RJ 10115 (91931) April 8, 1998
Information Management

Research Report

KEY TO EFFECTIVE RETRIEVAL: EFFECTIVE CATALOGING AND BROWSING

Dulce Ponceleon
Savitha Srinivasan
Dragutin Petkovic
Dan Zivkovic

IBM Research Division
Almaden Research Center
650 Harry Road
San Jose, CA 95120-6099

RJ 10115 (91931) April 8, 1998
Information Management

KEY TO EFFECTIVE RETRIEVAL: EFFECTIVE CATALOGING AND BROWSING

Dulce Ponceleon
Savitha Srinivasan
Dragutin Petkovic
Dan Zivkovic

IBM Research Division
Almaden Research Center
650 Harry Road
San Jose, CA 95120-6099

ABSTRACT:

Multimedia data is an increasingly important information medium today. Providing intelligent access for effective use of this information continues to offer challenges in digital library research. As computer vision, image processing and speech recognition research continue to progress, we examine the effectiveness of these fully automated techniques in architecting effective video retrieval systems. We present semi-automated techniques that combine manual input, and video and speech technology for automatic content characterization integrated into a single system we call *CueVideo*. CueVideo integrates voice and manual annotation, attachment of related data, visual content search technologies(QBIC™), and novel multiview storyboard generation to provide a system where the user can incorporate the type of semantic information that automatic techniques would fail to obtain.

Key To Effective Video Retrieval: Effective Cataloging And Browsing

Dulce Ponceleon, Savitha Srinivasan, Dragutin Petkovic, Dan Zivkovic
IBM Corporation/ Research Division
650 Harry Road
San Jose Ca 95120-6099, USA
Tel: 1-408-927-1927
E-mail: dulce, savitha, petkovic, danz @ almaden.ibm.com

ABSTRACT
Multimedia data is an increasingly important information medium today. Providing intelligent access for effective use of this information continues to offer challenges in digital library research. As computer vision, image processing and speech recognition research continue to progress, we examine the effectiveness of these fully automated techniques in architecting effective video retrieval systems. We present semi-automated techniques that combine manual input, and video and speech technology for automatic content characterization integrated into a single system we call *CueVideo*. CueVideo integrates voice and manual annotation, attachment of related data, visual content search technologies(QBIC$_{tm}$), and novel multiview storyboard generation to provide a system where the user can incorporate the type of semantic information that automatic techniques would fail to obtain.

KEYWORDS: digital library creation, cataloger, video annotation, speech recognition, video segmentation, video search and browse, and multiview storyboard.

INTRODUCTION
Vast amounts of multimedia information including video is becoming prevalent as a result of advances in multimedia computing technologies and high-speed networks. Video is rapidly becoming the most popular media, due to its high information and entertainment powers. Applications that benefit from video are education and training, marketing support, medical, entertainment etc. However, there are two basic impediments to enabling effective digital libraries of videos. The first is cataloging which includes video digitization, compression and annotation, and the second is the lack of fast and effective search and browse techniques for this massive video content over the network. While much of the research has focused on search and browse, cataloging has often been overlooked. We believe that cataloging should be performed by relying on minimal human input, together with an intuitive user interface that leverages advances in image and video processing, as well as speech recognition to speed up and simplify the process to the extent possible.

For our discussion, we define metadata to be any data descriptors that "tell us something" about video content that can be used as index terms for video browsing to help locate the desired material and deliver it in a manageable format. The international community has recognized the need to standardize on video descriptors. A new member of the MPEG standards family, denoted "Multimedia Content Description Interface", i.e. MPEG-7[16] has begun in 1996 and its definition is in the works. MPEG-7 will result into a standard set of descriptors that can be used to specify various types of multimedia. Note that metadata can include visual material that helps browsing, such as keyframes, storyboards, audio clips etc. We define attachments to be a group of related information consisting of different types of metadata and/or unstructured data that may be accessed from the retrieval system. In the context of speech recognition applications, we define a vocabulary to be the set of words or phrases that can be translated by a speech engine and constitutes one of the resources that the engine uses to process spoken input. The speech recognition system matches the acoustics from the speech input to words in the vocabularies; therefore, only words in the vocabulary are capable of being recognized.

In this paper we present our new CueVideo system that provides cataloging and search/browse in an integrated manner. The novelties of our approach are several: we use speech recognition technology to enable user annotations of video; we combine image content search[8,14] (QBIC tm) together with user input to create novel multiview storyboards as an effective tool to browse long videos. The rest of the paper is structured as follows: we first articulate the challenges in cataloging and retrieval of video in a precise manner, we then describe the concepts and some aspects of the implementation of the CueVideo system and conclude with a description of future work.

