Creating interactive virtual humans: some assembly required

J. Gratch, J. Rickel, Elisabeth André, J. Cassell, E. Petajan, N. Badler

Angaben zur Veröffentlichung / Publication details:

Gratch, J., J. Rickel, Elisabeth André, J. Cassell, E. Petajan, and N. Badler. 2002. "Creating interactive virtual humans: some assembly required." *IEEE Intelligent Systems* 17 (4): 54–63. https://doi.org/10.1109/mis.2002.1024753.

Creating Interactive Virtual Humans: Some Assembly Required

Jonathan Gratch, *USC Institute for Creative Technologies*
Jeff Rickel, *USC Information Sciences Institute*
Elisabeth André, *University of Augsburg*
Justine Cassell, *MIT Media Lab*
Eric Petajan, *Face2Face Animation*
Norman Badler, *University of Pennsylvania*

Science fiction has long imagined a future populated with artificial humans—human-looking devices with human-like intelligence. Although Asimov's benevolent robots and the Terminator movies' terrible war machines are still a distant fantasy, researchers across a wide range of disciplines are beginning to work together toward a more modest goal—building virtual humans. These software entities look and act like people and can engage in conversation and collaborative tasks, but they live in simulated environments. With the untidy problems of sensing and acting in the physical world thus dispensed, the focus of virtual human research is on capturing the richness and dynamics of human behavior.

The potential applications of this technology are considerable. History students could visit ancient Greece and debate Aristotle. Patients with social phobias could rehearse threatening social situations in the safety of a virtual environment. Social psychologists could study theories of communication by systematically modifying a virtual human's verbal and nonverbal behavior. A variety of applications are already in progress, including education and training,[1] therapy,[2] marketing,[3,4] and entertainment.[5,6]

Building a virtual human is a multidisciplinary effort, joining traditional artificial intelligence problems with a range of issues from computer graphics to social science. Virtual humans must act and react in their simulated environment, drawing on the disciplines of automated reasoning and planning. To hold a conversation, they must exploit the full gamut of natural language processing research, from speech recognition and natural language understanding to natural language generation and speech synthesis. Providing human bodies that can be controlled in real time delves into computer graphics and animation. And because an agent looks like a human, people expect it to behave like one as well and will be disturbed by, or misinterpret, discrepancies from human norms. Thus, virtual human research must draw heavily on psychology and communication theory to appropriately convey nonverbal behavior, emotion, and personality.

This broad range of requirements poses a serious problem. Researchers working on particular aspects of virtual humans cannot explore their component in the context of a complete virtual human unless they can understand results across this array of disciplines and assemble the vast range of software tools (for example, speech recognizers, planners, and animation systems) required to construct one. Moreover, these tools were rarely designed to interoperate and, worse, were often designed with different purposes in mind. For example, most computer graphics research has focused on high fidelity offline image rendering that does not support the fine-grained interactive control that a virtual human must have over its body.

In the spring of 2002, about 30 international researchers from across disciplines convened at the University of Southern California to begin to bridge this gap in knowledge and tools (see www.ict.usc.edu/~vhumans). Our ultimate goal is a modular architecture and interface standards that will allow researchers in this area to reuse each other's work. This goal can only be achieved through a close multidisciplinary collaboration. Towards this end, the workshop gathered a collection of experts representing the range of required research areas, including

- Human figure animation
- Facial animation
- Perception
- Cognitive modeling
- Emotions and personality
- Natural language processing
- Speech recognition and synthesis
- Nonverbal communication
- Distributed simulation
- Computer games

Here we discuss some of the key issues that must be addressed in creating virtual humans. As a first step, we overview the issues and available tools in three key areas of virtual human research: face-to-face conversation, emotions and personality, and human figure animation.

Face-to-face conversation

Human face-to-face conversation involves both language and nonverbal behavior. The behaviors during conversation don't just function in parallel, but interdependently. The meaning of a word informs the interpretation of a gesture, and vice versa. The time scales of these behaviors, however, are different—a quick look at the other person to check that they are listening lasts for less time than it takes to pronounce a single word, while a hand gesture that indicates what the word "caulk" means might last longer than it takes to say, "I caulked all weekend."

Coordinating verbal and nonverbal conversational behaviors for virtual humans requires meeting several interrelated challenges. How speech, intonation, gaze, and head movements make meaning together, the patterns of their co-occurrence in conversation, and what kinds of goals are achieved by the different channels, are all equally important for understanding the construction of virtual humans. Speech and nonverbal behaviors do not always manifest the same information, but what they convey is virtually always compatible.[7] In many cases, different modalities serve to reinforce one another through redundancy of meaning. In other cases, semantic and pragmatic attributes of the message are distributed across the modalities.[8] The compatibility of meaning between gestures and speech recalls the interaction of words and graphics in multimodal presentations.[9] For patterns of co-occurrence, there is a tight synchrony among the different conversational modalities in humans. For example, people accentuate important words by speaking more forcefully, illustrating their point with a gesture, and turning their eyes toward the listener when coming to the end of a thought. Meanwhile listeners nod within a few hundred milliseconds of when the speaker's gaze shifts. This synchrony is essential to the meaning of conversation. When it is destroyed, as in low bandwidth videoconferencing, satisfaction and trust in the outcome of a conversation diminishes.[10]

Regarding the goals achieved by the different modalities, in natural conversation speakers tend to produce a gesture with respect to their *propositional* goals (to advance the conversation *content*), such as making the first two fingers look like legs walking when saying "it took 15 minutes to get here," and speakers tend to use eye movement with respect to *interactional* goals (to ease the conversation *process*), such as looking toward the other person when giving up the turn.[7] To realistically generate all the different verbal and nonverbal behaviors, then, computational architectures for virtual humans must control both the propositional and interactional structures. In addition, because some of these goals can be equally well met by one modality or the other, the architecture must deal at the level of goals or functions, and not at the level of modalities or behaviors. That is, giving up the turn is often achieved by looking at the listener. But, if the speaker's eyes are on the road, he or she can get a response by saying, "Don't you think?"

