Intelligent Access to Digital Video: Informedia Project

Howard D. Wactlar
Takeo Kanade
Michael A. Smith
Scott M. Stevens
Carnegie Mellon University

Information retrieval is an increasingly complex process, due to digital integration of video, audio, and text resources. An experimental project will explore the challenges posed by these digital video libraries.

The Informedia Digital Video Library project[1] will establish a large, on-line digital video library featuring full-content and knowledge-based search and retrieval. Intelligent, automatic mechanisms will be developed to populate the library. Search and retrieval from digital video, audio, and text libraries will take place via desktop computer over local-, metropolitan-, and wide-area networks. Initially, the library will be populated with 1,000 hours of raw and edited documentary and education videos drawn from video assets of WQED/Pittsburgh, Fairfax County (Virginia) Public Schools, and the Open University (United Kingdom). To assess the value of video reference libraries for enhanced learning at different ages, we will deploy the library at Carnegie Mellon University and local schools, from elementary school through high school.

Our approach applies several techniques for content-based searching and video-sequence retrieval. Content is conveyed in both the narrative (speech and language) and the image. Only by the collaborative interaction of image, speech, and natural-language understanding technology can we successfully populate, segment, index, and search diverse video collections with satisfactory recall and precision.

This collaborative interaction approach uniquely compensates for problems of interpretation and search in error-ridden and ambiguous data sets. We start with a highly accurate, speaker-independent, connected speech recognizer that automatically transcribes video soundtracks. A language-understanding system then analyzes and organizes the transcript and stores it in a full-text information retrieval system. This text database permits rapid retrieval of individual video segments that satisfy an arbitrary query on the basis of the words in the soundtrack and in associated annotations and credits. Image and language understanding lets us locate and delineate the corresponding "video paragraph" context through combined source information about camera cuts, object tracking, speaker changes, timing of audio and/or background music, and change in content of spoken words. Controls let the user interactively request corresponding video paragraphs to full volumes, browse the results, intelligently "skim" the returned content, and reuse the stored video objects in different ways. Figure 1 illustrates a typical user retrieval display.

The data and network architecture we have implemented provides a distributed data multilevel hierarchy and enables networking on commercial data services. To protect data rights in intellectual property and to provide security and privacy, we've incorporated network billing, variable pricing, and access control.

All digital libraries share common technical and sociological issues, attributes, features, and challenges.[2] The digital video library exacerbates many of these problems. Moreover, it generates new research challenges across diverse disciplines, beginning with automated techniques to derive

0018-9162/96/$5.00 © 1996 IEEE

semantic content directly from source material in the absence of metadata describing it. The machine-cognition-technology approach to library creation—integrating speech, image, and language understanding—confronts each such area with additional constraints and requirements, thereby necessitating novel solutions. Finally, special user interface issues relate to the creation of visual and textual abstracts, skimming, and extraction of video data for reuse.

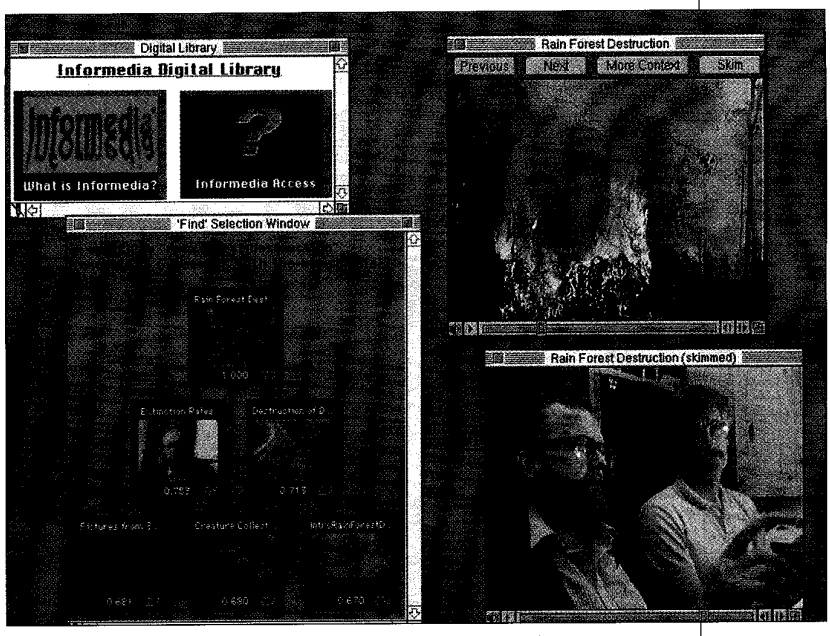

Figure 1. Typical Informedia digital library user display screen.

