Editor: Glorianna Davenport
MIT Media Lab

Smarter Tools for Storytelling: Are They Just Around the Corner?

Glorianna Davenport
MIT Media Lab

Achieving the "ultimate" multimedia machine has tantalized us for many years. From a hardware perspective, we might be approaching the illusionary crock of gold at the end of the rainbow. Most PCs are now sold with audiovisual and accelerated graphics capabilities. Scanners, printers, camcorders, and video capture boards are all available at reasonable prices, and affordable networking extends the computer's reach to distant resources. Inexpensive storage devices capable of holding a useful amount of video are finally reaching the market.

In addition to the right hardware, the ultimate multimedia machine requires magical software that will let us paint our dreams into our journals, design postcards with movies of the kids in them that can be shipped to and read by Moms anywhere, search the world effortlessly for information or other products that tickle our fancy, and previsualize large-scale entertainment productions. Perhaps the software could even "trick" us into performing useful work as we interactively play!

State of the art—1996

Current software is anything but magical. Our generation seems condemned to point and click through every sort of application, regardless of the process it represents. At least we are kinder to our children—their video games embrace a variety of controllers: joysticks, specialized button boxes, data gloves, pressure-sensitive floor mats. To the machine, it might all be surrogate pointing and clicking; but to the child, these special controllers allow for deeper, more intuitive engagement with the story.

Research in voice and gesture recognition continues to progress. Some exciting work has taken place in sensor technology for music applications[1] and in vision algorithms that can, for instance, read sign language.[2] However, no one has unearthed a more general, universal language for gesture, and none may be forthcoming. We face the sticky problem of predetermining a gesture set for any given application and possibly forcing the user to learn to flap out an arbitrary set of signals to the computer, like a flagman sending semaphore messages.

Mum's the word on biofeedback, except in medical applications and expensive, high-tech programs such as the space shuttle—there's still no "biosensor" port in the back of my PowerBook. Meanwhile, work on machine-augmented aesthetics remains in its infancy. In fact, we have progressed only marginally in our formal understanding of aesthetics over the past two decades. In the past, rule-based systems proved intractable for design; there are simply too many vague rules-of-thumb that artists mix intuitively and in different proportions. Statistical modeling has produced some successful research in design-by-example. But despite Russell Kirsch's program passing the Turing test by generating Diebenkorn-like fakes with Diebenkorn himself present,[3] progress toward consumer software that learns and applies aesthetic styles of presentation has been slow.

Tools to keep in mind

Recently, Roger Schank and Robert Abelson published a summary article, "Knowledge and Memory: The Real Story."[4] This article probes the relationship of story to personal memory as well as to social and cultural experience. The authors argue that knowledge and story are inextricably connected, and that stories are partial, structured memories of observed and articulated reality. If their theory is correct, the development of constructive storytelling tools seems like a worthy goal.

These storytelling tools must invite use by the broadest spectrum of people. They cannot balk when faced with real-world complexity, ambiguity, personalization, or the sweeping scope of distributed resources. They must simultaneously feature control, fluidity, and a temporal memory.

1070-986X/96/$5.00 © 1996 IEEE

Even the simplest of smart tools would benefit greatly by incorporating some type of storytelling mechanism and story memory. These could help the tool understand the process it is going through, remember what it has learned from previous encounters, anticipate the user's needs, explore alternative solutions to problems, and offer its support and guidance wherever appropriate.

One problem that seems to inhibit tool development is the popularity of the current generation of tools—Photoshop, Premiere, Inventor, Hypercard, Director, and various word processors. While they lack the satisfying tactile and fluid nature of their mechanical, atom-based counterparts, these programs do offer authors a useful and somewhat effective set of controls for manipulating text, images, sound, and indexes. These tools remain relatively single-purpose, however, employing a point-and-click paradigm and having little or no data substructure.

Indeed, visionary development on the tool front seems to have all but stalled on problems related to search functionality rather than acknowledging that it is the "feel" of tools that will essentially shape the creative work of tomorrow. I use the word "feel" advisedly. While computation is based on rational models that we can calculate, the expressive arts capture intuitive impressions of emergent story. These intuitions are never totally isolated from the process, the media, or, indeed, the motivation for expression.

Direct manipulation and the pencil

Consider a simple tool: the pencil. The pencil invites direct manipulation by the human hand. The pencil responds fluidly to conceptual and intuitive control. Over time, the pencil has proven itself reasonably well matched to human desire and capability.

Can we imagine a "smart" pencil? Would we be happy, for example, with a pencil that helped a child draw a snowman, possibly shading the snowballs, or making them rounder, or generating arms that look like branches? Does the act of image-making or storytelling become more fulfilling when such a tool is used? Will the story captured in the image become stronger or better? Does the value of the image improve?

