Cover Feature

Lessons Learned from Building a Terabyte Digital Video Library

The Informedia Project has integrated artificial intelligence techniques in creating a digital library providing full-content search and discovery in the video medium.

Howard D. Wactlar
Michael G. Christel
Yihong Gong
Alexander G. Hauptmann
Carnegie Mellon University

The Informedia Project at Carnegie Mellon University has created a terabyte digital video library in which automatically derived descriptors for the video are used for indexing, segmenting, and accessing the library contents. The project, begun in 1994, was one of six funded by the US National Science Foundation, the US Defense Advanced Research Projects Agency, and the National Aeronautics and Space Administration, under the US Digital Library Initiative.

Digital video presented a number of interesting challenges for library creation and deployment: the way it embeds information, its voluminous file size, and its temporal characteristics. In the Informedia Project, we addressed these challenges by

- automatically extracting information from digitized video,
- creating interfaces that allowed users to search for and retrieve videos based on extracted information, and
- validating the system through user testbeds.

We met these objectives during the course of the project, learning many lessons along the way.

VIDEO PROCESSING

Our library consisted of two types of video: news video (from the Cable News Network) and documentary video (from the British Open University, QED Communications, the Discovery Channel, and a number of US government agencies, including NASA, the National Park Service, and the US Geological Survey).

As of May 1998, the library contained more than 1,000 hours of news and 400 hours of documentary video, with additional video being added daily. The news video often included tags which marked story boundaries within longer broadcasts. At first, we added these story boundaries manually for the documentary video; in our subsequent experiments we looked at generating story boundaries automatically. The stories, or video segments, averaged a few minutes in length, so that the entire library contained more than 40,000 video segments.

We used artificial intelligence techniques to create metadata, the data that describes video content. We found that all the AI techniques we used were applicable to both the news corpus and the documentary corpus. These techniques included speech recognition, image processing, and information retrieval. We learned that by integrating across these three areas we were able to compensate for limitations in accuracy, coverage, and communicative power.

Speech processing

We analyzed the audio component of the video with the CMU Sphinx speech recognition system.[1] This created a complete transcript for text-based retrieval from the speech and aligned existing imperfect transcripts to the video. The tightly synchronized transcript was subsequently used in the library interface for quickly locating regions of interest within relevant video segments.

In creating transcripts, CMU Sphinx's word error rate is inversely proportional to the amount of processing time devoted to the task. Processing time varies from real time, which offers relatively poor transcription (more than half the words were wrong), to several hundred times real time, which offers optimal performance. Our standard processing is around 35 times real time, and it offers very accurate transcrip-

tion. CMU Sphinx, however, can run on parallel machines, providing excellent word recognition in two to three times real time.

In our initial experiments on news data in 1994, word error rate—the sum of erroneous word insertions, deletions, and substitutions—was 65 percent. The latest version of the speech recognition system, CMU Sphinx III, which was trained on 66 hours of news broadcasts, provides a word error rate of about 24 percent. We have been able to lower that to 19 percent through the use of general language models interpolated with the current "news of the day" as obtained from the Web sites of CNN, Reuters, and the Associated Press. In the interpolated news-language model, the speech recognition vocabulary reflects the names and places of today's news and adjusts probabilities of word pairs and triplets that reflect expressions used in today's news.

Information retrieval

In Informedia, we needed to provide accurate searching of audio content despite relatively high word error rates. Our initial experiments suggested that information retrieval would still be effective despite the transcription mistakes by the speech recognition system. With word error rates up to 30 percent, information retrieval precision and recall were degraded by less than 10 percent. This is a reflection of the redundancy in language, where key words repeat multiple times within a document and different key terms describe the same thing. Redundancy means that even if the speech recognizer misses one or two instances, the document will still be considered appropriately relevant to a user's query.

Other experiments showed that transcript errors due to words missing from the speech recognizer's vocabulary could be mitigated by using a phonetic transcription together with a word transcription of the document.[2,3] We were also able to improve information retrieval by having the speech recognizer produce multiple alternate candidate transcription hypotheses rather than a single best guess.[4]

From these experiments we learned that even relatively high word error rates in the speech recognition nevertheless permit relatively effective information retrieval. Using standard information retrieval techniques, the state of the art in speech recognition was adequate for information retrieval from broadcast news type video. Additional techniques specialized for speech transcripts could further improve the retrieval results.