Assembling library content

Without suitable indexing, a collection of video material cannot serve as an information resource. Our goal of full-content search/retrieval in the Informedia library requires an automatically generated index pointing to meaningful, small clips within the videos (adjustable "video paragraphs" of two to five minutes) and yielding alternate representations and abstraction levels. Davis notes that a physical segmentation of the video data imposes a fixed segmentation of the content and a potential separation from its original context.[3] Because this may limit subsequent use of the library, our approach logically segments the library data with video paragraph markers and indices but keeps the video data intact in its original context. Our multimodal approach to generating the index and the abstractions poses difficult challenges for each of the speech, image, and language understanding technologies that we incorporate.

Speech understanding for automated transcript generation

Even though much of broadcast television is closed-captioned, most of the nation's video and film assets are not. More importantly, typical video production generates 50 to 100 times more content than what is broadcast and is thus not captioned. We therefore combine automatically generated transcripts, containing tolerable errors, with captioning (where available) for the analysis, indexing, and retrieval of multimedia data.

Unlimited-vocabulary, speaker-independent, connected-speech recognition is an incompletely solved problem. However, recent results in domain-specific applications demonstrate the promise and potential of being able to automatically transcribe spoken language with an unlimited vocabulary. Currently, our Sphinx-II system recognizes, with 90-percent accuracy in benchmark evaluations, speaker-independent, continuously spoken speech with a vocabulary of more than 60,000 words.[4] Several sources of error and variability occur in the video transcription task that must be resolved. These include

- *Music and noise mixed with speech*. FFT spectrogram data can be used to determine high-energy areas outside the human speech bandwidth. Neural-net-feature detectors of other noise types appear promising.
- *Segmentation of long fragments*. In video productions, the begin and end points for utterances are not marked. Using energy profiles for algorithms to detect breaks between utterances will help.
- *Inappropriate language models*. Adaptive language models must be incorporated that automatically change, based upon recognition likelihood in the first pass. Hints from the title, as well as from ancillary notes and annotations, may help in selecting alternative models.
- *Errorful closed-captioned data and scripts*. The use of forced alignment with language model modifications and the accounting for spontaneous speech not in the captions or script will together significantly reduce error over straight transcript alignment.
- *Acoustic modeling*. New models must be trained for noise and music, and each type must be recognized separately. Specialized audio parsers for noise, laughter, and other distinct acoustic phenomena have been developed that will enable detection and retrieval of these sounds from the audio content.[5]
- *Identification of speaker change*. Speaker gender change is straightforward. Neural nets and various pitch-dependent techniques will provide the functionality.
- *Speech recognition for keyword retrieval*. Focusing on language models for keyword recognition may improve overall accuracy of query-based retrieval where relevant subject matter is sought. Absolute correctness of the derived transcript, however, may be less important in the library search than in a man-machine conversational application.

For digital video transcription, processing time can be traded for higher accuracy. The system doesn't have to operate in real time, which permits the use of larger, continuously expanding dictionaries and more computationally intensive language models and search algorithms.

Image processing for classification, segmentation, and retrieval

Image understanding plays a critical role in Informedia for organizing, searching, and reusing digital video. When the digital video library is formed, the first requisite capability is video segmentation (or paragraphing) into a group of frames. Part of this task can be achieved with content-free image statistics such as color histograms, DCT (discrete cosine transform) coefficients, shape, and texture measures. Scene transition effects such as fades, dissolves, and cuts can also be automatically detected.[6] Although queries are expected for subject matter (comprising both image and textual content), subsequent refinement of the query might be visual, referring to image content. Examples are searches for "similar scenery" or "comparable buildings."