To design a smart tool, we must look beyond its basic functionality and understand how the "expert" tool might affect an "expert" user of that tool. Using a pencil engages us both physically and psychologically. A concept of the picture forms in the artist's mind, and she propels her pencil across the writing surface, leaving controlled marks. Emotional neurons, thinking neurons, motor neurons, flesh and bones, gravity, the "feel" of the pencil in the hand, the resistance of the pencil point moving over the paper, and other aspects of the physical world make using a pencil a rich sensory experience. As the artist works, rich tactile and visual feedback allow her to better monitor and control her activity.

As humans, we have developed two expressive modes especially well suited to "pencil-aided design": the creation of visual images, and the drawing of letters grouped into symbol strings of words. The human, rather than the pencil, must adapt herself to output text or image.

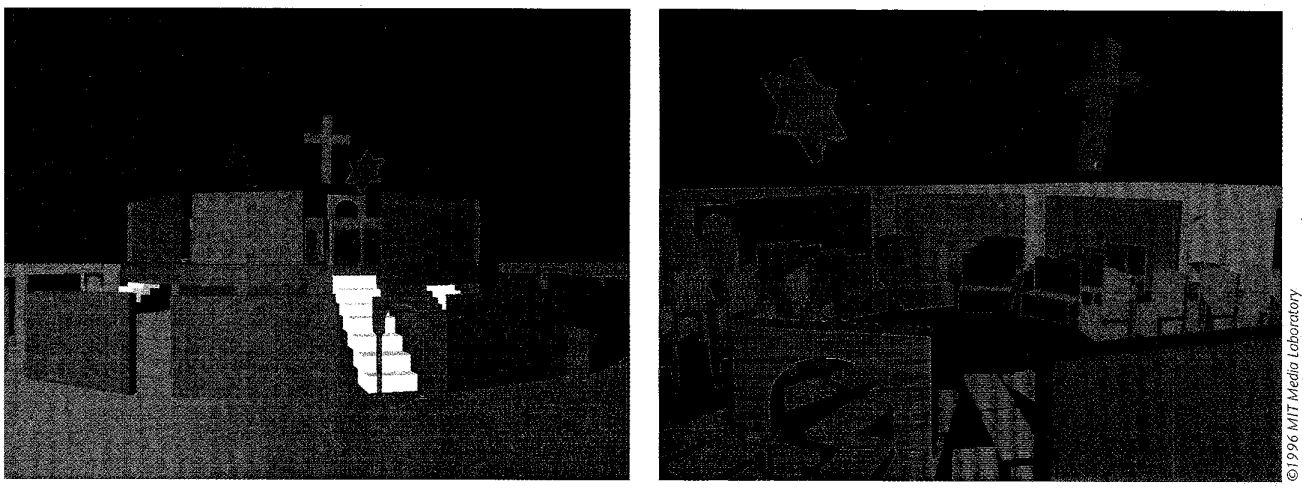
The artist benefits from the rigors of discipline and repetition, both in the conceptualization of her art and in its physical realization. In an expert's hand, the pencil becomes one with the artist. However, this fluidity does not mean that the process lacks difficulty. Becoming an expert is hard work. While Michaelangelo and Picasso, among others, have been judged drawing "experts," we cannot borrow their expertise. Nor can we assume that Picasso might pick up a pencil that made aesthetic decisions for him.

Given my druthers, I cannot imagine being satisfied with a pencil that guided my hand and allowed me to draw like da Vinci, unless it also imitated the "feel" of the pencil in his hand. The fact is, there is a joy in drawing that grows out of the feeling that we are gaining control and that our hand is mastering a subject. True rewards are only as deep as our own expertise.

Knowing what to do when, and when not to

The problem of the "smart pencil" is analogous to the problem of the "smart camera." Some people assume that a smart camera is one that can act autonomously in 3D space. Many years ago, one of my contemporaries proposed that my boss of the time, the well-known observational filmmaker Richard Leacock, would enjoy sitting with his subjects while an autonomous camera moved, "Leacock-style," to capture the situation.

I was stunned at the suggestion that a "smart camera" might replace Leacock's extraordinary eye, or that the filmmaker would have any interest in remaining in the situation once the camera was taken from his shoulder. One has only to watch a Leacock movie to understand that this filmmaker revels in the "joy of shooting." As Leacock moves his camera to frame the action, he is shaping his perception of what is taking place in



Two frames from "Wacky New Age Temple." The first is a wide-angle shot designed to emulate the viewer's approach to the temple. As the viewer draws near and passes each arch, sounds representing the different faiths grow louder. The second frame, with a smaller focal angle, shows the viewer's perspective as she enters the temple. The religious sounds become more intense and finally converge as the viewer comes to the temple's center.

the presence of the camera. By intense concentration on what is happening, Leacock uses the camera to reveal the situation to himself. Later, when the story is edited, his memories become the world's memories.

