Applications of Video-Content Analysis and Retrieval

Nevenka Dimitrova
Philips Research

Hong-Jiang Zhang
Microsoft Research

Behzad Shahraray
AT&T Labs Research

Ibrahim Sezan
Sharp Laboratories of America

Thomas Huang
University of Illinois at Urbana–Champaign

Avideh Zakhor
University of California at Berkeley

Managing multimedia data requires more than collecting the data into storage archives and delivering it via networks to homes or offices. We survey technologies and applications for video-content analysis and retrieval. We also give specific examples.

The advances in the data capturing, storage, and communication technologies have made vast amounts of video data available to consumer and enterprise applications. However, interacting with multimedia data, and video in particular, requires more than connecting with data banks and delivering data via networks to customers' homes or offices. We still have limited tools and applications to describe, organize, and manage video data. The fundamental approach is to index video data and make it a structured media. Manually generating video content description is time consuming—and thus more costly—to the point that it's almost impossible. Moreover, when available, it's subjective, inaccurate, and incomplete.

This conundrum has attracted researchers from various disciplines, each with their own algorithms and systems. In addition, the MPEG group recently issued MPEG-7 as a standard to provide normative framework for multimedia content description. However, in contrast, there are few convincing stories we can tell about successful applications of the research results. It seems that the excitement enjoyed by many researchers from both academia and industries has yet to generate significant impact in the marketplace. Is there any significant application that can benefit from our research? Can we solve the video retrieval problem as we originally claimed? Are we falling into the same hype as artificial intelligence (AI) once did? We believe the answer is no. However, we need to reexamine our research strategies and methodologies, and most importantly, users' need for technologies for content-based video retrieval.

To address these questions, we held a panel discussion chaired by Hong-Jiang Zhang at the International Workshop on Very Low Bit-Rate Video Coding (VLBV 98), with researchers in the video coding and content analysis field. This article summarizes the evolving views from the panelists and audience, as well as the continued online discussion regarding state-of-the-art technologies, directions, and important applications for research on content-based video retrieval.

Content-based video retrieval

We perceive a video program as a document. Video indexing should be analogous to text document indexing, where we perform a structural analysis to decompose a document into paragraphs, sentences, and words, before building indices. When someone authors a book, they create a table of contents for browsing the content's order and a semantic index of keywords and phrases for searching by content. Similarly, to facilitate fast and accurate content access to video data, we should segment a video document into shots and scenes to compose a table of contents, and we should extract keyframes or key sequences as index entries for scenes or stories. Therefore, the core research in content-based video retrieval is developing technologies to automatically parse video, audio, and text to identify meaningful composition structure and to extract and represent content attributes of any video sources.

A typical scheme of video-content analysis and indexing, as proposed by many researchers,

1070-986X/02/$17.00 © 2002 IEEE

involves four primary processes: feature extraction, structure analysis, abstraction, and indexing. Each process poses many challenging research problems. In what follows, we briefly review these challenging research issues and the algorithms developed so far to address them.

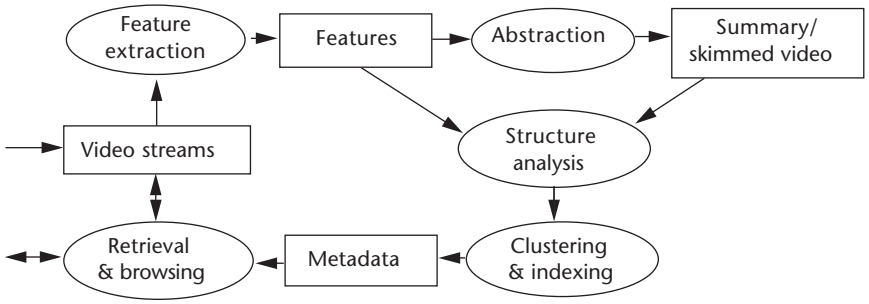


Figure 1. Process diagram for video-content analysis and indexing.

Feature extraction for content analysis

A critical process in content-based video indexing is feature extraction, which we show in Figure 1. The effectiveness of an indexing scheme depends on the effectiveness of attributes in content representation. However, we can't map easily extractable video features (such as color, texture, shape, structure, layout, and motion) easily into semantic concepts (such as indoor and outdoor, people, or car-racing scenes). In the audio domain, features (such as pitch, energy, and bandwidth) can enable audio segmentation and classification.

Although visual content is a major source of information in a video program, an effective strategy in video-content analysis is to use attributes extractable from multimedia sources. Much valuable information is also carried in other media components, such as text (superimposed on the images, or included as closed captions), audio, and speech that accompany the pictorial component. A combined and cooperative analysis of these components would be far more effective in characterizing video program for both consumer and professional applications. The Informedia system,[1] AT&T's Pictorial Transcripts system,[2–5] and Video Scout[6] are examples of such approaches.

Structure analysis

Video structure parsing is the next step in overall video-content analysis and is the process of extracting temporal structural information of video sequences or programs. This process lets us organize video data according to their temporal structures and relations and thus build table of contents. It involves detecting temporal boundaries and identifying meaningful segments of video. Many effective and robust algorithms for video parsing have been developed[7–11] for segmenting a video program into its temporal composition bricks. Ideally, these composition bricks should be categorized in a hierarchy similar to film storyboards. The top level consists of sequences or stories, which are composed of sets of scenes. Scenes are further partitioned into shots. Each shot contains a sequence of frames recorded contiguously and representing a continuous action in time or space. With such structural information, we can automatically build a video program's table of contents.[12]

An important step in the process of video structure parsing is that of segmenting the video into individual scenes. From a narrative point of view, a scene consists of a series of consecutive shots grouped together because they're shot in the same location or because they share some thematic content. The process of detecting these video scenes is analogous to paragraphing in text document parsing, but it requires a higher level of content analysis. There are two approaches for automatically recognizing program sequences: one based on film production rules,[13] the other based on a priori program models.[10] Both have had limited success because scenes or stories in video are only logical layers of representation based on subjective semantics, and no universal definition and rigid structure exists for scenes and stories.