However, there are still many open questions related to information retrieval from video documents. We need to verify that research results obtained with relatively small amounts of spoken data will apply as well to very large collections. There was initial evidence that retrieval from heterogeneous sources (such as a corpus of some perfect text documents and some error-laden speech transcripts) would be more adversely affected by speech errors. Finally, errors in the automatic partitioning of video streams into video segments may also adversely affect information retrieval effectiveness. If a single relevant news story is inadvertently broken up into two pieces, with some key terms in each piece, the resulting relevance rank for the individual halves of the story may be substantially lower than for the complete news story segmented properly.

Even relatively high word error rates in the speech recognition nevertheless permit relatively effective information retrieval.

Image processing

The ultimate goal for image processing in the context of digital video libraries is the ability to fully characterize the scene and all objects within it and to provide efficient, effective access to this information through indices and similarity measures. Other research groups have made progress toward this goal by restricting the domain, for example, to a particular sport or to news broadcasts from a particular agency.[5-7] Programs can then look for certain cues, like the appearance of scoreboards or news anchors. The Informedia Project, however, focused on general-purpose processing without narrowing the candidate videos to a particular domain.

MPEG-1 compression was used for the Informedia library, with video at 30 frames per second and individual frame resolution at 352×240 pixels. Our image processing techniques worked with the compressed video frames to reduce processing time requirements.

The relatively low resolution and compression artifacts adversely affected the quality of the metadata produced through image processing, such as face and text detection. By taking advantage of the temporal redundancy in video, we minimized the effects of such errors. For example, text that appears on one frame of video is likely to remain visible in subsequent frames of video for at least the next few seconds.

In the Informedia Project, we used image processing to

- partition each video segment into *shots*, defined as a set of contiguous frames representing a continuous action in time or space,[5,6] and choose a representative frame (key frame) for each shot;
- identify and index features to support image similarity matching; and
- create metadata derived from imagery for use in text matching.

Shots and key frames. To enable quick access to content within video segments and to provide visual summaries, we decomposed each video segment into

Face detection was adequate in the case of full frontal shots and when faces were well lit.

shots. Each shot was bounded by frames in which significant color shifts occurred.[5] Shot detection was improved by incorporating optical flow characteristics for tracking camera and object motion. For example, a camera pan was left intact as a single shot.

By default, a shot's middle frame became its *key frame* representing that section of video. We found we could improve the choice of key frame by applying cinematic heuristics. For example, when there is camera motion in a shot, the point where it stops is probably the most informative and significant, so this point becomes the key frame. Similarly, low-intensity images are likely less important, so they are excluded as key frames.

Face and color detection. In support of image matching and richer visual summaries, we weighted the key frame selection to use images with faces wherever possible. We also provided library users with a face-searching capability based on this technique.

Face detection was adequate in the case of full frontal shots (where both eyes were distinguishable) and when faces were well lit. In news stories, then, it did well for anchorpersons in the news studio but often not for subjects of stories in the field. With improved face detection, the other key players besides just the anchorperson could be identified, resulting in more representative key frames for the video and more complete face matching across the whole library.

Library users could also match images based purely on color. However, matching on the basis of color histograms often failed to deliver relevant images. The color histogram technique failed in part because it uses a primitive representation of image color properties and in part because it cannot accomplish partial mapping (matching images in finer granularities). The inability to use partial mapping meant that users could not retrieve images on the basis of regions of interest: For example, they could not retrieve all the video that contained a green lawn without regard to the rest of the imagery.

We developed a new method for image retrieval based on human perceptual color clustering.[8] This perceptual color clustering method first clusters color images into a few prominent colors and uniform regions based on human color perception. It then creates a multidimensional feature vector for each of the prominent colors and uniform regions. Finally, it uses these feature vectors to form indices of the original color images.

