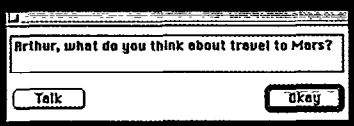
Scott Stevens, Michael Christel, and Howard Wactlar

Improving Access to Digital Video



ast 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.

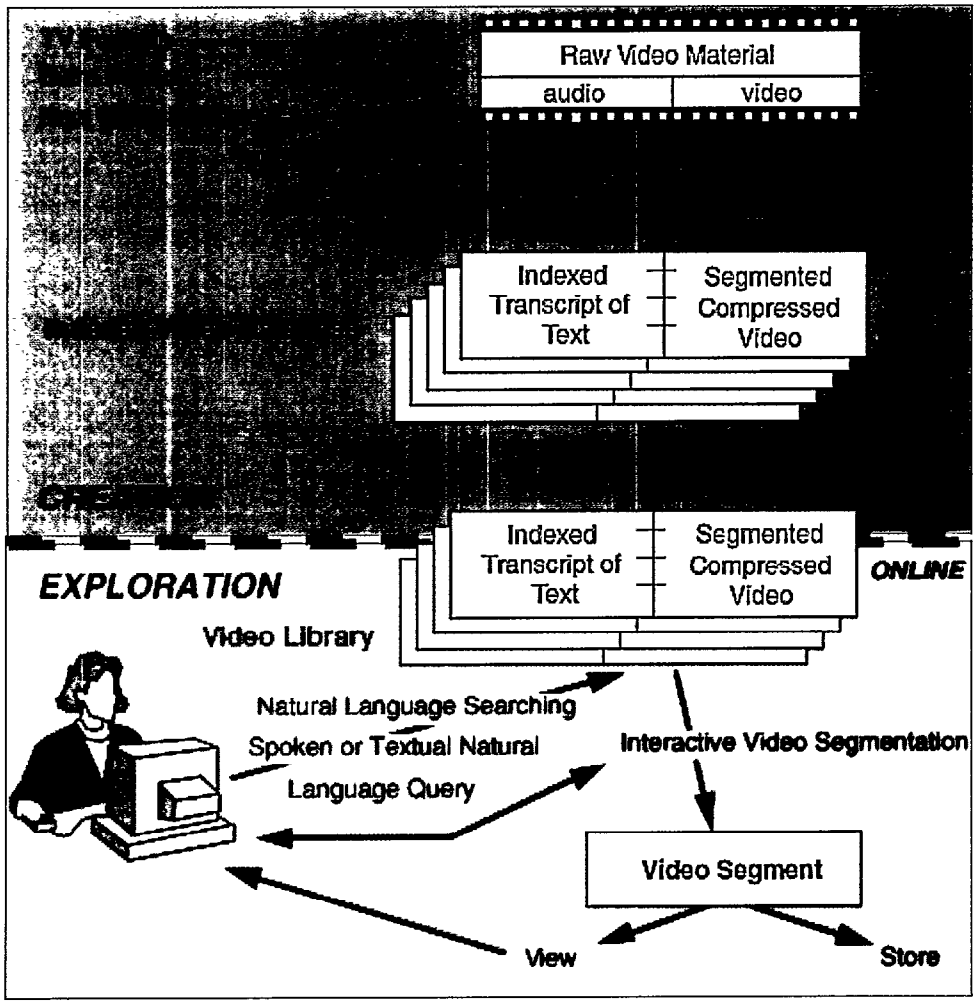


Figure 1
Informedia™ System Overview

The Informedia Project is developing these new technologies for data storage, search, and retrieval, and in collaboration with WQED Pittsburgh 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 (vision and image understanding)
- Center for Machine Translation (information retrieval and natural language processing)
- Software Engineering Institute (information modeling and digital video user interface development)

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 and the British Open University'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 90% success rate on standardized tests for a 20,000-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 video's will be segmented into smaller subsets for faster access and retrieval of relevant information. Some segmentation is possible via the time-based transcript generated from the audio information. Segmenting video clips via visual content is also being performed based on work at CMU's Image Understanding Systems Laboratory. Rather than manually reviewing a file frame-by-frame around an index entry point, machine vision methods that interpret image sequences are used to automatically locate beginning and end points for a scene or conversation. This segmentation process can be improved through contextual information supplied by the tran-

script and language understanding. Figure 1 gives an overview of the InforMedia system.

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 is, in addition, applying 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 text. 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 semantically 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. Informedia's integration of speech recognition, natural language, and image understanding technologies creates a natural, literally invisible first-order user interface for searching large corpora of digital video. Nonetheless, significant user interface issues remain. Three principal issues with respect to searching for information are: how to aid users in the identification of desired video when multiple objects are returned; how to let the user adjust the size of the video objects returned; and how to let the user quickly skim the video objects to locate sections of interest. With respect to reuse of video objects tools that go beyond editing of video are required and include expert assistance in visual and temporal organization of video. Solutions to these problems require an intimate understanding digital video and the development of new modes of interfaces based on this model.

