TECHNIQUES FOR THE CREATION AND EXPLORATION OF DIGITAL VIDEO LIBRARIES

Michael Christel, Scott Stevens,
Takeo Kanade, Michael Mauldin, Raj Reddy, and Howard Wactlar

1.1 Introduction

The Information Age is fully upon us. A recent article noted that there are perhaps 50 million people using the Internet on a regular basis, and that "the current growth rate is about 15% per month (!) and this could well continue until almost all of those in the 'developed world' are connected" [Fenn94, p. 30]. In addition, the digital domain consists not only of text but increasingly of other media representations, from graphics images to audio to motion video. As the amount of information and number of users exponentially escalate, more attention focuses on the basic problems of information management: How do you digitize information? How can you then visualize it and find what you need? How do you use and manipulate it effectively? How is it stored and managed? The proliferation of technical articles and special issues addressing these questions underscore their importance; see for example the special issue on content-based retrieval [Narasimhalu95] or digital libraries [Fox95]. This chapter will survey some of that work, especially that which relates to the treatment of video and the use of digital video libraries for education.

With the growth and popularity of multimedia computing technologies, users are able to store greater amounts of information and retrieve data more quickly than ever before. Advances in data compression, storage, and telecommunications have enabled video to become an important data type for the future. However, it is not enough to simply store and play back complete video movies as in commercial video-on-demand services. New techniques are needed to organize and search these vast data collections, retrieve the most relevant selections, and effectively reuse them.

Emerging techniques for digital video libraries will allow independent, self-motivated access to information for self-teaching, exploration, and research. The potential impact on training and education delivery is critical, considering that U.S. schools and industry together spend between $400 and $600 billion per year on education and training, an activity that is 93% labor-intensive, with little change in teacher productivity ratios since the 1800s [Perelman90]. Digital video libraries offer the potential to deliver vicarious field trips to places that are too dangerous or expensive to visit in person. Digital video libraries allow virtual guest speakers and topic experts to deliver talks and be interviewed in the classroom and at home, and provide virtual access to rare, unique, expensive, or dangerous materials in a safe, comfortable educational setting. Exploring Antarctica can be done without the need for a winter coat; the results of combining volatile chemicals can be witnessed without fear of bodily harm. Marchionini and Maurer outline more promises of digital libraries for education in their paper [Marchionini95], but caution that for this promise to be tapped, the information embedded within the digital library must become

easy to find, manage, and use. This chapter will discuss the challenges introduced when video is included as a primary component of a digital library.

Video poses unique problems because of the difficulties in representing its contents. It is well known that if you take a page from a book and electronically scan it into a raster image, the raster image will use a significantly greater number of bits than would an ASCII representation of the original text [Srihari94]. While page description languages may be more efficient, if the page contains many images, then a raster image may be the only choice for representation. Video is not only imagery, but consists of 30 images per second. The adage “a picture is worth a thousand words” was never more appropriate. Detailed descriptions of video images can be many thousands of words and even a short video clip description can be massive. But the alternative of no description leaves even the shortest video clip a black box, giving the user no way to know what is within it short of viewing it.

The problems of creating a digital video library such as gathering video, representing its contents, and segmenting it appropriately will be discussed in the next section. In order to utilize and explore the library, a user must be able to effectively retrieve and browse its holdings, as well as perhaps reuse the materials in a different context. These issues will be addressed as well. The concluding sections will discuss technological solutions to the problems posed, and then present the work and strategy of the digital video library project at Carnegie Mellon University, where such technologies are being integrated in establishing a terabyte, one thousand hour digital video library testbed.

1.2 ISSUES WITH VIDEO IN A DIGITAL VIDEO LIBRARY

1.2.1 Building a Video Database

Digital video takes a tremendous amount of space. A single high quality, uncompressed video channel would require a bandwidth of 200 million bits per second. Such bandwidth requirements are not practical today or perhaps ever, so the quality of the video may be reduced and compression schemes used to make possible the inclusion of video into digital libraries. For example, the MPEG algorithm for video compression was designed to deliver good quality at a very high compression ratio and random access to various points within the sequence. It is a scalable algorithm allowing more quality at the expense of requiring greater bandwidth. The MPEG1 SIF resolution will work for standard CD-ROM bandwidth requirements (1.2 Megabits per second), allowing 352 X 240 resolution at 30 frames per second or 352 X 288 resolution at 25 frames per second, thus delivering VHS quality NTSC/PAL video. MPEG and other digital video compression techniques are dealt with in detail in a special issue on the topic [CACM91].

Even before the video can be digitized and placed into the library, a number of intellectual property rights issues need to be resolved. As discussed by Pamela Samuelson and others [Samuelson95, Samuelson93], new legal rules will likely be established and evolve as

consumers and publishers move fully into the electronic age where copying is simple, accurate, and cheap. In the coming years, the U.S. Library of Congress will play a leading role in the resolution of problems of copyright and intellectual property rights with respect to digital libraries [Becker95]. For now, these problems can be dealt with in the following ways:

- only include public domain resources in the digital library, or resources for which you have proper permissions (the approach taken by the Library of Congress Digital Library effort to date [Becker95])
- make arrangements with resource providers for remuneration and proper attribution; then control access to the digital video library so that owners and retailers of information can be paid when their materials are accessed. NetBill is one such electronic commerce mechanism enabling a market economy in information and providing all of the services necessary to account for intellectual property delivered via a network [Sirbu95].

A third consideration in the creation of a digital library is enabling access to the information. Even with MPEG1 compression, a thousand hours of video will take approximately a terabyte of storage, and so it is highly unlikely that user workstations will have the complete library stored locally at their machines! Rather, a key element of on-line digital video libraries will be the communication fabric through which media servers and satellite (user) nodes are interconnected. Traditional modem-based access over voice-grade phone lines is not adequate for this multimedia application, as evidenced by the difficulty in trying to move VHS-quality video between arbitrary sites on the Internet. The ideal fabric has the following characteristics:

- communication should be transparent to the user. Special-purpose hardware and software support should be minimized in both server and slave nodes.
- communication services must be cost effective, implying that link capability (bandwidth) be scalable to match the needs of a given node. Server nodes, for example, will require the highest bandwidth because they are shared among a number of satellite nodes.
- the deployment of a custom communication network should be avoided. The most cost-effective, and timely, solution will build on communication services already available or in field-test.

A number of commercial video-on-demand networks have been deployed in trials across the U.S. These networks may prove suitable as well for access to digital video libraries. Network issues concerning the delivery of digital video are covered in recent conferences dealing with multimedia systems [ICMCS94, ACM94].

A complete discussion of compression, networking, or intellectual property issues could easily consume the remainder of this chapter, and so they will not be dealt with further here. There are network topologies capable of supporting MPEG1 video delivery, and intellectual property rights can be accounted for with schemes such as NetBill. The remainder of this chapter will deal with the tasks of indexing, segmenting, retrieving, and reusing video once it has been acquired for the library.

1.2.2 Indexing the Video Contents

A library cannot be very effective if it is merely a collection of information without some understanding of what is contained in that collection. Without that understanding it could take hundreds of hours of viewing to determine if an item of interest is in a 1000 hour video library. Obviously, such a library would not be used very often. Marchionini and Maurer reflect on information accessible via the Internet [Marchionini95, p. 72]:

> It has often been said that the Internet is starting to provide the largest library humankind has ever had. As true as this may be, the Internet is also the messiest library that ever has existed.

Information is found best on the Internet when the providers augment the information with rich keywords and descriptors, provide links to related information, and allow the contents of their pages to be searched and indexed. There is a long history of sophisticated parsing and indexing for text processing in various structured forms, from ASCII to PostScript to SGML and HTML. However, how does one represent video content to support content-based retrieval and manipulation?

An hour-long motion video segment clearly contains some information suitable for indexing, so that a user can find an item of interest within it. The problem is not the lack of information in video, but rather the inaccessibility of that information to our primarily text-based information retrieval mechanisms today. In fact, the video likely contains an overabundance of information, conveyed in both the video signal (camera motion, scene changes, colors) and the audio signal (noises, silence, dialogue). A common practice today is to log or tag the video with keywords and other forms of structured text to identify its contents. Such text descriptors have the following limitations:

- Manual processes are tedious and time consuming.
- Manual processes are seriously incomplete. Even if full transcripts of the audio track are entered, other information about the video will almost surely be left out, such as the identity of persons and objects in each scene.
- Transcripts are inaccurate, with mistypings and incorrect classifications often introduced.
- Text descriptors are biased by whatever predetermined structures are used to classify the video contents. For example, if you have a classification of “inside or outside”, how do you tag a scene of people in a cave?
- Cinematic information is complex and difficult to describe, especially for non-experts. For example, in an establishing shot that zooms from a wide angle to a close-up, determining the point when the scene changed is open to interpretation.
- Text descriptors are biased by the ambiguity of natural language. For example, one indexer may decide to label a particular video segment as occurring in a city street. Another may decide to label the same segment as occurring in a New York alley. These different tags have implications for later browsing and retrieval of the video.

1.2.3 Breaking the Video into Segments

Anyone who has retrieved video from the Internet realizes that because of its size a video clip can take a long time to move from one location to another, such as from the digital video library to the user. Likewise, if a library consists of only 30 minute clips, when users check one out it may take them 30 minutes to determine whether the clip met their needs. Returning a full one-half hour video when only one minute is relevant is much worse than returning a complete book, when only one chapter is needed. With a book, electronic or paper, tables of contents, indices, skimming, and reading rates permit users to quickly find the chunks they need. Since the time to scan a video cannot be dramatically shorter than the real time of the video, a digital video library must be efficient at giving users the material they need. To make the retrieval of bits faster, and to enable faster viewing or information assimilation, the digital video library will need to support:

- partitioning video into small-sized clips
- alternate representations of the video

Video Paragraphing

Just as text books can be decomposed into paragraphs embodying topics of discourse, the video library can be partitioned into video paragraphs. The difficulties arise in how this partitioning is to be carried out. Does the author of the video information supply paragraph tags marking how a larger video should be subsetted into smaller clips? This is routinely accomplished in text through chapters, sections, subheadings, and similar conventions. Analogous structure is contained in video through scenes, shots, camera motions, and transitions. Manually describing this structure in a machine readable form would place a tremendous burden on the video author, and in any case would not solve the partitioning problem for pre-existing video material created without paragraph markings.

