The Digital Video Library System:Vision and Design

(published, Digital Library'94,college Station, TX,47-52)

Susan Gauch*, Ron Aust, Joe Evans*, John Gauch*,
Gary Minden*, Doug Niehaus*, James Roberts*

*Electrical Engineering and Computer Science
sgauch, evans, jgauch, gminden, niehaus, roberts@eecs.ukans.edu
School of Education
aust@kuhub.cc.ukans.edu
University Of Kansas, Lawrence, KS 66045

Abstract

The digital libraries of the future will provide electronic access to information in many different forms. Recent technological advances make the storage and transmission of digital video information possible. This paper will describe the design of a Digital Video Library System (DVLS) suitable for storing, indexing, searching, and retrieving video and audio information and providing that information across the Internet or the evolving National Information Infrastructure. To be an effective library, users need to be able to find the video segments they want. Realizing this goal will require ground-breaking research into automatic content-based indexing of videos that will significantly improve the users' ability to access specific segments of interest with videos. In our approach, videos, soundtracks and transcripts will be digitized, and information from the soundtrack and transcripts will be used to automatically index videos in a frame by frame manner. This will allow users to quickly search indices for multiple videos to locate segments of interest, and to view and manipulate these segments on their remote computer. While this technology would be applicable to any collection of videos, we will target educational users, providing teachers with the ability to select segments of nature and/or current events videos which complement their lessons.

Keywords: video libraries, indexing, information retrieval, education

1. The Vision

How does a teacher find a video clip of the Challenger explosion? Or a video of Alan Shepard's first sub-orbital rocket launch and capsule recovery? Or a video of Vice President Gore announcing the National Information Infrastructure? Or a video showing how to unjam the office copier?

We plan to develop technologies necessary to provide desktop access to video segments stored in remote digital libraries, specifically automatic video indexing and digital video delivery via computer networks. We intend to focus on four primary areas: (1) the acquisition, digitization, and storage of video; (2) the indexing of video using scripts, manual techniques, and speech recognition applied to the sound track; (3) the retrieval of appropriate video clips using information retrieval techniques; and (4) the access to the video library and distribution of video via the Internet. In short, we intend to implement a Digital Video Library System, or DVLS.

The exponential growth of the Internet over the past decade has fundamentally altered the way researchers and educators look at information. New information is created and shared among the millions of Internet users at an inspiring rate. As a result, there is an increasing expectation that the information we need is "out on the net somewhere", all we need to do is find it. Unfortunately, the supply of information is increasing more rapidly than our ability to support effective searching of this tremendous resource. To combat this problem before the Internet implodes, we must make fundamental advances in how this information can be captured, stored, searched, filtered, and displayed.

Textual information such as news postings and technical reports make up a large number of the items accessible by the Internet, but they represent a smaller portion of the "volume" of data available. Non-textual information such as sound recordings, images, video, and scientific data require considerably more storage per item. For example, a page of text contains only 4,000 characters, while a single image requires roughly 300 KB, and a minute of uncompressed digitized video requires over 500 MB. With the increasing availability of input and output devices on the multimedia workstations, the supply and demand for non-textual information sources is likely to grow exponentially in the near future.

Current technology to support the search and retrieval of non-text information is far behind text based systems. Items from archives are typically selected by name, copied from the archive to the user's site, and examined. Although simple, this approach has a number of serious drawbacks. If the names of items are not chosen well, certain items may never be accessed, and other items may be incorrectly retrieved. Because non-text items such as images and videos can be 100 to 100,000 times larger than typical text items, the retrieval of useless data puts a tremendous strain on the communications capabilities of the Internet. To reduce this strain, and increase the availability of valuable video data, we plan to develop and evaluate new technologies for digital video libraries which support intelligent content-based searching and retrieval of video information.

Our plan is to build a digital video library storing approximately 100 hours of short (2-5 minute) video segments which can be searched using full-text queries. To support the storage, retrieval, and transmission of this enormous quantity of digital video, a high performance video storage system must be constructed which utilizes state-of-the-art compression and communication techniques. To support text-based video searching, automatic indices must be constructed based on textual information extracted from the video. To support remote access to the DVLS by students and educators, a variety of graphical user interfaces must be developed. A number of important research questions need to be addressed in each of these areas while building the DVLS.

2. Related Work

Technological advances of several kinds are converging to transform the ways in we generate, store, and use information. Digital libraries are being built which store a wide variety of information and information types: page images of technical journal articles [Lesk, 1991; Hoffman and O'Gorman, 1993], nucleic acid sequence data [Burks, 1991], geographic information [Pissinou, 1993], computer science technical literature [Brunei and Cross, 1993] to name a few.

