CyberBELT: Multi-Modal Interaction with a Multi-Threaded Documentary

Joshua Bers, Sara Elo, Sherry Lassiter, David Tamés
MIT Media Laboratory
20 Ames Street
Cambridge, MA 02139, USA
tel: 1.617.253.9401
elo@media.mit.edu

ABSTRACT
CyberBELT allows a viewer to interact with a multi-threaded documentary using a multi-modal interface. The viewer interacts with the documentary by speaking, pointing and looking around the display. The viewer selects the threads of the story to follow or lets the system navigate through the story. Feedback from the viewer evolves the story to present concepts she is interested in. We discuss the suitability of combining multi-modal interaction and multi-threaded narrative.

KEYWORDS
multi-modal interaction, interactive documentary, information exploration, dynamic story-telling system

SCENARIO
'I walk into the room. On a wall-size display, Seymour Papert tells with shiny eyes: 'Cybernetics helps us learn about life.' Then Evelyn Fox Keller, Jay Forrester, Oliver Selfridge and Slavan Gerovitch appear here and there on the display. As I let my gaze wonder, the characters unveil text seducing my gaze. I look at Evelyn, smiling wisely, and the text appears: 'Evelyn Fox Keller gives her point of view: Cybernetics embraces the complexity found in nature.' As if knowing I had looked at her, she begins to talk to me. I am immersed in a sea of conversations, words, and images...'

INTRODUCTION
The scenario above is a futuristic view of the interaction with the CyberBELT system and the documentary on the history of Cybernetics. Needless to say, the interaction is more cumbersome with current technology than depicted in the scenario. The viewer needs to wear an uncomfortable eye-tracker, a microphone and a data glove. Despite the discomfort of the armor, the multi-modal interaction allows the viewer to use her whole body to control the flow of the documentary and to convey her preferences and interests to the system. This work in progress brings together, to our knowledge for the first time, multi-modal interaction with an multi-threaded documentary. We show that a multi-modal interaction suits well the exploration of threads of a narrative.

MULTI-MODAL INTERACTION
Previous work
For a 'full-body' communication with the CyberBELT system, the viewer uses three technologies: a speech recognizer, an eye-gaze tracker, and data gloves. Previous work in multi-modal interfaces includes the early 'Put That There' project [1] where the viewer used speech and gesture to manipulate objects on a display. The system resolved ambiguous references to objects by combining the information from speech and gesture. Work by Starker [2] used eye-tracking information to reveal detail about objects or groups of objects gazed at by the viewer. All three technologies have been integrated in applications such as [3].

Multi-Modality in CyberBELT
In CyberBELT the viewer uses the three modalities to the extent that she wishes. She takes on an active or a passive role: she can select threads of the story and explore the different themes at her own pace or she can let the system navigate. When presented with a collage of video icons representing the possible threads to follow, she selects one by saying "go there" and pointing or looking at the appropriate icon. If she dislikes a clip she can interrupt it to return to the previous collage of clips by saying "back", to go to the next collage by saying "skip", or to go to the next chapter by saying "next". If the viewer remains passive and does not select a thread, CyberBELT moves on to the next chapter.

CyberBELT uses the eye gaze of the viewer at two phases of the interaction. When the viewer is looking at a collage of icons to select the next thread, the system reveals explanatory text under any icon she is looking at. It explains how the thread will relate to the previous clip and reveals a brief quote. The system reveals text until the viewer selects a thread or until the selection period times out and the story moves to the next chapter. CyberBELT also uses eye tracking information to gauge the viewer's level of attention while watching a clip. If the viewer's gaze wonders outside

the playing clip, she is not attentively watching the clip; we assume she is not interested.

While speaking and pointing are active feedback, eye-gaze tracking provides passive feedback. Eye-tracking is a powerful and non-invasive way to monitor where the user is focusing her attention.

MULTI-THREADED NARRATIVE

Previous Work

Working with digital video permits flexible manipulation of clips and personalization of a narrative. Non-linear video stories complement classical linear film as they allow the viewer to explore multiple intertwined threads and characters. Aspen [4], an early interactive project, was a surrogate travel experience through the city of Aspen. The viewer navigated through photos and video clips of the city by touching the screen or moving a joystick. The transition between two clips was allowed only if the corresponding sites were adjacent in the real city. In Portraits of People Living with Aids [5] the viewer explores the different themes in the data base of interview clips. While the transitions between clips are predefined and static, the documentary is dynamic because viewers can record comments addressed to the characters. In other interactive documentaries the sequencing of clips is dynamic. For example, Train of Thought [6] uses filters to select scenes from a video data base and fills a story template.