DESIGN CHALLENGE

Effective use of video in various applications is impeded by the difficulty of cataloging and managing video data. For example, media industries indicate that an hour long footage of video in the field can take up to tens hours to be fully cataloged and archived into their system. Similarly, archival of video clips to enable efficient reuse is a time consuming, tedious and inefficient process. A customer with 20000 video clips of a minute each is faced with spending at least 5 to 10 minutes per video clip between viewing, extracting keyframes and manual annotation for the purpose of cataloging the video clip.

Video as an information medium offers unique challenges. While there are commonalties between video and still images, video distinguishes itself with key features such as size of the data (several megabits/sec in compressed form), the time element and the presence of audio. Treating video simply as a collection of still images is not adequate[7] and does not convey all the information present. In contrast, even though the literature contains approaches to video abstractions[9,19], it is reasonable to say that there is not a straightforward and commonly accepted definition of a video summary[17]. Users such as editors, educators, and advertisers are unwilling to settle for simple playback to view the content of video and require faster and more effective ways to search and browse video collections.

Let us examine the effectiveness of fully automated video content characterization. One aspect of cataloging deals with annotation (mainly keywords and text) of the video, which constitutes part of the metadata. Certain type of information about the video such as time, author, venue etc. is often not encoded in the video content and therefore, cannot be extracted automatically. Moreover, the type of information being provided by the annotation can be very domain specific and may vary for different

applications and users. For example, the attributes of a particular video clip that a movie editor may wish to highlight are completely different from the attributes that a video footage enterprise may wish to highlight. Automated text extraction and audio indexing techniques at this time prove to be inadequate to address this issue. Another aspect of the cataloging process deals with the creation of logical groupings of segments of the video along with related information. The logical grouping of video segments may be driven by the specific application domain. For example, editorial cues may be an important aspect in highlighting a collection of video segments for a movie editor, where as related technical excerpts may form a relevant grouping for an education video. Video information being temporal and spatial in addition to being unstructured, poses real problems for automatic extraction of such semantic information. The nature of video content as in an indoor video with a single speaker versus an outdoor sports video impacts the retrieval, and therefore, the cataloging process. Yet another variable may be the average length of the videos in the library. Technology used must be flexible enough to adapt to a variety of video footage from a two-hour feature movie, an hour training course, a five minute interview, to a few seconds news report. The cataloging techniques and retrieval criteria for video clips of different lengths may vary considerably. Therefore, we believe that a combination of automated content characterization techniques and manual segmentation and characterization are necessary to address these issues.

Video segmentation is commonly used as a first step to automatically analyze video content. In video production a shot corresponds to the segment of video captured by a continuous camera recording. Shot-boundary detection algorithms[2,22,24] are used to partition the video into elemental units called shots. Each shot or group of shots is annotated with text and keywords, one or more frames are extracted and treated as still-images to apply visual content search technology. We use the term storyboard[14] to denote a collection of visual proxies, such as thumbnails of representative frames, that convey the story in a highly efficient manner. A storyboard can be two orders of magnitude smaller than the corresponding compressed MPEG-1[11] video. This greatly reduces the bandwidth required to deliver the information and makes the storyboard a suitable feedback mechanism for the internet. Such storyboards also offer "visual index" to videos, namely, *at a glance* they offer the user basic information on what is happening in the video and ability to decide whether to play it, and what part of it to play. In this way, using video storyboards enables our powerful "brain-eyes" system to speed up the search and browse.

The challenge, therefore, is that of designing a cataloging and browsing system that uses a balanced combination of automated techniques and human effort to populate, segment and index the content of a digital library, enabling innovative search and browse interfaces to retrieve video collections. Towards this goal, we are developing CueVideo, a system which combines manual and automatic techniques for video cataloging, and offers flexible ways of browsing catalogued video. We combine several mature research technologies in a novel manner. We use IBM's Speech Recognition technology[10] to exploit the use of speech recognition for annotation in our cataloging interface. We have chosen a look and feel similar to that of a traditional video editing system and have based the search and browse on the creation of a storyboard as a collection of representative frames, linearly arranged, with variety of viewing or filtering options, referred to as a multiview storyboard. We also use image content search technology[8] to help organize the storyboard. CueVideo offers the ability to manually enter metadata and group the extracted keyframes, either by typing or speaking to it. An intelligent video retrieval system must take into account that the user's needs for browsing may vary with time and with the application. By

combining different cues, such as speech annotations, storyboard editing and QBIC, our system allows grouping of representative frames into user-defined logical groups. This in turn enables search and browse beyond the shot level as well as enables multiple views of the video without modifying the raw material. In this manner, we aim to optimize the human-computer participation by letting computers do what they do best and facilitating human interaction to provide the semantics that computers cannot capture.