With regular libraries, the user goes to the information. In the digital realm, the information is delivered to the user; requiring easy to use, easy to learn user interfaces [Fox, 1993], and information servers which can interface with a wide range of client technologies [Kahle and Morris, 1993]. The ability of users to manipulate retrieved information has fundamentally changed the relationship between the information producer and consumer [Rawlins, 1993], prompting attention to both the legal and social aspects of this process [Garrett & Lyon, 1993].

A recent development is the emerging ability to digitize and manipulate video and audio information. In addition to teleconferencing, this has a wide range of commercial applications. For example, the AP wire service is beginning to transmit digitized video clips as well as text over its existing network [Broadcasting and Cable, 1993]. Twentieth Century Fox and Sony are digitizing news reels from the thirties and forties [Business Week, 1993a], which will be a unique educational resource. Digital video is also been utilized in marketing research firm reports [CD-ROM Professional, 1993] and in marketing products over the ECnet which links manufacturers and suppliers [Computer World, 1993]. Finally, digital post production is becoming standard in the film industry, which is continuing to push the state of the art for manipulating video images [Zorpette, 1993].

Large scale collections of video data are also getting attention. For example, AT&T envisions a huge digital library storing a wide range of data, including movies for viewing on demand, interactive presentations, educational materials, marketing presentations, and news [Business Week, 1993b]. To make this dream a reality requires research in the basic technologies necessary to implement digital video libraries. Recent efforts have been made in developing the individual components necessary for handling multimedia data [Nicolaou 1990; Rangan 1993], and building software systems and operating systems designed to handle multimedia data [Fox 1991; Jeffay 1992].

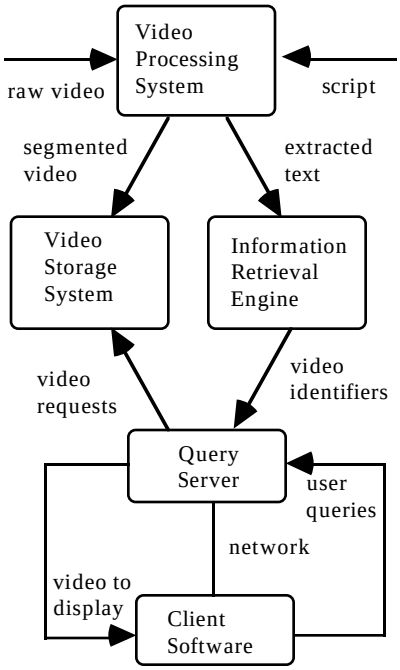
What is needed is the technology to treat collections of digital video segments as a library which can be automatically indexed and searched based on the contents of the video. Given the limited descriptive ability of current computer vision systems [Haralick and Shapiro, 1992], and the improving accuracy of connected speech recognition systems [Takebayashi, 1991], the most sensible approach for automatically indexing video is to extract textual descriptions of the video directly from the audio track. The Video Mail Retrieval Using Voice project at The University of Cambridge represents one effort in this direction [James, 1994]. This group is attempting to extract video indexing terms from the sound track and written contents of video mail.

3. The Design

3.1 System Overview

The DVLS is a complex system composed of the following five primary components.

- Video Storage System (VSS). The VSS stores video segments for processing and retrieval purposes. Since our objective is to provide intelligent access to portions of a video rather than entire videos, the VSS must be capable of delivering numerous short video segments simultaneously.
- Video Processing System (VPS). The VPS consists of video processing programs to manipulate, compress, compact, and analyze the video and audio components of a video segment. In particular, the VPS contains a component to recognize keywords from the sound track of video segments.
- Information Retrieval Engine (IRE). The IRE is used to store indices extracted from video segments and other information about the video segments, such as source, copyright, and authorization. The IRE will be capable of supporting both free-text and Boolean queries.
- Client. The Client is a graphical user interface which resides on the user's computer. It includes interfaces for conducting structured and free text searching, hypertext browsing and a simple video editor.
- Query Server (QS). The QS processes video queries from the remote Client and communicates with the IRE and VSS to enable users of the digital library to extract video data and create multimedia representations of the information of interest.



As can be seen in Figure 1, these components are tightly interrelated and support three very different DVLS functions: (1) the creation of the DVLS archive [Section 3.2], (2) the processing of video in the DVLS to build automatic indices [Section 3.3], and (3) the access of the DVLS by users of the testbed system [Section 3.4].