> With the untidy problems of sensing and acting in the physical world thus dispensed, the focus of virtual human research is on capturing the richness and dynamics of human behavior.

Constructing a virtual human that can effectively participate in face-to-face conversation requires a control architecture with the following features:[4]

- *Multimodal input and output.* Because humans in face-to-face conversation send and receive information through gesture, intonation, and gaze as well as speech, the architecture should also support receiving and transmitting this information.
- *Real-time feedback.* The system must let the speaker watch for feedback and turn requests, while the listener can send these at any time through various modalities. The architecture should be flexible enough to track these different threads of communication in ways appropriate to each thread. Different threads have different response-time requirements; some, such as feedback and interruption, occur on a sub-second time scale. The architecture should reflect this by allowing different processes to concentrate on activities at different time scales.
- *Understanding and synthesis of propositional and interactional information*. Dealing with propositional information—the communication content—requires building a model of the user's needs and knowledge. The architecture must include a static domain knowledge base and a dynamic discourse knowledge base. Presenting propositional information requires a planning module for presenting multi-sentence output and managing the order of presentation of interdependent facts. Understanding interactional information—about the processes of conversation—on the other hand, entails building a model of the current state of the conversation with respect to the conversational process (to determine who is the current speaker and listener, has the listener understood the speaker's contribution, and so on).
- *Conversational function model.* Functions, such as *initiating a conversation* or *giving up the floor,* can be achieved by a range of different behaviors, such as looking repeatedly at another person or bringing your hands down to your lap. Explicitly representing conversational functions, rather than behaviors, provides both modularity and a principled way to combine different modalities. Functional models influence the architecture because the core system modules operate exclusively on functions, while other system modules at the edges translate input behaviors into functions, and functions into output behaviors. This also produces a symmetric architecture because the same functions and modalities are present in both input and output.

To capture different time scales and the importance of co-occurrence, input to a virtual human must be incremental and time stamped. For example, incremental speech recognition lets the virtual human give feedback (such as a quick nod) right as the real human finishes a sentence, there-

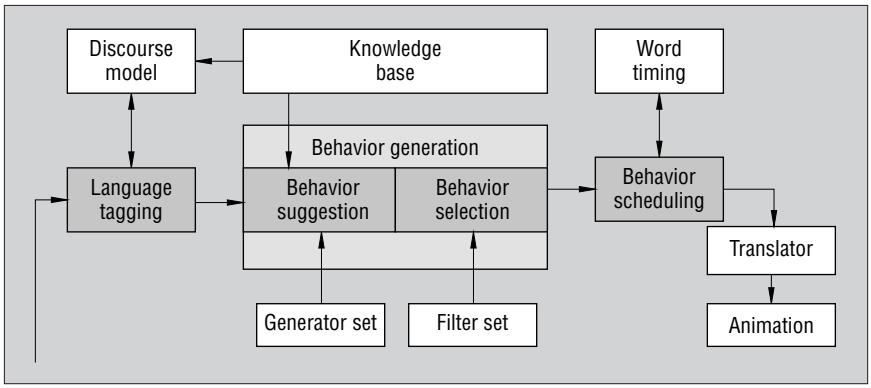

Figure 1. Behavior Expression Animation Toolkit text-to-nonverbal behavior module.

fore influencing the direction the human speaker takes. At the very least, the sytem should report a significant change in state right away, even if full information about the event has not yet been processed. This means that if speech recognition cannot be incremental, at least someone speaking or finished speaking should be relayed immediately, even in the absence of a fully recognized utterance. This lets the virtual human give up the turn when the real human claims it and signal reception after being addressed. When dealing with multiple modalities, fusing interpretations of the different input events is important to understand what behaviors are acting together to convey meaning.[12] For this, a synchronized clock across modalities is crucial so events such as exactly when an emphasis beat gesture occurs can be compared to speech, word by word. This requires, of course, that the speech recognizer supply word onset times.

Similarly, for the virtual human to produce a multimodal performance, the output channels also must be incremental and tightly synchronized. Incremental refers to two properties in particular: seamless transitions and interruptible behavior. When producing certain behaviors, such as gestures, the virtual human must reconfigure its limbs in a natural manner, usually requiring that some time be spent on interpolating from a previous posture to a new one. For the transition to be seamless, the virtual human must give the animation system advance notice of events such as gestures, so that it has time to bring the arms into place. Sometimes, however, behaviors must be abruptly interrupted, such as when the real human takes the turn before the virtual human has finished speaking. In that case, the current behavior schedule must be scrapped, the voice halted, and new attentive behaviors initiated—all with reasonable seamlessness.

Synchronicity between modalities is as important in the output as the input. The virtual human must align a graphical behavior with the uttering of particular words or a group of words. The temporal association between the words and behaviors might have been resolved as part of the behavior generation process, as is done in SPUD (Sentence Planning Using Description),[8] but it is essential that the speech synthesizer provide a mechanism for maintaining synchrony through the final production stage. There are two types of mechanisms, event based or time based. A text-to-speech engine can usually be programmed to send events on phoneme and word boundaries. Although this is geared towards supporting lip synch, other behaviors can be executed as well. However, this does not allow any time for behavior preparation. Preferably, the TTS engine can provide exact start-times for each word prior to playing back the voice, as Festival does.[13] This way, we can schedule the behaviors, and thus the transitions between behaviors, beforehand, and then play them back along with the voice for a perfectly seamless performance.

On the output side, one tool that provides such tight synchronicity is the Behavior Expression Animation Toolkit system.[11] Figure 1 shows BEAT's architecture. BEAT has the advantage of automatically annotating text with hand gestures, eye gaze, eyebrow movement, and intonation. The annotation is carried out in XML, through interaction with an embedded word ontology module, which creates a set of hypernyms that broadens a knowledge base search of the domain being discussed. The annotation is then passed to a set of behavior generation rules. Output is scheduled so that tight synchronization is maintained among modalities.