Perhaps the paragraph boundaries can be inferred from whatever parsing and indexing is done on the video segment. Some video, such as news broadcasts, have a well-defined structure which could be parsed into short video paragraphs for different news stories, sports, and weather. Techniques monitoring the video signal (discussed later in the chapter) can break the video into sequences sharing the same spatial location, and these scenes could be used as paragraphs.

Davis cautions, however, that physically segmenting a video library into clips imposes a fixed segmentation on the video data [Davis94]. The library is decomposed into a fixed number of clips, i.e., a fixed number of small video files, which are separated from their original context and may not meet the future needs of the library user. A more flexible alternative is to logically segment the library by adding sets of video paragraph markers and indices, but keeping the video data intact in its original context so that:

- annotations can be added later to enrich the description of the video content as more knowledge is acquired about the original material
- the original material can be retrieved easily and without redundancy in whole by the user if desired

- the clip to return to the user can be based dynamically on user and query characteristics, with richer annotations allowing more numerous possible segmentations of the video data.

A basic tenet of MIT's Media Streams is that what we need are *"representations which make clips, not representations of clips"* [Davis94, p. 121]. In order for a digital video library to be logically segmented as such, the system must be capable of delivering a subset of a movie (rather than having that subset stored as its own movie) quickly and efficiently to the user. Video compression schemes will have to be chosen carefully for the library to retain the necessary random access within a video to allow it to be logically segmented.

Alternate Representations for Video Clips

In addition to trying to size the video clips appropriately, the digital video library can provide the users alternate representations for the video, or layers of information. Users could then cheaply (in terms of data transfer time, possible economic cost, and user viewing time) review a given layer of information before deciding upon whether to incur the cost of richer layers of information or the complete video clip. For example, a given half hour video may have a text title, a text abstract, a full text transcript, a representative single image, and a representative one minute "skim" video, all in addition to the full video itself. The user could quickly review the title and perhaps the representative image, decide on whether to view the abstract and perhaps full transcript, and finally make the decision on whether to retrieve and view the full video.

These layered approaches to describing video are implemented in a number of systems [Stevens94, Zhang95, Rao95], and will be returned to in the discussions on specific techniques. The problems are similar to the indexing problem: how should the alternate representations or descriptors be generated? How can they be as complete and accurate as possible, and can tools alleviate the labor and tediousness involved in their creation?

1.2.4 Retrieving and Browsing Video

The utility of the digital video library can be judged on the ability of the users to get the information they need from the library easily and efficiently. The two standard measures of performance in information retrieval are *recall* and *precision*. Recall is the proportion of relevant documents that are actually retrieved, and precision is the proportion of retrieved documents that are actually relevant. These two measures may be traded off one for the other, i.e., returning one document that is a known match to a query guarantees 100% precision, but fails at recall if a number of other documents were relevant as well. Returning all of the library's contents for a query guarantees 100% recall, but fails miserably at precision and filtering the information. The goal of information retrieval is to maximize both recall and precision.

In many information systems, precision is maximized by narrowing the domain considerably, extensively indexing the data according to the parameters of the domain, and allow-

ing queries only via those parameters. This approach is taken by many CD-ROM data sets. For example, a CD-ROM on animals might fully index the data by genus, species, habitat, diet, gestation periods, growth rate, estimated population, and other biological and environmental factors. The data becomes very useful for its given purpose, e.g., an encyclopedia/browser on animals, but this approach has a few limitations:

- Data could really only be added if it falls within the boundaries of the domain established by the predefined indices. For example, if information about countries were to be added to this animals CD-ROM, new indices would have to be added as well.
- Access to the data is limited by the predefined indices. Continuing with the animals CD-ROM, the user may not be able to find all birds that are blue if color is not one of the attributes which were indexed. If the user is looking for examples of a hunt, a video clip showing a coyote chase down a roadrunner may not be able to be located if the indices only describe the clip as coyote and roadrunner without mention of the hunt.

Researchers of multimedia information systems have raised concerns over the difficulties in adequately indexing a video database so that it can be used as a general purpose library, rather than say a more narrow domain such as a network news archive [Davis94, Zhang95]. For general purpose use, there may not be enough domain knowledge to apply to the user's query and to the library index in order to return only a very small subset of the library to the user matching just the given query. For example, in a soccer-only library, a query about goal can be interpreted to mean a score, and just those appropriate materials can be retrieved accordingly. In a more open context, goal could mean a score in hockey or a general aim or objective. A larger set of results will need to be returned to the user, given less domain knowledge from which to leverage.

In attempting to create a general purpose digital video library, precision may have to be sacrificed in order to ensure that the material the user is interested in will be recalled in the result set. The result set may then become quite large, so the user may need to filter the set and decide what is important. Three principle issues with respect to searching for information are:

1. how to let the user quickly skim the video objects to locate sections of interest
2. how to let the user adjust the size of the video objects returned
3. how to aid users in the identification of desired video when multiple objects are returned

Collapsing Playback Rate

Browsing can help users quickly and intelligently filter a number of results to the precise information they are seeking. However, browsing video is not as easy as browsing text. Scanning by jumping a set number of frames may skip the target information completely. On the other hand, accelerating the playback of motion video to, for instance, twenty times normal rate presents the information at an incomprehensible speed.

Playing audio fast during the scan will not help. Beyond 1.5 or 2 times normal speed, audio becomes incomprehensible since the faster playback rates shift frequencies to inau-

dible ranges [Degen92]. Digital signal processing techniques are available to reduce these frequency shifts, but at high playback rates, these techniques present unintelligible sound bytes much like the analog videodisc scan.

To convey almost any meaning at all, video and audio must be played at a constant rate, the rate at which they were recorded. While, a user might accept video and audio played back at 1.5 times normal speed for a brief time, it is unlikely that users would accept long periods of such playback rates. In fact, studies show that there is surprisingly significant sensitivity to altering playback fidelity [Christel91]. Even if users did accept accelerated playback, the information transfer rate would still be principally controlled by the system.

The difference between video or audio and text or images is that video and audio have constant rate outputs that cannot be changed without significantly and negatively impacting the user's ability to extract information. Video and audio are a constant rate, continuous time media. Their temporal nature is constant due to the requirements of the viewer/listener. Text is a variable rate continuous medium. Its temporal nature is manifest in users, who read and process the text at different rates.

While video and audio data types are constant rate, continuous-time, the information contained in them is not. In fact, the granularity of the information content is such that a one-half hour video may easily have one hundred semantically separate chunks. The chunks may be linguistic or visual in nature. They may range from sentences to paragraphs and from images to scenes. If the important information from a video can be retrieved and the less important information collapsed, the resulting “skim” video could be browsed quickly by the user and still give him or her a great deal of understanding about the contents of the complete video clip. This introduces the issue of deciding what is important within a video clip and worthy of preservation in a “skim” video.

Returning Small Pieces

Another approach to letting the user browse and filter through search results more efficiently is to return smaller video clips in the result set. There are about 150 spoken words per minute of “talking head” video. One hour of video contains 9,000 words, which is about 15 pages of text. Even if a high playback rate of 3 to 4 times normal speed was comprehensible, continuous play of audio and video is a totally unacceptable browsing mechanism. For example, assume that a desired piece of information is halfway through a one hour video file. Fast forwarding at 4 times normal speed would take 7.5 minutes to find it. Returning the optimally sized chunk of digital video is one aspect of the solution to this problem.

If the user issues a query and receives ten half-hour video clips, it could take them hours to review the results to determine their relevance, especially given the difficulties in collapsing video playback as mentioned above. If the results set were instead ten two minute clips, then the review time by the user is reduced considerably. In order to return small, relevant clips the video contents need to be indexed well and sized appropriately, tasks whose problems were discussed at the start of this section.

Information Visualization

Users often wish to peruse video much as they flip through the pages of a book. Unfortunately, today's mechanisms for this are inadequate. Tools have been created to facilitate sound browsing which present graphical representations of the audio waveform to the user to aid identification of locations of interest. However, this has been shown to be useful only for audio segments under three minutes [Degen92]. When searching for a specific piece of information in hours of audio or video, other mechanisms will be required.

The results from a query may be too large to be effectively handled with conventional presentations such as a scrollable list. To enable better filtering and browsing, the features deemed important by the user should be emphasized and made visible. What are these features, though, and how can they be made visible, especially if the digital video library is general purpose rather than specialized to a particular domain? These questions return us back to the problem of identifying the content within the video data and representing it in forms that facilitate browsing, visualization, and retrieval. Researchers at Xerox PARC's Intelligent Information Access and Information Visualization projects note that the information in digital libraries should not just be retrieved but should allow for rich interaction, so that users can tailor the information into effective and memorable renderings appropriate to their needs [Rao95]. If such rich interaction can be achieved, it can be used to browse not only query result sets but the contents of the full library itself, allowing for another access mechanism to the information.

1.2.5 Reusing Video Resources

Just viewing video from digital video libraries, while useful, is not enough. Once users identify video objects of interest, they will need to be able to manipulate, organize, and reuse the video. Demonstrations abound where students create video documents by the association of video clips with text. While excellent steps in the right direction, the reuse of video is more than simply editing a selection and linking it to text.

Today, very good stand-alone tools exist to edit digital video in the commercial market. However, there are currently no tools to aid in the creative design and use of video as there are for document production. One reason is the intrinsic, constant rate temporal aspect of video. Another is the complexities in understanding the nature and interplay of scene, framing, camera angle, and transition. To be able to effectively write, we spend years learning formal grammar. The language of film is both rich and complex, and deep cinematic knowledge, the grammar of video, cannot be required of users. Tools providing expert assistance in cinematic knowledge need to be developed in order for the digital video library to reach its reuse potential.