To develop the needed solutions initial studies are focusing on presentation and control interfaces:

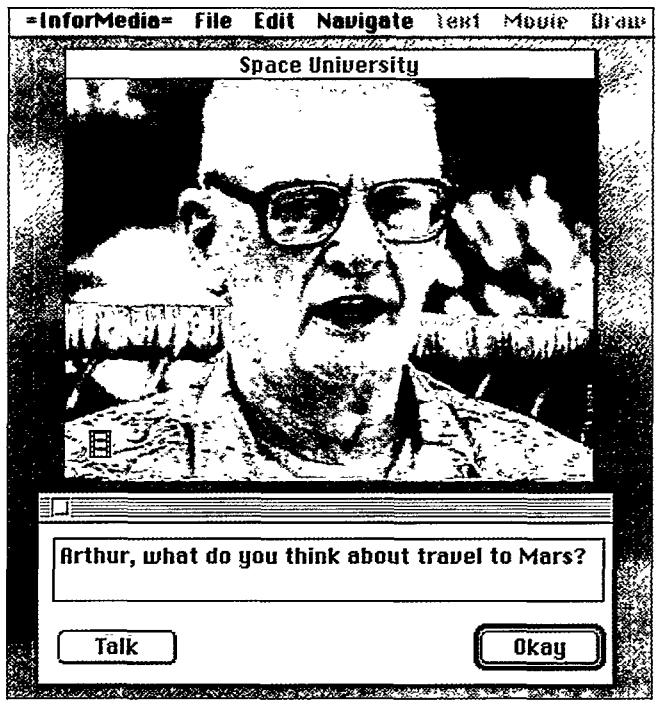
Parallel presentation. When a search contains many hits, the system will simultaneously present icons, intelligent moving icons, imicons, and full motion sequences along with their text summarization. To develop heuristics for imicon creation, empirical studies are being performed to determine the number of unique scenes needed to represent a video chunk; the effect of camera movements and subject movements on the selection of images to represent each scene; and the best rate of presentation of images. Users will likely react differently to a screen populated by still images than the same number of moving images. Therefore studies will also be used to identify the optimal number and mix of object types.

Context-sizing. This simulated slide switch enables the user to adjust the ""size" (duration) of the retrieved video/audio segments for playback. Here, the "size" may be time duration, but more likely it will be abstract chunks where

The Informedia Project goes beyond simply searching the transcript text and is, in addition, applying natural-language understanding for knowledge-based search and retrieval.

information complexity or type will be the determining measure. This research will investigate the appropriate metaphors to use when the "size" the user is adjusting is abstract content. Here, empirical studies will be used to help determine typical visual "paragraphs" for different materials. For example, it is well know that higher production value video has more shot changes per minute than, for example, a video taped lecture. And although it is visually richer, it may be linguistically less dense. These studies will help determine unique balance of linguistic and visual information density appropriate for different types of video information. Here we will research what it means, from both interface development and a search methods, to permit the user to say "I want more background on each subject returned."

Figure 2
User as Interviewer



Skimming. This simulated analog rotary-dial will interactively control the rate of playback of a given retrieved segment, at the expense of both informational and perceptual quality. One could also set this dial to skim by content, e.g., visual scene changes. Video segmentation will aid this process. By knowing where scenes begin and end the Informedia system will perform high speed scans of digital video files by presenting quick representations of scenes. This can be an improvement over jumping a set number of frames, since scene changes often reflect changes in organization of the video much like sections in a book. Empirical studies will be conducted to determine the rate of scene presentation that best enables users searches and the differences, if any, of image selection for optimal scans compared to image selection for the creation of imicons.

Once users identify video objects of interest they will need to be able to manipulate, organize, and reuse the video. To effectively reuse video assets, the user will need to combine text, images, video and audio in new and creative ways. To become competent writers, we spend years learning formal grammar. The language of film is both rich and complex and deep cinematic knowledge, the grammar of video, cannot be required of users. To aid the user, Informedia can use cinematic knowledge to 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 as in Figure 2. 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 as an active interviewer in search of information.

