Coordination and Fusion in Multimodal Interaction

Mark Maybury
Information Technology Center
The MITRE Corporation
202 Burlington Road
Bedford, MA 01730, USA
maybury@mitre.org
www.mitre.org/resources/centers/it

When we converse with one another, we utilize an array of media to interact, including spoken language, gestures, and drawings. We exploit multiple sensory systems or modalities of communication including vision, audition, and taction. Providing machines with the ability to interpret multimedia input and generate coordinated multimedia output promises benefits including:

- More *efficient* interaction -- enabling faster task completion with less work.
- More *effective* interaction -- doing the right thing at the right time, tailoring the content and form of interaction to the context of the user, task, and dialogue.
- More *natural* interaction – supporting fused spoken, written, and gestural interaction, as found in human-human communication.

Our research has focused on intelligent systems that exploit multiple media and modes.

Multimedia Input Analysis

Whereas traditional interfaces support sequential and unambiguous input from devices such as keyboard and conventional pointing devices (e.g., mouse, trackpad), intelligent multimodal interfaces (see www.mitre.org/resources/centers/it/maybury/tutorial.html) relax these constraints and typically incorporate a broader range of input devices (e.g., spoken language, eye and head tracking, three dimensional gesture). For example, they support asynchronous, ambiguous, and inexact input by applying more sophisticated analysis of input. These systems allow the resolution of multimedia references, for example enabling the user to say "Put that there" while gesturing to a map, by correlating eye and hand gestures with the deictic expressions “that” and “there” (Burger and Marshall 1993). Intelligent interfaces can also exploit models of the media, user, discourse, and task and automatically detect and correct errors.

Multimedia Output Generation

Traditional interfaces draw upon canned presentations (e.g., windows, menus, dialogue boxes). In contrast, automated interface and presentation generation systems reason about communication plans and intentions, select content to achieve given communicative goals, design the presentation, allocate and coordinate information across media (e.g., typed or spoken language, graphics, gesture), realize media, and lay them out. In earlier research, we designed communicative actions for automated multimedia generation (Maybury 1991).

Multimedia Information Access

We have also focused on the ability to provide content-based access to multimedia information sources (e.g., text, audio, video, maps). We have investigated key tasks such as multistream segmentation, indexing, extraction, summarization, visualization, navigation and retrieval. Our advanced news on demand system, MITRE’s Broadcast News Navigator, includes content-based processing of integrated text, images, audio, and video (www.mitre.org/resources/centers/it/g061/bnn/mmhomeext.html).

Evaluation

Benchmarking, hypothesis testing and repeatable experiments are fundamental to any scientific endeavor. In an empirical study assessing the performance of analysis using multimedia vs. monomedia presentations of news (Merlino and Maybury 1999), we discovered that mixed media presentations can reduce task time and increase task quality. To advance further, we need to move toward community-based evaluation using standard multimedia corpora and tasks.

REFERENCES

1. Burger, J., and Marshall, R. 1993. The Application of Natural Language Models to Intelligent Multimedia. In (Maybury 1993), 167-187.
2. Maybury, M. T. 1991. Planning Multimedia Explanations using Communicative Acts, Proceedings of AAAI-91, 61-66. AAAI/MIT Press, Menlo Park, CA.
3. Maybury, M. T. (ed.) 1993. *Intelligent Multimedia Interfaces.* Menlo Park: AAAI/MIT Press. (www.aaai.org:80/Press/Books/Maybury1)
4. Maybury, M. T. (ed.) 1997. *Intelligent Multimedia Information Retrieval.* Menlo Park: AAAI/MIT Press. (www.aaai.org:80/Press/Books/Maybury2/)
5. Maybury, M. and Wahlster, W. (eds.) 1998. *Readings in Intelligent User Interfaces*. Morgan Kaufmann: Menlo Park, CA. (www.mkp.com/ books_catalog/1-55860-444-8.asp)
6. Merlino, A. and Maybury, M. 1999. An Empirical Study of the Optimal Presentation of Multimedia Summaries of Broadcast News. In Mani, I. and Maybury, M. (eds.) *Automated Text Summarization*, MIT Press. pp. 391-401.