For example, the contraposition of a high quality, visually rich presentation edited together with a selection from a college lecture on the same material may be inappropriate. However, developing a composition where the lecture material is available for those interested, but not automatically presented, may create a richer learning environment.

As another example, permitting a student to interview an important historical or contemporary figure would provide a more interesting, personal, and exploratory experience than watching a linear interview. Creating such an synthetic interviewee is possible with existing video resources. Broadcast productions typically shoot 50 to 100 times as much material as they actually broadcast. WQED interviewed Arthur C. Clarke for its recent series *Space Age*. Approximately two minutes of the interview were broadcast, but over four hours were taped. While few would want to sit through four hours of an interview, many would like to ask their own questions. It would be especially interesting and motivating if the character responded in a fashion that caused the viewer to feel as if the answer was "live", i.e., specifically and dynamically created in response to the question.

Similar synthetic interviews have been hand-crafted [Stevens89, Christel92]. For typical users to create such an interview, new tools will be needed. The nature and form of such tools for creating synthetic interviews and facilitating other manners of reuse will likely evolve as digital video libraries come on-line.

1.3 TECHNIQUES ADDRESSING DIGITAL VIDEO LIBRARY ISSUES

The previous section introduced a few techniques during the presentation of issues in order to provide some context and clarity to the discussion. This section will examine these and other techniques in more detail, organized as follows:

- using supplemental information which already may exist in other forms to help describe the video contents more completely
- organizing descriptive information for more efficient browsing, retrieval, and reuse
- taking advantage of the information in the audio which accompanies most video to more fully describe the video contents
- applying successes in text-based natural language processing to the domain of digital video libraries
- deepening descriptions for video contents and improving library access by incorporating image processing techniques
- improving ways to browse and visualize video
- supporting reuse of library materials, especially for education and training purposes

1.3.1 Gathering Text Descriptions

The creators of a digital video library will begin with more than just a set of videotapes. The videos may have close-captioned text associated with them, and they will more than likely have titles and production credits. There may also be more detailed production notes with some video source material, outlining the composition of shots into scenes and scenes into the full video. Marketing material, teachers' guides, and critics' reviews may be available for some videos. Close-caption recorders and OCR technology can be used to

convert this information into an electronic text representation [Srihari94], suitable for processing and augmentation by the other techniques described in this chapter. Even if no other automation techniques are used, a human indexer would produce more accurate and complete text indices, or tags, for the video if given this supplemental information rather than just a title or nothing at all.

If the syntactic structure of a particular class of videos is well known and stable, that structure could easily be parsed into an electronic text representation for further processing. For example, *CNN Headline News* presents the top stories at the top of the hour, sports at twenty minutes after the hour and fifty minutes after the hour, and so forth. If the time for the top of the hour is known, then the corresponding video there could be tagged with the appropriate date and "top stories for this date"; the sports clips could be tagged with "sports stories and scores." Movie previews, sports highlight shows, and some talk shows all possess a high degree of syntactic structure which could be parsed into text to supplement that video's description within the library.

1.3.2 Structuring Descriptions

Given that a large body of text (and perhaps other information like image characteristics, to be discussed shortly) can be accumulated to describe the video contents, it needs to be structured in at least three ways:

- the text needs to be associated with the video it describes; a title will describe a large chunk of video, while text from a close-caption will describe only a few seconds' worth. By associating the text annotations closely with the video, they can be used to retrieve more precise, shorter duration video clips, as well as being used to build clips matching the user's needs more closely
- it needs to be kept in separate fields, i.e., its semantics of origin must be preserved, so that a user interested in filtering out production notes information and looking only at titles can do so, or a visualization technique that lets the user browse according to information in the close-caption track, production notes, or other attribute can do so. Preserving as much semantic information as possible also allows a query asking for Kevin Costner as director to distinguish videos where he is an actor, director, narrator, or host.
- it also needs to be layered, to support browsing and the user's needs. If users wish to quickly determine whether a result is likely to have promise, perhaps they only need to browse through the titles, or a text abstract. Perhaps a movie preview will suffice in letting the user decide whether he or she wants to retrieve the complete movie from the library. The Scatter/Gather paradigm [Rao95] uses titles and terms of importance in a cluster of documents as a high level interface the user can browse with before focusing in on particular documents.

1.3.3 Using the Audio Information

Much of the information conveyed in the audio for a given movie is captured in its close-caption text. Even though much of broadcast television is close-captioned, many other video and film assets are not. More importantly, typical video production generates 50 to

100 more data that is not broadcast and therefore not captioned. Clearly, effective use and reuse of all these video information assets within digital video libraries will require automatic generation of transcripts in order to make the information in the audio more accessible. Speech recognition technology can be applied to automatic transcript generation, but a number of problems need to be addressed in this effort. These problems are discussed within the context of a specific speech recognizer in Section 1.4.3.

The audio conveys other information besides just dialog. Researchers have made progress in identifying pauses and silence [Arons93], as well as specialized audio parsers for music, laughter, and other highly distinct acoustic phenomena [Hawley93]. This information can supplement the other structured descriptors, and some such as pauses may be especially useful to identify natural start and stop times for video paragraphing as well as allowing for a degree of compression in presenting a "skim" video.

1.3.4 Natural Language Processing

Natural language processing can be used in several ways to improve the utility of a digital video library:

- improving the focus of a user's query, as well as allowing that query to be a straighforward description rather than requiring a complex query language [TREC93]
- organizing the other descriptive information into semantic networks and hierarchical structures, so that the whole library can be browsed more conveniently through models like Scatter/Gather [Rao95]
- correcting other representations, such as an automatically produced transcript

However, natural language is inherently ambiguous, so that both the query and the library natural language descriptors can be misinterpreted. Probabilistic matching can be used to return a rank-ordered result list rather than one "exact" match as a way to deal with these problems. The user can set the threshold limits on how large a set to return, thereby having direct control over precision and recall. By allowing all results to be returned, no matter how low they scored given a particular query, recall is increased at the expense of precision. A user setting a high threshold so that only the top few of the ranked set of results are returned will increase precision but perhaps sacrifice recall. The Center for Intelligent Information Retrieval utilizes probabilistic matching in this manner via the INQUERY retrieval engine [Croft95], and Carnegie Mellon University's digital video library project uses this same model with the Pursuit search engine, as will be discussed later in this chapter. The TileBars interface allows the user to see why the results were ranked as they were for a query consisting of term sets by indicating the relative length of result documents, the frequency of term sets in the document, and the distribution of the term sets with respect to the document and one another [Hearst95].

By analyzing the transcript, production notes, treatment, and whatever other text information exists to describe a video, natural language processing can be used to determine the subject area and theme of the narrative. This understanding can be used to generate head-

lines or summaries of each video segment for icon labelling, tables of contents, browsing, and indexing.

1.3.5 Image Processing

Research in image databases which allow for visual query is becoming popular. However, video information is temporal, spatial, often unstructured, and massive. As a result, a complete solution of automatic extraction of semantic information or a "general" vision recognition system is not feasible at this point. Current efforts in image databases, in fact, are mostly based on indirect image statistics methods. With few exceptions, they fail to exploit language information associated with images or to deal with three dimensional events.

Image statistics methods compute primitive image features and their time functions, such as color histograms [Swain91, Gong92], coding coefficients, shape [Kato92, Satoh92] and texture measures [Kato92], and use them for indexing, matching and segmenting images. Transitions between scenes such as fades, cuts, and dissolves [Zhang93, Hampapur95] can be identified through analysis of the video signal, with new algorithms running more efficiently as they work on the compressed video stream [Zhang95b]. Image analysis can also be used to determine camera motion (pans and tilts) and lens zooms [Akutsu94]. These are all practical and powerful approaches for some applications, but obviously deal with only images, but not their content.

Image processing can be used to add further description concerning a particular video. Identifying camera pans and zooms, edit effects like fades, cuts, and dissolves, can be useful for segmenting, or "paragraphing", the video into a group of frames when video library is formed. Each group can be reasonably abstracted by a representative frame. Part of this task can be done by content-free methods that detect big image changes, for example, key frame detection by changes in the DCT (discrete cosine transform) coefficient used in common video compression algorithms like MPEG.

However, a more efficient digital video library needs content-based video paragraphing methods, and image processing by itself cannot determine all of the information. Some information system developers parse video in a particular domain, such as news footage, to supplement the image analysis with more structure and semantics, while others use human indexers to document video content, including space, time, weather, characters, objects, character actions, object actions, relative position, screen position, and cinematography. The digital video library user is interested in subject or content retrieval, not just "image" retrieval. The subject consists of both image content and textual content (from audio and other sources); the combination specifies the content. Any textual information attached is useful to quickly filter video segments locating potential items of interest. But subsequent query is usually visual, referring to image content. For example, "Find video with similar scenery," "Find the same scene with different camera motion," "Find video with the same person," and so on. Again, we notice that part of the capability can be realized by content-free methods, such as histogram comparison, but real solutions lie in content-based image search which presents a long-term challenge to the field of computer vision research.

1.3.6 Browsing and Visualization Techniques

Browsing and visualizing the contents of the digital video library can be achieved through a variety of strategies, only a sample of which are listed here.

Viewing it All

The traditional information visualization approach is to attempt to view the whole library, or some piece of it such as a large number of results returned for a given query, at once by presenting in parallel representations for the items to be viewed. This bird's eye view approach implies that there is some way to represent context, to show the relationship between different items in the library. The user can emphasize some portion of the information while retaining most or all of the context. For example, the Envision system displays search results as a matrix of icons [Heath95]. The user selects which information to emphasize, e.g., creation year, producer, and animals in scene, and uses those attributes for the icon color, shape, size, and label, as well as perhaps for the matrix x and y axes. The Perspective Wall and Document Lens are other "fish-eye" browsing techniques which show more detail for one section of information while keeping its context still visible but with less detail [Rao95].

