the gestural language of a riveting speaker, leaving the exact control of that skill up to the avatar itself.

13. THE FUTURE OF AUTOMATED AVATARS

At the beginning of this chapter, the point was made that introducing automation to avatar animation was a way to bring them alive, but then we saw that making the automation purposeful and driven by a model of face-to-face communication could significantly augment the social experience. However, it is unlikely that the creators of virtual worlds would be able to spend considerable resources on developing the social skills for these new kinds of agent-based avatars. Therefore, what has to happen for this new generation to take hold is that such skills need to be packaged and made available as plug-ins or engines that can be dropped into any virtual world without too much effort.

Consider how the availability of 3rd party physics engines has made the automated animation of physical behavior ubiquitous in virtual worlds today. The next revolution will likely happen in the area of life-like social animation, based on the rules of face-to-face and social behavior. Linking an avatar up to a set of virtual sensors that perceive the social environment and a social behavior generator might become part of the virtual world development process. This packaging is the subject of an ongoing research project at CADIA (Pedica and Vilhjalmsson 2010).

Another shift we may experience, along with the greater emphasis on skillful and complex animation of avatar abilities, is a greater demand for customized behavior that sets the avatar apart from other avatars. Such customization would still have to comply with the general behavior model, but there is room for endless variety. We can expect this based on how much time and effort users spend on customizing their avatar appearance. If given the choice, users are likely to tweak how their avatar walks, greets and makes a point. There is an obvious issue with complexity, but future user interfaces can help manage that by making behavior customization intuitive with the right level of abstraction.

When you enter a virtual world in the future, you will be struck by the richness of movement exhibited by the avatars around you, but perhaps more importantly, you will notice that the movement reveals patterns of coordination, giving you an instant read on the social situation evolving around you. Just like in real life.

14. BIBLIOGRAPHY

Adalgeirsson, S. and Breazeal, C. (2010) "MeBot: a robotic platform for socially embodied presence". In Proceeding of the 5th ACM/IEEE international conference on Human-robot interaction (HRI '10). ACM, New York, NY, USA, 15-22.

Argyle, M. and Cook, M. Gaze and Mutual Gaze. Cambridge University Press, Cambridge, 1976.

Argyle, M., Ingham, R., Alkema, F. and McCallin, M. The Different Functions of Gaze. Semiotica., 1973

Bavelas, J.B., Chovil, N., Coates, L. and Roe, L. Gestures Specialized for Dialogue. Personality and Social Psychology, 21 (4). 394-405. 1995

Cassell, J., Bickmore, T., Billinghurst, M., Campbell, L., Chang, K., Vilhjalmsson, H. and Yan, H., Embodiment in Conversational Interfaces: Rea. in CHI, (Pittsburgh, PI, 1999), ACM, 520-527.

Cassell, J, Bickmore, T., Campbell, L., Vilhjalmsson, H., and Yan, H. "More Than Just a Pretty Face: Conversational Protocols and the Affordances of Embodiment", in Knowledge Based Systems 14: 55-64., 2001

Cassell, J. and H. Vilhjalmsson ,"Fully Embodied Conversational Avatars: Making Communicative Behaviors Autonomous." Autonomouse Agents and Multi-Agent Systems 2(1): 45-64. 1999

Chovil, N. Discourse-Oriented Facial Displays in Conversation. Research on Language and Social Interaction, 25 (1991/1992). 163-194.

Colburn, A. R., Cohen, M. F., Drucker, S. M. , LeeTiernan, S. , and Gupta, A. "Graphical Enhancements for Voice Only Conference Calls," Microsoft Corporation, Redmond, WA, Technical Report MSR-TR-2001-95, October 1, 2001 2001.

Doherty-Sneddon, G., A. H. Anderson, et al. (1997). "Face-to-Face and Video-Mediated Communication: A Comparison of Dialogue Structure and Task Performance." Journal of Experimental Psychology: Applied 3(2): 105-125.

Donath, J. (1995). "The Illustrated Conversation." Multimedia Tools and Applications 1(March): 79-88.

Duncan, S. On the structure of speaker-auditor interaction during speaking turns. Language in Society, 3. 161-180., 1974

Garau, M., Slater, M. , Bee, S. , and Sasse, A, "The Impact of Eye Gaze on Communication using Humanoid Avatars," presented at CHI 2001, Seattle, WA, 2001.