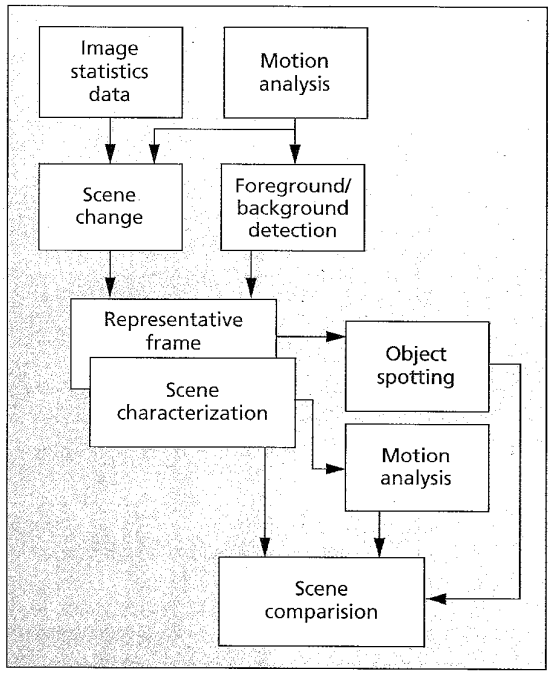

Figure 2. Informedia image-understanding video processing overview.

Video information is temporal, spatial, often unstructured, and massive. As a result, a complete solution—automatic extraction of semantic information or a general vision recognition system—is not yet feasible. Our overall approach focuses on the interrelated problems of segmentation, object detection, characterization, and similarity matching. Figure 2 depicts the various image-processing analyses that, when performed in the system, enable appropriate data characterizations, both content-free and content-based, for Informedia segmentation and search. The technical obstacles and problem approaches are summarized below.

- *Comprehensive image statistics for segmentation and indexing.* This initial segmentation can be done in a content-free manner with image statistics by detecting fast changes in them. A simple histogram difference measure is robust and efficient enough to provide accurate segmentation for detecting scene changes. An example of this is shown in the top graph of Figure 3. Once a video is identified, we extract image features like texture, color, and shape from video as attributes. While these are "indirect statis-

