HyperPlex: a World of 3D Interactive Digital Movies

Flavia Sparacino, Christopher Wren, Alex Pentland, Glorianna Davenport
The Media Laboratory,
Massachusetts Institute of Technology
Room E15-384, 20 Ames Street, Cambridge MA 02139, USA
flavia@media.mit.edu, cwren@media.mit.edu
sandy@media.mit.edu, gid@media.mit.edu

Abstract

We present a new environment for browsing a visual landscape inhabited by digital movies that live, interact and play in a graphical virtual world. The movies are modeled as autonomous agents which have their own sensors and goals and which can interpret the actions of the participant and react to them. Our environment allows one—or more—people to interact with the HyperPlex world through the use of vision techniques. No goggles, gloves or wires are needed: interaction takes place with the use of computer vision techniques that analyze the image of the person. An extension of this system to a multi-user game is currently being considered.

1 Introduction

The HyperPlex system assembles work from three different research groups: Vision and Modeling, Interactive Cinema, and Autonomous Agents.

Research in the Vision and Modeling group at the Media Laboratory allows interaction without the use of cumbersome device or cables. When users are in front of a big display screen, the use of a mouse and a keyboard to issue commands to the system is extremely limiting. It is much better if interaction takes place by the use of gesture and voice recognition systems (multi-modal interaction). The user becomes a 3D-mouse that can point at or highlight different portions of the screen, and give simple commands [Darrell *et al.*, 1994].

Research at the Interactive Cinema group has created a video browsing environment for large databases of digital video [Davenport, 1993], by making the spatial dimension of the displayed information (i.e., where things appear on the screen) as important as its temporal dimension (i.e., what comes after what in a video sequence). The conceptual characteristics of each video clip are associated with physical locations, and users browse the database by steering through this "conceptual space." The challenge that this research group has been faced with is the construction of storyteller systems that handle variable, non-linear, multi-threaded narratives.

Research by the Autonomous Agents group has considered how computer programs ("agents") should respond to user's activity given their internal motivation, past history and a perceived environment with its attendant opportunities, challenges and changes [Maes, 1995]. Moreover, the pattern and rhythm of the chosen activities should be such that it neither dithers between multiple activities nor persists too long in a single activity. It should be capable of interrupting a given activity if a more pressing need or if an unforeseen opportunity arises.

Finally, navigation in such a graphical virtual environment implies considering not only the construction of a visually compelling virtual world but also providing the users with visual cues that help them orient themselves in the simulated reality.

2 The HyperPlex

The HyperPlex consists of a many-dimensional virtual building inhabited by digital movie clips and other visual objects (photographs, text, graphics). The user can navigate around the building exploring the different rooms in each floor and moving from one floor to another through virtual doors, corridors, and elevators. Each part of the building is associated with a particular cluster of topics. For example, one set of rooms might concern a particular set of people, places, politics, and time; in addition, nearby rooms would share some common theme (e.g., adventure, special effects, fun, or romantic).

Movies live in this space and can appear at many different locations in the building according to the subject/information of their component clips (subsegments) (see fig. 1). Each movie clip appears in the form of a window on the screen showing a keyframe from the clip. Keyframe windows can move around in a manner that tries to reflect the "personality" of each clip (see [Lasseter, 1987; Johnson, 1995]); keyframe windows also react to the user's gestures and voice in a manner characteristic of the clip.

The behavior of each visual object is modeled using Blumberg's computational model of action-selection [Blumberg, 1994a]. In Blumberg's framework an agent's set of activities is organized as a loose hierarchy with the top of the hierarchy representing more general activities