In contrast, shots are actual physical basic layers in video, whose boundaries are determined by editing points or where the camera switches on or off. Fortunately, analogous to words or sentences in text documents, shots are a good choice as the basic unit for video-content indexing, and they provide the basis for constructing a video table of contents. Shot boundary detection algorithms that rely only on visual information contained in the video frames can segment the video into frames with similar visual contents. Grouping the shots into semantically meaningful segments such as stories, however, usually isn't possible without incorporating information from the video program's other components. Multimodal processing algorithms involving the processing of not only the video frames, but also

the text, audio, and speech components that accompany them have proven effective in achieving this goal.[12]

Video abstraction

Video abstraction is the process of creating a presentation of visual information about a landscape or the structure of video, which should be much shorter than the original video. This abstraction process is similar to extraction of keywords or summaries in text document processing. That is, we need to extract a subset of video data from the original video such as keyframes or highlights as entries for shots, scenes, or stories. Abstraction is especially important given the vast amount of data for a video program of even a few minutes' duration. The result forms the basis not only for video content representation but also for content-based video browsing. Combining the structure information extracted from video parsing and keyframes extracted in video abstraction, we can build a visual table of contents of a video program.

Several terms and corresponding methods exist for abstracting video content, including skimming, highlights, and summary. A video skim is a condensed representation of the video containing keywords, frames, visual, and audio sequences. Highlights normally involve detection of important events in the video. A summary means that we preserved important structural and semantic information in a short version of the video represented via key audio, video, frames, and/or segments.

Keyframes play an important role in the video abstraction process. Keyframes are still images, extracted from original video data, that best represent the content of shots in an abstract manner. Often we use keyframes to supplement the text of a video log.[2] The representational power of a set of keyframes depends on how they're chosen from all frames of a sequence. Not all image frames within a sequence are equally descriptive, and the challenge is how to automatically determine which frames are most representative. An even more challenging task is to detect a hierarchical set of keyframes such that a subset at a given level represents a certain granularity of video content, which is critical for content-based video browsing. Researchers have developed many effective algorithms,[14] although robust keyframe extraction remains a challenging research topic.

Automatically extracting video highlights is an even more challenging research topic, because it requires more high-level content analysis. Rui et al.[15] have developed a way to reduce the usual three-hour baseball game—which normally includes long and uneventful close-ups—to 10 minutes of the most exciting highlights from each inning. They use audio features such as excited speech and baseball hits. The algorithm to determine the exciting parts weighs all the probabilities for different features using a support vector machine (SVM).

Li et al.[16,17] have recently reported another sports-related technology, where they propose algorithms for automatically detecting all segments containing interesting events of a particular game. Interesting events are specific to a particular sport. The proposed event detection algorithms use two types of prior knowledge to extract semantics from broadcast sports video: domain and production knowledge. Domain knowledge in sports means the definition of key events that are important for a particular sport, such as a play in American football, a hit in baseball, and a goal and its set up in soccer. Production knowledge refers to techniques used to produce the broadcast video. These techniques help viewers follow the game in an entertaining, informative, and captivating manner. They direct viewers' attention via well-accustomed production patterns that viewers expect, such as scene transitions after plays, replays, dynamic broadcaster logos sandwiching replay segments, certain camera angles that are sport or event specific, and scoreboard overlays.

The event detection algorithms in Li et al.[16,17] represent these two types of knowledge in terms of rule bases where rules are expressed in terms of low-level visual and aural features that are automatically computed from the media. They automatically detect all key events in baseball, American football, and sumo wrestling broadcast programs. Their algorithms[16] detect every play in a baseball broadcast video and every bout in a sumo match broadcast. One algorithm[17] automatically detects every play in an American football broadcast video. Plays include segments of the game where the ball is put into play and actively played, including pitches, hits, base steals, and home runs in baseball, and running or passing plays and field goals in football. By detecting an event, we mean detecting the start and end points of video segments containing the event. These algorithms use the methods Pan et al.[18,19] discuss for automatically detecting replay segments.

Li et al.[20] presented a prototype system demon-

strating the results of these technologies (referred to as High-Impact Sports). The prototype system used an MPEG-7-compliant XML description format for the event segments and an MPEG-7 browser that provided novel user interface paradigms offering summarized viewing or play-by-play nonlinear navigation. A play summary typically provides three to six times compaction in the viewing time for baseball and American football, depending on the particular game. The compaction ratio for sumo wrestling can be as high as 20 times.

It should be noted that the High-Impact Sports technology, unlike the approach in Rui et al.,[15] detects every single play without necessarily attempting to prioritize the events. It therefore isn't limited to generating short highlights. It serves the needs of a sports production studio or an avid sports fan that may want to see all the plays (or prefers to be in control of selecting the exciting plays). It facilitates efficient digital media management in a production environment. For sports fans, this technology provides the opportunity to catch a missed game during its regular broadcast and to consume even more sports.

A successful skimming approach involves using information from multiple sources, including sound, speech, transcript, and video image analysis. The Informedia project[1] is a good example of this approach, which automatically skims documentary and news videos with textual transcriptions by first abstracting the text using classical text skimming techniques and then looking for the corresponding parts in the video. This method creates a skim video, which represents a short synopsis of the original. The goal was to integrate language and image understanding techniques for video skimming by extracting significant information, such as specific objects, audio keywords, and relevant video structure. The resulting skim video is much shorter, where compaction is as high as 20 to 1, and yet retains the original segment's essential content. Another example[21] combines audio, video, speech, and text to process TV news programs. This approach results in the segmentation of the program into individual stories. The system selects a few representative images and keywords to represent each story's contents. The textual information provided by closed captions or derived from the audio track using speech recognition plays an important role in this process. When transcriptions of the program are generated by automatic speech recognition (ASR), satisfactory results may not be achievable using a keyword-driven approach to videos, where soundtrack contains more than just speech, such as movies. Sundaram and Chang[22] proposed a solution to the skimming problem based on two important questions:

- What's the relationship between the visual complexity of a shot and its comprehension time?
- How does syntactical structure in the video affect its comprehension?

They introduced a framework for determining visual skims and formulated the problem of skim generation as a general utility maximization problem with constraints. This is an important step because it allows for a principled way to impose additional constraints and make tradeoffs between them.

Summarization is another challenging topic because of the need to explore the content's structure. Liu and Kender[23] explore the type of scene changes to signal transitions between semantic units in the domain of documentaries. Their approach was to look for evidence for shot composition rules by means of Hidden Markov Models. They found that the best approach is one that trains the HMM with labeled subsequences that have approximately equal elapsed time, rather than subsequences with an equal number of shots, or subsequences with shots aligned to some semantic event. Agnihotri et al.[24] proposed summarization of video programs using the transcript. Their process involves cue extraction, categorization, classification, and a summarizer. Given a paragraph, the categorization process finds the underlying topic. Each of the 20 categories is an aggregation of a set of keywords related to that particular class. The summarizer exploits the underlying temporal structure and domain knowledge as well as textual cues in the transcript.