RELATED WORK

Efforts to provide video browsing date back to the early nineties. Those systems did not use semantic information in the video but extracted keyframes at evenly-spaced time intervals and displayed them in chronological order[13]. By 1993, the use of image processing techniques lead to more sophisticated and content-based solutions[23]. These solutions involve segmenting the video using shot-boundary detection algorithms, and then selecting one or more frames[24] from every shot. Initially, video analysis meant almost exclusively image analysis. In 1994 the literature reports efforts to analyze video as a rich media and to leverage from advances in several areas such as computer vision, natural language processing, speech recognition and text retrieval[2,22]. The proliferation of retrieval systems is just a sign that this is a very active area of research. A comprehensive survey of techniques, algorithms and tools addressing content-based analysis and visual content representation has been published[1]. We list some of the existing solutions to intelligent retrieval of digital video, and summarize their distinctive characteristics.

The Informedia project[9] constitutes a pioneer and successful system for library creation and exploration that integrates different sources of information, such as, video, audio, and close-caption text, present in video data. Speech recognition is used to create time-aligned transcript of the spoken words, and to segment the video into "paragraphs". Significant images and words from a paragraph are extracted to produce a short summary of the video, a video skim, that allows effective searching and browsing. The transcription element and the novel integration of several cues makes this project a landmark in video retrieval. The video annotation system Vane[5] constitutes a semi-automated tool that facilitates the creation of large video databases. It is a domain-independent approach and it uses the Standard Generalized Markup Language (SGML) as the model of metadata collection. The Video: Visual Information Search, Transformation, and Analysis (VideoVISTA) project[4] constitutes a comprehensive ongoing research and development on automatic video annotation, segmentation, searching and browsing. One of the distinctive elements in this system is the capability of displaying video in a nonhierarchical fashion, a video graph. This graph captures the dynamics of the video, and the transitions among different segments in the video. The Virage Media Management System[3] provides a compelling solution to the cataloging problem. The system has two components: the video cataloguer and the media manager and browser. The cataloger supports time-based and content-based keyframe sampling, text-based annotation, and metadata extraction such as audio profiling (speech, music, noise or silence), close-caption text, and time stamping. WebClip[12] and VideoQ[6] constitute good examples of components of a video retrieval systems tailored to the world wide web environment.

Most digital library projects incorporate varying extents of cataloging in their solutions, while much of the emphasis is on the library search and exploration phase. In contrast with some of the systems

mentioned, CueVideo uses speech recognition for annotation, makes related media available through attachments, and supports user-defined grouping of shots and/or units of higher granularity. Our novel multiview storyboard addresses the issue of incorporating different subjective views of the same underlying raw data. Additionally, our systems leverage from a powerful and mature tool such as QBIC to enrich the exploration phase supporting image-content querying and viewing. We are integrating the above features into one system and continue to experiment with a variety of video content, with our primary focus being footage of technical talks presented within IBM.

THE CUEVIDEO SYSTEM

We describe the CueVideo system, with emphasis on the cataloging phase and the novel features incorporated in it. The domain of our prototype consists primarily of videos of technical talks and presentations of general interest to IBM's Almaden Research community. The cataloging system enables users to browse and search this talks database over the intranet.

Cataloging Phase

Figure 1 summarizes the various aspects of the cataloging process tightly integrated into one system. We combine components for digitization, compression, creation of logical groupings, annotation and multiview storyboard generation, resulting in representing the metadata in a generic data schema that satisfies retrieval in different application domains.

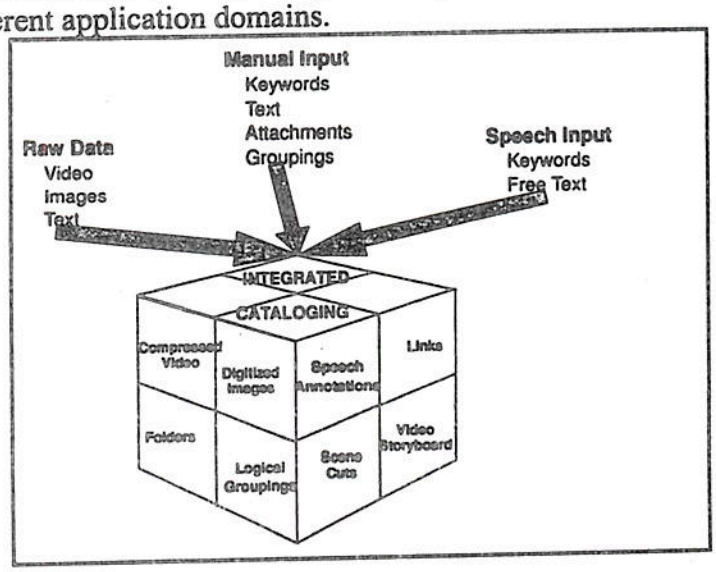

Figure 1: What is Cataloging?

The logging process integrates automatic video segmentation with the ability to characterize the content manually by creating user-defined groupings based on the application domain. Figure 2 illustrates some of the elements involved in a typical cataloging session with our system. Initially the video is segmented into shots and a number of frames are extracted to summarize each shot. In this example, three shots have been detected, and annotated through speech. Video segmentation results in an automatically generated storyboard which contains all the representative frames extracted (five representative frames for the three shots in Figure 2). Additionally, related media is attached to these shots, enabling queries on the content of such attachments. Attachments can include foils, text files, relevant graphs or pictures and ultimately references to another video. Specifically, for our application domain of technical talks, the

value of attaching related foils, the speaker's publications, related internal projects, and talks is immense.