Emotions and personality

People infuse their verbal and nonverbal behavior with emotion and personality, and modeling such behavior is essential for building believable virtual humans. Consequently, researchers have developed computational models for a wide range of applications. Computational approaches might be roughly divided into communication-driven and simulation-based approaches.

In communication-driven approaches, a virtual human chooses its emotional expression on the basis of its desired impact on the user. Catherine Pelachaud and her colleagues use facial expressions to convey affect in combination with other communicative functions.[14] For example, making a request with a sorrowful face can evoke pity and motivate an affirmative response from the listener. An interesting feature of their approach is that the agent deliberately plans whether or not to convey a certain emotion. Tutoring applications usually also follow a communication-driven approach, intentionally expressing emotions with the goal of motivating the students and thus increasing the learning effect. The Cosmo system, where the agent's pedagogical goals drive the selection and sequencing of emotive behaviors, is one example.[15] For instance, a congratulatory act triggers a motivational goal to express admiration that is conveyed with applause. To convey appropriate emotive behaviors, agents such as Cosmo need to appraise events not only from their own perspective but also from the perspective of others.

The second category of approaches aims at a simulation of "true" emotion (as opposed to deliberately conveyed emotion). These approaches build on *appraisal* theories of emotion, the most prominent being Andrew Ortony, Gerald Clore, and Allan Collins' cognitive appraisal theory—commonly referred to as the OCC model.[16] This theory views emotions as arising from a valenced reaction to events and objects in the light of agent goals, standards, and attitudes. For example, an agent watching a game-winning move should respond differently depending on which team is preferred.[3]

Recent work by Stacy Marsella and Jonathan Gratch integrates the OCC model with coping theories that explain how people cope with strong emotions.[17] For example, their agents can engage in either *problem-focused* coping strategies, selecting and executing actions in the world that could improve the agent's emotional state, or *emotion-focused* coping strategies, improving emotional state by altering the agent's mental state (for example, dealing with guilt by blaming someone else). Further simulation approaches are based on the observation that an agent should be able to dynamically adapt its emotions through its own experience, using learning mechanisms.[6,18]

Appraisal theories focus on the relationship between an agent's world assessment and the resulting emotions. Nevertheless, they are rather vague about the assessment process. For instance, they do not explain how to determine whether a certain event is desirable. A promising line of research is integrating appraisal theories with AI-based planning approaches,[19] which might lead to a concretization of such theories. First, emotions can arise in response to a deliberative planning process (when relevant risks are noticed, progress assessed, and success detected). For example, several approaches derive an emotion's intensity from the importance of a goal and its probability of achievement.[20,21] Second, emotions can influence decision-making by allocating cognitive resources to specific goals or threats. Plan-based approaches support the implementation of decision and action selection mechanisms that are guided by an agent's emotional state. For example, the Inhabited Market Place application treats emotions as filters to constrain the decision process when selecting and instantiating dialogue operators.[3]

In addition to generating affective states, we must also express them in a manner easily interpretable to the user. Effective means of conveying emotions include body gestures, acoustic realization, and facial expressions (see Gary Collier's work for an overview of studies on emotive expressions[22]). Several researchers use Bayesian networks to model the relationship between emotion and its behavioral expression. Bayesian networks let us deal explicitly with uncertainty, which is a great advantage when modeling the connections between emotions and the resulting behaviors. Gene Ball and Jack Breese presented an example of such an approach. They constructed a Bayesian network that estimates the likelihood of specific body postures and gestures for individuals with different personality types and emotions.[23] For instance, a negative emotion increases the probability that an agent will say "Oh, you again," as opposed to "Nice to see you!"

Figure 2. Pelachaud and colleagues use a MPEG-4 compatible facial animation system to investigate how to resolve conflicts that arise when different communication functions need to be shown on different channels of the face.

Recent work by Catherine Pelachaud and colleagues employs Bayesian networks to resolve conflicts that occur when different communicative functions need to be shown on different channels of the face, such as eyebrows, mouth shape, gaze direction, head direction, and head movements (see Figure 2).[14] In this case, the Bayesian network estimates the likelihood that a face movement overrides another. Bayesian networks also offer a possibility to model how emotions vary over time. Even though neither Ball and Breese nor Pelachaud and colleagues took advantage of this feature, the extension of the two approaches to dynamic Bayesian networks seems obvious.

While significant progress has been made on the visualization of emotive behaviors, automated speech synthesis still has a long way to go. The most natural-sounding approaches rely on a large inventory of human speech units (for example, combinations of phonemes) that are subsequently selected and combined based on the sentence to be synthesized. These approaches do not, yet, provide much ability to convey emotion through speech (for example, by varying prosody or intensity). Marc Schröder provides an overview of speech manipulations that have been successfully employed to express several basic emotions.[25] While the interest in affective speech synthesis is increasing, hardly any work has been done on conveying emotion through sentence structure or word choice. An exception includes Eduard Hovy's pioneering work on natural language generation that addresses not only the goal of information delivery, but also pragmatic aspects, such as the speaker's emotions.[26] Marilyn Walker and colleagues present a first approach to integrating acoustic parameters with other linguistic phenomena, such as sentence structure and wording.[27]

Obviously, there is a close relationship between emotion and personality. Dave Moffat differentiates between personality and emotion using the two dimensions duration and focus.[28] Whereas personality remains stable over a long period of time, emotions are short-lived. Moreover, while emotions focus on particular events or objects, factors determining personality are more diffuse and indirect. Because of this obvious relationship, several projects aim to develop an integrated model of emotion and personality. As an example, Ball and Breese model dependencies between emotions and personality in a Bayesian network.[23] To enhance the believability of animated agents beyond reasoning about emotion and personality, Helmut Prendinger and colleagues model the relationship between an agent's social role and the associated constraints on emotion expression, for example, by suppressing negative emotion when interacting with higher-status individuals.[29]

Another line of research aims at providing an enabling technology to support affective interactions. This includes both the definition of standardized languages for specifying emotive behaviors, such as the Affective Presentation Markup Language[14] or the Emotion Markup Language (www.