Indexing for retrieval and browsing

The structural and content attributes extracted in feature extraction, video parsing, and abstraction processes, or the attributes that are entered manually, are often referred to as metadata. Based on these attributes, we can build video indices and the table of contents through, for instance, a clustering process that classifies sequences or shots into different visual categories or an indexing structure. As in many other database systems, we

MPEG-7

Having realized the importance of content management, the Moving Pictures Expert Group (MPEG) started the standardization activity on content description. Formally called the Multimedia Content Description Interface, MPEG-7 provides a standardized description of various types of multimedia information, as well as descriptions of user preferences and usage history pertaining to multimedia information. The normative part of the standard focuses on a framework for encoding the descriptors and description schemes. The standard doesn't comprise the extraction of descriptors (features) or specify search engines that will use the descriptions. Instead, the standard enables the exchange of content between different content providers along the media value chain. In addition, it enables the development of applications that will use the MPEG-7 descriptions without specific ties to a single content provider. More information about MPEG-7 is available at the MPEG homepage (http://mpeg.telecomitalialab.com/).

MPEG-7 became an international standard in December 2001. This will have an impact on availability of additional information along with the images and video segments for many applications. This fact will help focus the needs for research on content analysis topics, which aren't available in the MPEG-7 description schemes.

need schemes and tools to use the indices and content metadata to query, search, and browse large video databases. Researchers have developed numerous schemes and tools for video indexing and query. However, robust and effective tools tested by thorough experimental evaluation with large data sets are still lacking. Therefore, in the majority of cases, retrieving or searching video databases by keywords or phrases will be the mode of operation. In some cases, we can retrieve with reasonable performance by content similarity defined by low-level visual features of, for instance, keyframes and example-based queries.

Often in queries of video clips, we want to quantify queries based on particular attributes that involve objects and subregions within the viewable image. Some support for automated object extraction could help us with these queries. The object-oriented compression scheme standardized by MPEG-4 provides an ideal data representation for supporting such indexing and retrieval schemes. It will also simplify the task of video structure parsing and keyframe extraction, because many of the necessary content features (such as object motion) are readily available. Zhang et al.[25] proposed a framework to use such content information in video content representation, abstraction, indexing, and browsing. Similarly, Chang et. al.[26] have applied object-based representation in indexing and retrieving video clips.

Although we tend to think of indexing for supporting fast retrieval of video clips, browsing is equally significant for video data, because the volume of video data requires techniques to present information landscape or structure to give a quick overview. By browsing, we mean a casual and quick access of content. The visual table of contents built based on structure information and keyframes provides an ideal representation for content-based video browsing. Browsing tools built based on such representation are especially useful, given that it still isn't feasible to automatically build semantic content-based indexing of video programs. However, browsing shouldn't be viewed as only a compromised tool for video indexing. Rather, it's an effective alternative and a complementary step for searching video data. Often users want quick access to relevant video data, although the process may initially lack any specific goal or focus. Browsing may suitably address those needs. Furthermore, browsing is also intimately related to and essential for video retrieval. It can help formulate queries, making it easier for the user to just ask around in the process of figuring out the most appropriate query to pose. Applications of such browsing tools include video editing and composition, where we often browse through a large number of relevant video clips before determining the final cut list.

We can also use MPEG-7 for multimedia indexing. We discuss this further in the sidebar "MPEG-7."

Application models: The user's perspective and research methodologies

We're facing a barrier similar to the one that the AI research community faced for many years: machine understanding of visual content. This is one of the major reasons that we haven't seen many convincing stories about successful applications of our research results, especially in the marketplace. In reviewing the past success in developing algorithms for video structure parsing, abstraction and content analysis, and in examining our research strategies and methodologies, we need to address several issues.

First, although a technology's success will be ultimately judged by its usefulness in the targeted applications, we should distinguish between the long-term research objective and short-term applications. Long-term research will result in general solutions to many applications. On the other hand, short-term applications will educate

the users for the potential of this technology while providing focus for some hard technical problems for the research community. The ultimate goal for the long-term research is to provide a link between the extracted low-level features and the high-level semantic description that humans perceive without significant effort. Nevertheless, many working systems exist that serve as proof to the effectiveness of even partial and domain-specific content descriptors for selective retrieval of visual information based on low-level feature extraction.

In designing these applications and products in the area of multimedia content analysis, we must keep the user in mind at all times, because users at different levels view a technology's usefulness differently. We can broadly classify users into two extremes:

- nontechnical consumers and
- trained, technical, professional corporate users who regularly use the products.

The requirements for each of these classes are different, so we should use different technologies to address their needs. Professional and consumer applications of video indexing technologies can both absorb functionalities, which stem from single modality processing, and extract even low-level features.

For technology-savvy users working with content analysis, indexing, searching, and authoring tools on a daily basis, it makes sense to design systems that require more user sophistication. For example, major news agencies and TV broadcasters own large video archives. If we develop automated indexing, analysis, and search products for these applications, it's conceivable to have trained individuals to retrieve and access required multimedia information, much the same way as today's trained professional librarians and information specialists retrieve information based on textual data. Under these circumstances, we can expect the operator or user to search images via textures, color histograms, or other low-level feature analysis that we wouldn't expect a typical consumer to be able or willing to cope with.

At the other extreme, for consumers, the products and applications need to be extremely simple for them to be viable in the marketplace. As an example, increasing consumer access to electronic imaging devices such as still digital cameras and digital camcorders has resulted in an explosion in the volume of data being generated. For consumers to annotate, index, handle, process, and access their data, products must be designed with simple yet useful functionalities. This would preclude searching techniques that are, for example, based on color histograms. Instead, consumers might want to find all the pictures in which uncle Joe is with the baby by defining once and for all, who uncle Joe and baby are pictorially. Clearly, from a technical point of view, this is harder to solve than searching color histograms. At the moment, no technically robust solutions exist for this problem. Generally speaking, the low-level features that most indexing and query systems are based on might prove to be nonviable in the consumer market in their raw form. We should conceal algorithms for low-level feature extraction from the user. For example, a video segmentation and filtering algorithm can only give the final visual table of contents of home video with simple interaction for quick overview and access. The intermediate step—setting thresholds—is prohibitively beyond the consumer's horizon.

Professional and educational applications

Professional activities that involve generating or using large volumes of video and multimedia data are prime candidates for taking advantage of video-content analysis techniques. Here we discuss several such applications.