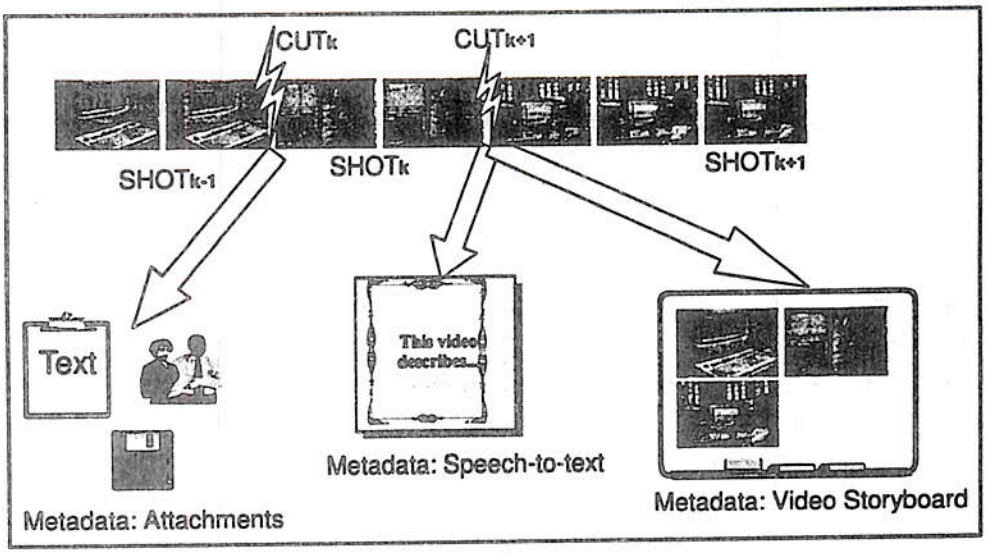

Figure 2: Some Details of Cataloging

While the storyboard is a starting point, it may not always be adequate to search and browse the entire video, since videos can have very different characteristics, duration being one of them. Viewing a thousand-keyframe storyboard of a two-hour documentary is not very valuable. Likewise movies that exhibit a lot of motion, such an MTV movie, or a educational clip on fractals will tend to render storyboards with more detail than desired. Additional views with higher level granularity are needed. To address this, our system supports manual grouping of representative frames. Shots can be grouped into scenes, scenes into stories, stories into segments, and so on. User-defined groupings enables addressability at the scene, story and segment levels resulting into manageable storyboard displays. Figure 2 shows a storyboard view of a scene, composed of three shots, with one keyframe per shot. In the context of technical talks, it may be desirable to have several views of the same raw material. For example, a view for a naive audience that highlights only the general ideas and skips in-depth descriptions, a view for an expert audience that eliminates the introductory material, history or common definitions, or a view with just the mathematical notations. These views are driven by the interest and background of the user. Additionally, annotation at high level groupings enables the search and browse component to perform full-text search beyond the shot level.

We now describe the design and implementation of the key features in the cataloging phase.

The Primary Abstraction

Since the cataloging system is designed to address different application domains, we create an abstraction which is representative of a record in the specific domain. The primary abstraction we attempt to catalogue is a technical talk. The system represents each talk as a combination of metadata in a relational database and the associated digital content on the network file system. We use the title of the talk as the unique key and create a minimal record when it is first added to the system. In addition, we create the necessary tables to store the associated logical groupings, annotations and video storyboard. With this generic schema, we envision being able to accommodate other application domains such as

sports videos or consumer items where the theme object may be a sports highlight or a merchandise, but the relevant information is represented using the same data model.

Metadata Creation: Text Annotation using Speech Recognition

We use real-time speech recognition technology to provide an efficient means of entering the metadata associated with a talk. Speech recognition may be used in one of two ways in an application. The first is aimed at text-entry or document creation and is referred to as dictation. The second is targeted for transaction processing and data entry systems and is referred to as command and control or navigation. We use both, dictation and navigation to speed up the manual annotation process for the creation of the metadata.

Speech as an input modality for data entry offers unique user interface challenges. The reasons for this are somewhat apparent; in the context of a graphical user interface(GUI) application, there is a need for distinguishing between voice commands that are directed towards filling in the content of a particular field versus voice commands that must act upon the contents of a particular field. For example, Delete is a command that must act upon the contents of an edit control field. Having to deal with these modes in an application makes speech a cumbersome and unnatural means of input. We counter this problem in our annotation interface by eliminating modes such as the mode where a user issues commands that must act upon the content of a field. We create a *modeless* interface with a single vocabulary which contains the choices for all the fields, where the choices for each field are represented in a separate list. Based on what the user speaks, we replace the previous selection for a particular field with the new selection in the appropriate list.