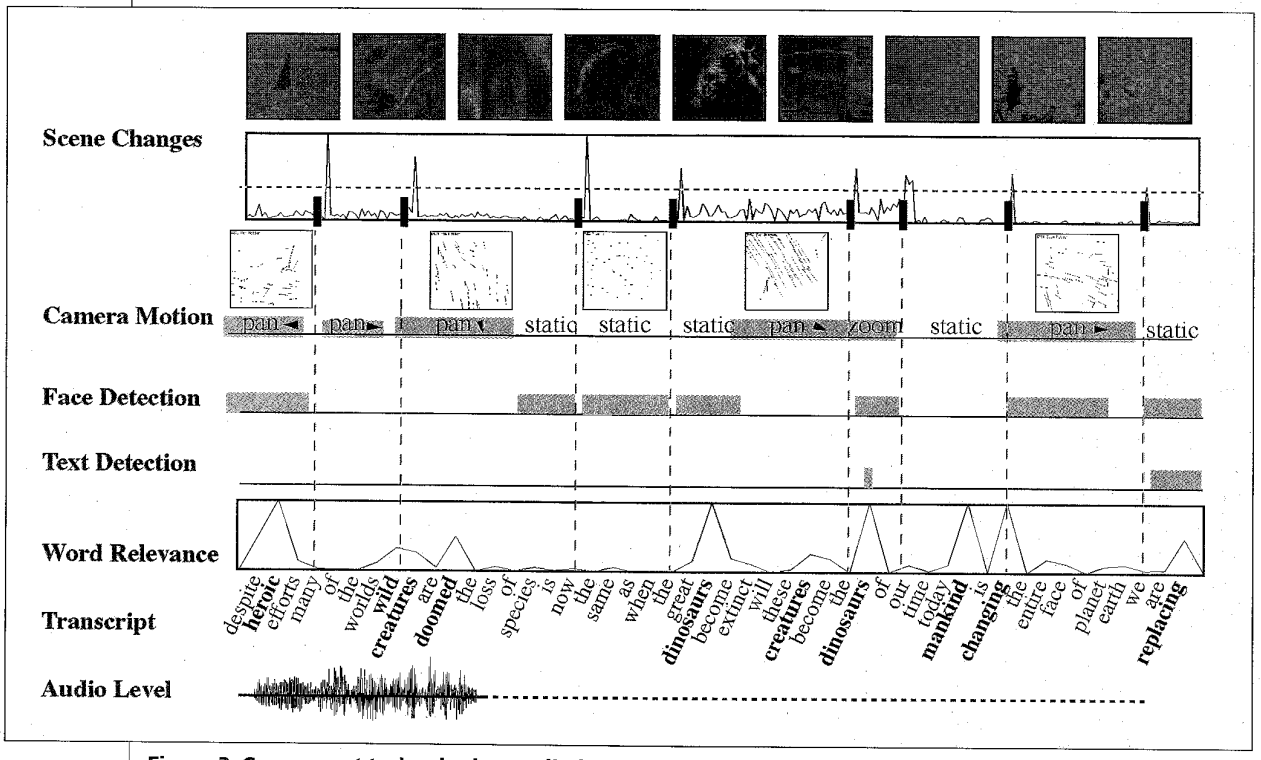

Figure 3. Component technologies applied to segment video data.

tics" to image content, they have proved quite useful in quickly comparing and categorizing images, and these attributes will be used for retrieval.

- *Concurrent use of image and speech/language information.* In addition to image properties, other cues, such as speaker changes, timing of audio and/or background music, and change in the content of spoken words can be used for reliable segmentation.
- *Camera and object motion in 2D.* An especially useful kind of visual segmentation is based on the computer's interpreting and following smooth camera motions such as zooming, panning, and forward camera motion. Using the Lucas-Kanade gradient descent method for optical flow,[7] we can track individual regions from one frame to the next and create a vector representation for all associative camera motion. Optical flow for a variety of camera motion is shown for the scenes in Figure 3. A different (but equally important) kind of video segment is defined not by camera motion but by motion or action of the objects being viewed. Object motion typically exhibits flow fields in specific image regions. Camera motion is characterized by flow throughout the entire image.

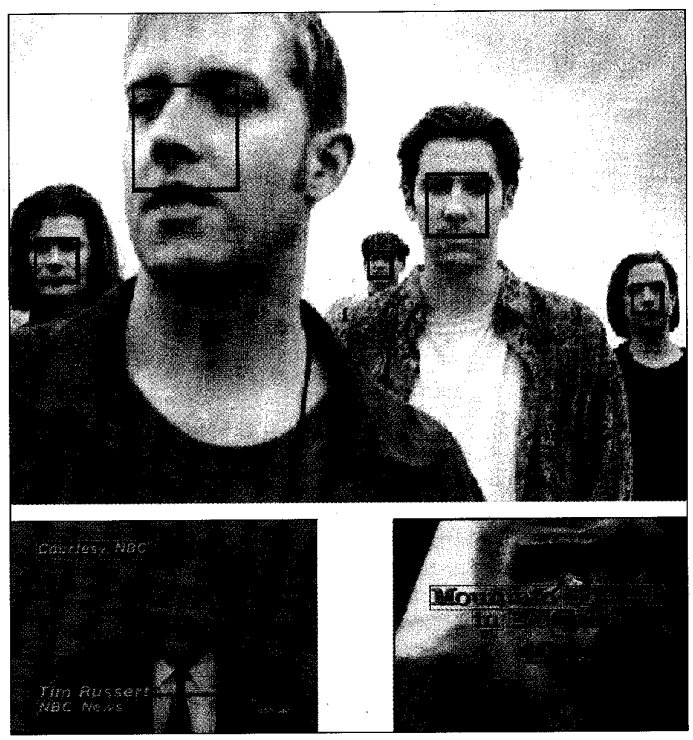

Figure 4. Face and text detection results.

OBJECT PRESENCE. A powerful technique segments video by the appearance of a particular object or combination of objects. Human content is a particularly important and common case of object-presence detection, as is a human interacting within an environment. The human-face detection system used for our experiments is based on the method of neural-net arbitration developed by Rowley et al.[8] Its current performance level detects over 90 percent of more than 300 faces contained in 70 images, with approximately 60 false detections.

Another essential detection technique is that of textual information appearing in the video but not repeated in the audio. By detecting the clustered and often high-contrast structure of printed characters, we can extract regions from video that contain text.[9] For example, out of 75 images processed, we can currently detect 86 percent of the regions containing text while producing only 12 false detections. Once text is extracted, optical character recognition can be applied and the resulting data added to the searchable text. Examples of face and text detection are shown in Figure 4.

OBJECT AND SCENE IN 3D. Because video represents mostly 3D shape and motion, adding a 3D understanding capability to the image understanding analyses will enlarge the system's scope. The "factorization" approach can potentially reconstruct 3D information from a 2D video data sequence.