Automated authoring of Web content

Media organizations and TV broadcasting companies have shown considerable interest in presenting their information on the Web. A survey conducted in 1998 by the Pew Research Center for the People and the Press indicated that the number of Americans who obtained their news on the Internet was growing at an astonishing rate. This survey indicated that 36 million people got their news online at least once a week. This number had more than tripled in a two-year period. (The full survey results are available at http://people-press.org/reports/).

The process of generating Web-accessible content usually involves using one of several existing Web-authoring tools to manually compose documents consisting of text, images, and possibly audio and video clips. This process usually consumes considerable amounts of time. When

Image courtesy of NBC news

Figure 2. A snapshot from the Pictorial Transcripts system.

Image courtesy of NBC news

Figure 3. A snapshot from the DVL system at AT&T.

other representations of the same content are already composed for presentation in the video form, we can use such presentations to repurpose the content for the Web, thereby reducing work. Analysis of the already composed video content through image and video understanding, speech transcription, and linguistic processing can serve to create alternative presentations of the information suitable for the Web. We can automatically convert large video archives to digital libraries. We can also automatically augment these Web presentations with related and supplementary information and thereby create a richer source of information than the original video programs. An example of such an automated authoring system is the Pictorial Transcripts system.[2–5]

Pictorial Transcripts uses video and text analysis techniques to convert closed-captioned video programs to Hypertext Markup Language (HTML) presentations with still frames containing the visual information accompanied by text derived from the closed captions. A content-based sampling method[14] performs the task of reducing the video frames into a small set of images that represent the visual contents of each scene in a compact way. This sampling process is based on detecting cuts and gradual transitions, as well as a quantitative analysis of the camera operations. Linguistic analysis of closed-caption text refines the text, generates textual indices, and creates links to supplementary information. Figure 2 shows a sample screen for the Pictorial Transcripts that uses the output of the content-based sampling and the processed text from the closed captions. We can easily retrieve such a compact presentation over low-bandwidth communications networks. When more bandwidth is available, the presentation can include audio and video information. In this case, the still images and text serve as a pictorial and textual index into the audio and video media components. Users can search and browse a digital video library (DVL) created in this way using pictorial information as well as textual information extracted from closed captions, or recognized speech, to retrieve selective pieces of video from a large archive (see Figure 3).

The AT&T DVL system employs additional media processing techniques to improve the organization and presentation of the video information. The system can use speech processing to correct misalignment between the audio track and the closed-caption text. When closed-caption text isn't available, it can employ a large vocabulary automatic speech recognizer (LVASR) to generate a transcript of the program from the audio track (see Figure 4). The quality of the automatically generated transcripts is determined by several factors, such as the quality of speech, background noise, vocabulary size, and language models. Although we can obtain high-quality results under favorable conditions, the accuracy of such automatically generated transcripts is generally below those generated manually and therefore isn't suitable for direct presentation to the users. Nevertheless, these automatically generated transcripts provide a viable alternative to the closed-caption text for information retrieval purposes. When sufficient bandwidth is available, it can deliver video presentations (such as

TV programs) with the same quality as the original productions. The real-time transport protocol (RTP) and the specific payload types defined by the Internet Engineering Task Force (IETF) have already made it possible to deliver high-quality MPEG-2 encoded video over IP networks. In the short term, this will only be feasible over private local IP networks. In the long term, however, this will let us create searchable and browsable TV.[6,12]

Searching and browsing large video archives

Another professional application of automated media content analysis is in organizing and indexing large volumes of video data to facilitate efficient and effective use of these resources for internal use. Major news agencies and TV broadcasters own large archives of video that have been accumulated over many years. Besides the producers, others outside the organization use the footage from these archives to meet various needs. These large archives usually exist on numerous different storage media, ranging from black-and-white film to magnetic-tape formats.

Traditionally, the indexing information used to organize these large archives has been limited to titles, dates, and human-generated synopses. We generally use this information to select video programs possibly relevant to the application at hand. We ultimately discern the material's relevance by viewing the candidate programs linearly or nonlinearly. Converting these large archives into digital form is a first step in facilitating the search process. This in itself is a major improvement over the old methods. We must address several practical issues, however, to make such an undertaking feasible and economical. These large video libraries create a unique opportunity for using intelligent media analysis techniques to create advanced searching and browsing techniques to find relevant information quickly and inexpensively. Intelligent video segmentation and sampling techniques can reduce the visual contents of the video program to a small number of static images. We can browse these images to spot information and use image similarity searches to find shots with similar content and motion analysis to categorize the video segments. Higher-level analysis can extract information relevant to the presence of humans or objects in the video. Audio event detection and speech detection can extract additional information to help the user find segments of interest.

Despite the limitations in the robust and efficient extraction of information from the constituent media streams, existing methods are effective in reducing manual labor.[27]

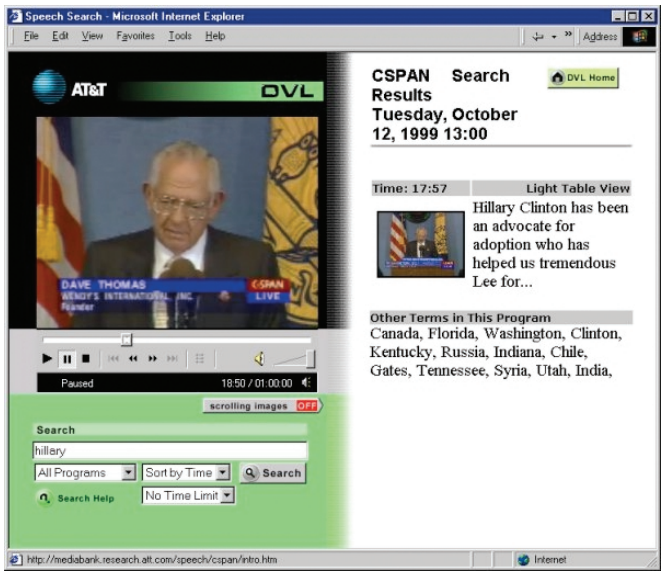


Image courtesy of C-SPAN

Figure 4. Search based on automatic speech recognition.

Easy access to educational material

The availability of large multimedia libraries that we can efficiently search has a strong impact on education. Students and educators can expand their access to educational material. The Telecommunications Act of 1996 has acknowledged the significance of this. It has special provisions for providing Internet access to schools and public libraries. This holds the promise of turning small libraries that contain a small number of books and multimedia sources into ones with immediate access to every book, audio program, video program, and other multimedia educational material. It also gives students access to large data resources without even leaving the class.