VOCABULARY NAME	GUI FIELD	WORDS IN VOCABULARY
AnnotationVocab	Category	Talks, Presentation ...
AnnotationVocab	Location	Auditorium, Lab, Outdoors...
AnnotationVocab	Subject	Storage, Computers, Science...

Figure 3: Annotation Vocabulary Design

The table in figure 3 illustrates the design of a vocabulary for annotating a talk. The *Category*, *Location* and *Subject* fields share a common vocabulary such that the choices for each field are enabled at all times. The user does not need to, but may, speak the navigational words such as *Category*, *Location* or *Subject*. The choices for each field are enabled at all times and the cataloging system replaces a previous selection with the current voice input. Figure 4 shows the corresponding GUI window associated with the speech annotation in the cataloging system.

We also use the dictation mode of speech recognition for the creation of unstructured, free form text so as to allow recording of any information which was not captured using the constrained vocabulary interface. Dealing with the correction of recognition errors is an important issue for document creation systems; however, we use dictation only to capture the free form, textual metadata where the accuracy of the transcript is not critical. We therefore alleviate the need for an error correction interface; we simply use this transcript as yet another field on which to perform a full-text search when the video retrieval query is issued.

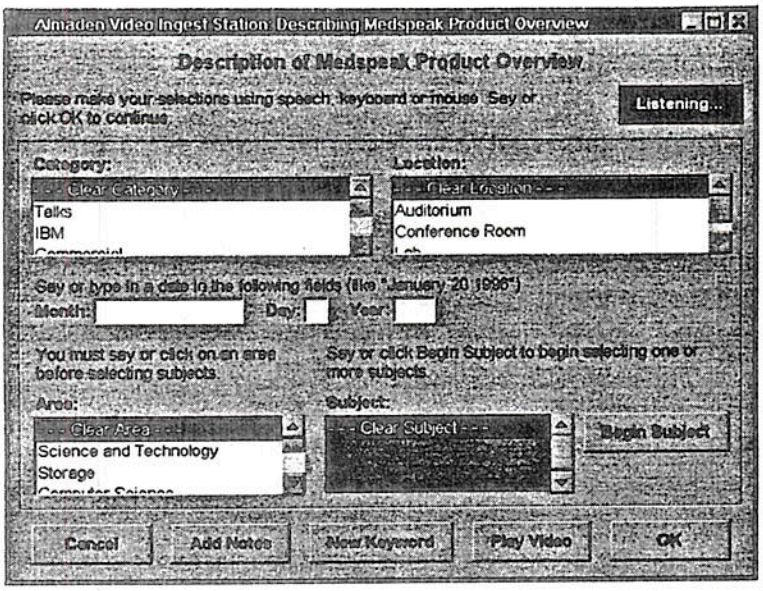

Figure 4: Annotation Window

Metadata Creation: Attachments

Our cataloging system provides the ability to create attachments associated with the primary abstraction, in our case, a talk. Different types of attachments such as text documents, images and word processing/presentation documents are supported. This also gives us the ability to perform content based searches on certain types of attachments such as text, images and presentations.

Metadata Creation: Multiview Storyboard Creation

We describe here the different elements involved in the generation of a multiview storyboard. Video segmentation constitutes the first step of video processing and shots are the fundamental unit of video manipulation: cataloging, processing, and representation. In the preprocessing phase, representative frames are extracted from every shot and a storyboard is automatically generated. This storyboard

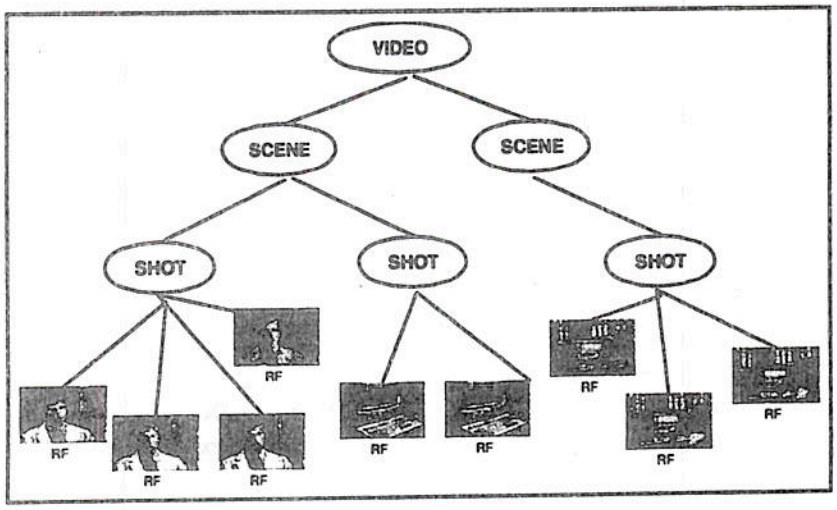

Figure 5: Storyboard Hierarchy

exhibits the finest granularity. At the same time, during cataloging the user can use speech or text to annotate the shots. In a post processing phase, different storyboard views can be created by grouping shots. User-defined groupings can also be performed through speech during cataloging.