Figure 3. PeopleShop and DI-Guy are used to create scenarios for ground combat training. This scenario was used at Ft. Benning to enhance situation awareness in experiments to train US Army officers for urban combat. Image courtesy of Boston Dynamics.

vhml.org), as well as the implementation of toolkits for affective computing combining a set of components addressing affective knowledge acquisition, representation, reasoning, planning, communication, and expression.[30]

Human figure animation

By engaging in face-to-face conversation, conveying emotion and personality, and otherwise interacting with the synthetic environment, virtual humans impose fairly severe behavioral requirements on the underlying animation system that must render their physical bodies. Most production work involves animator effort to design or script movements or direct performer motion capture. Replaying movements in real time is not the issue; rather, it is creating novel, contextually sensitive movements in real time that matters. Interactive and conversational agents, for example, will not enjoy the luxury of relying on animators to create human time-frame responses. Animation techniques must span a variety of body systems: locomotion, manual gestures, hand movements, body pose, faces, eyes, speech, and other physiological necessities such as breathing, blinking, and perspiring. Research in human figure animation has addressed all of these modalities, but historically the work focuses either on the animation of complete body movements or on animation of the face.

Body animation methods

In body animation, there are two basic ways to gain the required interactivity: use motion capture and additional techniques to rapidly modify or re-target movements to immediate needs,[31] or write procedural code that allows program control over important movement parameters.[32] The difficulty with the motion capture approach is maintaining environmental constraints such as solid foot contacts and proper reach, grasp, and observation interactions with the agent's own body parts and other objects. To alleviate these problems, procedural approaches parameterize target locations, motion qualities, and other movement constraints to form a plausible movement directly. Procedural approaches consist of kinematic and dynamics techniques. Each has its preferred domain of applicability; kinematics is generally better for goal-directed activities, and slower (controlled) actions and dynamics is more natural for movements directed by application of forces, impacts, or high-speed behaviors.[33] The wide range of human movement demands that both approaches have real-time implementations that can be procedurally selected as required.

Animating a human body form requires more than just controlling skeletal rotation angles. People are neither skeletons nor robots, and considerable human qualities arise from intelligent movement strategies, soft deformable surfaces, and clothing. Movement strategies include reach or constrained contacts, often achieved with goal-directed inverse kinematics.[34] Complex workplaces, however, entail more complex planning to avoid collisions, find free paths, and optimize strength availability. The suppleness of human skin and the underlying tissue biomechanics lead to shape changes caused by internal muscle actions as well as external contact with the environment. Modeling and animating the local, muscle-based, deformation of body surfaces in real time is possible through shape morphing techniques,[35,36] but providing appropriate shape changes in response to external forces is a challenging problem. "Skin-tight" texture mapped clothing is prevalent in computer game characters, but animating draped or flowing garments requires dynamic simulation, fast collision detection, and appropriate collision response.[37,38]

Accordingly, animation systems build procedural models of these various behaviors and execute them on human models. The diversity of body movements involved has led to building more consistent agents: procedural animations that affect and control multiple body communication channels in coordinated ways.[11,24,39,40] The particular challenge here is constructing computer graphics human models that balance sufficient articulation, detail, and motion generators to effect both gross and subtle movements with realism, real-time responsiveness, and visual acceptability. And if that isn't enough, consider the additional difficulty of modeling a specific real individual. Computer graphics still lacks effective techniques to transfer even captured motion into features that characterize a specific person's mannerisms and behaviors, though machine-learning approaches could prove promising.[41]

Implementing an animated human body is complicated by a relative paucity of generally available tools. Body models tend to be proprietary (for example, Extempo.com, Ananova.com), optimized for real time and thus limited in body structure and features (for example, DI-Guy, BDI.com, illustrated in Figure 3), or constructions for particular

animations built with standard animator tools such as Poser, Maya, or 3DSMax. The best attempt to design a transportable, standard avatar is the Web3D Consortium's H-Anim effort (www.h-anim.org). With well-defined body structure and feature sites, the H-Anim specification has engendered model sharing and testing not possible with proprietary approaches. The present liability is the lack of an application programming interface in the VRML language binding of H-Anim. A general API for human models is a highly desirable next step, the benefits of which have been demonstrated by Norman Badler's research group's use of the software API in Jack (www.ugs.com/products/efactory/jack), which allows feature access and provides plug-in extensions for new real-time behaviors.

Face animation methods

A computer-animated human face can evoke a wide range of emotions in real people because faces are central to human reality. Unfortunately, modeling and rendering artifacts can easily produce a negative response in the viewer. The great complexity and psychological depth of the human response to faces causes difficulty in predicting the response to a given animated face model. The partial or minimalist rendering of a face can be pleasing as long as it maintains quality and accuracy in certain key dimensions. The ultimate goal is to analyze and synthesize humans with enough fidelity and control to pass the Turing test, create any kind of virtual being, and enable total control over its virtual appearance. Eventually, surviving technologies will be combined to increase accuracy and efficiency of the capture, linguistic, and rendering systems. Currently the approaches to animating the face are disjoint and driven by production costs and imperfect technology. Each method presents a distinct "look and feel," as well as advantages and disadvantages.

Facial animation methods fall into three major categories. The first and earliest method is to manually generate keyframes and then automatically interpolate frames between the keyframes (or use less skilled animators). This approach is used in traditional cell animation and in 3D animated feature films. Keyframe and morph target animation provides complete artistic control but can be time consuming to perfect.