Of course, the autonomous "smart camera" envisioned by that colleague is not yet available in stores. However, a bit of the groundwork needed to support such lofty ambitions is now being laid. Camera manufacturers are finally beginning to take the idea of a data track seriously—but perhaps not seriously enough. Similar to an audio track, a data track allows a stream of data to become married with the captured video and audio images. Among other blessings, it will let users log shots automatically in the camera while shooting. "Camera on" markers, time-and-date stamps, camera parameters and settings (such as *f*-stop and lens focus), Global Positioning System (GPS) location coordinates, and other shot-describing data can be recorded to assist in editing, postproduction, and other uses.

Soon, a host of other add-on devices will vie for space on the data track. Some sensors, such as those that can satisfactorily track the ever-changing position and orientation of a hand-held camera, will generate a stream of data—which might even merit a second data track.

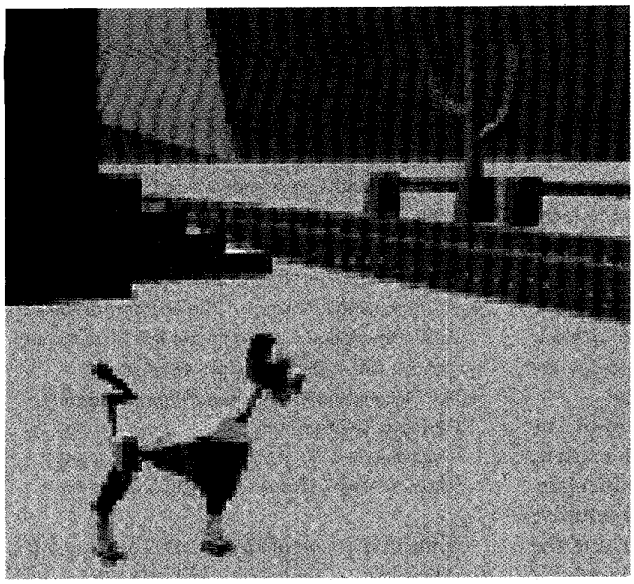
Image processing can be used to segment the image stream or to generate aesthetic framing decisions; it is already used for image stabilization in consumer camcorders. Furthermore, vision techniques have become quite accurate in reading gesture. These examples suggest that a machine-reasoning system could be developed to follow the action in simple scenes. Active "panning and scanning" within a higher-resolution shot might be one way to accomplish this. While we might use autonomous framing in a wide shot, however, it is difficult to imagine the rule set that could direct the camera to zoom and truck in close.

A smart camera (and its robotic mount) might someday successfully mimic the shooting style and compositional sense of an expert camerawoman by exploiting statistical models of her past performance. However, the smart camera will have a harder time figuring out exactly what it should be shooting in the current scene. The human camerawoman loves discovering which details in a scene are especially interesting or significant. She follows and anticipates the action. She knows where the camera should be.

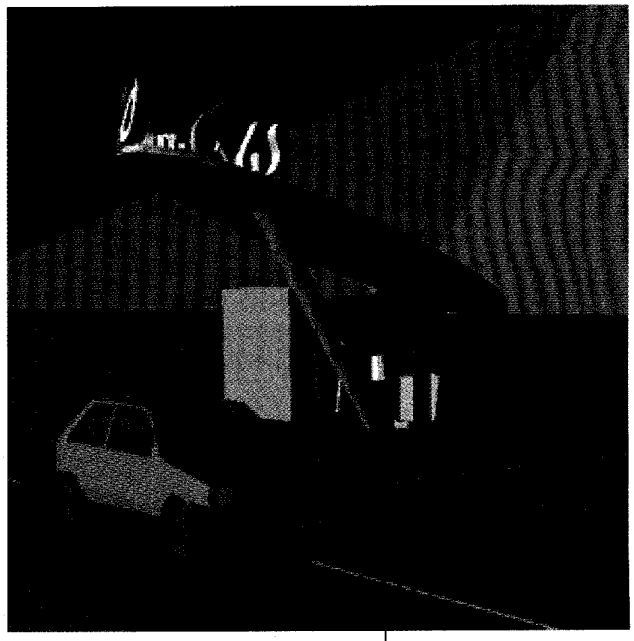
The virtual 3D camera

Virtual reality and 3D simulation environments position the viewer as a camera in a world that is more or less complexly rendered. Until recently, the speed of rendering complex structures was too slow to deliver a meaningful, real-time interactive experience. But there were also underlying problems. The strength of current 3D virtual reality is not the identification with a human character but the identification with a sensual experience. What range of effects might we use to make the story more "feelable"?