3.2 Creating the Video Archive

The first step in building the DVLS is acquiring digital video. We plan to obtain Nature programs from WNET, Nova programs from WGBH and news stories from CNN and WGBH in analog form on professional quality magnetic media. This data will be digitized using a video-rate frame grabber and recorded on fast magnetic disks in the DVLS. There are a number of important design and implementation issues in this data acquisition and storage phase which require careful consideration.

The volume of data produced by digitizing an hour of video is tremendous. An hour of video contains 30 x 60 x 60 = 108,000 individual frames. Each frame of color video requires 640 x 480 x 2 = 614,400 bytes. Hence, an hour of raw video requires 66 GB of storage. In addition, an hour of audio digitized assuming a 10 KHz bandwidth will require 317 MB. Obviously, data compression is essential. We plan to store video with limited compression to retain "original quality" video rather than using higher compression schemes which yield "VCR quality" or

"cable quality" video. We will compare three lossy video compression techniques for this purpose: (1) differential pulse code modulation (DPCM) with compression rates of roughly 5:1, (2) Joint Photographic Expert Group (JPEG) compression with rates near 10:1, and (3) Motion Picture Expert Group (MPEG) compression with rates near 15:1. On average, we can expect to store one hour of video in 6.6 GB, so the 100 hour DVLS will require roughly 660 GB of disk storage.

In addition to bulk storage, the DVLS must be capable of storing, retrieving and transmitting digital video at 30 frames per second. Our initial estimate of usage patterns is that the testbed system will support 20 simultaneous users collectively accessing 500 minutes of video per hour. This translates to a total bandwidth requirement of 25-35 MBps. To support this high rate of data access and transmission, we have designed a video storage system consisting of a number of high performance processors with high capacity, high bandwidth disk arrays, connected by a high capacity Asynchronous Transfer Mode (ATM) local area network. In particular, each video storage module (VSM) will consist of:

- A Digital DECStation 3000 Model 600S Alpha workstation with three external fast SCSI-2 buses, and two 155 Mbps ATM network interfaces. The fast SCSI-2 buses have a maximum bandwidth of 20 MBps for a total maximum video segment bandwidth of 60 MBps. Each ATM network interface has a maximum bandwidth of 20 MBps and the DECStation 3000 Model 600S have a maximum I/O bus bandwidth of 100 MBps.
- Three disk arrays consisting of seven 2.1 GB disk drives for a total of 14.7 GB per bus and 44.1 GB per video storage module. The disk drives have a media transfer rate between 2.7 MBps and 5.5 MBps and a bus transfer bandwidth of 10 MBps. To store 100 hours of digital video, we will start with two VSMs in year 1. We expect disk capacity to double by year 2, so by adding four VSMs in year 2, and three VSMs in year 3, we will have roughly 700 GB of digital video online.

The VSMs will be interconnected using a local ATM network based on Digital Equipment Corporation's experimental AN2 network, loaned to the University of Kansas by Digital's System Research Center in Palo Alto as part of the ARPA sponsored MAGIC Gigabit Testbed. The AN2 connects processors to switches at 155 Mbps. The capacity of an AN2 switch is 12.8 GBps. The design of the network insures that multiple traffic streams can flow simultaneously and that, lacking failure, no data is discarded. The AN2 interfaces with other processors in the local area and is connected to the MAGIC wide area gigabit testbed.

We feel that by retaining the "original quality" video and audio data, we will be able to address a number of important issues relating to the development and use of digital libraries. First, by retrieving high quality video and audio, we can present a better product to the locally connected user, and evaluate the effectiveness of the DVLS for creating "production quality" composite videos. Second, by adaptively applying video and audio compression techniques which are suitable to the transmission mechanism between the server and the client, we can evaluate the effect of varying video and audio quality on delivery and use of the DVLS. Finally, by combining DVLS clients

with different video access and manipulation interfaces, we can examine heterogeneous system design issues, and usage patterns of a DVLS with multiple levels of video quality access.

3.3 Indexing the Video Library

Preprocessing to Support Video Search Before we can begin to index video segments, each video stored in the DVLS must be segmented into short meaningful scenes. Although this task is relatively easy for humans to perform, automatic image analysis to detect scene transitions in a video is an open problem. As a first step, we plan to develop and evaluate a number of image difference metrics to detect the large temporal changes which coincide with camera transitions. To segment each video, we will manually select the subset of camera transitions which mark scene transitions. The time stamps associated with each video segment will be recorded in a database for indexing purposes.

Next, we need to segment the audio track into utterances. Although the field of speech recognition has made significant progress in recent years, we do not expect to have 100% success in segmenting the audio track into words nor in performing word recognition to obtain a textual transcript of each video scene. Here is where the scripts and transcripts of each video are invaluable. By scanning these documents and performing OCR, we will have a second representation of what is being said in the video. By fusing this information with the text extracted via speech recognition, we hope to have an accurate transcription for each of the video segments in the DVLS. This will be a valuable contribution to the technologies necessary to support digital libraries.