Goffman, E. Forms of Talk. University of Pennsylvania Publications, Philadelphia, PA, 1983.

Goodwin, C. Conversational Organization: Interaction between speakers and hearers. Academic Press, New York, 1981.

Inoue, T., k.-i. Okada, et al. (1997). "Integration of Face-to-Face and Video-Mediated Meetings: HERMES." Proceedings of ACM SIGGROUP'97: 385-394.

Isaacs, E. A. and J. C. Tang (1994). "What video can and cannot do for collaboration: a case study." Multimedia Systems 2: 63-73.

Isaacs, E. A. and J. C. Tang (1997). "Studying video-based collaboration in context: From small workgroups to large organization." Video-Mediated Communication. K. Finn, A. Sellen and S. Wilbur, Lawrence Erlbaum Associates, Inc.: 173-197.

Kendon, A. Conducting Interaction: Patterns of behavior in focused encounters. Cambridge University Press, New York, 1990.

Kendon, A. On Gesture: Its Complementary Relationship With Speech. in Siegman, A.W. and Feldstein, S. eds. Nonverbal Behavior and Communication, Lawrence Erlbaum Associates, Inc., Hillsdale, 1987, 65-97.

Kurlander, D., T. Skelly, et al. (1996). "Comic Chat." Proceedings of SIGGRAPH'96: 225-236.

McNeill, D. Hand and Mind. The University of Chicago Press, Chicago and London, 1992.

Nardi, B. A. and S. Whittaker (2002). "The Place of Face-to-Face Communication in Distributed Work." Distributed work: New ways of working across distance using technology. P. Hinds and S. Kiesler. Cambridge, MA, MIT Press.

Neale, D. C. and M. K. McGee (1998). "Making Media Spaces Useful: Video Suppor and Telepresence." HCIL-98-02. Hypermedia Technical Report. Human-Computer Interaction Laboratory, Virginia Tech

Pedica, C. and Vilhjalmsson, H. (2010). "Spontaneous Avatar Behavior for Human Territoriality", in Applied Artificial Intelligence, Volume 24 Issue 6, July 2010, 575-593 Taylor and Francis, Inc. Bristol, PA, USA

Rosenfeld, H.M. Conversational Control Functions of Nonverbal Behavior. in Siegman, A.W. and Feldstein, S. eds. Nonverbal Behavior and Communication, Lawrence Erlbaum Associates, Inc., Hillsdale, 1987, 563-601.

Seif El-Nasr, M., Isbister, K., Ventrella, J., Aghabeigi, B., Hash, C., Erfani, M., Morie, J.F. and Bishko L. (2011), "Body Buddies: Social Signaling through Puppeteering" in HCI 14, Vol. 6774, Springer, p. 279-288.

Shami, N.S., Cheng, L., Rohall, S., Sempere, A. and Patterson, J. (2010) "Avatars Meet Meetings: Design Issues in Integrating Avatars in Distributed Corporate Meetings", in Proceedings of the 16th ACM international conference on Supporting group work (GROUP '10). ACM, New York, NY, USA, 35-44.

Ventrella, J. (2011) Virtual Body Language. EyebrainBooks.

Vertegaal, R. and Ding, Y., "Explaining Effects of Eye Gaze on Mediated Group Conversations: Amount or Synchronization," presented at CSCW 2002, New Orleans, LA, 2002.

Vilhjalmsson, H. "Animating Conversation in Online Games", in M. Rauterberg (ed.), Entertainment Computing ICEC 2004, Lecture Notes in Computer Science 3166, pp. 139-150, Springer, 2004

Vilhjalmsson, H. "Augmenting Online Conversation through Automated Discourse Tagging", 6th annual minitrack on Persistent Conversation at the 38th Hawaii International Conference on System Sciences, January 3-6, 2005, Hilton Waikoloa Village, Big Island, Hawaii, IEEE, 2005

Whittaker, S. and B. O'Conaill (1997). "The Role of Vision in Face-to-Face and Mediated Communication." Video-Mediated Communication. K. Finn, A. Sellen and S. Wilbur, Lawrence Erlbaum Associates, Inc.: 23-49.

Yankelovich, N., Simpson, N., Kaplan. J. and Provino, J. (2007) "Porta-person: telepresence for the connected conference room" CHI Extended Abstracts 2007: 2789-2794