To retrieve images using this technique, the user specifies colors and regions of interest in the sample image. The system retrieves relevant images by first constructing feature vectors of the user-specified colors and regions, and it then uses these feature vectors as search keys to perform a nearest-neighbor search on the database. An experimental evaluation showed that this new image retrieval method surpassed color histogram-based methods.[8]

Video OCR (VOCR). We used text superimposed on the video (VOCR) to identify key frames and enrich searching.[9] The VOCR process involves three steps.

1. Identify video frames with probable text. Text captions in video can typically be characterized as a horizontal rectangular structure of clustered sharp edges. A horizontal differential filter highlights vertical edge features, with postprocessing to eliminate extraneous fragments and connect character segments that may have become detached. The resulting clusters are analyzed for their horizontal and rectangular text-like appearance. Clusters satisfying this analysis are identified as text regions in a frame.
2. Filter the probable text region across the multiple video frames containing it, producing a higher quality image than if a single frame were used.
3. Use commercial OCR software to process the final filtered image.

Tests run with MPEG-1 videos of broadcast news showed that text regions were identified correctly 90 percent of the time, with 83.5 percent of characters subsequently recognized correctly and a word recognition rate of 70 percent.[9] We could improve word accuracy further by basing corrections on dictionaries, thesauri, and the same "news of the day" information as was used to create interpolated language models for speech recognition.

Future challenges. Image processing has not yet reached the point where it can automatically characterize all objects and their interrelationships within video segments.[7] Associated challenges include:

- Evolving from color-based image similarity matching to content-based image similarity matching. A long-term goal is to recognize both a close-up of a single soccer player and an aerial shot of the whole soccer field as similar shots based on content, despite differences in low-level image features like color, texture, and shape.
- Effectively segmenting general-purpose video into stories, and further breaking down these stories into meaningful shots. Segmentation will likely benefit from improved analysis of the video corpus, analysis of video structure, and application of cinematic rules of thumb.

We believe these challenges can be met by applying techniques across various modalities—for example,

using language models together with transcript information to improve video segmentation. Correlating human faces in video images with human names in transcripts could facilitate finding appearances of individuals in video sequences. Other correlation could enable labeling of video segments with who, when, and where information.

THE INFORMEDIA INTERFACE

The Informedia interface was designed to provide users with quick access to relevant information in the digital video library. A basic element of the design was the provision of alternative browsing capabilities—or *multimedia abstractions*—to help users decide which video they wanted to see. In the Informedia Project, abstractions included headlines, thumbnails, filmstrips, and skims.

As Cox et al. state, "powerful browsing capabilities are essential in a multimedia information retrieval system"[7] because the underlying speech, image, and language processing are imperfect and produce ambiguous, incomplete metadata. The search engine introduces additional ambiguity, while the users themselves contribute more error through poor query formulation or simply not knowing what to ask for. The results from a query to a digital video library are hence imprecise, with users left to filter out the useful information. Multimedia abstractions have been invaluable for assisting in this task.

The Informedia Project began with a conceptual prototype suggesting the utility of certain multimedia abstractions and interface features. The prototype was refined into an operational interface that was then used at the Winchester Thurston School, Pittsburgh, by high school students and teachers in computer laboratory and traditional library settings. The system was later delivered to the US Defense Advanced Research Projects Agency as a prototype application for "News-on-Demand" and also used at Carnegie Mellon University for demonstrations and controlled experiments.

We tracked usage data primarily by automatically logging mouse and keyboard input actions, supplemented with user interviews. We conducted formal empirical studies to measure the effectiveness of particular multimedia abstractions. During the fall of 1997, students in a graduate course in human-computer interfaces at Carnegie Mellon University conducted a number of evaluations on the system, including contextual inquiry, heuristic evaluation, cognitive walk-throughs, and think-aloud protocols. The Informedia system was also demonstrated at numerous multimedia, user interface, and digital library conferences.

Data from these instruments and forums collectively forms the basis for the discussion that follows.

Figure 1. Query input and result set interface for the Informedia digital video library.

