INFORMEDIA DIGITAL VIDEO LIBRARY

Michael Christel, Scott Stevens, and Howard Wactlar

Carnegie Mellon University
Pittsburgh, PA 15213
mac@sei.cmu.edu

Vast digital libraries of information will soon be available on the nation's Information Superhighway as a result of emerging technologies for multimedia computing. These libraries will profoundly impact the conduct of business, professional, and personal activity. However, it is not enough to simply store and play back video (as in currently envisioned commercial video-on-demand services); to be most effective, new technology is needed for searching through these vast data collections and retrieving the most relevant selections.

The Informedia Project is developing these new technologies for data storage, search, and retrieval, and in collaboration with QED Communications is embedding them in a video library system for use in education, training, and entertainment. The Informedia Project leverages efforts from many Carnegie Mellon University computing research activities, including:

- Sphinx-II speech recognition
- Image Understanding Systems Laboratory
- Center for Machine Translation (information retrieval)
- Software Engineering Institute (information modeling)

The Informedia Project is developing intelligent, automatic mechanisms that provide full-content search and retrieval from digital video, audio, and text libraries. The project integrates speech, image, and language understanding for the creation and exploration of such libraries. The initial library will be built using WQED's video assets.

LIBRARY CREATION

The Informedia system uses Sphinx-II to transcribe narratives and dialogues automatically. Sphinx-II is a large vocabulary, speaker-independent, continuous speech recognizer developed at Carnegie Mellon. With recent advances in acoustic and language modeling, it has achieved a 95% success rate on standardized tests for a 5000-word, general dictation task. By relaxing time constraints and allowing transcripts to be generated off-line, Sphinx-II will be adapted to handle the video library domain's larger vocabulary and diverse audio sources without severely degrading recognition rates.

In addition to annotating the video library with text transcripts, the videos will be segmented into smaller subsets for faster access and retrieval of relevant information. Some of this segmentation is possible via the time-based transcript generated from the audio information. The work at CMU's Image Understanding Systems Laboratory focuses on segmenting video clips via visual content. Rather than manually reviewing a file frame-by-frame around an index entry point, machine vision methods that interpret image sequences can be used to automatically locate beginning and end points for a scene or conversation. This segmentation process can be improved through the use of contextual information supplied by the transcript and language understanding. Figure 1 gives an overview of the InforMedia system.

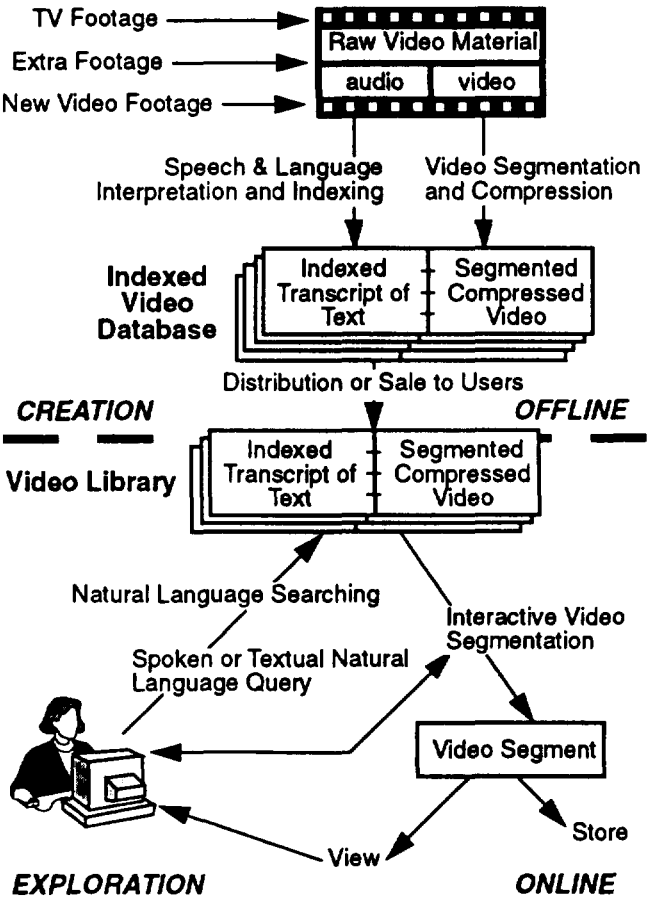

Figure 1. Informedia™ System Overview

LIBRARY EXPLORATION

Finding desired items in a large information base poses a major challenge. The Informedia Project goes beyond simply searching

the transcript text and will, in addition, apply natural-language understanding for knowledge-based search and retrieval. One strategy employs computational linguistic techniques from the Center for Machine Translation for indexing, browsing, and retrieving based on identification of noun phrases in a written document. Other techniques from the Center include statistical weighting, term selection heuristics, and natural-language processing. More complex than individual words, these linguistic units provide a conceptually richer domain for subsequent processing.