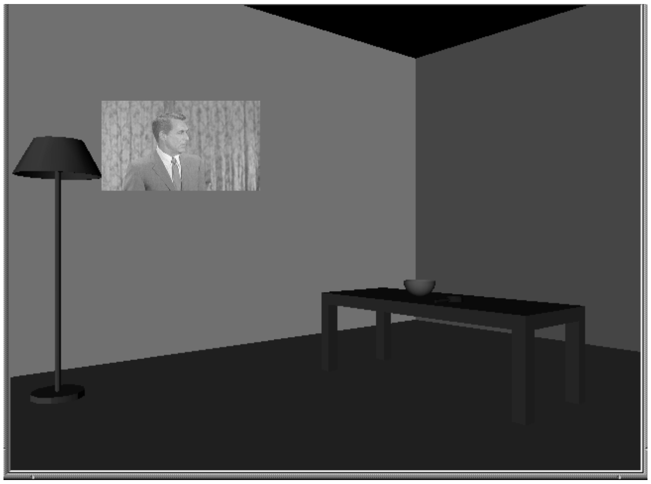

Figure 1: Movie clip competing for attention in the magic room

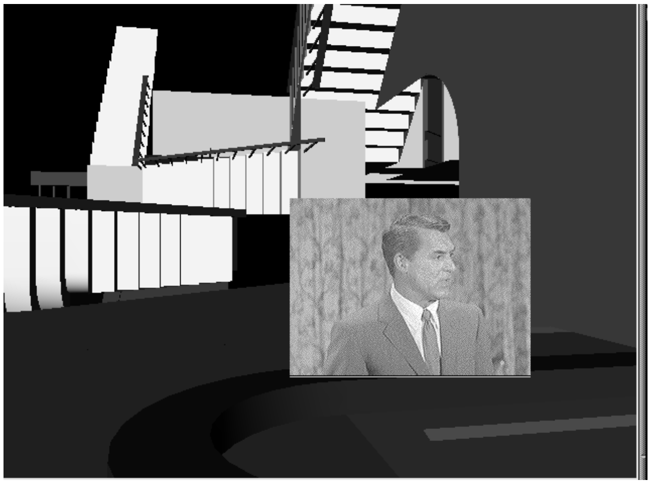

Figure 2: Movie clip in an Escher-esqe environment.

and the leaves representing more specific activities. Activities compete on every time step for the the control of the agent that engages in a single activity at a time. A movie's goal is to play itself and to compete with other movies to catch the user's attention. Movies form a community where movies associated with similar concepts collaborate with each other, whereas movies pertaining to distant concepts compete to play or to have a central position on the display screen.

The user can "call a movie", "grab a movie", "play a movie", "play a movie again", "send a movie away", "send a movie to another user as a postcard", "take a movie", "ask more info about a movie" (that comes in the form of text), "stop a movie that is playing"—like using a smart gesture-driven VCR—or just let the movies organize themselves dynamically on the screen and play as a result of the interaction amongst themselves and the graphical world they are immersed in.

3 A Dynamic Display

Practically all current multimedia applications are point-and-click applications where the visual display is static and nothing happens until the user clicks on the "right spot". The user may be involved either in an exploration or in a role-playing type of game but the objects that appear in the display can only be turned on and off—according to the user's interests. We call this type of display static because the objects it contains—text, graphic, photographs or video clips—always appear in the same position on the screen and their appearance is always triggered by the same type of action. The main drawback of static point-and-click displays is that the behavior of the user is reduced to clicking on all the the possible active spots and as a consequence the user may lose interest in the game or application after the space has been explored completely.

Dynamic displays offer instead a more compelling organization of the visual material layout. Each object has a notion of where to go, stay, its size, and movement according to the context in which it appears (background). Moreover the motor behavior of the objects is a function of their "personality" i.e. content and goals. The background is a dynamic object itself as it can attract or push away selectively certain objects in specific regions of the space and transform itself according the type and content of the visual material it holds (see fig. 2). Objects relate to themselves as a community of agents where hierarchy is built locally in space and time when two or more objects inhabit the same space and it is weighted by the user's past interactions with the display. As a result of the interaction of its objects amongst themselves and with the user, the display becomes a dynamic environment where the visual material re-arranges itself over time in a meaningful manner and is animated by its own internal dynamic and not uniquely by the intervention of the user. The display is responsible for **where** objects appear over time and **how**. **When** the objects are to appear is the result of the user's input together with the intern dynamic of the display and **what** visual material is shown depends on the orchestrator of the environment (the storyteller) that takes into consideration all of the above elements.

Blumberg's behavior based autonomous agents' model [Blumberg, 1994b] offers an ideal framework to develop dynamic displays. It provides tools to build animated characters for interactive virtual environments that are not only capable of autonomous action but respond also to the user's input.

Dynamic displays may also offer one possible answer to the aesthetic question raised by [Youngblood, 1989]:

> How has the corpus of aesthetic strategies inherited in a medium like photography or film transferred over to electronic media and especially to the code?

They allow: *image transformation* ("if mechanical cinema is the art of transition, electronic cinema is the art of transformation"), *parallel event streams* ("past, present and future can be spoken in the same frame at once"), *temporal perspective*, and *the image as object* ("with code it becomes a trivial matter to remove the image from the frame and treat it as an object, an image plane...").