Building Search Indices Videos are typically produced in relatively long segments (30 minutes to 2 hours) whereas many educational applications prefer short clips for conveying concepts or use in constructive activities that combine several video clips from different sources. To individually retrieve clips from the DVLS, the video and audio segments must be individually indexed. At present, content analysis of digital images via computer vision is not up to the task of providing indexing information for the 10 million frames in a 100 hour DVLS. However, there are many sources of information from which to build these search indices:

- speech recognition of the audio track
- transcripts
- closed-captions
- manually assigned keywords
- video/audio segment source (title, start-time, date, length...)
- video images characteristics (contrast, brightness, colors, ...)
- audio sound characteristics (background noise, volume, ...)

The DVLS will need to be able to do free-text searching on indices built from speech recognition, transcripts and closed-caption. In addition, Boolean searching will be provided on structured data from keywords, video segment sources and video image characteristics. However, not all information sources will be available for all videos. One of the important research questions for this project is the comparison of the effectiveness of the various indexing schemes. To run experiments, we will need to provide search capabilities on any combination of the available indexing sources. To do this, the query processor will need to be able to select which search indices to use when processing a given query, and to search based on multiple, disparate indices.

3.4 Accessing the Digital Video Library System

User Interface Our goal is to support searching of the digital library using both text-based queries and video-based queries. For example, a text-based query might ask for video sequences which contain scenes of monkeys or some other animal of interest. Once a collection of scenes are identified using our video indexing scheme, the user can view these scenes and identify the subset which are of most interest. Then, a video-based query could be used to ask for more scenes which are related to a specified video segment.

We will develop graphical user interfaces for accessing the DVLS which are appropriate in a K-12 environment. Based on the processing power of the video display workstation and the speed of data communication available from the digital library to the user, queries could return one or more of the following:

- a textual description of the set of video segments retrieved (e.g. video title, start and end time, transcript of script or audio track),
- the audio segments which coincide with the video segment,
- a small number of frames from each video segment, e.g. the first frame, one frame per minute, one frame per camera transition,
- a very small ("postage stamp") version of the video sequence,
- a time sampled version of the full resolution video for "fast forward" mode display, or
- a full resolution version of the video.

Remote Access Initially, we plan to support between 10 and 100 simultaneous users accessing the digital library over the Internet. Because users will typically extract multiple scenes from random locations within multiple videos, we will need to devise a different data retrieval and transmission scheme from developers of "video-on-demand servers". In particular, we can not assume the simultaneous transmission of the same video to multiple users, a constant stream rate, nor easily anticipate future requests.

An important aspect of the planned work is to develop models of user interaction patterns. Our initial hypothesis is that users will formulate and issue a query to the Digital Video Library System. The response is likely to be text and images describing the retrieved video segments. We then expect the user to retrieve each video segment in rapid succession. We expect to transmit short segments of full resolution video to 5 to 10 users simultaneously, and "low density" query information to an additional 20-30 users simultaneously.

One of the specific requirements is to support a variety of user classes connected in different ways to the DVLS. In particular, we propose to handle: (1) "local" users who are connected to the DVLS by our 155 Mb/s ATM network, (2) "nearby" users who are connected over radio links at 1.5 Mb/s, and (3) "distant" users who are connected to the DVLS via the Internet or 14.4 Kb/s modems. To support this mix of users, video compression rates and communication protocols must vary accordingly.

For local users, no additional video compression is necessary for communication. Full size "original" quality can be transmitted over the ATM network. Nearby users will have video images which are reduced in size to 320 x 240 pixels and MPEG compressed at roughly 23:1 rate. We are anticipating very little loss in subjective video quality. For distant users, we will need to reduce the image size again to 160 x 120, reduce the frame rate to 10 images per second, and apply 30:1 compression using MPEG. The user interface must be adjusted accordingly for "postage stamp" views of video, and non-interactive transfer of larger size videos to local video servers.

3.5 Evaluation

The Digital Video Library System will be evaluated on three fronts: (1) the effectiveness of the chosen system architecture; (2) the quality of the audio-based indexing and video-based segmentation; and (3) the usefulness of a video library for educational purposes. This activity will be an ongoing activity involving the developers of the DVLS and our educational partners in local and remote schools.