Multi-Threaded Narrative in CyberBELT

Like Trains of Thought CyberBELT dynamically selects clips to fill a pre-defined narrative structure, however, the viewer's choices affect the ongoing story. We divided the story into chapters to assure a coherent progression from one theme to another and to give a broad view on the topic even to the viewer who follows the shortest thread and watches only one clip from each chapter. While the overall structure is pre-defined, the sequencing of clips within a chapter is dynamic.

As full content representation of video is a complex task [7], we annotated the clips with the speaker, the topics of the conversation and the point of view of the speaker with regard to the topic. The annotated clips create a semantic web where the proximity of two clips is proportional to the number of annotations in common. The annotations permit CyberBELT to present the viewer with a collage of clips that exhibit contrast with the previous clip. Two clips that have annotations in common but mismatch on one annotation make for an interesting contrast. For example, after seeing a clip where Evelyn Fox Keller regrets the lack of influence of Cybernetics on molecular biology, the viewer selects among clips on the influence of Cybernetics on other disciplines, from opposing points of view, or by different speakers.

As mentioned earlier, the documentary evolves with the feedback from the viewer. Initially, all clips are equally weighted, or have equal likelihood to get proposed to the user. As the viewer selects or interrupts clips and as her gaze is monitored, the system alters the weights of the clips. When the weight of a clip changes the weights of all other clips change to the degree that they have annotations in common. Thus, the weights model the viewer's preferred concepts, not preferred clips. For example, if the viewer frequently selects clips with Jay Forrester, the likelihood of all Forrester clips in the data base increases. At subsequent decision points in the documentary CyberBELT is more likely to propose threads with Jay Forrester.

The clip weights, or the model of the viewer, can be saved at the end of a session and reloaded at the beginning of another. By loading another viewer's weights, one can watch a documentary on Cybernetics that reflects the other viewer's preferences. The system could also model the preferences of a group of viewers.

CONCLUSION: SUITABILITY OF MULTI-MODAL INTERACTION IN MULTI-THREADED NARRATIVE

Multi-modal interaction is a suitable way to interact with a multi-threaded documentary. It facilitates controlling the documentary and modeling the viewer's interests. With speech and gesture the viewer can express in a natural way her desires to the system. Whereas a mouse interface could replace speaking and pointing, no technology could replace eye-tracking to provide passive feedback about the user's focus of attention. Eye-tracking enables CyberBELT to know what to show next to the viewer and which items on the display to expand with an explanation. Eye-tracking is particularly useful in a multi-threaded documentary where the viewer's options are spatially distributed on the display.

On the other hand, an interactive documentary is a well suited application for multi-modal interaction. A multi-threaded documentary presents a complex data space to the viewer. The viewer traverses the space by exploration. She considers different options without knowing ahead of time exactly where she wants to go. Since speed is not important in an exploratory mode, an immersive slow-moving experience through a multi-modal interface is ideal.

In the future we plan to have people watch the multi-threaded documentary on Cybernetics and to record their reactions.

References

1. Bolt, Richard A. 'Put That There: Voice and Gesture at the Graphics Interface'. Computer Graphics, 14(3), 262-270, 1980.
2. Starker, India and Bolt, Richard A. 'A Gaze-Responsive Self-Disclosing Display' in CHI-90 Proceedings, Association of Computing Machinery, 1990.
3. Koons, David B., Sparrell, Carlton J. and Thorisson, Kristinn R. 'Integrating Simultaneous Input from Speech, Gaze and Hand Gestures'. In Intelligent Multi-Media Interfaces. Ed. M. T. Maybury. AAAI/MIT Press, 1993.
4. Mohl, Robert. 'Cognitive Space in the Interactive Movie Map: An Investigation of Spatial Learning in Virtual Environments'. Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, Massachusetts, 1982.
5. Reed, Hazen. 'Portraits of People Living with Aids, an Interactive Documentary'. In CHI-94 Proceedings, Association of Computing Machinery, 1994.
6. Halliday, Mark. 'Digital Cinema, An Environment for Multi-Threaded Stories'. Master's Thesis, Massachusetts Institute of Technology, Cambridge, Massachusetts, 1985.
7. Davis, Marc. ' Content Representation for Video'. Proceedings of AAAI-94, 1994.