Natural-language processing

Library search and retrieval, precision, and recall can be improved through natural-language processing to understand and expand the user's query and to associate it with correct but inexact matches from the library's content. This lets us go beyond limited keyword matching in our library search. Natural-language processing in Informedia is applied to both query processing and library creation. It serves four principal functions—spoken and typed free-form query processing, ranked retrieval, automated transcript correction, and summarization for use in title generation and video abstract creation (for example, skims). The latter two pertain to special functions for the library-creation process. Our retrieval engine, based on the Pursuit engine embedded in the Lycos Web browser, is of a class that implements probabilistic matching to return a rank-ordered result list. By varying relative thresholds, either precision or recall can be adjusted by the user.

The following goals for Informedia's natural-language processing stem from the system's use of spoken language and automated speech recognition for both query and data.

- *Provide* multiple types of similarity matching. Several kinds of similarity can be implemented and adjusted—prefix, synonym, string, phonetic, and conceptual.
- *Tolerate* errors in speech recognition of the query.
- *Correct* errors in speech recognition-generated transcripts.
- *Parse* both fluent and ungrammatical spoken language.
- *Provide* phonetic matching to both query and transcript.

- *Apply* data extraction techniques to spoken language.
- *Offer* broad-domain semantic matching.

EXPLORING THE LIBRARY

Library exploration includes search, retrieval, display, and reuse. This complicates matters for user interface alternatives, data and network architectures, and charging for content access mechanisms.

Video skimming through integrated processing

Users of any information-retrieval system often want to quickly review the results of their query to judge each item's relevance or interest. For text, the delivery is static, and the user applies personal techniques to select and skip content. Simply speeding up video and audio delivery (beyond twice normal speed) eliminates the audio comprehension and distorts much of the image beyond visual recognition. In addition, displaying video frames at fixed intervals might cause important video content to be skipped. As a result, devising a method for conveying the essence of a video segment's content in a fraction of the normal display time is a significant challenge.

Through combined techniques from language and image understanding, we have developed video skims of the original video at varying compression ratios.[9] This compact video is created with significant image and audio regions to produce a synopsis of the original, which can also be used to select a single representative frame for each scene. These frame icons are useful when only a single image is needed to describe a segment.

We apply *term-weighting* techniques to identify the most relevant keywords and phrases[10] in the transcribed audio track (as shown in the bottom graph of Figure 3). We automatically examine the time-corresponding video for scene changes and breaks, relevant objects, and motion analysis. We examine the audio level for additional clues to detect transitions between speakers and topics, which often correspond to low energy or silence in the signal.

Having segmented the video, we statistically compute the relative importance of each scene's image content. Image significance is characterized through desirable camera motion and object presence. Through optical flow analysis, we can determine which images in a scene contain the most desirable motion. A film producer will often use static frames preceding or following camera motion as the focus of a given scene. Objects such as human faces and text can be identified in video and used as a basis for significance during skim creation. For example, statistical numbers are not usually spoken but are included in the captions for viewer inspection. The "talking head" image common in interviews and news clips illustrates a clear example of video production focusing on an individual of interest.

The unsynchronized audio and video are now integrated into an effective skim of the original content. In Figure 5 we show the keywords and significant images selected for skim creation, and the corresponding skim video. Keywords will not always align with the selected frames. The audio data can cross multiple frames depending on keyword length. The word "dinosaurs" consumes 1.13 seconds (34 frames), so frames from adjacent scenes are also selected. Scenes with human faces are important; however, the same frames with text captions contain more