Indexing and archiving multimedia presentations

Intelligently indexing multimedia presentations is another area where content-based analysis can play a major role. Existing video compression and transmission standards have made it possible to transmit presentations to remote sites. We can then store these presentations for on-demand replay. Different media components of the presentation can be processed to characterize and index it. Such processing could include analyzing the speaker's gestures, slide transition detection, extracting textual

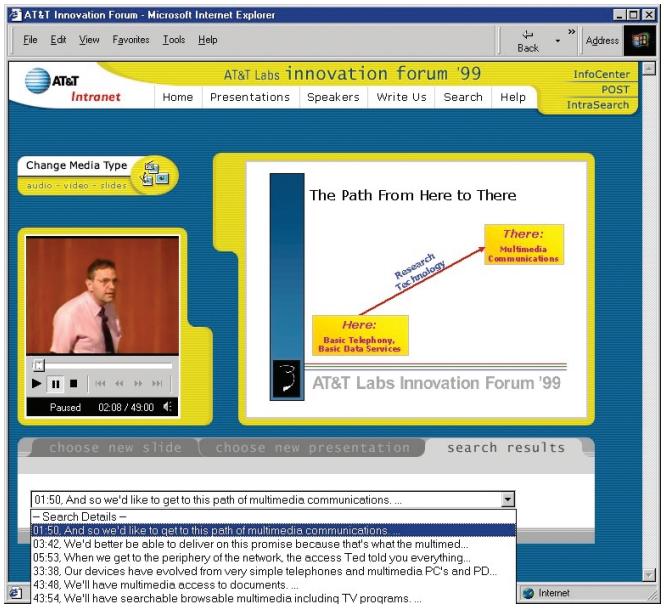


Figure 5. Sample screen of the system for indexing and archiving multimedia presentations.

information by performing optical character recognition (OCR) on the slides, speech recognition, speaker identification and discrimination, and audio event detection. The information extracted by this processing generates powerful indexing capabilities that would enable content-based retrieval of different segments of a presentation. Users can search an archive of presentations to find information about a topic.

Figure 5 shows a sample screen from a technical presentation with the speaker in the left window. The slides synchronize with the talk. This is done using a specialized scene change detection algorithm to find the slide transitions. An offline-generated transcript of the talk synchronizes with the video using speech processing and searches and jumps to the correct points in the talk (as the search results window shows at the bottom of Figure 5).

Indexing and archiving multimedia collaborative sessions

Multimedia collaborative systems can also benefit from effective multimedia understanding and indexing techniques. Communication networks give people the ability to work together despite geographic distances. The multimedia collaborative sessions involve real-time exchange of visual, textual, and auditory information. The information retained is often limited to the collaboration's end result and doesn't include the steps that were taken or discussions that took place. We can set up archiving systems to store all the information together with relevant synchronization information. Content-based analysis and indexing of these archives based on multiple information streams enable the retrieval of segments of the collaborative process. Such a process lets users not only access the end result but also the process that led to those results. When the communication links used for the collaborative session are established by a conferencing bridge, we can use the available data in the indexing process, thereby reducing the processing required to identify each stream's source.

Consumer domain applications

Video-content analysis research is geared toward large video archives. However, the widest audience for video-content analysis is consumers. We all have video content pouring through broadcast TV and cable. Also, as consumers, we own unlabeled home video and recorded tapes. To capture the consumer's perspective, the management of video information in the home entertainment area will require sophisticated yet feasible techniques for analyzing, filtering, and browsing video by content.

The methods dealing with video information in the consumer domain will have different requirements. In large archives, we store data in files, and we can access it repeatedly and slower than in real time. Therefore, the algorithms for content extraction can operate at rates slower than 30 fps. In consumer devices, however, the content may be available only during real-time display (recording or playback). Consequently, we can analyze video data only in real time. In large archives, we assume that the workstation for video management has considerable power (and possibly hardware support) to run video-content analysis algorithms. In the consumer domain, recording and display devices are impoverished from the point of view of information processing—devices normally have limited memory and processor power. Therefore, the algorithms must run with all the constraints in real time. In addition, these algorithms have different kinds of accessory information available. Large video archives will probably have the information about the author, actors, and storyboard. However, in the consumer domain, we can probably expect metadata to be available in the broadcast stream or in the electronic program guide. Researchers are developing standards in

this area, such as digital video brocasting service information (DVB-SI) and MPEG-7, which will add descriptions and accessory data to the stored and streamed content.

Consumer devices can use many additional features for video cataloging, advanced video control, personalization, profiling, and time-saving functions. Information filtering functions for converging PCs and TVs will add value to video applications beyond digital capture, playback, and interconnect. We can use these consumer applications in video editing, cataloging applications, enhanced access, and filtering applications.

Video overview and access

An example of a video home library application that performs VHS tape cataloging functions is Video Indexing for True Access and Multimedia Information Navigation (Vitamin).[28] In this system, the video-content analysis process extracts visual information, which it then archives and presents to the user as a visual table of contents for the analyzed video. The purpose is to later use this information for retrieval purposes in a master index. This prototype has an archival and a retrieval module to perform these two functions.

During archiving, the system performs abrupt scene change and static scene detection based on a comparison of discrete cosine transform (DCT) coefficients of subsequent frames (MPEG-1 and MPEG-2). However, from the user's perspective, not all the keyframes are important or necessary to convey the video's visual contents. We apply a keyframe filtering method to reduce the number of keyframes. We reduce the number of frames by filtering out noisy, blurry, unicolor, and repetitive frames. For example, in a dialogue scene, it's likely that both speakers will be shown several times, requiring two frames to represent the dialogue scene.

During the keyframe selection process, we use frame signatures to compare keyframes and to detect a particular frame's content. A keyframe signature representation is derived for each grouping of similarly valued DCT blocks in a frame. By using different thresholds, we can control the number of filtered keyframes. The signatures are also used for spotting certain patterns, which might correspond to objects of interest. Selected keyframes are structured in a temporal hierarchy, which is flattened to aid in optimal retrieval, even from slow storage devices. The system presents this hierarchy to the user in a visual table of contents (see Figure 6.) The user can browse and navigate through the visual index and fast-forward or rewind to a certain point on the videotape or MPEG file.

Figure 6. A snapshot from the visual table of contents user interface in a home library application.

Video content filtering

In the consumer domain, some products already perform video-content analysis and filtering functions—for example, VCRs with the automatic commercial skip feature. Most systems work by detecting black frames and changes in activity.