Headlines

Figure 1 shows the Informedia interface following a query on the Northern Ireland peace treaty vote. The display at the bottom shows thumbnail images for video segments returned as matches for the query. When the user positions the mouse arrow over a thumbnail, the interface pops up a headline for the segment.

Early Informedia implementations created headlines for segments based on the "most significant" words in their text (for example, in the transcript and VOCR-produced text). These words were determined by *term frequency-inverse document frequency* (tf-idf) statistics. If the word occurred often in the text associated with the video segment and infrequently in the whole library's text, then that word became part of the headline. These headlines were deficient in the following ways:

- They did not read well, being an assemblage of words usually not occurring together naturally.
- They did not use semantics provided by the segment belonging to a certain class of video, such as network news or animal documentaries.

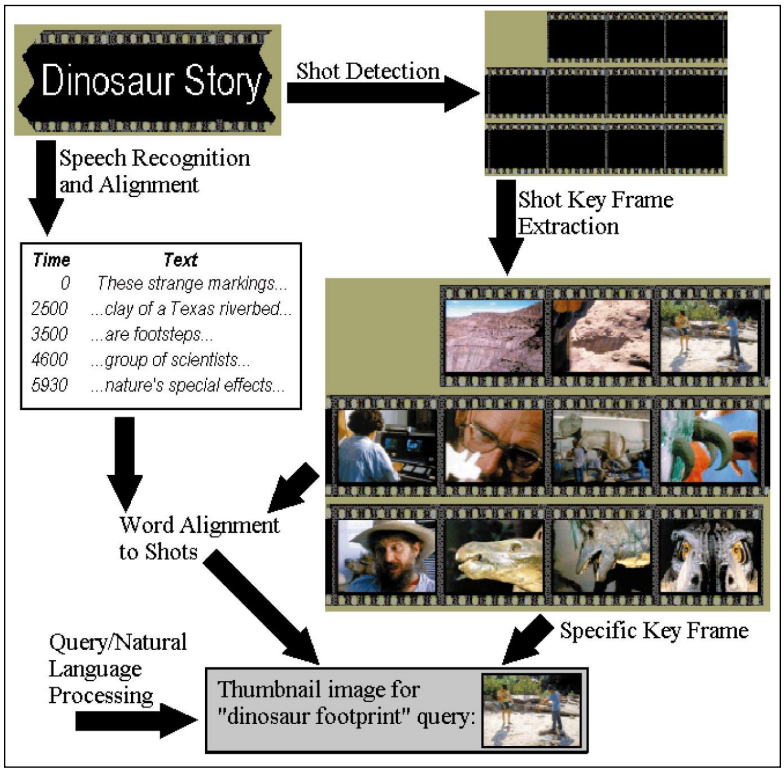

Figure 2. Using speech, image, and language processing to derive a thumbnail best matching the user's query.

- They did not offer additional summary information, such as segment size or date.

The new headlines, shown in Figure 1, addressed these shortcomings:

- Phrases, instead of words, were weighted using tf-idf statistics and became the component pieces for the headlines.
- Headlines for news segments used the heuristic that the most significant information is given immediately, with details and illustrations following it.
- Headlines included segment size and the copyright date. If the user sorted results by date, the date appeared first in the headline.

Thumbnails

Figure 1 shows thumbnails for 12 segments. A formal experiment was conducted to test whether the results layout shown in Figure 1 had any benefits over presenting the results in a text menu, where the headline text would be displayed for each result. The experiment, conducted with high school scholarship students, showed that a version of the pictorial menu had significant benefits for both performance time and user satisfaction: Subjects found the desired information in 36 percent less time with certain pictorial menus over text menus.

Most interestingly, the manner in which the thumbnail images were chosen was critical. If the thumbnail for a video segment was taken to be the key frame for the first shot of the segment, then the resulting pictorial menu of "key frames from first shots in segments" produced no benefit compared to the text menu. Only when the thumbnail was chosen based on usage context was there an improvement. When the thumbnail was chosen based on the query, by using the key frame for the shot producing the most matches for the query, then pictorial menus produced clear advantages over text-only menus.[10]

This empirical study validated the use of thumbnails for representing video segments in query result sets. It also provided evidence that leveraging multiple processing techniques improves digital video library interfaces. Thumbnails derived from image processing alone, such as choosing the image for the first shot in a segment, produced no improvements over the text menu. However, through speech recognition, natural language processing and image processing, improvements can be realized.