The second method is to synthesize facial movements from text or acoustic speech. A TTS algorithm, or an acoustic speech recognizer, provides a translation to phonemes, which are then mapped to visemes (visual phonemes). The visemes drive a speech articulation model that animates the face. The convincing synthesis of a face from text has yet to be accomplished. The state of the art provides understandable acoustic and visual speech and facial expressions.[42,43]

The third and most recent method for animating a face model is to measure human facial movements directly and then apply the motion data to the face model. The model can capture facial motions using one or more cameras and can incorporate face markers, structured light, laser range finders, and other face measurement modes. Each facial motion capture approach has limitations that might require postprocessing to overcome. The ideal motion-capture data representation supports sufficient detail without sacrificing editability (for example, MPEG-4 Facial Animation Parameters). The choice of modeling and rendering technologies ranges from 2D line drawings to physics-based 3D models with muscles, skin, and bone.[44,45] Of course, textured polygons (nonuniform rational b-splines and subdivision surfaces) are by far the most common. A variety of surface deformation schemes exist that attempt to simulate the natural deformations of the human face while driven by external parameters.[46,47]

MPEG-4, which was designed for high-quality visual communication at low bitrates coupled with low-cost graphics rendering systems, offers one existing standard for human figure animation. It contains a comprehensive set of tools for representing and compressing content objects and the animation of those objects, and it treats virtual humans (faces and bodies) as a special type of object. The MPEG-4 Face and Body Animation standard provides anatomically specific locations and animation parameters. It defines Face Definition Parameter feature points and locates them on the face (see Figure 4). Some of these points only serve to help define the face's shape. The rest of them are displaced by Facial Animation Parameters, which specify feature point displacements from the neutral face position. Some FAPs are descriptors for visemes and emotional expressions. Most remaining FAPs are normalized to be proportional to neutral face mouth width, mouth-nose distance, eye separation, iris diameter, or eye-nose distance. Although MPEG-4 has defined a limited set of visemes and facial expressions, designers can specify two visemes or two expressions with a blend factor between the visemes and an intensity value for each expression. The normalization of the FAPs gives the face model designer freedom to create characters with any facial proportions, regardless of the source of the FAPs. They can embed MPEG-4 compliant face models into decoders, store them on CD-ROM, download them as an executable from a Web site, or build them into a Web browser.

Integration challenges

Integrating all the various elements described here into a virtual human is a daunting task. It is difficult for any single research group to do it alone. Reusable tools and modular architectures would be an enormous benefit to virtual human researchers, letting them leverage each other's work. Indeed, some research groups have begun to share tools, and several standards have recently emerged that will further encourage sharing. However, we must confront several difficult issues before we can readily plug-and-play different modules to control a virtual human's behavior. Two key issues discussed at the workshop were consistency and timing of behavior.

Consistency

When combining a variety of behavioral components, one problem is maintaining consistency between the agent's internal state (for example, goals, plans, and emotions) and the various channels of outward behavior (for example, speech and body movements). When real people present multiple behavior channels, we interpret them for consistency, honesty, and sincerity, and for social roles, relationships, power, and intention. When these channels conflict, the agent might simply look clumsy or awkward, but it could appear insincere, confused, conflicted, emotionally detached, repetitious, or simply fake. To an actor or an expert animator, this is obvious. Bad actors might fail to control gestures or facial expressions to portray the demeanor of their persona in a given situation. The actor might not have internalized the character's goals and motivations enough to use the body's own machinery to manifest these

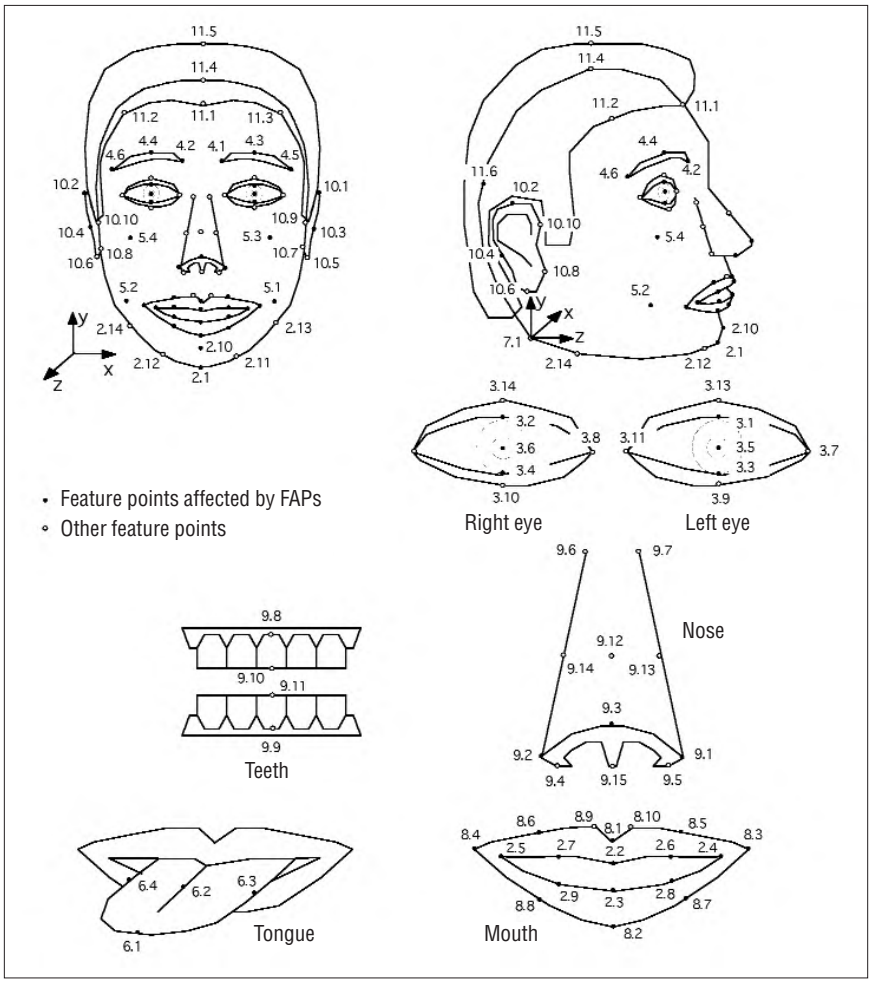

Figure 4. The set of MPEG-4 Face Definition Parameter (FDP) feature points.

inner drives as appropriate behaviors. A skilled animator (and actor) knows that all aspects of a character must be consistent with its desired mental state because we can control only voice, body shape, and movement for the final product. We cannot open a dialog with a pre-animated character to further probe its mind or its psychological state. With a real-time embodied agent, however, we might indeed have such an opportunity.