These techniques succeed by leveraging from existing structure in the information space relating any item to another. Creating rich, detailed structures for a general purpose digital video library will enable context to be shown more accurately and completely. If all the video entries share with one another is that they have titles, then the techniques above will fail. If, however, there are temporal, hierarchical, tabular, and other structures as well as many shared fields of descriptors, then visualizing the library or a given subset of it with emphasis on certain attributes becomes possible.

The human visual system is adept at quickly, holistically viewing an image, or a page of text, and finding a desired piece of information while ignoring unwanted information (noise). This has been viewed as a general principle of selective omission of information [Resnikoff89] and is one of the factors that makes flipping through the pages of a book a relatively efficient process. Even when the location of a piece of information is known a priori from an index, the final search of a page is aided by this ability.

Building on these principles, a digital video library could take advantage of the special abilities of the human vision system and present many video snippets in parallel. When a search produces multiple hits, as will usually be the case, the library could present numerous sequences simultaneously in separate windows. The simplest sequence, a single image extracted from the video, could use the first image with valid (i.e. non-blank) data as determined by image processing techniques. A slightly more complex representation would be motion icons, micons [Brondmo90]. As implemented by Brondmo, micons are short motion sequences extracted from the first few seconds or minutes of the video they are to represent.

Both still iconic and miconic representations of video information can easily mislead a user. For example, a search for video sequences related to transportation of goods during

the early 1800s may return 20 relevant items. If the first 20 seconds of several sequences are “talking head” introductions, icons and micons will provide no significant visual clue about the content of the video: the information after the introduction may or may not be interesting to the user. However, intelligent moving icons, imicons, may overcome some of these limitations. Image segmentation technology can create short sequences that more closely map to the visual information contained in the video stream [Stevens94, Zhang95b]. Since the human visual system is adept at quickly finding a desired piece of information, the simultaneous presentation of intelligently created motion icons will let the user act as a filter to choose high interest material.

Building Your Own Bridges

Rather than following someone else’s links through an information space, the user can select a body of text in the library and use that text, rather than a predetermined link, as the basis of the search. This “click-to-search” interface is used in the Dienst digital library architecture [Lagoze95]. *Snippet Search* can extend the utility of using current results to jump to new results by showing not just matched terms but also the context around matched terms when displaying results [Rao95]. Heuristics guarantee that other relevant words are shown with the matched terms, so a subsequent query using a snippet would include those relevant words in addition to the matched term(s) in that snippet.

For example, a search on coyote may produce a snippet “coyote hunting in the chapparal” in the results, and the user selects this snippet as the basis for a new search. The subsequent results present more information on the diminishing population of coyotes in California due to the reduction in chapparal wild environments as more land gets developed. Users not only determine which paths to follow; they in effect are clearing their own trails through the information by formulating their queries in these ways.

Expanding/Collapsing

Previously in the chapter multiple representations were discussed as a way to facilitate browsing. A user might only want to scan through titles, or may want to expand to see abstracts, expand further and see representative images, or “skim” videos, or movie previews, or expand completely to see the complete video. Each successive representation adds more detail, but also adds more processing time for the user.

The Hierarchical Video Magnifier [Mills92] was one of the first ways to collapse video data and see more of its contents at once in reduced time rather than having to play through the video sequentially. Zhang, Low, and Smoliar extend this work by selecting the representative images to show when collapsing a video via video parsing and key frame extraction based on MPEG frame difference statistics [Zhang95b]. If the representative images are shown all at once on a screen with adequate resolution, the user can holistically examine the contents of the video. Alternately, the user can play back the images temporally in much less time than required for the full video.

Collapsing a video, or presenting a video "skim", reduces the rate of playback at the expense of informational and perceptual quality. The simplest skim might present a subset of video frames distributed uniformly in time, e.g., every hundredth frame. An improved technique would present representations of all the visual scene changes [Arman94]. A further improvement would be to present images for when significant content information is being communicated. But then, as with other techniques, the stumbling block for collapsing a video becomes content-based understanding. What are the representative images to use? Can this be determined through an analysis of solely the video signal? As discussed earlier, content is conveyed in more than just the video signal. Can representative "skim" features be determined through a combination of analyzing the video signal, audio signal, and natural language processing of all the text information, including the transcribed audio? Perhaps so for a given purpose, but "collapse" and "expand" are dependent on the user's context as well, which may vary for a general purpose digital video library.

Knowledge-based skims were first described by Stevens [Stevens92]. As an example, one user may examine a video on coyotes chasing roadrunners and may want to collapse it for quicker browsing. This user is interested in the predator/prey story, and so he or she will want to see the initiation of the hunt, the tracking, the chase, and the resolution all represented in the "skim." A different user may be interested in cinematic principles for nature films, and would want to see set-up shots of the coyote and roadrunner, background scene changes, changes in background music and narration, and other details of interest to a film researcher. A video may not have only one but perhaps many collapsed representations reflecting different perspectives on its contents. Equally challenging, a digital video library should allow the user to expand out from a video and say "I want more background on this subject" with the expansion being biased by the user's perspective.

1.3.7 Building for Reuse

While much of this chapter has focused on the "video" aspect of digital video libraries, the "digital" format implies that the information is more malleable and can change and evolve. One of the simplest ways to enable reuse of video assets while expanding the library's utility would be an annotation feature, by which critics, experts in the field, teachers, or perhaps anyone can add notes and markings to the video materials. Other users could access this information to get the materials reserved by their teacher for an assignment, to follow the recommendations of a given critic, or to filter the information based on person X marking it as a good source, because person X's opinions are respected by the user.

Now if the videos all have rich descriptors associated with them, they can be used as building blocks to construct other videos, presentations, and simulations [Christel92]. Such rich descriptions may include a semantically structured generalization space of categorical descriptors and an episodically structured relational space of temporal analogical descriptors, as generated by human indexers with Media Streams [Davis94]. The following techniques must mature in the coming years so that the materials in digital video libraries can be reused not only by adding annotations but by composing clips into new works:

- standardize the language used to describe video contents

- continually improve the tools for indexing using this language, automating the indexing process wherever possible to reduce the tremendous labor investment, increase accuracy, and reduce any personal bias of human indexers
- supplement the user's search and retrieval interface to the library with a library exploration interface containing tools for composition of results which enforce good cinematic principles

1.4 THE INFORMEDIA DIGITAL VIDEO LIBRARY PROJECT

The Informedia Digital Video Library Project (IDVL) at Carnegie Mellon University is and on-going research project begun in 1994, but leveraging two decades of related CMU research. Central to the project is the establishment of a large, on-line digital video library. To accomplish this, the project is developing intelligent, automatic mechanisms to populate the library and allow for full-content and knowledge-based search and retrieval via networked desktop computers. This section will discuss the techniques being employed by the project as a concrete example illustrating how the problems discussed earlier can be addressed in a digital video library system.

1.4.1 A User's Perspective

Imagine a high school student sitting at a multimedia workstation running the IDVL. Her class project is to create a multimedia composition on how world culture has been changed by communications satellites. Groping for a beginning she begins speaking to the monitor, "I've got to put something together on communication satellites. What is there?"

> *Transparent to the user the system has just performed highly accurate, speaker independent, continuous speech recognition on her query. It then used sophisticated natural language processing to understand the query and translate it into retrieval commands to locate relevant portions of digital video. The video is searched based on transcripts from audio tracks that were automatically generated through the same speech recognition technology. The appropriate selection is further refined through scene sizing developed by image understanding technology.*

Almost as soon as she has finished her question, the screen shows several icons, as in Figure 1-1. These icons consist of either still images or short motion clips representing the video segment matching the student's query, complemented with text forming an extended title (as shown for the top ranked result) and possible abstracts of the information contained in the video.

> *Making this possible, image processing helped select representative still images for icons and sequences from scenes for intelligent moving icons. Speech recognition created transcripts which are used by natural language technologies to summarize and abstract the selections. The processed transcripts can also aid the image processing in determining which images to keep as representatives of the whole video's content. Rather than returning only one search query result, the ambiguous nature of the query*

and data set are taken into account and probabilistic searching is used, returning a set of ranked results to the student. Even if the top-ranked result does not match the student's needs, it is likely that one from the top set of results will be appropriate.