Our shot detection algorithm works directly on an MPEG-1 compressed video and handles both abrupt shot changes and some gradual transitions. Shot detection can be perform in the uncompressed or compressed domain. In the latter, the technique is similar to that proposed by Yeo and Liu[21], where fast compressed shot detection is achieved by extracting and processing spatially reduced images, *DC images*. While DC images are small and of low resolution, global image features are still well preserved. The storyboard proposed here corresponds to a collection of representative frames, linearly arranged, that faithfully capture the content of the movie. Our storyboard supports a variety of viewing options that render higher granularity. Figure 5 shows a typical hierarchical partitioning of a video. It is composed of two scenes: the first one with two shots and the second one with a single shot. At the finest granularity, the representative frames(RF) per shot are extracted automatically. We have chosen thumbnails to be the visual proxies for representative frames. Storyboards that use a mixture of thumbnails and mosaic views[20] have been reported. In some cases, our storyboard is not simply a collection of *one* frame per shot, because that may not be enough to communicate the story. Preliminary versions of the storyboard generation incorporate some heuristics that follow simple semantic editing rules. For instance, most shots have been structured by the producer to progress smoothly from the beginning, to the middle, and to the end. Therefore, it is not unreasonable to provide a strategy for keyframe extraction that is based on the *duration* of the shot. This simple approach, although not perfect, is well suited to typical feature movies. Our experiments show that for a typical feature film, we get a storyboard with approximately one keyframe per two to five seconds. Consequently, the storyboard can represent several minutes of video on a single HTML page. The compression rate i.e. MPEG-1/Storyboard for feature films is approximately a hundred. Our experiments on IBM commercials, where shots are short, show that the storyboard can fit in a single HTML page, but the compression rate is approximately fifty. The current version of our storyboard has been incorporated as an extension in a database product.

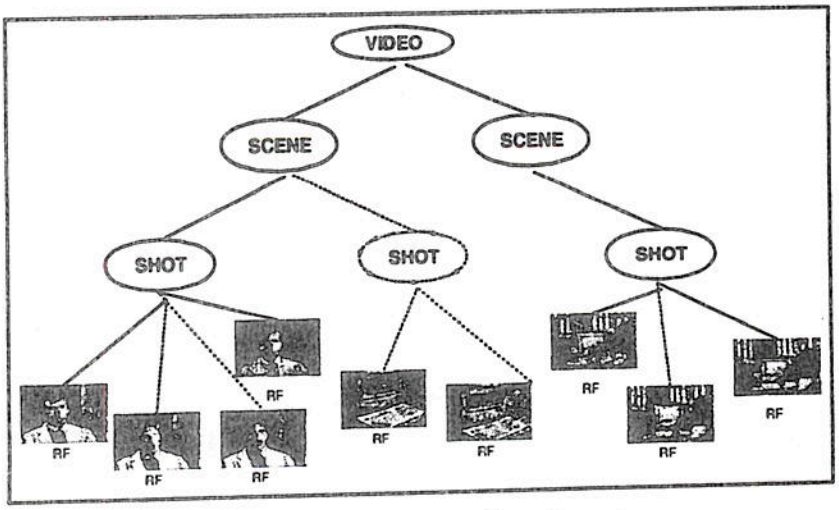

Figure 6: Multiview Storyboard

Figure 7: Storyboard Window

Different application domains can be addressed by the multiview storyboard generation. Figure 6 shows the conceptual creation of a storyboard view by the user. Only those keyframes connected by a continuous line, are selected to create a new view. Keyframes with dotted lines are no longer visible within the new view. Such grouping of keyframes into user-defined logical scenes provides addressability at coarser levels than the shot level.

Once we edit/change an existing storyboard, the new version can be considered as a different view Figure 7 shows a screen shot which displays the results of our automatic video segmentation in a storyboard on a sample video clip. It also shows how the user can generate a new view for that storyboard, by keeping only the representative frames of interest. Given the automatically generated storyboard, our system provides a user interface where it is possible to select keyframes of interest by selecting the enable option. Analogously, the user can select the disable option and eliminate a keyframe from the storyboard view. These changes are only reflected in the new storyboard view, the underlying raw footage and automatically generated storyboard are preserved. In figure 7, keyframes with a darker border are not included in the new view. Notice it is also possible to play the video. We plan to extend our system to support segmented playing of the video. That is, by selecting the beginning keyframe and end keyframe the user can play only the segment of video between the two selected entry points in the storyboard.