They also satisfy some of the criteria for creating compelling interactive systems proposed by Brenda Laurel in *A Taxonomy of Interactive Movies* [Laurel, 1989]:

> ...you need to think about intelligent animation— characters who know what they look like and how they move ...it means generating (or retrieving) and then manipulating backgrounds as the action is pieced together...

4 Content Orchestration and Game Design

Content orchestration is strictly dependent on the organization of the visual material on the display. However we make use of two distinct levels of representation: the plot level and the presentation level [Galyean, 1994]. At the plot level the orchestrator of the environment records the past events and plans ahead while the user's input continuously influences the presentation. Although the user is engaged in exploring the HyperPlex, the interactive form of our system is not uniquely "navigational" [Laurel, 1989]. Where the user goes does affect the world and the user's previous choices of the content also determines the subsequent material presented. The "narrative form" of the HyperPlex is reinforced by the personalization of the visual objects in the display that try to catch the user's attention in order to attain their goal of being seen. Moreover we're planning to have a version of the HyperPlex where many users explore the environment at the same time. Each user would "connect" to the world from its own location and engage in a game with the other users present at the same time. An interesting solution to having many users sharing the same environment that we would like to explore is the one proposed by [Ishii *et al.*, 1994]. The key design idea of their collaborative medium is to make use of translucent overlay to combine the workspaces of the different users together with the users' image. Other solutions that involve interactive control of the camera in the virtual world are also being taken into consideration [Drucker *et al.*, 1992].

The design of the user interface is based on the analysis of popular computer games [Crawford, 1990]. The art of game design comes in constructing a set of different possible interactions with the environment. "The difference between the New Hollywood and the Old is that computer games are 'interactive cinema' in which the game player takes on the role of the protagonist" [Friedman, 1995]. However most of the adventure or explorative type of computer games currently on the market do not have a "content orchestrator" or story-telling system. They import a narrative structure from popular stories and reduce the narrative trajectory of the user to a succession of enchanted worlds to explore [Fuller and Jenkins, 1995]. Our approach aims to orchestrate the presentation of the visual material in the HyperPlex with respect to the content and the temporal structure [Davenport *et al.*, 1993], all embedded in a game. Users communicate with each other through magic mirrors, going to meeting rooms and can exchange messages and objects having a virtual dog traveling through the different users' environment.

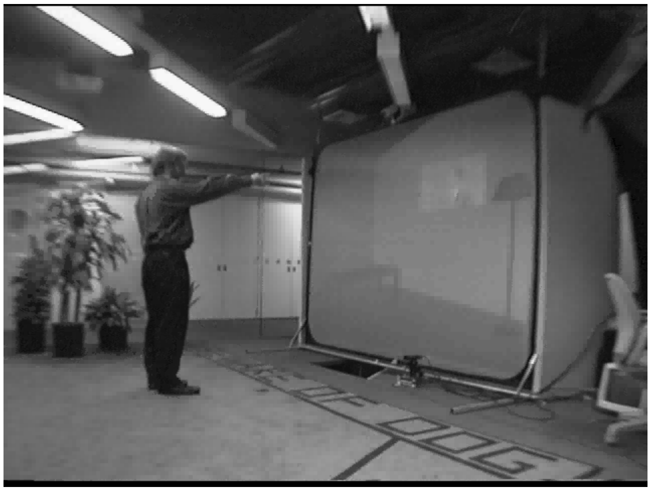

Figure 3: User interacting with a movie clip.

5 Human-Computer Communication for Dynamic Displays

When navigating in a virtual environment and meeting both virtual and real characters the use of a keyboard or mouse to give an input to the system can be heavily limiting. First of all, in certain interface tasks using a gesture, rather than clicking or choosing from a menu, can give the user a better feeling of the responsiveness of the system. Also there are tasks that can be given uniquely with gesture input [Kurtenbach and Hulteen, 1990]. In our system where the user is constantly interacting with a big screen so as to create an immersive type of environment the use of (audio)visual gestures to interact with the environment becomes a key element of the human-computer communication system (see fig. 3).

6 Interaction and Gesture Recognition

The interactive environment interface is built to be entirely non-invasive. The use of a computer vision system to measure the user eliminates the need to harness the user with many sensors and wires. A large display format allows an immersive experience without the need for head-mounted displays and opens the environment up to multiple users [Russell *et al.*, 1995].