When people watch a TV program, such as *Seinfeld*, they immediately recognize the nonprogram segments in the broadcast. This is because they recognize the broadcast's switch in context. Some characteristics of commercials are therefore naturally different than the TV program. Such special characteristics include rapid scene changes, repetitive nature, use of text with different sizes, transitional monochrome frames into and out of the commercial break, and absence of the station logo. Based on these characteristics, suitable methods for advertisement isolation are cut rate, average keyframe distance, black frame, static frame rate, similar frame distance, text location detection, logo detection, audio analysis, and memorizing (autolearning) advertisements. The commercial breaks are usually preceded and terminated by a series of black frames.

We can often detect commercials by determining when a high number of cuts per minute

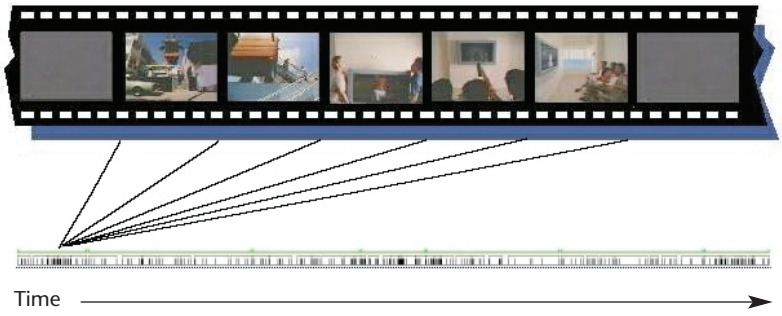


Figure 7. Keyframes from an area with a high density of cuts representing a commercial break.

occur in conjunction with the identification of the black-frame series. Specifically, by using the relative times between cuts, we can determine the cut density. However, action movies may also have prolonged scenes with a large number of cuts per minute. For a more reliable commercial isolation, we can analyze a total distribution of the cuts and black frames in the source video in an attempt to deduce their frequency and length.

We analyze the keyframes for color uniformity and similarity to previously selected keyframes. We use the DCT coefficients to create a frame signature for the keyframe clustering process. During this process, we use the frame signatures to compare keyframes and detect a particular frame's content. In addition, we analyze the video cut rate and determine the duration of the cut-rate change. We also identify the black frame boundaries and analyze the total distribution of the cut-rate change in the program to deduce the frequency and length of the cuts. This can help us determe the likelihood of commercial segments.

Figure 7 shows the keyframes extracted from a commercial break. The vertical lines at the bottom represent the cuts, and a high density of cuts-per-unit time (cut rate) may represent a commercial break, as Figure 7 depicts. The cut rate alone produces many false positives. To reduce the number of false positives, we can examine the false-positive sections of the commercials for the presence or absence of text. All in this example had a significant amount of text. The text varies significantly in the position on the TV screen as well as size. The type of text present in some of these areas was scene text (for example, text on cars, helicopters, or police vehicles), buildings with names, or a product logo.

For consumer applications, we must detect commercials on a constrained platform.[29] We've also developed methods that use features produced during MPEG compression. We developed an algorithm that uses features as triggers and verifiers to detect commercials. Our algorithm first looks for a triggering event, such as a black frame, to mark a potential commercial start. Once a potential start is found, it uses other characteristic features to verify the commercial break. We've achieved a recall of 93 percent and a precision of 95 percent when station logos and trailers are excluded and a precision of 99 percent when station logos and trailers are regarded as a part of the commercial break.

Enhanced access to broadcast video

Intelligent access and enhanced search tools for broadcast content is an important area where content-based analysis can have a strong contribution. The existing high-definition TV, video compression, and transmission standards have made it possible to transmit a large number of high-quality TV channels to the consumer over the broadcast networks and the Internet. However, the only available tools to grapple with the growing number of TV channels are the scrolling program guide (at least in the US) and the old-fashioned paper guide. The program guide is envisioned under the analog paradigm of passively receiving the entertainment. The scrolling isn't interactive and is valid for only a limited time. Within the Society of Motion, Picture, and Television Engineers (SMPTE), TVAnytime, MPEG, and DVB, there are ongoing efforts to provide auxiliary information to the regular broadcast stream. Current personal video recorders on the market use an electronic program guide for an interactive selection of programs to watch or store. There could be layers of additional personalization of content where new video-content analysis algorithms could be employed. The combined information extracted by this video-content processing will generate powerful indexing capabilities that would enable content-based retrieval of different segments of TV programs online or offline.

We developed Video Scout, a content-based retrieval system for personalizing TV at a subprogram level.[6] Users make content requests in their user profiles and then Scout begins recording TV programs. In addition, Scout actually watches the TV programs it records and personalizes program segments. Scout analyzes the visual, audio, and transcript data to segment and index the programs. When viewing full programs, users see a high-level overview as well as topic-specific starting points. (For example, users

can quickly find and play Dolly Parton's musical performance within an episode of *Late Night with David Letterman*.) In addition, users can access video segments organized by topic (such as finding all the segments on Philips Electronics that Scout has recorded from various financial news programs). We divide Scout's interface into two sections—program guide and TV magnets. The program guide lets users interact with whole TV programs that they can segment in different ways. TV magnets (see Figure 8) let users access their profiles and video clips organized by topic. Users navigate the interface on a TV screen using a remote control.

The archiving module employs a three-layered, multimodal integration framework to segment, analyze, characterize, and classify segments. The multimodal segmentation and indexing incorporates a Bayesian framework that integrates information from the audio, visual, and transcript (closed-caption) domains. This framework uses three layers to process low-, mid-, and high-level multimedia information. The retrieval module relies on users' personal preferences to deliver both full programs and video segments. In addition to using electronic program guide metadata and a user profile, Scout lets users request specific topics within a program. For example, users can request the video clip of the US President speaking from a half-hour news program. The high-level layer generates semantic information about TV program topics used during retrieval.

Advanced access to broadcast video must take into account user preferences captured in a user profile. Ferman et al.[30] proposed automatic user profiling and filtering agents. The profiling agent, based on fuzzy reasoning, automatically generates the users' profile on the basis of their usage history. Given program description metadata, the filtering agent filters programs on the basis of the user's profile. The agents developed by Ferman et al.[30] can generate MPEG-7 and TV-Anytime compliant usage history and user preference descriptions as well as filtering programs that are described by MPEG-7 and TV-Anytime compliant descriptions.

