Emotional Personae and Directorial Modeling for Interactive Entertainment

Scott M. Stevens
and
Michael G. Christel

Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA 15213

(In *Proceedings of the Workshop on Artificial Intelligence and Interactive Entertainment*
AAA[-92, The Tenth National Conference on Artificial Intelligence, San Jose, CA)

This work sponsored by the U.S. Department of Defense.

Beginning in 1987 the Advanced Video Technologies Project (AVT) at the Software Engineering Institute of Carnegie Mellon University began developing a radically new paradigm for interactive multimedia and behavioral modeling [1]. The goal of the AVT Project was to create a captivating, virtual world. Critical to this world was modeling highly realistic simulated personae that reacted on an emotional level both with other personae in the world and with the user. Since scenes in the AVT world are generated dynamically, modeling a director's visual knowledge is critical to the success of this virtual world.

The user is placed in a virtual building where they may walk through rooms and interact with personae. The user may talk to the video personae through a natural language interface or, instead, may simply sit back, look, and listen. Of course, as in real life, personae will sooner or later ask the user to participate. This is a high fidelity simulation of interpersonal interactions, in contrast to the traditional multimedia/interactive video paradigm of "play a video sequence, stop, wait for user input, branch;" which is often labeled "interrupted video."

In our world, called "A Cure for the Common Code," video, audio, and image objects are fine-grained, knowledge-rich entities. The fine granularity and embedded knowledge provides for the ability to create a highly interactive learning environment.

"A Cure..." was designed around the principle of an intrinsically motivating fantasy. A considerable body of research has been performed looking at what makes "Nintendo" type games so captivating [2]. Important motivating design aspects include: challenge (goals with uncertain attainment), curiosity (sensory and cognitive), and fantasies (especially intrinsic fantasies where the fantasy depends crucially on the task).

We believe these are important attributes for creating exciting interactive entertainment. There is a fundamental belief that the universe, whether it is the universe of a mystery, love story, or an academic discipline, is knowable. The unknown produces cognitive curiosity evoked by the need for completeness, consistency, and parsimony. When one is presented with just enough information that their existing schemes (knowledge) seem incomplete, inconsistent, or unparsimonious, there is an innate desire to "get one's intellectual house in order."

Decades ago, the pioneering film maker and theoretician Sergei Eisenstein used the work of Jean Piaget in the formulation of his theory of film. Intrinsically motivating fantasy is also closely tied to Piaget. For Piaget people assimilate new knowledge in terms of their existing mental schemes. If the new knowledge is too far removed from their existing schemes the interactive entertainment audience, game player, learner, or researcher cannot accommodate the knowledge.

A problem all too frequently encountered is that an experience may be so unexpected, so far beyond the audience's existing mental schemes (models), that the incompleteness, inconsis-

tency, and chaos go unnoticed. Alternately, expectations may be exactly matched by a virtual world, which is equally demotivating. "A Cure..." attempts to walk this fine line between the unimaginable and the expected by presenting the optimal level of information complexity for the user.

"A Cure..." begins with an introductory, motivating "hook" where the user is exposed to his or her possible roles and goals. The user can select to be any of the four roles involved in the group process activity which is the focus of "A Cure..." The system intelligently simulates the remaining roles not chosen by the user, and can model behavior for these personae ranging from dumb and inappropriate to ideal for the given group process activity. This is achieved in part through a rule-based expert system that was developed to model the personae in emotional, temporal, and contextual dimensions. The expert system defines the "personalities" and controls the dialogue between the personae and the user.

The rule base consists of over 8,000 lines of OPS/83 code and was developed from analyses of taped group meetings, analyses of normal conversations, and a review of the group process literature, including the work of Bales [3, 4]. It makes decisions in areas such as who should speak, the tone they should speak with, the content of what is to be spoken, and who is the persona speaking to.

The rule base uses ten hours of audio, two hours of video with audio, and several thousand still images to dynamically compose scenes for use during the group process simulation. The audio and video alone consist of over five thousand objects. All of these objects are organized in a four dimensional multimedia object base.

The first dimension is the topic under discussion (context space). The second dimension is the speaker (persona). The third dimension is the specificity within the topic. (This is related to the temporal aspect being modeled in the conversation. As people speak on a topic, they tend to get more specific and build on what was said previously, much as a text builds on previously chapters.) The fourth dimension is affect (emotion). The affect space includes aggressiveness, defensiveness, talkativeness, and comic wit.

Abstract information about objects in AVT varies in granularity from scene headers describing information that is globally constant for a set of frames, to frame headers, information that is local to a single frame. Figure 1 depicts the scene header information for the video objects available to the rule base. Along with traditional "file control block" information, embedded in each object is information on camera angle and field of view (objective points of view), character, topic, specificity within topic, tone, to whom is the scene addressed, opinion, emotional subject, discussion resolution status, pointers to other topics, and gesticulation (all subjective points of view). The objects under the subjective points for both scene and frame (in the image object base) header define the dialogue element's location in the four dimensional space. The

rule base determines the points in this space which are needed to compose the appropriate dialogue.

Traditional "file control block" information
- Name of scene
- Size in bytes
- *Creation date of scene
- Media type (audio, video, audio and video)
- Media-specific descriptors
  - Audio (sampling rate, filtering, etc.)
  - Video (frame rate, position, compression algorithm, etc.)