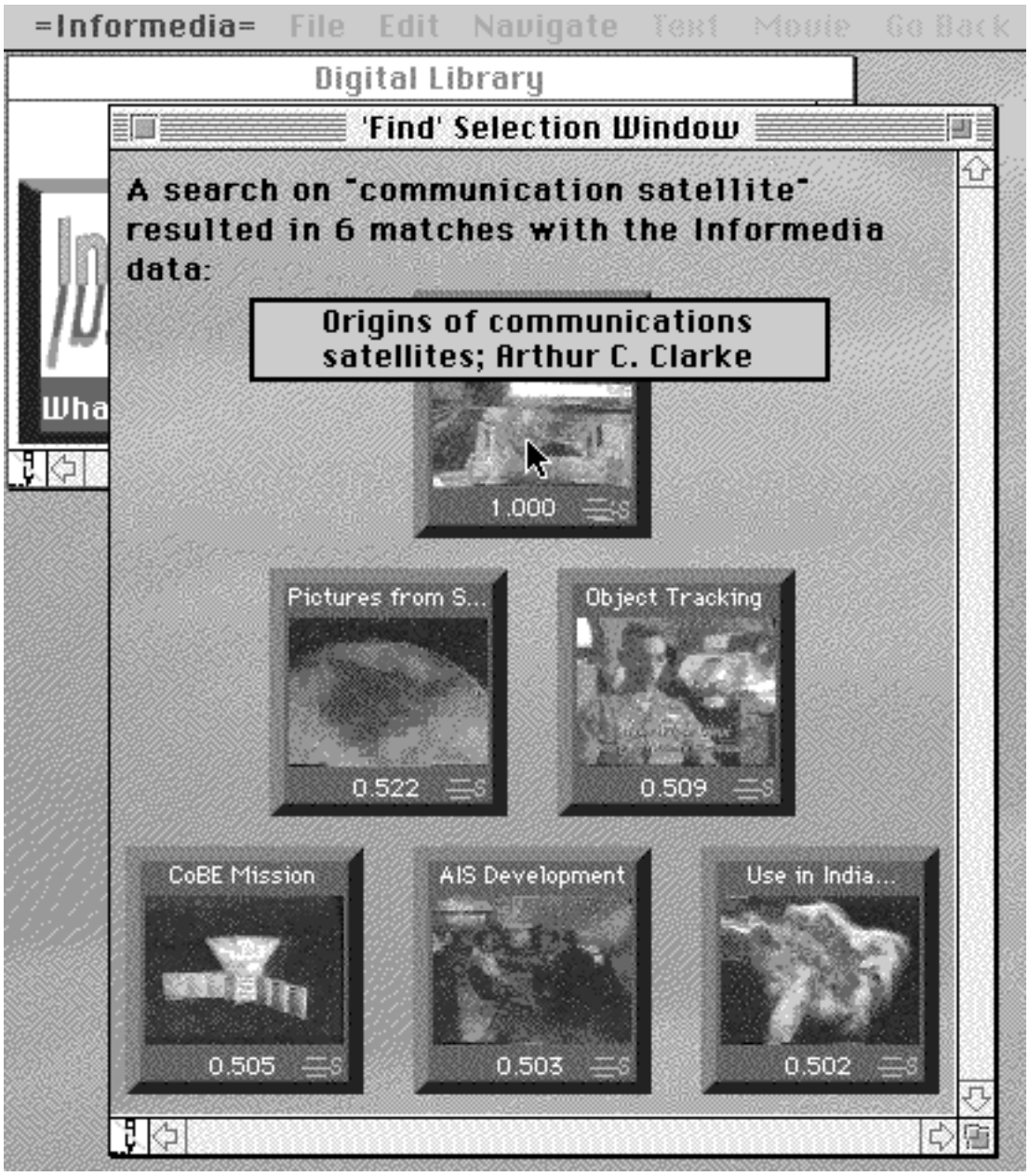

FIGURE 1-1. Presentation of ranked search results following query in IDVL

Through either a mouse or a spoken command, the student requests the top icon. The screen fills with a video of Arthur C. Clarke describing how he did not try to patent communications satellites, even though he was the first to describe them. Next the student requests the sixth choice, and sees villages in India that are using satellite dishes to view educational programming.

Asking to go back, Arthur C. Clarke reappears. Now, speaking directly to Clarke, she wonders if he has any thoughts on how his invention has shaped the world. Clarke starts talking about his childhood in England and how different the world was then (see Figure 1-2). Using a skimming control she finds a particularly relevant section to be included in her multimedia composition.

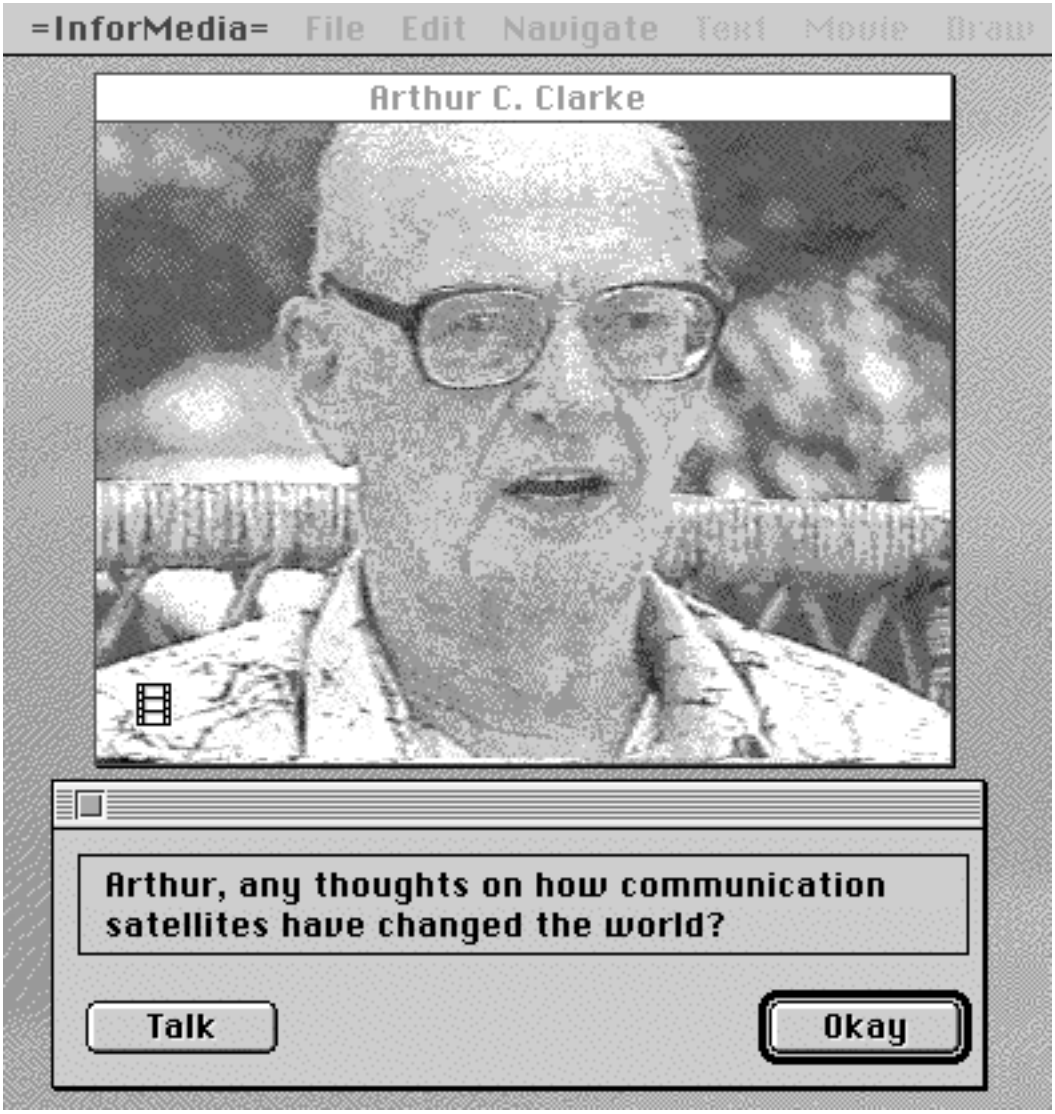

FIGURE 1-2. Playing back video from the IDVL

Beyond the requisite search and retrieval, to give the student such functionality requires image understanding to intelligently create scenes, database techniques to trim the search space dynamically to only Clarke's opinions, and the integration of speech, image, and natural language processing for the skimming control.

The next day she gives her teacher access to her project. More than a simple presentation of a few video clips, the student has created a video laboratory that can be explored and whose structure is itself indicative of the student's understanding.

Helping this student be successful are tools for building multimedia objects that include assistance in the language of cinema, appropriate use of video, and structuring composition. Behind the scenes the system has created a profile of how the video was used, distributing that information to the library's accounts. Assets for which the school has unlimited rights are tracked to understand curricular needs. Accounts for assets that the school has restricted, pay-per-use rights are debited.

A digital video library such as the one needed to make this scenario a reality entails the integration of diverse technologies with rich content. The distinguishing feature of the IDVL's technical approach is the integrated application of speech, language and image understanding technologies for efficient creation and exploration of the library. Using a high-quality speech recognizer, the sound track of each videotape is converted to a textual transcript. A language understanding system analyzes and organizes the transcript and stores it in a full-text information retrieval system. Image understanding techniques are used for segmenting video sequences by automatically locating boundaries of shots, scenes, and conversations. Exploration of the library is based on these same techniques. The user interface is instrumented to investigate user protocols and human factor issues peculiar to manipulating video segments. A network billing server is incorporated to ensure privacy and security, and to study the economics of charging strategies. The remainder of this section will describe these technologies as they are being applied to and developed for the Informedia Digital Video Library.

1.4.2 Library Contents

The Informedia Digital Video Library is being populated with 1000 hours of both raw and highly produced, edited video. The video is from three primary sources:

- a vast library of science programs, documentaries, and original source materials from WQED (Pittsburgh's PBS station)
- the BBC's educational video course material developed for the British Open University
- the Fairfax County (VA) public schools' Electronic Field Trip series

The library is being deployed initially at Carnegie Mellon University and Winchester-Thurston, an independent Pittsburgh K-12 school. The first deployment is using MPEG1 compression, requiring about 10 Megabytes per source video minute to achieve VHS quality playback (352 x 240 x 30Hz). The primary media-server file system requires one terabyte (10^12 bytes) of storage and when full populated will comprise over 1000 hours of video.

This collection incorporates not only the broadcast programs themselves, but also the unedited source materials from which they were derived. Such background materials enrich the library significantly, as reference resources and for uses other than those originally intended. They also enlarge it greatly: typical WQED sources run 50 to 100 times longer than the corresponding broadcast footage. Some of this material duplicates what was broadcast or is flawed (e.g., noisy background, misspoken dialogue). But an order of

magnitude more useful, unique material remains that simply could not fit into the time allocated for the original broadcast.

This particular combination of video resources enables Informedia users to retrieve the same subject matter material presented at varying levels of complexity, ranging from the popular example-based presentation often used in PBS documentaries through elementary and high school presentations from Fairfax Co., to the more advanced college-level treatment by the Open University. The self-learner at any level can iterate on the search in order to build understanding and comprehension through multiple examples and decreasing (or increasing) depth, complexity, and formalism.

Much like printing and binding books is an off-line, background process, so too is IDVL library creation model. Although it may it may take several hours to compress, transcribe, index, and segment one hour of video, exploration and retrieval is real-time. This model of off-line creation of materials for on-line, real-time exploration is used in many other digital library and multimedia database efforts. As pointed out by Levy and Marshall [Levy95], this model then assumes that the video data is fixed rather than fluid and has a long useful lifetime, as opposed to, for example, a digital video news server that may utilize simpler techniques for indexing business news digital videos in real time. The Informedia Project assumes that real-time constraints on library creation can be relaxed in order to realize increased automation and deeper parsing and indexing for the activities of identifying what is in the library and breaking it into pieces. This model is shown in Figure 1-3.