One approach to remedying this problem is to explicitly coordinate the agent's internal state with the expression of body movements in all possible channels. For example, Norman Badler's research group has been building a system, EMOTE, to parameterize and modulate action performance.[24] It is based on Laban Movement Analysis, a human movement observation System. EMOTE is not an action selector per se; it is used to modify the execution of a given behavior and thus change its movement qualities or character. EMOTE's power arises from the relatively small number of parameters that control or affect a much larger set, and from new extensions to the original definitions that include non-articulated face movements. The same set of parameters control many aspects of manifest behavior across the agent's body and therefore permit experimentation with similar or dissimilar settings. The hypothesis is that behaviors manifest in separate channels with similar EMOTE parameters will appear consistent to some internal state of the agent; conversely, dissimilar EMOTE parameters will convey various negative impressions of the character's internal consistency. Most computer-animated agents provide direct evidence for the latter view:

- Arm gestures without facial expressions look odd.
- Facial expressions with neutral gestures look artificial.
- Arm gestures without torso involvement look insincere.
- Attempts at emotions in gait variations look funny without concomitant body and facial affect.
- Otherwise carefully timed gestures and speech fail to register with gesture performance and facial expressions.
- Repetitious actions become irritating because they appear unconcerned about our changing (more negative) feelings about them.

Timing

In working together toward a unifying architecture, timing emerged as a central concern at the workshop. A virtual human's behavior must unfold over time, subject to a variety of temporal constraints. For example, speech-related gestures must closely follow the voice cadence. It became obvious during the workshop that previous work focused on a specific aspect of behavior (for example, speech, reactivity, or emotion), leading to architectures that are tuned to a subset of timing constraints and cannot straightforwardly incorporate others. During the final day of the workshop, we struggled with possible architectures that might address this limitation.

For example, BEAT schedules speech-related body movements using a pipelined architecture: a text-to-speech system generates a fixed timeline to which a subsequent gesture scheduler must conform. Essentially, behavior is a slave to the timing constraints of the speech synthesis tool. In contrast, systems that try to physically convey a sense of emotion or personality often work by altering the time course of gestures. For example, EMOTE works later in the pipeline, taking a previously generated sequence of gestures and shortening or drawing them out for emotional effect. Essentially, behavior is a slave to the constraints of emotional dynamics. Finally, some systems have focused on making the character highly reactive and embedded in the synthetic environment. For example, Mr. Bubb of Zoesis Studios (see Figure 5) is tightly responsive to unpredictable and continuous changes in the environment (such as mouse movements or bouncing balls). In such systems, behavior is a slave to environmental dynamics. Clearly, if these various capabilities are to be combined, we must reconcile these different approaches.

One outcome of the workshop was a

number of promising proposals for reconciling these competing constraints. At the very least, much more information must be shared between components in the pipeline. For example, if BEAT had more access to timing constraints generated by EMOTE, it could do a better job of upfront scheduling. Another possibility would be to specify all of the constraints explicitly and devise an animation system flexible enough to handle them all, an approach the motion graph technique suggests.[48] Norman Badler suggests an interesting pipeline architecture that consists of "fat" pipes with weak uplinks. Modules would send down considerably more information (and possibly multiple options) and could poll downstream modules for relevant information (for example, how long would it take to look at the ball, given its current location). Exploring these and other alternatives is an important open problem in virtual human research.

Figure 5. Mr. Bubb is an interactive character developed by Zoesis Studios that reacts continously to the user's social interactions during cooperative play experience. Image courtesy of Zoesis Studios.

The future of androids remains to be seen, but realistic interactive virtual humans will almost certainly populate our near future, guiding us toward opportunities to learn, enjoy, and consume. The move toward sharable tools and modular architectures will certainly hasten this progress, and, although significant challenges remain, work is progressing on multiple fronts. The emergence of animation standards such as MPEG-4 and H-Anim has already facilitated the modular separation of animation from behavioral controllers and sparked the development of higher-level extensions such as the Affective Presentation Markup Language. Researchers are already sharing behavioral models such as BEAT and EMOTE. We have outlined only a subset of the many issues that arise, ignoring many of the more classical AI issues such as perception, planning, and learning. Nonetheless, we have highlighted the considerable recent progress towards interactive virtual humans and some of the key challenges that remain. Assembling a new virtual human is still a daunting task, but the building blocks are getting bigger and better every day. ■

Acknowledgments

We are indebted to Bill Swartout for developing the initial proposal to run a Virtual Human workshop, to Jim Blake for helping us arrange funding, and to Julia Kim and Lynda Strand for handling all the logistics and support, without which the workshop would have been impossible. We gratefully acknowledge the Department of the Army for their financial support under contract number DAAD 19-99-D-0046. We also acknowledge the support of the US Air Force F33615-99-D-6001 (D0 #8), Office of Naval Research K-5-55043/3916-1552793, NSF IIS99-00297, and NASA NRA NAG 9-1279. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of these agencies.