Navigation constraints now appear almost as conventions. For example, 3D game worlds are rendered using minimal detail, and the paths



In "Dogmatic," the viewer can look around freely, but is at a fixed location in each scene. This figure illustrates two sequential shots in the third scene, in which the viewer must look at Lucky to trigger the action, the car driving by. If the viewer is looking elsewhere, Lucky barks to gain her attention. The variable timing of the cut requires a variable music score. The entire system, therefore, adjusts to the story's needs and the viewer's action.



offered through them are limited. The rapid pace of action obscures the inexact mapping of the control device to the player. For me, games will come into their own when we combat physical forces—wind and weather—as well as enemy sprites.

In surrogate travel applications, the discrepancies between the control device and the participant's desires are often minimized by restricting navigation to buttons, such as the "turn left/ turn right/ forward/ reverse" buttons in Aspen (1979). (The makers of Myst (1993) added "zoom in" and "zoom out.") The use of buttons renders moot any issue about the physical attributes of the world, such as gravity. However, as movie people will advise, "travel rarely a story makes."

For a storyteller, 3D participatory environments pose a quandary. Since the viewer is the "I" of story, her role must be married to the navigational task. In a recent exercise, MIT Master's candidate Freedom Baird created a "Wacky New Age Temple." The story idea was that of pilgrimage. Traveling in another age, the viewer should feel as if she has traveled for a long time, has felt the ground under her feet, and has been drawn toward the light of the temple by the age-old sounds of faith.

One failing of Freedom's sketch, or perhaps one of its points, is that the pilgrimage is painless. The "I" of the story does not feel the weariness of travel, the contours of seasoned pathways beneath her feet, or the smoothness of the temple's polished stone. Furthermore, because the piece does not resonate experientially, a colleague who helped program the sound found it easier to fly over the space than to pass through an arch.

Three-dimensional VR requires that we go the extra mile to create sensation that approaches believability. Most virtual environments do not use changes in camera position to confront the viewer emotionally. Last year, Tinsley Galyean, an MIT PhD student, decided to explore this cinematic option in "Dogmatic," a four-act movie rendered on an SGI Reality Engine.

The system's guidance properties allowed Galyean to explore foreshortened time, transitions between shots, and the close-up as they might be applied to VR storytelling. The movie happens around the viewer, and interactions are motivated by other characters' actions, particularly the dog Lucky, an autonomous character whose behavioral engine was written by MIT's Bruce Blumberg. The real meat of this story is the viewer herself as she discovers that Lucky is a surprisingly aggressive form of "story tool." Ouch!

If we are ever to create dynamic, reconfigurable electronic stories that feel and are felt, we need to

transform our tools to engage both the human and the machine in satisfying and productive acts of collaboration. The point-and-click paradigm is an inadequate window into the vast range of human thoughts, feelings, desires, and actions. The use of sensors, force-feedback effects, and storytelling mechanisms are just a few of the techniques that will play important roles in building tools "with a sense of themselves." As we gain experience with cameras and other tools that are substantially self-aware, have knowledge of their own movements, and model aesthetic styles, we are building a library of reactive potential.

Gradually, these smart tools will gain the expertise and knowledge of their human masters, and their utility (and complexity) will increase. Perhaps in the next century, these more intuitive technologies will become the standard, and the gap between our computational mind and our storytelling memory will lessen. **MM**

References

1. J.A. Paradiso and N. Gershenfeld, "Musical Applications of Electric Field Sensing," tech. report, Physics and Media Group, MIT Media Lab., MIT, Cambridge, Mass., submitted for publication.
2. T. Starner and A. Pentland, "Visual Recognition of American Sign Language Using Hidden Markov Models," *Int'l Workshop on Auto. Face and Gesture Recognition*, Univ. of Zurich, 1995, pp. 189-194.
3. J.L. Kirsch and R.A. Kirsch, "The Anatomy of Painting Style: Description with Computer Rules," *Leonardo*, Vol. 21, No. 4, Dec. 1988, pp. 437-444.
4. R.C. Schank and R.P. Abelson, "Knowledge and Memory: The Real Story," in *Advances in Social Cognition*, Vol. VIII, R.S. Wyer, Jr., ed., Lawrence Erlbaum Assoc., Hillsdale, N.J., 1995, pp. 1-85.

Contact Davenport at Media Arts & Sciences, MIT, 20 Ames St., Cambridge, MA 02139, e-mail gid@media.mit.edu.