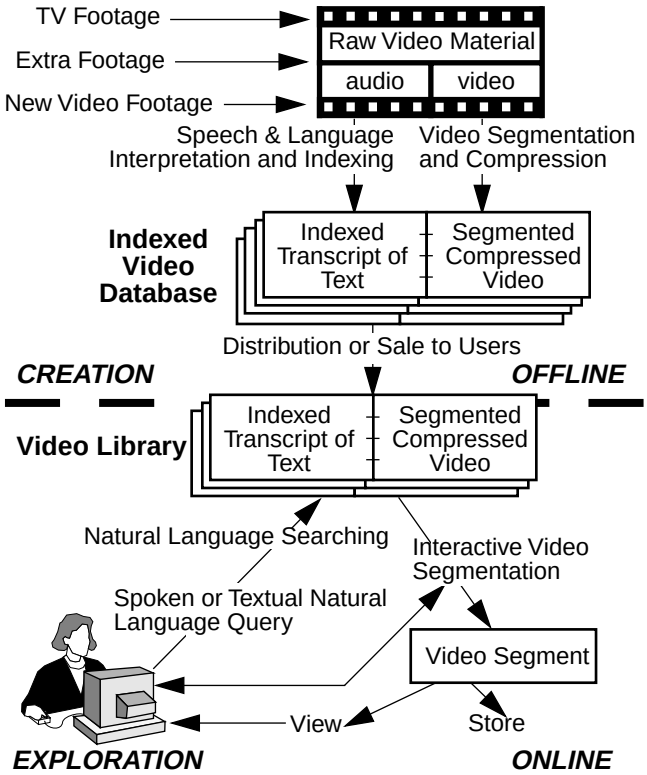

FIGURE 1-3. Overview of the Informedia Digital Video Library System

1.4.3 Automated Transcription via Speech Recognition

The Informedia system is using the Sphinx-II speech recognition system to transcribe narratives and dialogues automatically. Sphinx-II is a large-vocabulary, speaker-independent, continuous speech recognizer developed at Carnegie Mellon. In a 1992 ARPA speech recognition evaluation, Sphinx-II had the highest accuracy of all systems tested [Hwang93].

The current best system, Sphinx-II, uses a 20,000 word vocabulary to recognize connected spoken utterances from many different speakers. The task domain is recognition of dictation of passages from the Wall Street Journal. On a 150 MIPS DEC Alpha workstation the system operates in near real-time and on average makes one error out of eight words [Hwang94]. Although unlimited vocabulary, speaker-independent, connected speech recognition is an unsolved problem, recent advances in acoustic and language modeling have allowed Sphinx-II to achieve a 5% error rate on standardized tests for a 5000-word, general dictation task. Such performance promises a strong potential impact on automatic transcript generation, but there are a number of sources of error and variability arising naturally in a video transcription task. These include the following, listed with possible techniques to solve the problem:

- *Multiple Signal to Noise Ratio Problem:* Broadcast video productions, whether they are documentary style interviews or theatrical productions, have to recognize speech from multiple speakers standing in different locations. This results in speech signal quality with different signal to noise ratio properties. Further confounding the problem are the effects of different orientations of the speakers and reverberation characteristics of the room [Liu93]. Signal adaptation techniques have been developed which appear to automatically correct for such variability. However, such systems have not been tested with environments where nearly every other sentence has a different signal to noise ratio. The project is developing appropriate preprocessing and detection of the signal levels to be able to modify the current CDCN technology to solve this problem.
- *Multiple Unknown Microphone Problem:* Most current systems optimize recognition performance using close-talking, head-mounted microphones. With tabletop microphones, lapel microphones, and directional boom microphones traditionally used in broadcast video productions, the variability arising from differences in microphone characteristics and differences in signal to noise ratios will significantly degrade performance. Recent results by Stern and Sullivan indicate that dynamic microphone adaptation can significantly reduce the error without having to retrain the speech recognizer for the new microphone [Sullivan93].
- *Fluent Speech Problem:* In a typical video interview, people speak fluently. This implies many of the words are reduced or mispronounced. Lexical descriptions of pronunciations used in conventional systems for dictation, where careful articulation is the norm, do not work very well for spontaneous, fluent speech. At present the only known technique is for manual adaptation of the Lexicon using knowledgeable linguists. IDVL is using the rich data source provided by the library to formulate automatic pronunciation learning techniques to handle fluent speech phenomena.

- *Unlimited Vocabulary Problem:* Unlike the Wall Street Journal dictation task where the domain limits the size and nature of the vocabulary likely to be used in sentences, video transcriptions generally tend not to have such constraints. However, they do represent specific task domains. The Informedia Project's recent research in long distance language models appears to indicate twenty to thirty percent improvement in accuracy may be realized by dynamically adapting the vocabulary based on words that have recently been observed in prior utterances. In addition, most broadcast video programs have significant descriptive text available. These include early descriptions of the program called treatments, working scripts, abstracts describing the program, and captions. In combination, these resources are providing valuable additions to dictionaries used by the recognizer.

Fortunately for transcription of digital video, processing time can be traded for higher accuracy. And for the creation of the library, the system does not have to operate in real time. This permits the use of larger, continuously expanding dictionaries and more computationally intensive language models and search algorithms.

1.4.4 Improved Understanding through Natural Language Processing

Natural language queries allow straightforward description of the subject matter of the material desired. An initial query may be textual, entered either through the keyboard, mouse, or spoken words entered via microphone and recognized by the system. Subsequent refinements of the query, or new, related queries may relate to visual attributes such as: "find me scenes with similar visual backgrounds." Current retrieval technology works well on textual material from newspapers, electronic archives and other sources of grammatically correct and properly spelled written content. However, the video retrieval task, based upon searching errorful transcripts of spoken language, challenges the state of the art. Even understanding a perfect transcription of the audio would be too complicated for current natural language technology.

Natural language processing in the Informedia Digital Video Library consists of three principle tasks:

- *Query processing*: the user must be able to specify a subject or content area for search without having to resort to specialized syntax or complicated command forms.
- *Retrieval*: once the system has digested a user query, the corresponding text objects must be located, scored, and ranked according to user interest.
- *Display*: the video segments associated with each relevant text object must be located, and appropriate scene boundaries identified for each video object (visual sentence, paragraph or page) used to generate a menu of visual segments for user selection.

The video retrieval task challenges the state of the art in two ways:

- *Non-grammaticality:* Written texts, especially news articles, correspond closely to the strict rules of classroom English, whereas the utterances recorded on videotape contain false starts, meta-utterances, pauses, um's, grunts, deictic references to objects in the

visual plane, and other phenomena that are not handled by standard grammars of English. So even perfect transcripts of the audio would be more complicated than current natural language technology can reliably parse.

- *Noise:* Current speech recognition techniques do not provide perfect transcripts. Transcripts derived from Sphinx-II provide four out of five correctly recognized words. This level of error reduces the effectiveness of typical retrieval algorithms. For example, the audio for a video interview contained the phrase "...self fulfilling prophecies." Because Sphinx-II was run using a smaller dictionary that does not contain the words "prophecy" or "prophecies," Sphinx-II returns the closest phonetic match: "...self fulfilling profit seize." This is an understandable error considering the dictionary was derived from the Wall Street Journal.

IDVL natural language understanding research is focusing on two main lines of attack. First is the elaboration of current pattern sets, rules, grammars and lexicons to cover the additional complexity of spoken language by using large, data-driven grammars. This method uses regular expression approximations to the context-free grammars typically used for natural language. The working hypothesis is that extending this technique to an automatically recognized audio track will provide acceptable levels of recall and precision in video scene retrieval. Second is extending the basic pattern matching and parsing algorithms to be more robust, and to function in spite of lower level recognition errors by using a minimal divergence criterion for choosing between ambiguous interpretations of the spoken utterance.

The existing algorithm is being extended to match in phonetic space as well as textual. For example, when *prophecy* and *profit seize* are converted to phonetic space

prophecy	becomes:	P R AA1 F AH0 S IY0 Z
	and	
profit seize	becomes:	P R AA1 F AH0 T S IY1 Z

which deviate only by one insertion (T) and one change in stress (IY0 to IY1).

Other natural language understanding research in the project include:

- *Summarization:* by analyzing the words in the audio track for each visual paragraph, the Informedia system will attempt to determine the subject area and theme of the narrative.
- *Tagging:* using data extraction technology to identify names of people, places, companies, organizations and other entities mentioned in the sound track.
- *Transcript correction:* the most ambitious goal is to automatically generate transcripts of the audio with speech recognition errors corrected. Using semantic and syntactic constraints from NLP, combined with a phonetic knowledge base such as the Sphinx-II dictionary, some recognition errors should be correctable.

Still, even if there were perfect recall based on perfect transcripts, much information resides only in the video. To permit a more comprehensive library retrieval, image processing technology is integrated in the IDVL.

1.4.5 Further Indexing and Segmentation via Image Processing

Image processing plays a critical role in the Informedia system for organizing, searching, and reusing digital video. Traditional database search by keywords, where images are only referenced, not directly searched for, is not appropriate or useful for the digital video library. Rather, digital video images themselves must be segmented, searched for, manipulated, and presented for similarity matching, parallel presentation, context sizing, and skimming, while preserving image content.

The first capability required for digital video library creation is segmentation, or "paragraphing," of video into a cinematically (and often linguistically) meaningful group. Each group can be reasonably abstracted by a "representative frame," and thus can be treated as a unit for context sizing or for image content search. Part of this task can be done by content-free methods that detect big "image changes," for example, "key frame" detection by changes in the DCT coefficient in the compressed video.