References

1. W.L. Johnson, J.W. Rickel, and J.C. Lester, "Animated Pedagogical Agents: Face-to-Face Interaction in Interactive Learning Environments," *Int'l J. AI in Education*, vol. 11, Nov. 2000, pp. 47–78.
2. S.C.Marsella, W.L. Johnson, and C. LaBore, "Interactive Pedagogical Drama," *Proc. 4th Int'l Conf. Autonomous Agents*, ACM Press, New York, 2000, pp. 301–308.
3. E. André et al., "The Automated Design of Believable Dialogues for Animated Presentation Teams," *Embodied Conversational Agents*, MIT Press, Cambridge, Mass., 2000, pp. 220–255.
4. J. Cassell et al., "Human Conversation as a System Framework: Designing Embodied Conversational Agents," E*mbodied Conversational Agents*, MIT Press, Cambridge, Mass., 2000, pp. 29–63.
5. J.E. Laird, "It Knows What You're Going To Do: Adding Anticipation to a Quakebot," *Proc. 5th Int'l Conf. Autonomous Agents*, ACM Press, New York, 2001, pp. 385–392.
6. S. Yoon, B. Blumberg, and G.E. Schneider, "Motivation Driven Learning for Interactive Synthetic Characters," *Proc. 4th Int'l Conf. Autonomous Agents*, ACM Press, New York, 2000, pp. 365–372.
7. J. Cassell, "Nudge, Nudge, Wink, Wink: Elements of Face-to-Face Conversation for Embodied Conversational Agents," *Embodied Conversational Agents*, MIT Press, Cambridge, Mass., 2000, pp. 1–27.
8. J. Cassell, M. Stone, and H. Yan, "Coordination and Context-Dependence in the Generation of Embodied Conversation," *Proc. INLG 2000*, 2000.
9. M.T. Maybury, ed., *Intelligent Multimedia Interfaces*, AAAI Press, Menlo Park, Calif., 1993.
10. B. O'Conaill and S. Whittaker, "Characterizing, Predicting, and Measuring Video-Mediated Communication: A Conversational Approach," *Video-Mediated Communication: Computers, Cognition, and Work*, Lawrence Erlbaum Associates, Mahwah, N.J., 1997, pp. 107–131.
11. J. Cassell, H. Vilhjálmsson, and T. Bickmore, "BEAT: The Behavior Expression Animation

Jonathan Gratch is a project leader for the stress and emotion project at the University of Southern California's Institute for Creative Technologies and is a research assistant professor in the department of computer science. His research interests include cognitive science, emotion, planning, and the use of simulation in training. He received his PhD in computer science from the University of Illinois. Contact him at the USC Inst. for Creative Technologies, 13274 Fiji Way, Marina del Rey, CA 90292; gratch@ict.usc.edu; www.ict.usc.edu/~gratch.

Jeff Rickel is a project leader at the University of Southern California's Information Sciences Institute and a research assistant professor in the department of computer science. His research interests include intelligent agents for education and training, especially animated agents that collaborate with people in virtual reality. He received his PhD in computer science from the University of Texas at Austin. Contact him at the USC Information Sciences Inst., 4676 Admiralty Way, Ste. 1001, Marina del Rey, CA 90292; rickel@isi.edu; www.isi.edu/~rickel.

Elisabeth André is a full professor at the University of Augsburg, Germany. Her research interests include embodied conversational agents, multimedia interface, and the integration of vision and natural language. She received her PhD from the University of Saarbrücken, Germany. Contact her at Lehrstuhl für Multimedia Konzepte und Anwendungen, Institut für Informatik, Universität Augsburg, Eichleitnerstr. 30, 86135 Augsburg, Germany; andre@informatik.uni-augsburg.de; www.informatik.uni-augsburg.de/~lisa.

Justine Cassell is an associate professor at the MIT Media Lab where she directs the Gesture and Narrative Language Research Group. Her research interests are bringing knowledge about verbal and nonverbal aspects of human conversation and story telling to computational systems design, and the role that technologies—such as a virtual story telling peer that has the potential to encourage children in creative, empowered, and independent learning—play in children's lives. She holds a master's degree in literature from the Université de Besançon (France), a master's degree in linguistics from the University of Edinburgh (Scotland), and a dual PhD from the University of Chicago in psychology and linguistics. Contact her at cassell@media.mit.edu.

Eric Petajan is chief scientist and founder of face2face animation, and chaired the MPEG-4 Face and Body Animation (FBA) group. Prior to forming face2face, he was a Bell Labs researcher where he developed facial motion capture, HD video coding, and interactive graphics systems. Starting in 1989, he was a leader in the development of HDTV technology and standards leading up to the US HDTV Grand Alliance. He received a PhD in EE in 1984 and an MS in physics from the University of Illinois where he built the first automatic lipreading system. He is also associate editor of the IEEE Transactions on Circuits and Systems for Video Technology. Contact him at eric@f2f-inc.com.

Norman Badler is a professor of computer and information science at the University of Pennsylvania and has been on that faculty since 1974. Active in computer graphics since 1968, his research focuses on human figure modeling, real-time animation, embodied agent software, and computational connections between language and action. He received a BA in creative studies mathematics from the University of California at Santa Barbara in 1970, an MSc in mathematics in 1971, and a PhD in computer science in 1975, both from the University of Toronto. He directs the Center for Human Modeling and Simulation and is the Associate Dean for Academic Affairs in the School of Engineering and Applied Science at the University of Pennsylvania. Contact him at badler@seas.upenn.edu.

Toolkit," to be published in *Proc. SIGGRAPH '01*, ACM Press, New York, 2001.