The Informedia system is extending this technology for spoken language and applying it to correct and index the automatically-transcribed soundtracks. Other tasks will include identification of topics and subtopics in transcript collections, and a rich natural language retrieval interface. A second thrust is developing robust techniques for matching transcribed words and phrases that sound alike when spoken. This integrated approach will significantly increase the Informedia system's ability to locate a particular video segment quickly, despite transcription errors, inadequate keywords, and ambiguous sounds.

Along with improving query capabilities, the Informedia Project is researching better ways to present information from a given video library. Interface issues include helping the user identify desired video when multiple objects are returned, adjusting the length of video objects returned, and letting the user quickly skim video objects to locate sections of interest.

Cinematic knowledge can enhance the composition and reuse of materials from the video library. For example, the library may contain hours of interview footage with experts in a certain topic area. Rather than simply presenting a series of disassociated video windows in response to user queries, this interview footage could be leveraged to produce an interface in which the user becomes the interviewer. The natural language techniques mentioned above are used to parse the user's questions, and scenes from the interview footage are composed dynamically to present relevant answers. Such an interface is designed to engage the user into more fully exploring and interacting with the video library in search of information as an active interviewer.

LIBRARY DEMONSTRATION

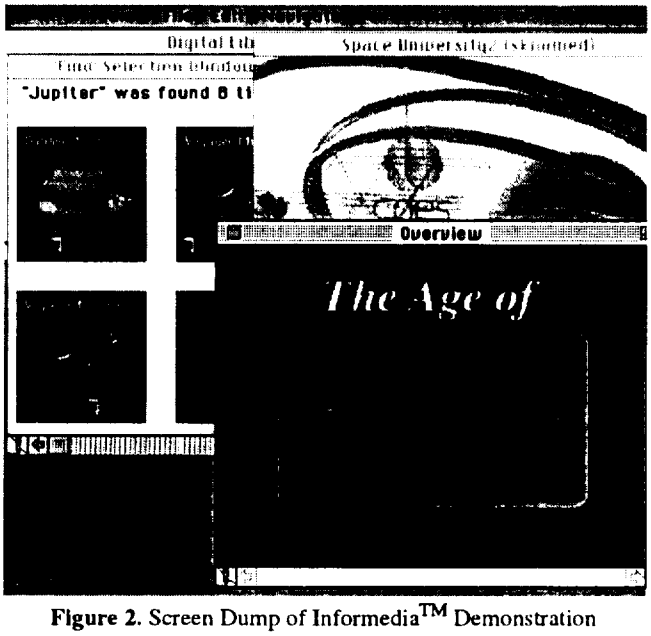

Figure 2. Screen Dump of Informedia™ Demonstration

The Informedia Project currently has a demonstration based on a small database (one gigabyte) of text, graphics, video, and audio material drawn from WQED's "Space Age" series. A sample display of this demonstration appears as Figure 2 for the reader's reference. The demonstration was carefully scripted to illustrate the following points (listed in temporal order, as they occur in the demonstration):

- parsing the user's input according to an appropriate grammar for that domain allows for more natural, less cumbersome queries
- natural language understanding of both a user's query and the video library transcripts enables the efficient retrieval of relevant information
- the location *within* a video object is identified relevant to a user query via the text transcript
- the video "paragraph", or size of the video object, is determinable based on language understanding of the transcript and image understanding of the video contents
- a larger video object can be "skimmed" in an order of magnitude less time, while coherently presenting all of the important information of the original object
- video clips can be reused in different ways, e.g., to create an interactive simulated interview, as shown in Figure 3

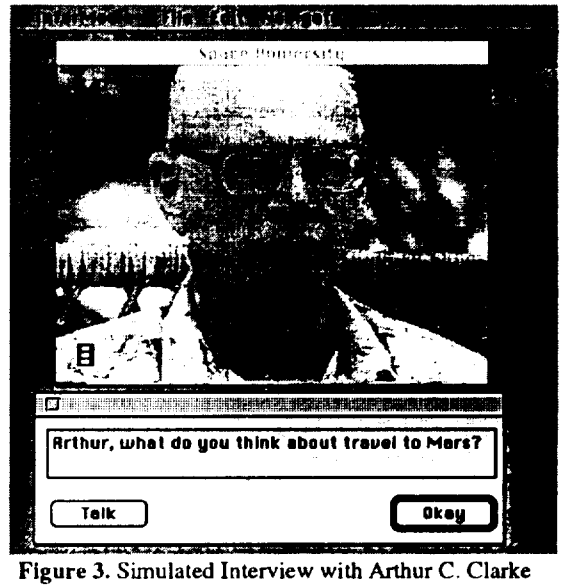

Figure 3. Simulated Interview with Arthur C. Clarke

The demonstration does not show Sphinx-II in action performing automatic transcription nor does it document the process of video segmentation or natural language parsing. It shows the benefits of such automatic indexing and segmentation, illustrating the accurate search and selective retrieval of audio and video materials appropriate to users' needs and desires. It shows how users can preview as well as scan video at variable rates of speed and presentation, akin to skimming written material. Finally, it demonstrates the concept of combining speech, language, and image understanding technologies to create entertaining educational experiences.