The vision system is composed of several layers. The lowest layer uses adaptive models to segment the user from the background. This allows the system to track users without the need for chromakey backgrounds or special garments. The models also identify color segments within the users silhouette (see fig. 4). This allows the system to track important features (hands) even when these features aren't discernible from the figure-ground segmentation. This added information may make it possible to deduce general 3D structure of the user: allowing better gesture tracking at the next leyer.

The next layer uses the information from segmentation and blob classification to identify interesting features: bounding box, head, hands, feet, and centroid. These

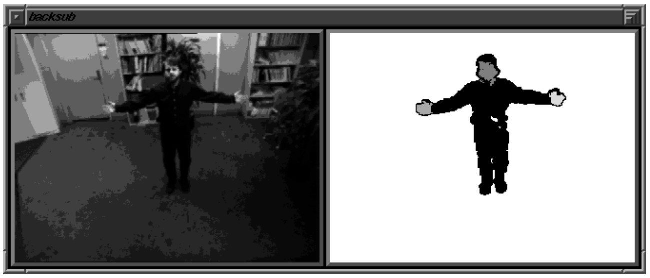

Figure 4: Backsub window showing a dithered sketch of the input video, figure/ground segmentation, and blob classifications (grey reagions within the foreground silhouette).

features can be recognized by thier characteristic impact on the silhouette (high edge curvature, occulsion) and *a priori* knowledge about people (heads are usually on top).

The highest layer then uses these features, combined with knowledge of the human body, to detect significant gestures. Audio processing included at the various levels will allow the system to use knowledge of human dialog to better recognize both audio and visual gestures.

These gestures become the input to the behavioral systems of the agents in the simulated environment. This abstraction allows the environment to react to the user on a higher, more meaningful and inflected level (see fig. 5). It can also allow us to avoid the distracting lag inherent in many other immersive systems.

7 Implementation

Each user station is comprised of several computers and other hardware. A two-processor, R4400-based SGI Onyx computer with a Reality Engine graphics board and Sirius video board generates the HyperPlex world (graphics, movies, and behavior) and displays the results on an 8' by 10' projection display. A video camera above the screen views the 15' by 15' workspace in front of the display. An R4400-based SGI *Indigo*^{2} computer with a Galileo video board uses the input from this camera to track the user and interpret gestures. The two machines communicate via TCP sockets.

Multiple user stations will communicate via TCP. Each station will maintain it's own copy of world state and will generate and display its own instantiation locally. Communication will involve changes to world state, audio from other users, and possibly low frame rate video images of other users.

8 Conclusion

The HyperPlex system is an interactive spatially-organized browser. The system is presented as a movie browser but this idea can be easily extended to a Virtual Museum, Electronic Exhibition Space or Photographic Magazine exploring system. Both *serious* and *game* versions of HyperPlex can be built according to the type of target user(s).