IDVL uses comprehensive image statistics for segmentation and indexing. Raw video materials are first segmented into video paragraphs so that each segment can be connected/integrated for indexing with transcribed text. This initial segmentation can be done in a relatively content-free manner by monitoring coding coefficients. Once a video paragraph is identified, image processing in the Informedia system extracts image features like texture, color, and shape from video as attributes. While these are "indirect statistics" to image content, they have been proven to be quite useful in quickly comparing and categorizing images.

Structural and temporal relationships between video segments are also extracted and indexed. One important kind of visual segmentation is based on the computer interpreting and following smooth camera motions such as zooming, panning, and forward camera motion. Examples include large panoramic scenes being surveyed, scenes in which the camera (and narration) zoom into an object to focus the viewer's attention on it, or scenes in which a camera is mounted on a vehicle in motion.

A more important kind of video segment is defined by motion or action of the objects being viewed rather than the motion of the camera. For example, in an interview, once a relevant segment has been located by speech recognition, the user may desire to see the entire clip containing the interview with this same person. This can be done by looking forward or backward in the video sequence to locate the frame at which this person appeared or disappeared from the scene. Such a single-object tracking is relatively easy and the Informedia image understanding sub-system is actually capable of tracking far more complicated objects. Further, a technique is being developed [Rehg94] to track high degree-of-freedom objects, such as a human hand (27 degrees of freedom), based on "deformable templates" [Kass87] and the Extended Kalman Filtering method. Such a technique provides a tool enabling IDVL to track and classify motions of highly articulated objects.

Segmenting video by the appearance of a particular object or a combination of objects is a powerful tool. While this is difficult for a general 3D object with arbitrary location and ori-

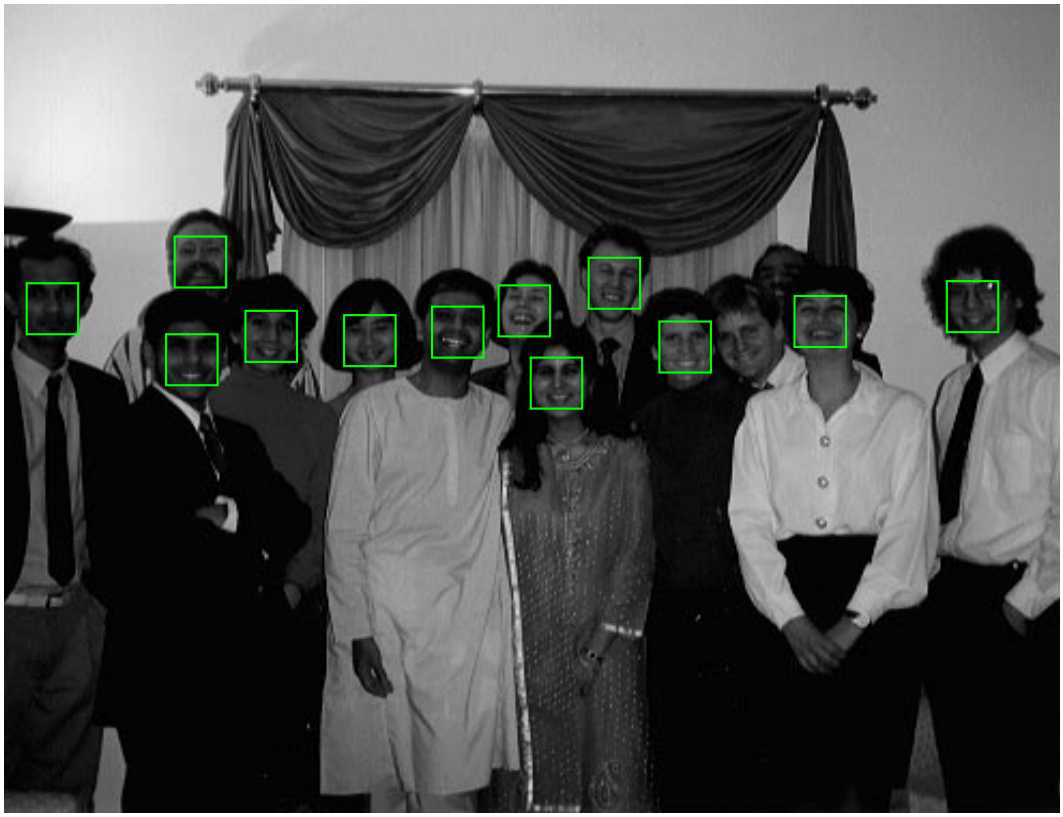

FIGURE 1-4. Output of face detection algorithm (boxed areas indicate areas identified as faces)

entation, the technique of the KL Transform [Lucas81] has proven to work to detect a particular class of object. Among object presence, human content is the most important and common case of object presence detection. IDVL uses a paired neural network technique that is proving highly reliable at the task of detecting human faces (see Figure 1-4).

Lastly, textual information such as names often appear in the video. Frequently, this information is not repeated in the audio. IDVL is applying vision methods to identify text in the video and isolate it from background noise. OCR technology is then applied to this data, transforming it to searchable text (see Figure 1-5).

Speech recognition, natural language processing, and image understanding all provide necessary components to IDVL. Furthermore, it is the integration of these technologies that makes possible a full content searchable digital video library. Yet without an adequate user interface, the library will be of little value.

1.4.6 User Interface for Exploring the Library

Three user interface techniques necessary for a successful digital video library were mentioned earlier: skimming, parallel presentation and context sizing. A description of the early IDVL implementation for these techniques follows.

FIGURE 1-5. Automatic extraction of text labels within video sources

Skimming

No matter how good a system's precision and recall, users wish to be able to quickly skim material to find items of interest. IDVL uses both image understanding and natural language processing to present the most meaningful information to the user. In creating a skim, image understanding techniques are used to select important, high interest segments of video. Scene changes (as marked by color histogram spikes characterizing big differences in adjacent frames), camera motion, object detection (e.g., the entrance and exit of a human face in the scene), and text detection (e.g., a title or name of a person being interviewed overlaid on the video) are used in the heuristics determining which video should be included in the skim. Using parallel criteria for linguistic information, natural language processing selects appropriate audio. For example, the term frequency-inverse document frequency weighting scheme can be used to determine word relevance, with other heuristics employed to further filter which audio to use, such as not repeating the same word within a certain time limit. Selected audio and video are then integrated into a skim of the original video. Early user tests suggest that "information compressions" of between 6 and 20 are both possible and useful (see Figure 1-6).

Parallel Presentation

When a search contains many hits, the system simultaneously presents icons, intelligent moving icons (imicons), and full motion sequences along with their text summarization.

FIGURE 1-6. IDVL techniques underlying skim generation and scene characterization

These objects are arranged in a pyramid that visually presents the most relevant objects at the top, as in Figure 1-1. Imicons are created with similar heuristics as skims. Since users react differently to a screen populated by still images than the same number of moving images, studies are being conducted to identify the optimal number and mix of object types.

Context-sizing

User are permitted to adjust the “size” (duration) of the retrieved video/audio segments for playback. Here, the “size” may be time duration, but it can also be based on scenes or information complexity. For example, it is well known that higher production value video has more shot changes per minute than, for example, a videotaped lecture. And although it is visually richer (finer grain shot sizes), it may be linguistically less dense. Ongoing studies in the Informedia project are helping determine unique balance of linguistic and visual information density appropriate for different types of video information.

1.4.7 Accounting and Economics

Commercialization of digital video information services cannot be realized without very low cost, auditable, private and secure data and billing services. Copyright owners need to be compensated when their intellectual property is distributed to users. Accordingly, the digital library must be supported by a system for authenticating users, verifying willing-

ness and ability to pay, authorizing access, recording charges, invoicing the user, receiving and processing payments, and managing accounts. IDVL is integrating NetBill, a generalized Internet billing service. IDVL's implementation will support the mechanisms necessary to provide adequate privacy protection, a wide range of pricing policies set by intellectual property owners, and restrictive access policies that dynamically limit the accessibility of certain collections to classes of users. For example, there may be content which is age-sensitive, with a school restricting access to only high school students.

1.5 CONCLUSION

The Informedia Project builds on the assumption that a video's contents are conveyed in both the narrative (speech and language) and the image. Only by the collaborative interaction of image, speech and natural language understanding technology can diverse video collections be successfully populated, segmented, indexed, and searched with satisfactory recall and precision. This approach compensates for problems of interpretation and search in error-full and ambiguous data environments.

Universal access to vast, low-cost digital information and entertainment will significantly impact the conduct of business, professional, and personal activity. Most of the major computer manufacturers, news media producers, publishers, cable and communication companies have involved themselves in one or more joint ventures to explore the technology and market potential of digital video information products and services. If the problems associated with having video in digital libraries can be overcome, then the resulting libraries will enable broad accessibility and reuse of a vast array of video assets. These include documentaries, news, and entertainment programs previously and continuously generated for public broadcast; educational programs for students, professionals, and life long learners; and vocational, military, and business training.

The greatest societal impact of digital video libraries will most likely be in K-12 education. The digital video library represents a critical step toward an educational future that we can hardly recognize today. Ready access to multimedia resources will bring to the paradigm of "books, blackboards, and classrooms" the energy, vitality, and intimacy of "entertainment" television and video games. The key, of course, is the access mechanism itself: easy and intuitive to use, powerful and efficient in delivering the desired video clip. The persistent and pervasive impact of such capabilities will revolutionize education, making it as engaging and powerful as the television students have come to love.

At the same time, the greatest commercial impact will be in creating organizational memories and in industrial/commercial training and education. When a company can deliver improved instruction at reduced cost and in less time, huge competitive advantages are realized. Whether the first applications are education, training, or entertainment, ubiquitous access to full content, searchable video libraries will ultimately transform the way we work, learn, and play.

REFERENCES

[ACM94] *ACM Multimedia 94 Conference Proceedings*, October 15-20, 1994, San Francisco, CA. New York: ACM Press.