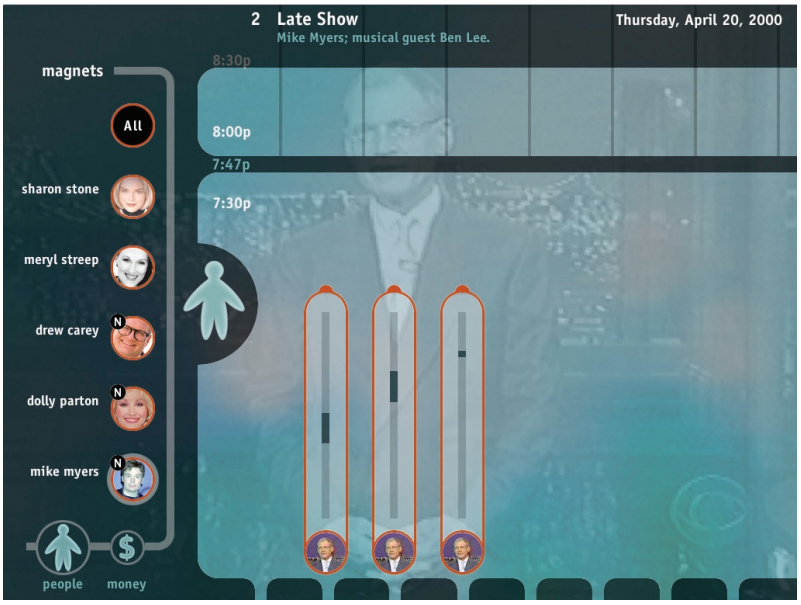


Figure 8. Content magnets attracting story segments in the content-based personal video recorder Video Scout application.

Conclusions

To keep things in perspective, it's important to distinguish between research activities, experiments, and real applications that have made, or are likely to make, the transition from research labs into the real world. Researchers and technologists, who are constantly reminded of the level of difficulty in meeting certain technological challenges, are more likely to be excited by new technologies that may not yet be ready for use. The targeted users are the ultimate judges of the technology's usefulness in meeting their needs. On the other hand, accepting new ways of doing things often involves a change in mindset. Users accustomed to performing a certain task by using existing tools and methods might have a tendency to resist new tools and methods. For example, most people who are accustomed to text-based information retrieval techniques may not feel as comfortable with the notion of performing image and video searches by nonlinguistic queries. Well-designed prototype applications can help bring about the necessary change in mindset for users to accept these applications. Such prototypes also serve the purpose of making technologists aware of users' requirements and preferences. **MM**

References

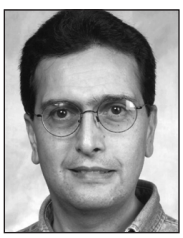
1. M.G. Brown et al., "Automatic Content-Based Retrieval of Broadcast News," *Proc. 3rd Int'l Conf. Multimedia* (ACM Multimedia 95), ACM Press, New York, 1995, pp. 35-43.
2. B. Shahraray and D.C. Gibbon, "Automatic Generation of Pictorial Transcripts," *Proc. SPIE Conf. Multimedia Computing and Networking 1995*, SPIE Press, Bellingham, Wash., 1995, pp. 512-518.
3. B. Shahraray and D.C. Gibbon, "Automated Authoring of Hypermedia Documents of Video Programs," *Proc. 3rd Int'l Conf. Multimedia* (ACM

Multimedia 95), ACM Press, New York, 1995, pp. 401-409.
4. B. Shahraray and D.C. Gibbon, "Efficient Archiving and Content-Based Retrieval of Video Information on the Web," *Proc. AAAI Symp. Intelligent Integration and Use of Text, Image, Video, and Audio Corpora*, 1997, pp. 133-136.
5. B. Shahraray, "Multimedia Information Retrieval using Pictorial Transcripts," *Handbook of Multimedia Computing*, B. Furht, ed., CRC Press, Boca Raton, Fla., 1999, pp. 345-359.
6. R.S. Jasinschi et al., "Integrated Multimedia Processing for Topic Segmentation and Classification," *Proc. IEEE Int'l Conf. Image Processing* (ICIP 2001), IEEE CS Press, Los Alamitos, Calif., 2001.
7. H.J. Zhang et al., "Video Parsing, Retrieval, and Browsing: An Integrated and Content-Based Solution," *Proc. Third Int'l Conf. Multimedia* (ACM Multimedia 95), ACM Press, New York, 1995, pp.15-24.
8. R. Zabih, K. Mai, and J. Miller, "A Robust Method for Detecting Cuts and Dissolves in Video Sequences," *Proc. 3rd Int'l Conf. Multimedia* (ACM Multimedia 95), ACM Press, New York, 1995.
9. H.J. Zhang et al., "Video Parsing Using Compressed Data," *Proc. SPIE 94 Image and Video Processing II*, SPIE Press, Bellingham, Wash., 1994, pp.142-149.
10. D. Swanberg, C.-F. Shu, and R. Jain, "Knowledge-Guided Parsing in Video Databases," *Proc. SPIE Conf. Storage and Retrieval for Image and Video Databases*, SPIE Press, Bellingham, Wash., 1993.
11. H.J. Zhang et al., "Automatic Parsing and Indexing of News Video," *Multimedia Systems*, vol. 2, no. 6, 1995, pp. 256-265.
12. Q. Huang et al., "Automated Generation of News Content Hierarchy by Integrating Audio, Video, and Text Information," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing* (ICASSP 99), IEEE CS Press, Los Alamitos, Calif., 1999.
13. P. Aigrain and P. Joly, "The Automatic Real-Time Analysis of Film Editing and Transition Effects and Its Applications," *Computers & Graphics*, vol. 18, no. 1, Jan./Feb. 1994, pp. 93-103.
14. B. Shahraray, "Scene Change Detection and Content-Based Sampling of Video Sequences," *IS&T/SPIE Symp. Digital Video Compression: Algorithm and Technologies*, SPIE Press, Bellingham, Wash., vol. 2419, 1995, pp. 2-13.
15. Y. Rui, A. Gupta, and A. Acero, "Automatically Extracting Highlights for TV Baseball Programs," *Proc. 8th Int'l Conf. Multimedia* (ACM Multimedia 2000), ACM Press, New York, 2000, pp. 105-115.
16. B. Li and M.I. Sezan, "Event Detection and Summarization in American Football Broadcast Video," *Proc. IS&T/SPIE Conf. Storage and Retrieval for Media Databases*, SPIE Press, Bellingham, Wash., 2002, pp. 202-215.
17. B. Li and M.I. Sezan, "Event Detection and Summarization in Sports Video," *Proc. IEEE Workshop on Content-Based Access to Video and Image Libraries*, IEEE CS Press, Los Alamitos, Calif., 2001, CD-ROM.
18. H. Pan, B. Li, and M.I. Sezan, "Automatic Detection of Replay Segments in Broadcast Sports Programs by Detection of Logos in Scene Transitions," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing* (ICASSP 2002), IEEE CS Press, Los Alamitos, Calif., 2002, CD-ROM.
19. H. Pan, P. van Beek, and M.I. Sezan, "Detection of Slow-Motion Replay Segments in Sports Video for Highlights Generation," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing* (ICASSP 2001), IEEE CS Press, Los Alamitos, Calif., 2001, CD-ROM.
20. B. Li et al., "Sports Program Summarization," *IEEE Computer Vision and Pattern Recognition (CVPR) Conf. Demonstration Session*, IEEE CS Press, Los Alamitos, Calif., 2001, CD-ROM.
21. D.C. Gibbon and J. Segen, "Video Based Detection of People from Motion," *Proc. Virtual Reality Systems Conf.*, 1993.
22. H. Sundaram and S.-F. Chang, "Constrained Utility Maximization for Generating Visual Skims," *Proc. IEEE Workshop on Content-Based Access of Image and Video Libraries* (CBAIVL-2001), IEEE CS Press, Los Alamitos, Calif., 2001.
23. T. Liu and J.R. Kender, "A Hidden Markov Model Approach to the Structure of Documentaries," *Proc. Computer Vision and Pattern Recognition: Workshop on Content-Based Access of Image and Video Libraries* (CVPR 00), IEEE CS Press, Los Alamitos, Calif., 2000.
24. L. Agnihotri et al., "Summarization of Video Programs Based on Closed Captioning," *Proc. SPIE Conf. Storage and Retrieval in Media Databases*, SPIE Press, Bellingham, Wash., 2001, pp. 599-607.
25. H.J. Zhang, J.Y.A. Wang, and Y. Altunbasak, "Content-Based Video Retrieval and Compression: A Unified Solution," *Proc. IEEE Int'l Conf. Image Processing*, IEEE CS Press, Los Alamitos, Calif., 1997.
26. S.F. Chang et al., "VideoQ: An Automated Content-Based Video Search System Using Visual Cues," *Proc. 5th Int'l Conf. Multimedia* (ACM Multimedia 97), ACM Press, New York, 1997, pp. 313-324.
27. D. Gibbon et al., "Browsing and Retrieval of Full Broadcast-Quality Video," *Proc. Packet Video Conf.*, 1999, CD-ROM.
28. N. Dimitrova, T.McGee, and H. Elenbaas, "Video Keyframe Extraction and Filtering: A Keyframe Is Not a Keyframe to Everyone," *Proc. ACM Conf. Knowledge and Information Management*, ACM