- Via speech recognition, the spoken dialogue words are tightly aligned to the video imagery.
- Through natural language processing, the query is compared to the spoken dialogue, and matching words are identified in the transcript and scored.
- Via word alignment, each shot can be scored for a query, and the thumbnail can then be the key frame from the highest scoring shot for a query.

Figure 2 illustrates these improvements. Result sets showing such query-based thumbnails as in the pictorial menu shown at the bottom of Figure 1 do produce advantages over text-only result presentations.

Filmstrips

Key frames from a segment's shots can be presented in sequential order as filmstrips, as shown at the top of Figure 3. The segment's filmstrip quickly shows that the segment contains more than a story matching the query "Mir collision," including an opening sequence and a weather report. The filmstrip helps identify key shots with bars color-coded to specific query words, in this case red for "Mir" and purple for "collision." When the user moves the mouse arrow over the match bars, a text window displays the matching word and other metadata.

Because key frame selection can be weighted to use images of faces, filmstrips of news stories typically show the anchorperson at points in the video seg-

ment when the story is introduced, when transitional comments are made before cutting to field footage or a new interviewee, and when the story is summarized before moving on to a different topic. The count and location of anchorperson images enables the user to infer the length, percentage of field footage, and flow of a news video segment—all via the filmstrip abstraction.

By investigating the distribution of match bars on the filmstrip, the user can determine the relevancy of the returned result and the location of interest within the segment. The user can click on a match bar to jump directly to that point in the video segment. Hence, clicking the mouse as shown in Figure 3 would start playing the video at this mention of Mir with the shot of the sitting cosmonauts. Similarly, the system provided "seek to next match" and "seek to previous match" buttons in the video player allowing the user to quickly jump from one match to the next.

In the example of Figure 3, these interface features allowed the user to bypass irrelevant video and seek directly to the Mir story. Users frequently accessed these and other features that revealed the workings of the search engine and explained why a result was returned, including details on the relative contributions of each query term.

Transcript text appears at the bottom of the video playback window. As the video plays, the text scrolls, with words highlighted as they are spoken. The interface worked well with showing a single type of tightly synchronized metadata: spoken transcripts. This interface feature might be useful to the hearing-impaired community and to those learning English.

As other metadata types were introduced, however, such as VOCR, user notes, topic labels, and headlines, having one text box or even multiple text boxes stacked beneath the video for displaying the metadata started to be confusing. If one text box was used, the problem was in how to show current scope: a type like a user note could have a duration of 20 seconds but each transcript word had a typical duration of a second or less. If multiple text boxes were used, one per metadata type, then not only did screen resolution dictate a maximum number of metadata types possible, but also the user's attention was severely fragmented watching over the multiple scrolling text boxes. A possible solution is the DIVA approach,[11] which shows temporal streams of metadata synchronized to the video as striped rectangular areas alongside the video presentation.

Figure 3. Filmstrip and video playback windows for a result from the query for "Mir collision."

Skims

We developed a temporal multimedia abstraction, the *video skim*, which is played rather than viewed statically. A skim incorporates both video and audio information from a longer source so that a two-minute skim, for example, may represent a 20-minute original video. The MoCA Project developed automatic techniques to create video skims that act as movie trailers, short appetizers for a longer video intended to attract the viewer's attention.[6] Our goal for video skims went beyond motivating a viewer to watch a full video segment; we sought to communicate the essential content of a video in an order of magnitude less time.