Objective point of view
- Camera angle
- Field of view
- -Audio attributes (e.g., attenuation and filtering)
- Image attributes (e.g., hue and saturation)

Subjective point of view
- Speaker (Persona)
- Tone
- Opinion
- Resolved status
- Gesticulation
- Topic and specificity within topic
- To whom the scene is addressed
- Personal focus (emotional subject)
- Bridge to another topic

Figure 1: Contents of AVT Video Dialogue Scene

One of our goals was to create a high fidelity virtual world. An obvious question is "how much fidelity is enough?" Put another way, how much can we rely on suspension of disbelief? All totaled, data on over 120 users of our virtual world have been analyzed. An interface experiment was conducted using a majority of these users, and is reported in detail in [5]. Two methods of navigating through the world were tested in this experiment, one a direct manipulation point and click map, and the other a surrogate travel interface where the user "walks" through the space and into the desired sub-world. Both groups of users liked, or disliked, the interfaces equally and used them in equivalent fashions to navigate the world. However, users with the surrogate travel interface came away from the experience with more positive opinions about the subject of the virtual world! While it can be argued that the surrogate travel interface was more cumbersome and slower, its users were more completely brought into the fantasy.

Another question to ask of virtual worlds is "do differing levels of fidelity matter with respect to frame rate?" Is one frame per second fast enough? Five frames? Thirty frames? We have shown that there are compelling reasons to believe that high frame rates are more than frills. In the experiment reported in [5], one group of users was presented with full motion video and audio in the various sub-worlds. Another group had the same experience except identical, but sequential still images with audio. Compared to the sequential still image users, the users with the full motion video retained more information and were able to identify personae's contributions better even though the information tested on was contained in the audio. In a second

experiment, a third group had intermediate frame rate (five frames per second) video, but an otherwise identical virtual world. Seven out of ten of the five frame per second users exited before they finished the experience! A typical and telling user response was "What did I do wrong in life that caused me to have to do this." Our perception of motion is obviously affected by slow frame rates. These studies show that understanding, attitudes, and memory are all affected by frame rate!

It is tempting to believe that "what we see is what we get" [6]:

> ... we *see a rectangular table as nearly rectangular even when the front edge is pushed quite close to our eyes. ...But the photograph does not so compensate, giving us an image of a man's foot, ...larger than his head if it is stretched out in front of him. When examining a photograph, our mind fails to compensate for this effect, since the photograph is a two-dimensional object... Being true to the mathematically real, photography is false to the psychologically real.*

We thus want the expert system to behave intelligently in presenting images, much like a director does today when shooting a scene, or when they edit it. But in our world, the system does it during playback. An interesting part of the rule base is called Hitchcock, our visual director of the system. With digital video, we can manipulate the images when we play them back. In fact, for the AVT system we must in order to compose scenes with different personae in them. Hitchcock draws from an object base of over 450 motion video sequences and two thousand still images to create animations and motion video sequences. Attributes of this object base include the persona, direction of gaze, the emotion shown, the gesticulation used, and the camera angle use in recording the image.

In the simplest case, Hitchcock must determine who the speaker is addressing and display appropriate images (i.e. the persona should frequently look at who he or she is talking to). If a secondary character is bored, Hitchcock may develop a scene with that persona slumped in a chair with eyes looking downward. Depending on scene length, dialogue, and a global history of the user's experience, Hitchcock may choose to generate close-up, medium, or wide shots. For example, if a persona has not been shown with a close-up for quite some time, and that persona has just had a sudden shift of emotion from passive to aggressive, then show a close-up during that persona's next scene for dramatic effect. More interestingly, if the user is dominating a conversation, the system may present a camera angle of the participants on the screen which is slightly lower. Years of experience and many studies have shown that images such as this tend to portray the viewer as more dominant [7, 8].

Since the AVT Project began, other researchers have discussed related schemes, defining cinematic primitives for use by multimedia systems [9]. Unfortunately, their current implementations use analog videodiscs as video sources and contain no behavioral modeling of personae [10]. Having the video and audio in digital form is crucial to our paradigm. The descriptions of a multimedia object base, being closely tied with the information, should remain with the information data rather than hidden within an application working on the information.

With our paradigm, variable granularity knowledge about a domain, content, image structure, and the appropriate use of content and image is embedded with the object. These often orthogonal descriptions of a multimedia object base are shown to promote usability, accessibility, and fidelity. This provides, for the first time, the ability to present disparate text, audio, images, and video intelligently in response to users' needs in a high fidelity, interactive environment.

1. Stevens, S.M. Intelligent Interactive Video Simulation of a Code Inspection. *Communications of the ACM* 32(7): 832-843, July 1989.

2. Malone, T.W. Toward a Theory of Intrinsically Motivating Instruction. *Cognitive Science 4:* 333-369, 1981.

3. Bales, R.F. *Interaction Process Analysis: A Method for the Study of Small Groups.* Addison Wesley, Cambridge, MA, 1951.

4. Bales, R.F. *Personality and Interpersonal Behavior.* Holt, Rienhart, & Winston, New York, 1970.

5. Christel, M.G. *A Comparative Study of Digital Video Interactive Interfaces in the Delivery of a Code Inspection Course.* Ph.D. Thesis, Georgia Institute of Technology, 1991.

6. Andrew, D. *The Major Film Theories.* Oxford University Press, London, 1976.

7. Kraft, R. The influence of camera angle on comprehension and retention of pictorial events. *Mem & Cogn.* 15(4): 291-307, 1987.

8. Kraft, R. Mind and media: The psychological reality of cinematic principles. In *Images, Information & Interfaces: Directions for the 1990's,* D. Schultz and C.W. Moody, Eds. Human Factors Society, New York, 1988, pp. 13-36.

9. Davenport, G., Aguierre Smith, T.G. & Pincever, N.C. Cinematic Primitives for Multimedia. *IEEE Computer Graphics & Applications* 11(4): 67-74, July, 1991.

10. Aguierre Smith, T.G. & Pincever, N.C. Parsing Movies in Context, *Proceedings of the USENIX Conference* held in Nashville, TN, Summer 1991.