Another storyboard view where the entries are sorted by a similarity measure on the content, based on QBIC technology, is also possible. This will result in a storyboard view that contains the thumbnails of videos that have similar content to a previously created storyboard. It provides one means of computing video similarity. This can be a powerful tool in an environment such as a brainstorming meeting, where the participants are interested on data that are "similar" to the topic in discussion. For example, an education video which consists of alternating shots of speakers and the white board may benefit by grouping the storyboard keyframes based on QBIC search. This will result in the white board shots being grouped together, and also all the shots for each speaker being grouped together.

In summary, the multiview storyboard offers the user several selective views of the video content. It can vary from a fine granularity view, i.e. automatically extracted keyframes including detailed information for each shot, to a higher granularity view i.e. user annotated logical groupings or sorting by visual similarity (using QBIC).

Retrieval Phase

The retrieval phase is envisioned as a web browser based interface through which end users may search, browse and view the content of a digital library. The user will have several options in displaying the storyboard, and also an option to play selected segments of the video. Preliminary experiments indicate that the current cataloging solution will provide acceptable performance in a low traffic, network file system based, intranet environment using a file based video player. The internet solution will necessarily use a streaming video player backed by an appropriate video server component, gated by a web server which controls access to all resources required by the browser.

System Implementation

We have initially focused on the cataloging interface of the CueVideo system, with the goal of establishing the infrastructure necessary to build a retrieval system for the internet. We are developing this system for Windows NT(tm) withTcl/Tk[15] as a scripting language for prototyping the cataloging user interface. We use DB2 as the relational database to store the metadata associated with the digitized videos. We used the network file system to store the MPEG-1 compressed video together with a file based video player for initial experimentation on the intranet. We use both, the command mode interface and the dictation mode interface of ViaVoice speech recognition technology to provide the speech annotation. We have experimented with different speech user interfaces in an attempt to provide an intuitive, easy-to-use annotation interface. Even without training the speaker, the accuracy of the command mode interface is greater than 90%. For the dictation mode interface, our experiments indicate that a recognition accuracy as low as 50% is still valuable in capturing important content words. We have integrated our video storyboard algorithms into the cataloging interface and are experimenting with the creation of different views of the storyboard.

CONCLUSION AND FUTURE WORK

Content-based video retrieval and browsing continue to pose challenging problems. Approaches that integrate different cues present in a video are essential to provide satisfactory solutions. Speech, sound and text are information sources as important as the visual data. We believe that solutions that leverage from the state-of-the-art technology in computer vision, and speech recognition etc. while providing tools to facilitate human input will render intelligent and precise access to digital video libraries. We also want to point out that summarizing video is a challenging user interface problem. Given the dynamic nature of video it is non trivial to create a good proxy that conveys its contents and is presented to the user in an intuitive manner.

Our video retrieval system proposes a novel way to integrate speech recognition technology and video processing into a powerful, appealing and reusable video retrieval solution. The use of speech recognition in the interface makes valuable text annotation possible. The designed speech interface makes it easy for untrained personnel to create the metadata associated with a video. The constrained

vocabulary interface captures the structured component of the metadata where as the dictation mode interface effectively captures the content words since many recognition errors are function words. This leads to an acceptable transcript for the purpose of full-text search despite low recognition accuracy. Our approach of combining automated video content characterization techniques empowered by human knowledge requires a knowledgeable domain expert to be involved in the cataloging process. While this certainly adds to the cost of cataloging, it enriches the retrieval and browsing experience enabling adaptable/tailored views of the content, facilitating repurposing of the data, and ultimately higher performance of the overall system.

Our work in this area has just begun, and there are many novel components we plan to research and incorporate as appropriate. Regarding speech recognition we will attempt automated word spotting to enable some form of text indexing and search based on the decoded audio track. We will also look into audio analysis to extract "audio events" such explosion, shots, loud noises etc. We plan to apply more advanced video segmentation to segment moving objects form background and enable search to be performed selectively, including QBIC search on objects only. In addition, we will allow the user to select region of interest and search for events happening only in those regions (important for applications such as surveillance). All selected keyframes will be labeled, thus offering "multiview" or selective browse capability for the user. We will also add new methods of selecting keyframes, rather than strict time sampling within the shot which has been used in the past. Ultimately, we will test the system with real users on several content classes such as education/training, technical presentations and feature films.

ACKNOWLEDGMENT

We would like to acknowledge Byron Dom for his contributions on early implementations of the scene cut detection algorithm. We also want to thank Michael Penner for sharing his expertise in building the cataloging prototype in CueVideo. Finally, we thank the T.J. Watson Research Center's Speech Research team for sharing the necessary technology, tools and experience.