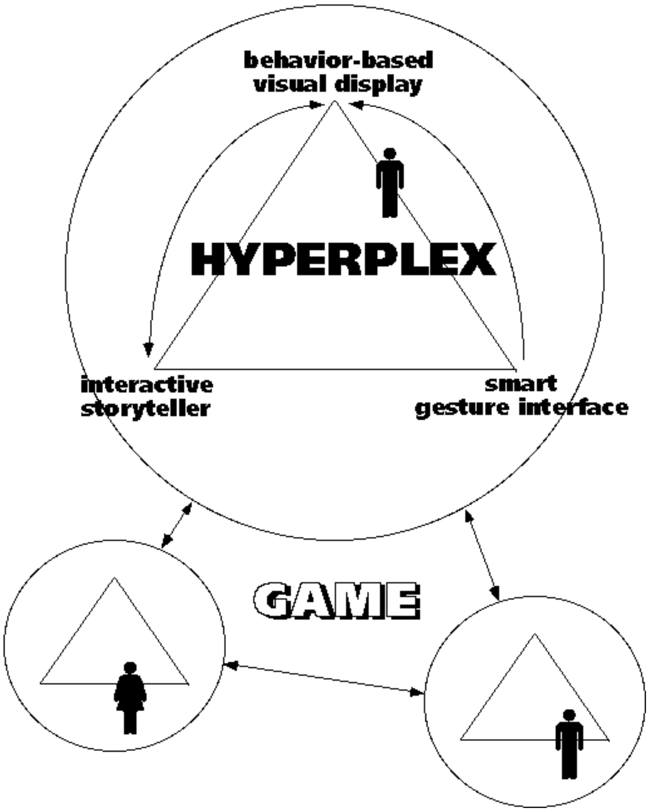

Figure 5: System architecture of the HyperPlex

Acknowledgments

The authors would like to thank Ken Russell for his tireless implemention efforts; Chris Graczyk and Greg Anderson for designing the environments;and Iki Nakagawa and Taketo Nakanishi for their work on the visual material. Tinsley Galyean for his suggestions on interactive story-telling. Thad Starner for being so ALIVE. Trevor Darrell, Ali Azarbayejani, and Dave Becker for their work on the multi-class backsub. Most of all we'd like to thank Bruce Blumberg for providing the behavior code, and for being so cool.

References

[Blumberg, 1994a] B. Blumberg. Action-selection in hamsterdam: Lessons from ethology. In *The Proceedings of the 3rd International Conference on the Simulation of Adaptive Behavior*, Brighton, August 1994.

[Blumberg, 1994b] B. Blumberg. Building believable animals. In *Proceedings of the 1994 AAAI Spring Symposium on Believable Agents*, 1994.

[Crawford, 1990] C. Crawford. *Lessons from Computer Game Design*, pages 103–111. Addison-Wesley, 1990.

[Darrell *et al.*, 1994] Trevor Darrell, Pattie Maes, Bruce Blumberg, and Alex Pentland. A novel environment for situated vision and behavior. In *Proc. of CVPR–94 Workshop for Visual Behaviors*, pages 68–72, Seattle, Washington, June 1994.

[Davenport *et al.*, 1993] G. Davenport, R. Evans, and Mark Halliday. Orchestrating digital micrmovies. *LEONARDO*, 26(4):283–288, 1993.

[Davenport, 1993] G. Davenport. My storyteller knows me: The challenge of interactive narrative. In *IDATE Conference:Investing in the Digital Image, Personal and Interactive Television*, pages 516–520, 1993.

[Drucker *et al.*, 1992] S.M. Drucker, T.A. Galyean, et al. Cinema: A system for procedural camera movements. In *1992 Symposium on Interactive 3D Graphics*, pages 67–70, Boston, MA, 1992. ACM SIGGRAPPH. Special Issue.

[Friedman, 1995] T. Friedman. *Making Sense of Software: Computer Games and Interactive Textuality*, chapter 4, pages 73–89. Sage, 1995.

[Fuller and Jenkins, 1995] M. Fuller and H. Jenkins. *Nintendo and New World Travel Writing: A Dialogue*, chapter 3, pages 57–72. Sage, 1995.

[Galyean, 1994] T. Galyean. Narrative guidance of interactivity, the link between plot and presentation. MIT Media Lab Ph.D. Thesis Proposal, 1994.

[Ishii *et al.*, 1994] H. Ishii, M. Kobayashi, and K. Arita. Iterative design of seamless collaboration media. *Communications of the ACM,*, 37(8), August 1994.

[Johnson, 1995] M.B. Johnson. Wavesworld: A testbed for 3d, semi-autonomous animated characters. http://wave.www.media.mit.edu/people/wave/PhDThesis/outline.html, Mar 1995. MIT Media Lab Ph.D. dissertation.

[Kurtenbach and Hulteen, 1990] G. Kurtenbach and E.A. Hulteen. *Gestures in Human-Computer Communication*, pages 103–111. Addison-Wesley, 1990.

[Lasseter, 1987] J. Lasseter. Principles of traditional animation applied to 3d computer animation. *Computer Graphics*, 21(4), July 1987.

[Laurel, 1989] B. Laurel. A taxonomy of interactive movies. *New Media News*, 3(1), 1989.

[Maes, 1995] P. Maes. Artificial life meets entertainment: Interacting with lifelike autonomous agents. *to appear in the Communications of the A.C.M., Special Issue on Novel Applications of A.I.*, Mar 1995.

[Russell *et al.*, 1995] Kenneth Russell, Thad Starner, and Alex Pentland. Unencumbered virtual environments. In *IJCAI-95 Workshop on Entertainment and AI/Alife*, 1995. submitted.

[Youngblood, 1989] G. Youngblood. Cinema and the Code, LEONARDO. *Computer Art in Context Supplemental Issue*, pages 27–30, 1989.