12. M. Johnston et al., "Unification-Based Multimodal Integration," *Proc. 35th Ann. Meeting Assoc. Compuational Linguistics* (ACL 97/EACL 97), Morgan Kaufmann, San Francisco, 2001, pp. 281–288.
13. P. Taylor et al., "The Architecture of the Festival Speech Synthesis System," *3rd ESCA Workshop Speech Synthesis*, 1998.
14. C. Pelachaud et al., "Embodied Agent in Information Delivering Application," *Proc. Autonomous Agents and Multiagent Systems*, ACM Press, New York, 2002.
15. J. C. Lester et al., "Deictic and Emotive Communication in Animated Pedagogical Agents," *Embodied Conversational Agents*, MIT Press, Cambridge, Mass., 2000, pp. 123–154.
16. A. Ortony, G.L. Clore, and A. Collins, *The Cognitive Structure of Emotions*, Cambridge Univ. Press, Cambridge, UK, 1988.
17. S. Marsella, and J. Gratch, "A Step Towards Irrationality: Using Emotion to Change Belief," to be published in *Proc. 1st Int'l Joint Conf. Autonomous Agents and Multi-Agent Systems*, ACM Press, New York, 2002.
18. M.S. El-Nasr, T.R Ioerger, and J. Yen, "Peteei: A Pet with Evolving Emotional Intelligence," *Proc. 3rd Int'l Conf. Autonomous Agents*, ACM Press, New York, 1999, pp. 9–15.
19. S. Marsella and J.Gratch, "Modeling the Interplay of Emotions and Plans in Multi-Agent Simulations," *Proc. 23rd Ann. Conf. Cognitive Science Society*, Lawrence Erlbaum, Mahwah, N.J., 2001, pp. 294–599.
20. A. Sloman, "Motives, Mechanisms and Emotions," *Cognition and Emotion, The Philosophy of Artificial Intelligence*, Oxford Univ. Press, Oxford, UK, 1990, pp. 231–247.
21. J. Gratch, "Emile: Marshalling Passions in Training and Education," *Proc. 4th Int'l Conf. Autonomous Agents*, ACM Press, New York, 2000.
22. G. Collier, *Emotional Expression*, Lawrence Erlbaum, Hillsdale, N.J., 1985.
23. G. Ball and J. Breese, "Emotion and Personality in a Conversational Character," *Embodied Conversational Agents*, MIT Press, Cambridge, Mass., 2000, pp. 189–219.
24. D. Chi et al., "The EMOTE Model for Effort and Shape," *ACM SIGGRAPH '00*, ACM Press, New York, 2000, pp. 173–182.
25. M. Schröder, "Emotional Speech Synthesis: A Review," *Proc. Eurospeech 2001*, ISCA, Bonn, Germany, 2001, pp. 561–564.
26. E. Hovy, "Some Pragmatic Decision Criteria in Generation," *Natural Language Generation*, Martinus Nijhoff Publishers, Dordrecht, Germany, 1987, pp. 3–17.

27. M. Walker, J. Cahn, and S.J. Whittaker, "Improving Linguistic Style: Social and Affective Bases for Agent Personality," *Proc. Autonomous Agents '97*, ACM Press, New York, 1997, pp. 96–105.

28. D. Moffat, "Personality Parameters and Programs," *Creating Personalities for Synthetic Actors*, Springer Verlag, New York, 1997, pp. 120–165.

29. H. Prendinger, and M. Ishizuka, "Social Role Awareness in Animated Agents," *Proc. 5th Conf. Autonomous Agents*, ACM Press, New York, 2001, pp. 270–377.

30. A. Paiva et al., "Satira: Supporting Affective Interactions in Real-Time Applications," *CAST 2001: Living in Mixed Realities*, FhG-ZPS, Schloss Birlinghoven, Germany, 2001, pp. 227–230.

31. M. Gleicher, "Comparing Constraint-based Motion Editing Methods," *Graphical Models*, vol. 62, no. 2, Feb. 2001, pp. 107–134.

32. N. Badler, C. Phillips, and B. Webber, *Simulating Humans: Computer Graphics, Animation, and Control*, Oxford Univ. Press, New York, 1993.

33. M. Raibert and J. Hodgins, "Animation of Dynamic Legged Locomotion," *ACM Siggraph*, vol. 25, no. 4, July 1991, pp. 349–358.

34. D. Tolani, A. Goswami, and N. Badler, "Real-Time Inverse Kinematics Techniques for Anthropomorphic Limbs," *Graphical Models*, vol. 62, no. 5, May 2000, pp. 353–388.

35. J. Lewis, M. Cordner, and N. Fong, "Pose Space Deformations: A Unified Approach to Shape Interpolation and Skeleton-driven Deformation," *ACM Siggraph*, July 2000, pp. 165–172.

36. P. Sloan, C. Rose, and M. Cohen, "Shape by Example," *Symp. Interactive 3D Graphics*, Mar., 2001, pp. 135–144.

37. D. Baraff and A. Witkin, "Large Steps in Cloth Simulation, *ACM Siggraph*, July 1998, pp. 43–54.

38. N. Magnenat-Thalmann and P. Volino, *Dressing Virtual Humans*, Morgan Kaufmann, San Francisco, 1999.

39. K. Perlin, "Real Time Responsive Animation with Personality," *IEEE Trans. Visualization and Computer Graphics*, vol. 1, no. 1, Jan. 1995, pp. 5–15.

40. M. Byun and N. Badler, "FacEMOTE: Qualitative Parametric Modifiers for Facial Animations," *Symp. Computer Animation*, ACM Press, New York, 2002.

41. M. Brand and A. Hertzmann, "Style Machines, *ACM Siggraph*, July 2000, pp. 183–192.

42. E. Cosatto and H.P. Graf, "Photo-Realistic Talking-Heads from Image Samples," *IEEE Trans. Multimedia*, vol. 2, no. 3, Mar. 2000, pp. 152–163.

43. T. Ezzat, G. Geiger, and T. Poggio, "Trainable Videorealistic Speech Animation, to be published in *Proc. ACM Siggraph*, 2002.

44. D. Terzopoulos and K. Waters, "Analysis and Synthesis of Facial Images Using Physical and Anatomical Models," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 15, no. 6, June 1993, pp. 569–579.

45. F.I. Parke and K. Waters, *Computer Facial Animation*, AK Peters, Wellesley, Mass., 1996.

46. T.K. Capin, E. Petajan, and J. Ostermann, "Efficient Modeling of Virtual Humans in MPEG-4," *IEEE Int'l Conf. Multimedia*, 2000, pp. 1103–1106.

47. T.K. Capin, E. Petajan, and J. Ostermann, "Very Low Bitrate Coding of Virtual Human Animation in MPEG-4," *IEEE Int'l Conf. Multimedia*, 2000, pp. 1107–1110.

48. L. Kovar, M. Gleicher, and F. Pighin, "Motion Graphs," to be published in *Proc. ACM Siggraph 2002*, ACM Press, New York.