REFERENCES

1. P. Aigrain, H-J. Zhang, D. Petkovic, Content-Based Representation and retrieval of Visual Media: A State-of-the-Art Review. In *Multimedia Tools and Applications, Vol. 3, 179-202, 1996.* Kluwer Academic Publishers.
2. F. Arman, R. Depommier, A. Hsu, and M.Y. Chiu, Content-based browsing of video sequences, *in ACM Multimedia* 94, pp. 97-103, Aug. 1994.
3. J. R. Bach et al., Virage image search engine: An open framework for image management, In *Proceedings of SPIE Storage and Retrieval for Still Images and Video Databases IV, Vol. 2670, pp. 76-87,* (San Jose, CA) February 1996. http://www.virage.com
4. R. M. Bolle, B. L. Yeo, and M. M. Yeung, Video Query: Beyond the Keywords, IBM Research Report RC 29586, October 1996.
5. M. Career, L. Ligresti, G. Ahanger, and T.D.C. Little, An Annotation Engine for Supporting Video Database Population. In *Mulitmedia Tools and Applications, Vol 5, 233-258, 1997.* Kluwer Academic Publishers.

6. S-F. Chang., W. Chen, H.J. Meng, H. Sundaram, and D. Zhong, VideoQ: An Automated content Based Video Search System Using Visual Cues. In *Proceedings of MM'97* (Nov 9 –Nov 13 pp. 313-324, Seattle, WA), 1997.
7. N. Dimitrova, The Myth of Semantic Video Retrieval. In *ACM Computing Surveys, Vol 27, No. 4,* December 1995.
8. M. Flickner et al., Query by image and video content: The QBIC system. In *IEEE Computer, Vol 28, No. 5, pp.23-32*, Sept., 1995. URL: wwwqbic.almaden.ibm.com
9. A. Hauptmann, M. J. Witbrock, Informedia News-On-Demand: Using Speech Recognition to Create a Digital Video Library.
10. IBM Corp., ViaVoice Dictation, ViaVoice Developers Toolkit, 1997.
11. ISO/IEC 11172:Coding of moving pictures and associated audio for digital storage media up to 1.5 Mbits/s, Part1: Systems; Part 2:Video; Part 3: Audio; Part 4: Conformance Testing, 1993.
12. H.J. Meng, D. Zhong, and S-F. Chang, WebClip: A WWW Video Editing/Browsing System. In *Proceedings of MM'97* (Nov 9 –Nov 13, Seattle, WA), 1997.
13. M. Mills, J. Cohen, and Y.Y. Wong. A magnifier tool for video data. In *Proc. of ACM Computer Human Interface (CHI)*, May 1992.
14. W. Niblack, X. Zhu, J.L. Hafner, T. Breuel, D. Ponceleon, D. Petkovic, M. Flickner, E. Upfal, S.I. Nin, S. Sull, B. Dom, B-L. Yeo, S. Srinivasan, D. Zivkovic and M. Penner, Udpates to the QBIC System, In *Proceedings of IS&T/SPIE Storage and Retrieval for Image and Video Databases VI.* (Jan 28 - Jan 30, San Jose, California), 1998.
15. J.K. Ousterhout, Tcl and Tk Toolkit, Addison-Wesley Publishing Company, 1994.
16. F. Pereira, MPEG-7: A standard for content-based audiovisual description, In *Proc. of Int. Conference on Visual Information Systems (VISUAL'97),* San Diego, California, December 1997.
17. S. Pfeiffer, R. Lienhart, S. Fischer, and W. Effelsberg, Abstracting Digital Movies Automatically. *In Journal of Visual Communication and Image Representations,* Vol 7, N 4, pp. 345-353, December, 1996.
18. M.A. Smith and M.G. Christel, Automating the Creation of a Digital Video Library. In *Proceedings of MM'95* Nov, San Francisco, CA, 1995.
19. M. Smith, and T. Kanade, Video skiming for quick browsing based on audio and image characterizations. *In Computer Science Technical Report*, Carnegie Mellon University, July 1995.
20. Y. Taniguchi, A. Akutsu, and Y. Tonomura, PanoramaExcerpts: Extracting and Packing Panoramas for Video Browsing. In *Proceedings of MM'97* (Nov 9 –Nov 13, Seattle, WA), 1997.
21. B. L. Yeo, and B. Liu, Rapid scene analysis on compressed video, *In IEEE Transactions on Circuits and Systems For Video Technology,* Vol 5, pp. 533-544, December 1995.
22. M. Yeung, B. L. Yeo, W. Wolf and B. Liu, Video Browsing using Clustering and Scene Transitions on Compressed Sequences. *In Multimedia Computing and Networking Proc. SPIE*, February, 1995.
23. H.J. Zhang, A. Kankanhalli, and S.W. Smoliar, Automatic partitioning of full-motion video, ACM/Springer Mulitmedia Systems, Vol. 1, No. 1, pp. 10-28, 1993.
24. H.J. Zhang, C.Y. Low, and S.W. Smoliar, Video parsing and browsing using compressed data, Multimedia Tools and Applications, Kluwer Academic Publishers, Vol. 5, pp. 533-544, Dec. 1995.