System Architecture Early in the project, we must validate that our architecture stores and delivers video efficiently and effectively. We will measure the tradeoffs between speed of delivery, target disk space requirements and quality of the video delivered. We will deliver the video at different levels of quality and collect data on which configurations provide acceptable video access and which do not. We must evaluate different compression choices and their effect on the quality of video stored, storage requirements, and the quality and speed of video we can deliver. We must compare the effects of different storage media for the videos in terms of price and performance. We must also collect usage statistics to examine what the mix of use is from the different classes of users and how well the DVLS handles the demands for diverse quality video access.

Automatic Indexing Searching strategies based on full-text indexing are effective with large libraries of written documents [Salton, 1986]. Development of automated mechanisms for indexing video on the bases of the audio track will certainly improve the ability to locate relevant clips within vast video libraries. However, the producers of video use elements of communications that are different from those used by authors of written documents. It may be that the audio track does not contain enough information to adequately index the video segments. However, even if the audio-based indexing is outperformed by the manual indexing, it will be important to demonstrate acceptable retrieval based on the audio. Because manual indexing is so labor intensive and costly, for many video collections the choice will be automatic indexing or no indexing at all.

Educational Merit Summative evaluation related to the educational goals will investigate factors that influence teacher and student interactions with the DVLS. Naturalistic inquiry will be used to gauge the teacher's impressions concerning such factors as: (1) adequacy of delivery speed across the different bandwidth capabilities, (2) educational merit across video types, (3) ease-of-use, (4) ability to find relevant video segments, and (5) influence of the DVLS on teaching strategies.

References

Broadcasting and Cable, *AP wire service article*, vol 123 n 39, September 27, 1993.

Brunei, D. J., B. T. Cross, et al. (1993). *What if there were desktop access to the computer science literature*? 21st Annual ACM Comp. Sci. Conf., Indianapolis, IN, ACM Press.

Burks, C., M. Cassidy, et al. (1991). GenBank. Nucleic Acid Res.

Business Week, *Fox and Sony article*, July 5, 1993, pg 98.

Business Week, *Video marketing article*, September 6, 1993, pg 78.

CD-ROM Professional, *Video marketing article*, vol 6 n 6 , November 1993, p 102.

Computer World, *ECnet article*, vol 27 n 49, December 6, 1993, p 35.

Fox, E., *Advances in Digital Multimedia Systems* , IEEE Computer, Vol. 24, No. 10, October 1991.

Fox, E. A., D. Hix, et al. (1993). *Users, User Interfaces, and Objects: Envision, a Digital Library*. JASIS 44(8): 474-479.

Garrett, J. R. and P. A. Lyons (1993). *Toward an Electronic Copyright Management System*. JASIS 44(8): 468-473.

Haralick and Shapiro, *Computer and Robot Vision*, Addison Wesley, 1992.

Hoffman, M. M., L. O'Gorman, et al. (1993). *The RightPages Service*. JASIS 44(8): 446-452.

James, D.A. and Young, S.J., *Wordspotting*, Proc. ICASSP, 1994, Adelaide.

Jeffay, Stone and Smith. *On kernel support for real-time multimedia applications* , Proc. of 3rd IEEE Workshop on Workstation Operating Systems, April 1992.

Kahle, B., H. Morris, et al. (1993). *Interfaces for Distributed Systems of Information Servers*. JASIS 44(8): 453-467.

Lesk, M. (1991). *The CORE Electronic Chemistry Library*. Proc. 14th Ann. Inter'l ACMSIGIR Conf. on R&D in Information Retrieval, Chicago, IL, ACM Press.

Nicolaou, C., *An architecture for real-time multimedia communication systems* , IEEE Journal on Selected Areas in Communications, Vol. 8, No. 3, April 1990.

Pissinou, N., K. Makki, et al. (1993). *Towards the Design and Development of a New Architecture for Geographic Information Systems*. CIKM-93, Washington, DC, ACM Press.

Rangan, P.V., and Vin, H.M., *Efficient storage techniques for digital continuous multimedia* , IEEE Trans. on Knowledge and Data Engineering: Special Issue on Multimedia Information Systems, August 1993.

Rawlins, G. J. E. (1993). *Publishing over the Next Decade*. JASIS 44(8): 474-479.

Salton, G. (1986). *Another Look at Automatic Text-Retrieval Systems*. Communications of the ACM 29(7), 648-656.

Takebayashi, Y., H. Tsuboi, H. Kanazawa, (1991). *A robust speech recognition system using word-spotting with noise immunity learning*, Proceedings of ICASSP 91, Toronto, 1991, 905-908.

Zorpette, G, *The latest box office draw: open post production* , IEEE Spectrum, vol 30 n 10, October 1993, p 14.