The Informedia Project's first version drew on a small database (three gigabyte) of text, graphics, video, and audio material drawn from WQED's "Space Age" series, distinguished lectures in computer science, and software engineering training lectures. A sample display appears as Figure 3. Early Informedia user feedback suggests:

- parsing the user's input according to an appropriate grammar for that domain allows for more natural, less cumbersome queries

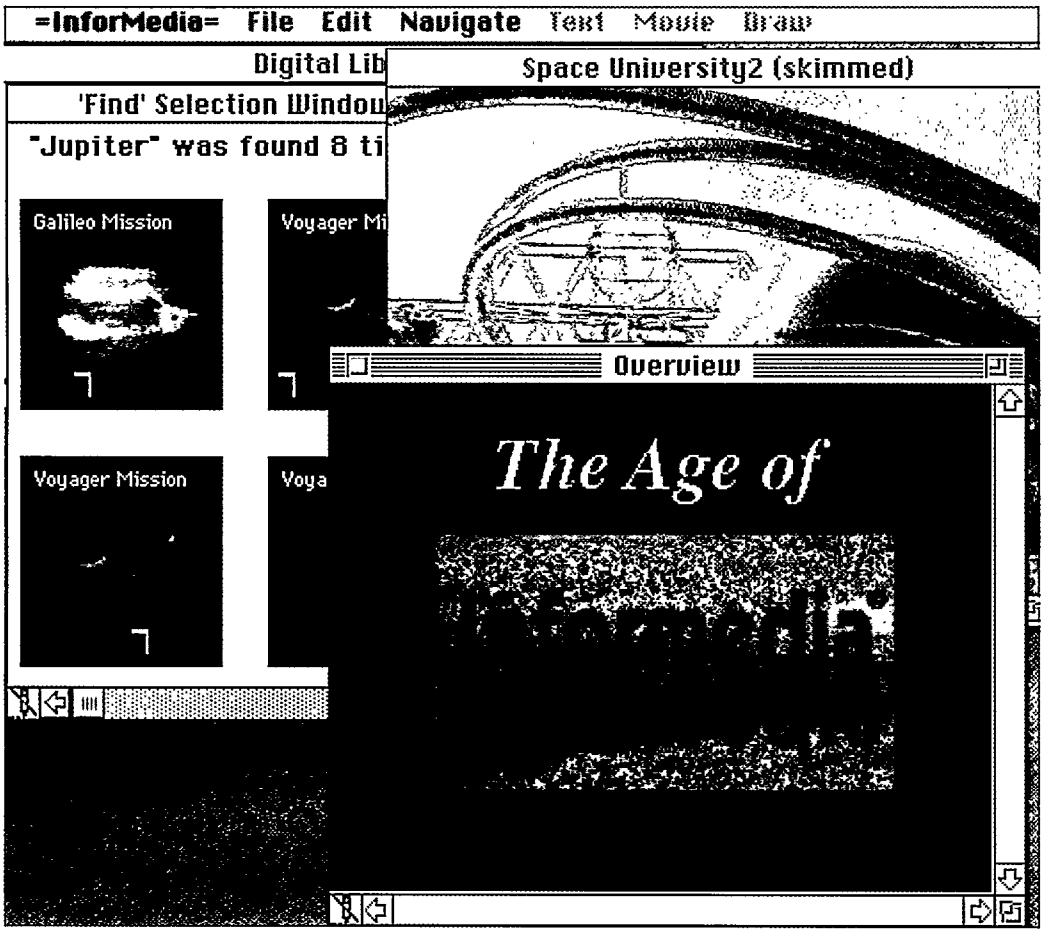


Figure 3
Sample
Display

Author's present address: Scott Stevens Senior Member of the Technical Staff, Software Engineering Institute, Carnegie Mellon University, Pittsburgh, PA 15213-3890
sms@sei.cmu.edu

- natural language understanding of both a user's query and the video library transcripts enables the efficient retrieval of relevant information
- locating information within a video object in response to a user query is appealing.
- the video "paragraph", or size of the video object, determinable based on language understanding of the transcript and image understanding of the video contents, provides fast, relevant information
- larger video objects can be "skimmed" in an order of magnitude less time, while coherently presenting the important information of the original object
- video clips can be reused in different ways, e.g., to create an interactive simulated interview, delivering information in a highly motivating form.

This anecdotal feedback shows the benefits of 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 practicality of combining speech, language, and image understanding technologies to create entertaining educational experiences.

Over the next several years the Informedia library project will establish a large, on-line digital video library by further developing these intelligent, automatic mechanisms to populate the library and allow for full-content and knowledge-based search and retrieval via desktop computer and metropolitan area networks. Critical to Informedia's success will be the design, development, and testing of the system's interface and the studies of how users interact with the digital video information, the access of which is made possible by Informedia.

Mike Christen
Member of Technical Staff, Software Engineering Institute, Carnegie Mellon University
mac@sei.cmu.edu

Howard Wactlar
Vice Provost for Research Computing and Associate Dean, School of Computer Science, Carnegie Mellon University
hdw@cs.cmu.edu

*** An earlier version of this paper appears in the ACM Multimedia'94 Video Program.**