Initial progress was slow in the development of video skims with this capability. Formal empirical tests in early 1997 showed no advantages to skims produced through image and speech processing versus trivial subsampled skims, where sections of audio and video were extracted from the longer source video at regular inter-

Intellectual property concerns will inhibit the growth of centralized digital video libraries.

vals. These early skims suffered from selecting words rather than phrases, much like headlines, and from ignoring information in the audio channel pertaining to silences. By taking advantage of audio characteristics and utilizing larger component pieces based on phrases, later skims were less choppy and better synchronized with the imagery. These later skims demonstrated significant performance and user satisfaction benefits compared to simple subsampled skims in follow-up empirical studies.[12] Further improvements can potentially be gained through:

- Heuristics appropriate for particular types of video and certain video uses. For example, in news broadcasts, the skim would emphasize the anchorperson's opening comments, which likely summarize the story.
- Creating skims on the fly, emphasizing video sections matching current context, such as a user's query.

OTHER LESSONS LEARNED

The Informedia system attempted to be general purpose, serving a wide range of users accessing wide-ranging video data. In retrospect, this approach may be more limiting rather than liberating. Many processing techniques, such as video skim creation, work well if heuristics can be applied based on the video belonging to a particular subclass. Users don't want to only find a video segment; they then want to do something with it—for example, integrating video into lesson plans and homework assignments. A library system needs to be user-centric and not just technology-driven.

Anecdotal feedback from users has been very positive. All age groups seem to find the digital library fun to explore. Users quickly learned how to use the interface, conducting searches and retrieving videos. Initially users were not satisfied with a corpus of a few hundred hours, but recognized the value of a news corpus that grew daily and one that held over a thousand hours of video, especially if the video were produced by multiple sources to represent multiple perspectives. In addition to wanting more data, users requested:

- Queries on multiple modalities—for example, searches for the word France and castles matching an image.
- Interface support for more than just query access—for example, hierarchical browsing access or other views of the whole corpus or views across large sets of query results.

In dealing with these requests, we have put the corpus under the control of a commercial database and added information visualization so that users can manipulate a thousand video segments at once using query terms and filters for relevance, date, and size. We now allow users to add their own notes as synchronized, searchable metadata to video segments, and have added automatically generated topics as another means for navigating the library.

When the Informedia Project began in 1994, we anticipated a quick maturation of video-on-demand services via consolidation among the communications, entertain- ment, and computer industries for enabling digital video access in the home. However, these services and interactive television failed to develop in time for use by the project as a digital video delivery infrastructure. The World Wide Web experienced tremendous growth in the same time span, but without support for VHS-quality video delivery.

The consequence of these shortfalls is that the terabyte digital video library amassed at Carnegie Mellon University remains accessible at VHS-quality playback rates only to those connected to its local area network or those with a network connection capable of sustaining MPEG-1 transfer rates.

One hope for the future is that video streaming technologies will continue to mature and network bandwidth to increase so that digital video libraries are accessible by users in a variety of settings, from schools to offices to homes around the globe.

Intellectual property concerns will inhibit the growth of centralized digital video libraries. While the Informedia library contained some public domain video, the majority of the contributions had access restrictions and could not be published to the Web, even if the video delivery infrastructure were in place. Often, the supplying agency would have liked to agree to broader access, but was not sure whether such rights could be granted because of lacking documentation and precedents on the rights held by producers, directors, and other contributors to the archived video.

One approach to these concerns involves support of federated digital video libraries, in which metadata descriptors and indices are provided in a common repository but source video material remains under the control of the various contributors. Access to the metadata, such as titles, abstracts, filmstrips, and perhaps even video skims, enables users to locate segments of interest. The full video segments are provided via the contributors, perhaps from widely distributed video servers, with the contributors managing the property rights of their own collections.

The Informedia Project began with the goal of allowing complete access to information content within multimedia sources. We believe the focus was accurate: A tremendous amount of potentially useful

information remained locked in the audio narrative, nonspeech audio, and video imagery. Through speech, image, and natural language processing, the Informedia Project has demonstrated that previously inaccessible data can be derived automatically and used to describe and index video segments. Much work remains, however, in refining this work, augmenting it with additional processing such as non-speech analysis, improving segmentation, and tailoring techniques for increased accuracy on subclasses of video and for specific user communities. ❖

Acknowledgments

The Informedia Project was supported by the US National Science Foundation, the US Defense Advanced Research Projects Agency, and the National Aeronautics and Space Administration, under NSF Cooperative Agreement No. IRI-9411299. A complete list of Informedia Project sponsors is found at http://www.informedia.cs.cmu.edu. We thank them for their support and also acknowledge the efforts of the many contributors to this project, including CMU students, staff, faculty, and visiting scientists.