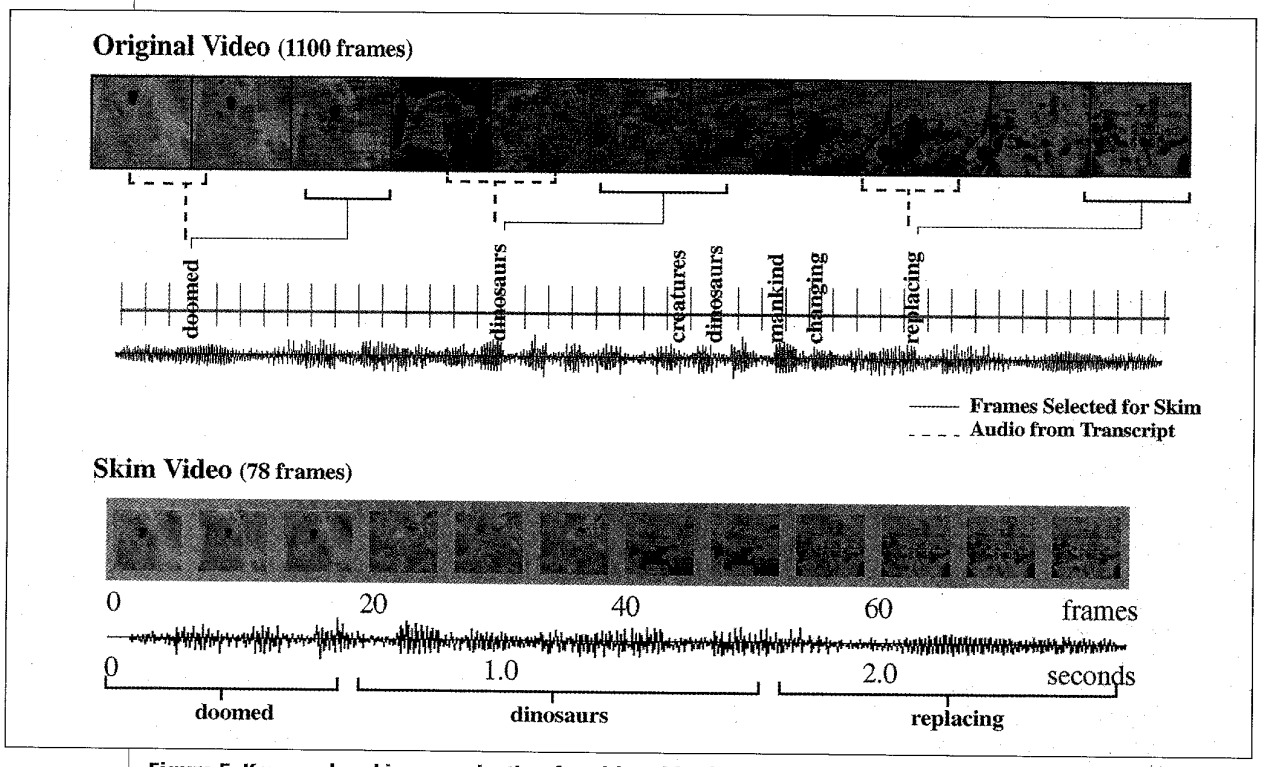

Figure 5. Keyword and image selection for video skim (14:1 compaction).

information. When possible, segments of shots that bound camera motion are used with scenes that contain pans or zooms. For example, the scene with the polar bear begins with a downward pan, showing only the lower portion of the animal. In the latter frames, camera motion has stopped and the camera focuses on the animal's face. The final representation is controlled by the user and can vary in size and content. We have found useful skims with time compression ratios ranging from 6:1 to 20:1. Table 1 lists the skim compaction results of various video segments.

Table 1. Skim compaction.

Video segments	Original (seconds)	Skim scenes
K'nex toy	61.0	7.13
Species destruction (half)	68.65	6.40
* Species destruction (full)	123.23	12.43
* Space university	166.20	28.13
* Rain forest	107.13	5.36
* Peru forest destruction	58.13	5.30
* Underwater exploration	119.50	5.67

*Manual skims

Figure 6 shows the complete skim for the video with associative frames and keywords for all scenes. Another representation for the significant image regions is the static skim. By displaying only a select group of frame icons from different scenes, the user can quickly interpret the content of a given segment. An extension to this form of skim will be the display of selected keywords or phrases along with the image frames.

Productive user interfaces

The user-interface requirements for a video library differ substantially from those for a text or image library due to the temporal nature of the retrieval data. Figure 1 illustrated a typical retrieval display. We believe several functions are essential for a successful digital video library interface, as we discuss next. The Informedia testbed will let us evaluate the relative effectiveness, sensitivity, and frequency of use of the alternative display methods and their user-adjustable parameters.

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. Users will likely react differently to a screen populated by still images than by moving images. Therefore, we will identify the optimal number and mix of object types through studies.

CONTEXT-SIZING. Users can adjust the "size" (duration) of the retrieved video/audio segments for playback. Here the "size" may be time duration, but it can also be based on scenes or information complexity. Users are also offered options with respect to increasing the context of a previously displayed segment by providing the preceding or following video paragraphs from the original work or the much larger video segment from which it was extracted. These controls were also pictured in Figure 1.