[Akutsu94] Akutsu, A. and Tonomura, Y. "Video Tomography: An efficient method for Camerawork Extraction and Motion Analysis," *Proc. of ACM Multimedia '94*, Oct. 15-20, 1994, San Francisco, CA, pp. 349-356.

[Arman94] Arman, F., Depommier, R., Hsu, A., and Chiu, M-Y. "Content-Based Browsing of Video Sequences," *Proc. of ACM Multimedia '94*, Oct. 15-20, 1994, San Francisco, CA, pp. 97-103.

[Arons93] Arons, B. "SpeechSkimmer: Interactively Skimming Recorded Speech," *Proc. of ACM Symposium on User Interface Software and Technology (UIST)'93*, Nov. 3-5, 1993, Atlanta, GA, pp. 187-196.

[Becker95] Becker, H. "Library of Congress Digital Video Library Effort," *Communications of the ACM* **38**, April 1995, p. 66.

[Brondmo90] Brondmo, H.P. and Davenport, G. "Creating and Viewing the Elastic Charles - a Hypermedia Journal," *Hypertext, State of the Art*, McAleese, R., Green, C. (eds.). Intellect Ltd., 1990.

[CACM91] *Communications of the ACM* **34**, special issue on digital multimedia, April, 1991.

[Christel91] Christel, M. *A Comparative Evaluation of Digital Video Interactive Interfaces in the Delivery of a Code Inspection Course*, Ph.D. Thesis, Georgia Institute of Technology, Atlanta, GA, 1991.

[Christel92] Christel, M. and Stevens, S. "Rule Base and Digital Video Technologies Applied to Training Simulations," *Software Engineering Institute Technical Review '92*. Pittsburgh, PA: Software Engineering Institute, 1992.

[Croft95] Croft, W. "NSF Center for Intelligent Information Retrieval," *Communications of the ACM* **38**, April 1995, pp. 42-43.

[Davis94] Davis, M. "Knowledge Representation for Video," *Proc. of AAAI '94*, 1994, Seattle, WA, pp. 120-127.

[Degen92] Degen, L., Mander, R., and Salomon, G. "Working with Audio: Integrating Personal Tape Recorders and Desktop Computers," *Proc. CHI '92*, May 1992, Monterey, CA, pp. 413-418.

[Fenn94] Fenn, B. and Maurer, H. "Harmony on an Expanding Net," *interactions* **1** (October 1994), pp. 26-38.

[Fox95] Fox, E., Akscyn, R. Furuta, R., and Leggett, J. (eds.), “Introduction,” special issue on digital libraries, *Communications of the ACM* **38**, April 1995, pp. 22-28.

[Gong92] Gong, Y. and Sakauchi, M. “A Method for Color Moving Image Classification Using the Color and Motion Features of Moving Images,” *ICARCV ‘92*, 1992.

[Hampapur95] Hampapur, A., Jain, R., and Weymouth, T. “Production Model Based Digital Video Segmentation,” *Multimedia Tools and Applications* **1** (March 1995), pp. 9-46.

[Hawley93] Hawley, M. *Structure out of Sound*. Ph.D. Thesis, Massachusetts Institute of Technology, 1993.

[Hearst95] Hearst, M.A. “Tilebars: Visualization of Term Distribution Information in Full Text Information Access.” *Proc. CHI ‘95*, May 1995, Denver, CO, pp. 59-66.

[Heath95] Heath, L., Hix, D., Nowell, L., Wake, W., Averboch, G., Labow, E., Guyer, S., Brueni, D., France, R., Dalal, K., and Fox, E., “Envision: A User-Centered Database of Computer Science Literature,” *Communications of the ACM* **38**, April 1995, pp. 52-53.

[Hwang93] Hwang, M.-Y., Huang, X., and Alleva, F. “Predicting Unseen Triphones with Senones”, *Proc. of IEEE Int’l Conf. on Acoustics, Speech, and Signal Processing (ICASSP-93)*, April 1993, Minneapolis, MN, vol. 2, pp. 311-314.

[Hwang94] Hwang, M.-Y., Thayer, E., and Huang, X. “Semi-Continuous HMMs with Phone Dependent VQ Codebooks for Continuous Speech Recognition,” *Proc. of ICASSP-94*, 1994.

[ICMCS94] *Proc. of the International Conf. on Multimedia Computing and Systems*, May 14-19, 1994, Boston, MA. Los Alamitos, CA: IEEE Computer Society Press.

[Kass87] Kass, M., Terzopoulos, D., Witkin, A. “Symmetry-Seeking Models and 3D Object Reconstruction”, *International Journal of Computer Vision* **1**, 1987, Netherlands, No. 3, pp. 211-221.

[Kato92] Kato, T. “Database Architecture for Content-Based Image Retrieval,” *SPIE: Image Storage and Retrieval Systems*, February 1992, San Jose, CA.

[Lagoze95] Lagoze, C. and Davis, J., “Dienst: An Architecture for Distributed Document Libraries,” *Communications of the ACM* **38**, April 1995, p. 47.

[Levy95] Levy, D. and Marshall, C. “Going Digital: A Look at Assumptions Underlying Digital Libraries,” *Communications of the ACM* **38**, April 1995, pp. 77-84.

[Liu93] Liu, F.H., Stern, R.M., Huang, X., and Acero, A., “Efficient Cepstral Normalization For Robust Speech Recognition,” *Proceedings of the Sixth ARPA Workshop on Human Language Technology,* M. Bates, Ed. Princeton, NJ: Morgan Kaufmann, 1993.

[Lucas81] Lucas, B.D., and Kanade, T. “An Iterative Technique of Image Registration and Its Application to Stereo,” *Proc. 7th Int’l Joint Conf. on Artificial Intelligence*, August, 1981, pp. 674-679.

[Marchionini95] Marchionini, G. and Maurer, H. “The Roles of Digital Libraries in Teaching and Learning,” *Communications of the ACM* **38**, April 1995, pp. 67-75.

[Mills92] Mills, M., Cohen, J., and Wong, Y.Y. “A Magnifier Tool for Video Data,” *Proc. CHI ‘92*, May 1992, Monterey, CA, pp. 93-98.

[Narasimhalu95] Narasimhalu, A. (ed.) Special section on content-based retrieval. *Multimedia Systems* **3**, No. 1, 1995.

[Perelman90] Perelman, L. “A New Learning Enterprise,” *Business Week* (Dec. 10, 1990).

[Rao95] Rao, R., Pedersen, J., Hearst, M., Mackinlay, J., Card, S., Masinter, L., Halvorsen, P.-K., and Robertson, G., “Rich Interaction in the Digital Video Library,” *Communications of the ACM* **38**, April 1995, pp. 29-39.

[Rehg94] Rehg, J. and Kanade, T. “Visual Tracking of High DOF Articulated Structures: an Application to Human Hand Tracking,” *Proc. ECCV94*, May 1994.

[Resnikoff89] Resnikoff, H. L. *The Illusion of Reality.* New York: Springer-Verlag, 1989.

[Samuelson93] Samuelson, P. and Glushko, R.J. “Intellectual property rights in digital library and hypertext publishing systems. *Harvard Journal of Law & Technology*, **6**, 237 (1993).

[Samuelson95] Samuelson, P. “Copyright and Digital Libraries,” *Communications of the ACM* **38**, April 1995, pp.15-21, 110.

[Satoh92] Satoh, T., Yamane, J., Yee-Hong, G., and Sakauchi, M. “A Multimedia Retrieval System Using Video Scene Description Language,” *Seisan Kenkyu* **44** (Japanese), No. 11 (November 1992), pp. 23-25.

[Sirbu95] Sirbu, M. and Tygar, J. “NetBill: An Internet Commerce System Optimized for Network Delivered Services,” *Proc. IEEE CompCon Conf.*, March, 1995. Available in electronic form at http://www.ini.cmu.edu/netbill/CompCon.html.

[Srihari94] Srihari, S., Lam, S., Hull, J., Srihari, R., and Govindaraju, V. “Intelligent Data Retrieval from Raster Images of Documents,” *Proceedings of the First Annual Conference on the Theory and Practice of Digital Libraries: Digital Libraries ’94*, June 19-21, 1994, Texas A&M University, College Station, TX. Available in electronic form at http://atg1.wustl.edu/DL94.

[Stevens89] Stevens, S. “Intelligent Interactive Video Simulation of a Code Inspection,” *Communications of the ACM*, July 1989, pp. 832-843.

[Stevens92] Stevens, S., “Next Generation Network and Operating System Requirements for Continuous time Media,” *Network and Operating System Support for Digital Audio and Video*, Ralph Herrtwich, Ed. New York, Springer-Verlag, Inc., 1992.

[Stevens94] Stevens, S., Christel, M., & Wactlar, H. Informedia: Improving Access to Digital Video. *interactions* **1** (October 1994), pp. 67-71.

[Sullivan93] Sullivan, T., and Stern, R. “Multi-Microphone Correlation-Based Processing for Robust Speech Recognition,” *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1993, Minneapolis, Minnesota, pp. 91-94.

[Swain91] Swain, M. and Ballard, D. “Color Indexing,” *International Journal of Computer Vision* **7**, November 1991, Netherlands, No. 1, pp. 11-32.

[TREC93] *Proceedings of the Second Text Retrieval Conference*, D. Harmon, editor, sponsored by ARPA/SISTO, August 1993.

[Zhang93] Zhang, H., Kankanhalli, A., and Smoliar, S. “Automatic partitioning of full-motion video,” *Multimedia Systems* (1993) **1**, pp. 10-28.

[Zhang95] Zhang, H., Tan, S., Smoliar, S., and Yihong, G. “Automatic parsing and indexing of news video,” *Multimedia Systems* (1995) **2**, pp. 256-266.

[Zhang95b] Zhang, H., Low, C., and Smoliar, S. “Video Parsing and Browsing Using Compressed Data,” *Multimedia Tools and Applications* **1** (March 1995), pp. 89-111.

ACKNOWLEDGMENT

This work is partially funded by the National Science Foundation, the National Space and Aeronautics Administration, and the Advanced Research Projects Agency.