References

1. M.J. Witbrock and A.G. Hauptmann, "Artificial Intelligence Techniques in a Digital Video Library," *J. Am. Soc. Information Science*, May 1998, pp. 619-632.
2. G.J.F. Jones et al., "Retrieving Spoken Documents by Combining Multiple Index Sources," *Proc. SIGIR 96*, ACM Press, New York, 1996, pp. 30-38.
3. M.J. Witbrock and A.G. Hauptmann, "Using Words and Phonetic Strings for Efficient Information Retrieval from Imperfectly Transcribed Spoken Documents," *Proc. DL '97*, ACM Press, New York, 1997, pp. 30-35.
4. A.G. Hauptmann et al., "Experiments in Information Retrieval from Spoken Documents," *Proc. DARPA Workshop on Broadcast News Understanding Systems*, Morgan Kaufmann, San Francisco, 1998, pp. 175-181.
5. H.-J. Zhang et al., "A Video Database System for Digital Libraries," *Lecture Notes in Computer Science 916*, Springer-Verlag, Berlin, 1995, pp. 253-264.
6. R. Lienhart, S. Pfeiffer, and W. Effelsberg, "Video Abstracting," *Comm. ACM*, Dec. 1997, pp. 55-62.
7. R.V. Cox et al., "Applications of Multimedia Processing to Communications," *Proc. IEEE*, May 1998, pp. 754-824.
8. Y.H. Gong, G. Proietti, and C. Faloutsos, "Image Indexing and Retrieval Based on Human Perceptual Color Clustering," *Proc. Computer Vision and Pattern Recognition*, IEEE CS Press, Los Alamitos, Calif., 1998, pp. 578-583.
9. T. Sato et al., "Video OCR for Digital News Archive," *Proc. Workshop on Content-Based Access of Image and Video Databases*, IEEE CS Press, Los Alamitos, Calif., 1998, pp. 52-60.
10. M. Christel, D. Winkler, and C.R. Taylor, "Improving Access to a Digital Video Library," *Human-Computer Interaction: INTERACT97*, Chapman & Hall, London, 1997, pp. 524-531.
11. W.E. Mackay and M. Beaudouin-Lafon, "DIVA: Exploratory Data Analysis with Multimedia Streams," *Proc. CHI '98*, ACM Press, New York, 1998, pp. 416-423.
12. M. Christel et al., "Evolving Video Skims into Useful Multimedia Abstractions," *Proc. CHI '98*, ACM Press, New York, 1998, pp. 171-178.

Howard D. Wactlar is the vice provost for research computing and associate dean of the School of Computer Science at Carnegie Mellon University. His research interests are multimedia, distributed systems, digital libraries, 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 the IEEE.

Michael G. Christel is a senior systems scientist in the Computer Science Department at Carnegie Mellon University and a charter member of CMU's Human-Computer Interaction Institute. His current research interests include multimedia interfaces, information visualization, and digital libraries. He received a BS in math and computer science from Canisius College, Buffalo, N.Y., and a PhD in computer science from the Georgia Institute of Technology.

Yihong Gong is a project scientist in the Robotics Institute at Carnegie Mellon University. His research interests include image and video analysis, multimedia database systems, and artificial intelligence. He received his BS, MS, and PhD degrees in electronic engineering from the University of Tokyo.

Alexander G. Hauptmann is a senior systems scientist in the Computer Science Department at Carnegie Mellon University. His research interests include speech recognition, speech synthesis, speech interfaces, and language in general. He received a BA and an MA in psychology from Johns Hopkins University, a diploma in computer science from Technische Universitaet Berlin, and a PhD in computer science from Carnegie Mellon.

Contact Wactlar, Christel, Gong, and Hauptmann at {wactlar, christel, ygong, hauptmann}@cs.cmu.edu.