SYNTHETIC INTERVIEWS. When sufficient data exists in the library in the form of interviews or news conferences with a single individual, it's possible to construct a simulated interview interface, whereby the user interacts virtually with the subject. This enables a more interesting personal experience than simply watching a linear interview by others. Comparable synthetic interviews have been handcrafted[11,12] that demonstrate this format's potential.

REUSE. Once users identify video objects of interest, they will need to be able to perform the difficult tasks of manipulating, organizing, and reusing the video. Even the editing task is difficult. To effectively reuse video assets, the user must combine text, images, video, and audio in new and creative ways. It is our intent to enable use of commercial video editors as well as to comply with standard object interfaces (for example, OLE), so that Informedia-created video segments can be incorporated into commercial applications. Effective video reuse is hindered by complexities in understanding the nature of cinematic production—interplay of scene, framing, camera angle, and transition. Building on previous work,[11,12] we plan to examine tools that provide expert assistance in cinematic knowledge, comparable to the successful function of templates in document production systems.

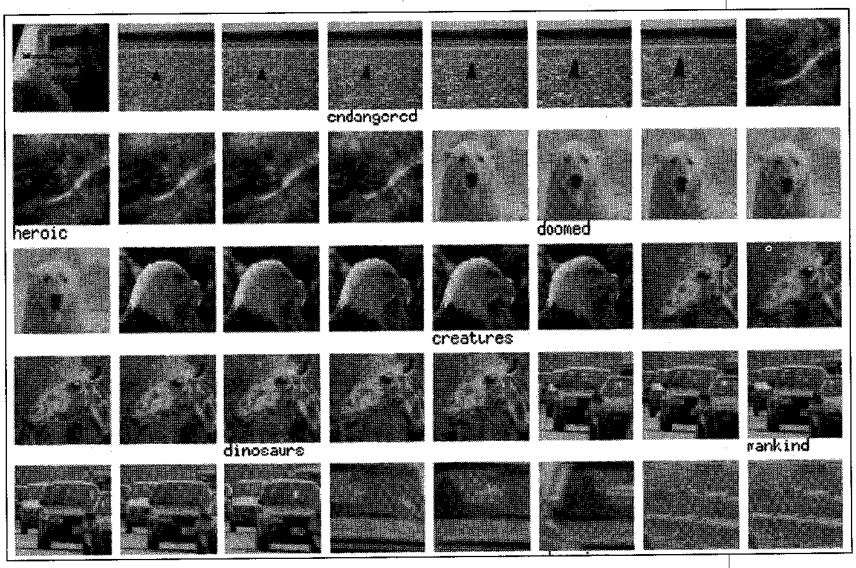

Figure 6. Skim video frames and audio keywords from "Destruction of Species," WQED, Pittsburgh.

WE HAVE FOCUSED THE WORK on two corpuses. One is based on science documentaries and lectures that have been experimentally deployed with corrected transcripts and segmentation at a local high school. The other is broadcast news content with partial closed-captions that is fully automatically processed and incorporated into the library. We have added a natural language, spoken query interface in the latter prototype. Future work will continue to improve the accuracy and performance of the underlying processing as well as explore performance issues related to Web-based access and interoperability with other digital video resources. Further information is available through http://informedia.cs.cmu.edu. ▌

Acknowledgments

This material is based on work supported by the National Science Foundation, ARPA, and NASA under NSF Cooperative Agreement No. IRI-9411299. Michael A. Smith is sponsored by AT&T Bell Laboratories.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