Press, New York, 1997, pp.113-120.
29. N. Dimitrova et al., "Real-Time Commercial Detection Using MPEG Features," to appear in *Proc. 9th Int'l Conf. Information Processing and Management of Uncertainty in Knowledge-Based Systems* (IPMU 2002), July 2002.
30. A.M. Ferman et al., "Content-Based Filtering and Personalization Using Structured Metadata," to appear in *Joint Conf. Digital Libraries*, 2002.

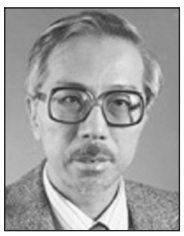
Nevenka Dimitrova is a research staff member at Philips Research. Her main research interests are in content information management, digital TV, content synthesis, video content navigation and retrieval, MPEG-7, and advanced multimedia systems. She has a BS in mathematics and computer science from the University of Kiril and Metodij, Skopje, Macedonia, and an MS and PhD in computer science from Arizona State University. She is on the editorial board of *IEEE MultiMedia*, *ACM Multimedia Systems Journal*, *ACM Transactions of Information Systems*, and *IEEE Communications Society* (an online magazine). She is a program chair of ACM Multimedia 2002 in the area of content processing.

Hong-Jiang Zhang is a senior researcher and assistant managing director at Microsoft Research Asia. He has a BS from Zhengzhou University, China, and PhD from the Technical University of Denmark, both in electrical engineering. He is a senior IEEE member and an ACM member. He currently serves on the editorial boards of five professional journals and a dozen committees of international conferences.



Behzad Shahraray is a division manager at AT&T Labs Research, where he heads the Multimedia Processing Research Department. His work focuses on multimedia indexing, multimedia data mining, content-based video sampling, and automated authoring of searchable and browsable multimedia content. He has MS degrees in electrical engineering and computer, information, and control engineering, as well as a PhD in electrical engineering from the University of Michigan, Ann Arbor. He's an IEEE and ACM member. He serves on the editorial board of the *International Journal of Multimedia Tools and Applications*.

Ibrahim Sezan is the director of the Information Systems Technologies Department at Sharp Laboratories of America. His research focuses on audio–visual content understanding and content summarization, automatic user profiling and content filtering, human visual system models, visually optimized information display algorithms for flat-panel displays, and smart algorithms for cameras. He has BS degrees in electrical engineering and mathematics from Bogazici University, Istanbul, Turkey. He has an MS in physics from the Stevens Institute of Technology, Hoboken, New Jersey, and a PhD in electrical computer and systems engineering from Rensselaer Polytechnic Institute, Troy, New York.



Thomas Huang is the William L. Everitt Distinguished Professor of Electrical and Computer Engineering at the University of Illinois at Urbana–Champaign (UIUC). He also serves at UIUC as a research professor at the Coordinated Science Laboratory, and he's the head of the Image Formation and Processing Group at the Beckman Institute for Advanced Science and Technology. He has a BS in electrical engineering from National Taiwan University, China, and an MS and ScD in electrical engineering from the Massachusetts Institute of Technology.

Avideh Zakhor is a professor in the Department of Electrical Engineering and Computer Sciences at the University of California at Berkeley. Her research interests include image and video processing, compression, and communication. She has a BS from the California Institute of Technology, Pasadena, and an MS and PhD from the Massachusetts Institute of Technology, all in electrical engineering. She is an IEEE fellow.

Readers may contact Nevenka Dimitrova at Phillips Research, 345 Scarborough Rd., Briarcliff Manor, NY 10510, email nevenka.dimitrova@philips.com.