References

1. M. Christel et al., "Techniques for the Creation and Exploration of Digital Video Libraries," in *Multimedia Tools and Applications, Vol. 2*, Borko Furht, ed., Kluwer Academic Publishers, Boston, 1995.
2. E. Fox et al., "Introduction," special issue on digital libraries, *Comm. ACM*, Apr. 1995, pp. 22-28.
3. M. Davis, "Knowledge Representation for Video," *Proc. AAAI*, AAAI Press/MIT Press, Cambridge, Mass., 1994, pp. 128-127.
4. M.Y. Hwang, E. Thayer, and X. Huang, "Semi-Continuous HMMs with Phone Dependent VQ Codebooks for Continuous Speech Recognition," *Proc. ICASSP*, IEEE Press, Piscataway, N.J., 1994.
5. M. Hawley, "Structure out of Sound," doctoral dissertation, MIT, Cambridge, Mass., 1993.
6. H. Zhang, C. Low, and S. Smoliar, "Video Parsing and Indexing of Compressed Data," *Multimedia Tools and Applications*, Mar. 1995, pp. 89-111.
7. B.D. Lucas and T. Kanade, "An Iterative Technique of Image Registration and Its Application to Stereo," *Proc. 7th Int'l Joint Conf. Artificial Intelligence*, Morgan Kaufmann, Los Altos, Calif., 1981, pp. 674-679.
8. H. Rowley, S. Baluja, and K. Kanade, "Human Face Detection in Visual Scenes," Tech. Report CMU-CS-95-158, Computer Science Dept., Carnegie Mellon Univ., Pittsburgh, 1995.
9. M. Smith and T. Kanade, "Video Skimming for Quick Browsing Based on Audio and Image Characterization," Tech. Report CMU-CS-95-186, Carnegie Mellon Univ., Pittsburgh, 1995.
10. M. Mauldin, "Information Retrieval by Text Skimming," doctoral dissertation, Carnegie Mellon Univ., Pittsburgh, 1989. (Also available as CMU Tech. Report CMU-CS-89-193.) Revised edition published as "Conceptual Information Retrieval: A Case Study in Adaptive Partial Parsing," Kluwer Academic Publishers, Boston, Sept. 1991.
11. S. Stevens, "Intelligent Interactive Video Simulation of a Code Inspection," *Comm. ACM*, July 1989, pp. 832-843.
12. M. Christel and S. Stevens, "Rule Base and Digital Video Technologies Applied to Training Simulations," *Software Eng. Inst. Tech. Review '92*, Software Eng. Inst., Pittsburgh, 1992.

Howard D. Wactlar *is the vice provost for research computing and associate dean of the School of Computer Science at Carnegie Mellon University. He was a founder of the Software Engineering Institute and director of the Information Technology Center, a research department focused on large-scale deployment and technology transfer. He is project director and was primary architect of the Informedia Digital Video Library. His research interests are multimedia, distributed systems, networking, and performance measurement.*

Wactlar received a BS in physics from MIT and an MS in physics from the University of Maryland. He is a member of IEEE.

Takeo Kanade *is the U.A. Helen Whitaker Professor of Computer Science and Director of the Robotics Institute. He has served on many government, industry, and university advisory or consultant committees, including the Aeronautics and Space Engineering Board of the National Research Council, NASA's Advanced Technology Advisory Committee, and the Advisory Board of the Canadian Institute for Advanced Research.*

Kanade received a PhD in electrical engineering from Kyoto University, Japan. He is an IEEE fellow, a founding fellow of the AAAI, and the founding editor of the International Journal of Computer Vision.

Michael A. Smith *is a doctoral candidate in the Electrical and Computer Engineering Department at Carnegie Mellon University. His research interests are image classification and recognition, and content-based image understanding. His research is an active component for the Informedia Digital Video Library project at Carnegie Mellon. He has published in the areas of pattern recognition, biomedical imaging, video characterization, and interactive computer systems.*

Smith received a BS in electrical engineering from North Carolina A&T State University and an MS in electrical engineering from Stanford University. He is a member of IEEE, Eta Kappa Nu, Tau Beta Pi, and Pi Mu Epsilon.

Scott M. Stevens *is a senior member of the technical staff at the Software Engineering Institute and a charter member of CMU's Human-Computer Interaction Institute. He is a principal investigator on the Informedia Digital Video Library Project, directing user interface research and development and testbed evaluation. He has been involved with multimedia research and development since the mid-1970s when he developed multimedia applications for an experimental system designed to distribute compressed interactive video into the home. Stevens is the general chair for the 1996 IEEE Computer Society International Conference on Multimedia Computing and Systems and the chair for the IEEE-CS Technical Committee on Multimedia Computing and Systems. He is also on the editorial board of the IEEE-CS Press Advances.*

Stevens received a BS and MS in physics from Northern Illinois University and a PhD in human-computer interaction from the University of Nebraska, Lincoln. He is a senior member of the IEEE Computer Society.

Address questions about this article to Wactlar, Carnegie Mellon University, School of Computer Science, 5000 Forbes Ave., Pittsburgh, PA 15213-3891; wactlar@cmu.edu.