To appear in Information Processing & Management.

The VISION Digital Video Library[1]

Susan Gauch, Wei Li, and John Gauch
Telecommunications and Information Systems Laboratory (TISL)
Department of Electrical Engineering and Computer Science
The University of Kansas

ABSTRACT

The goal of the VISION (Video Indexing for SearchIng Over Networks) project is to demonstrate the technology necessary for a comprehensive, on-line digital video library. We have developed automatic mechanisms to populate the library and provide content-based search and retrieval of video over computer networks. The salient feature of our approach is the integrated application of mature image or video processing, information retrieval, speech feature extraction and word-spotting technologies for efficient creation and exploration of the library materials. First, full-motion video is captured in real time with flexible qualities to meet the requirements of library patrons connected via a wide range of network bandwidths. Then, the videos are automatically segmented into a number of logically meaningful video clips by our novel two-step algorithm based on video and audio contents. A closed caption decoder has also been incorporated into the system to extract textual information to index the video clips by their contents. Finally, all information is stored in a full-text information retrieval system for content-based exploration of the library over networks of varying bandwidths.

[1]This paper is a revised version of "VISION: A Digital Video Library System," W. Li, S. Gauch, J. Gauch and K.M. Pua, presented at ACM Digital Libraries '96.

1. INTRODUCTION

As a result of the rapid development of multimedia computing technologies and high-speed network systems, vast amounts of multimedia information are becoming prevalent on the emerging national Information Superhighway. Organizing such a tremendous amount of data to provide intelligent access and effective use is a major topic of digital library research. Our previous experience with the UNITE project, which provides multimedia curriculum materials over computer networks (Gauch, 1994), shows digital libraries can have a profound impact on the conduct of education and training activities. According to education experts, learners should not only be involved in receiving information, they must actively participate in independent, self-motivated access to information.

The availability of audio/video recording and playback equipment at a relatively low cost has made it possible to incorporate video into constructive education strategies (Reinhardt, 1995). As a result, many schools have acquired extensive libraries of pre-recorded video. One problem which limits the use of this material is the lack of effective video indexing methods. Users of video archives often rely on title and keyword information to identify videos of interest. Valuable time is then wasted manually scanning the video to locate the portions which are most educational. Improved techniques for finding video segments of interest would therefore be quite beneficial. Digital video libraries capable of providing independent and self-directed ubiquitous access to information from K-12 schools, colleges, and government agencies will bring about a revolution in education and training (Fox 1995).

The VISION (Video Indexing for SearchIng Over Networks) digital video library prototype was developed at the Telecommunications and Information Sciences Laboratory of the University of Kansas as a testbed for evaluating automatic and comprehensive mechanisms for library creation and content-based search and retrieval of video across networks with a wide range of bandwidths (Gauch, et al., 1994). Our pilot system was populated with a collection of nature, science, and news videos from WGBH and CNN. These videos were automatically partitioned into short segments based on their content, and stored in a multimedia database. A client-server based graphical user interface was developed to enable users to remotely search this library and view selected video segments over networks of different bandwidths.

To implement our prototype, we had to address a number of challenging questions. Our approach was to integrate mature technologies and methods in image and video processing, information retrieval, and speech recognition in the VISION digital video library. First, full-motion videos are captured in real time with consideration of the video quality versus network bandwidth tradeoff. Then, they are automatically segmented into a number of logically meaningful video clips by our two-step algorithm, based on first the video and then the audio contents. A closed-caption decoder was then used to extract textual information to automatically

index the related video clips by their contents. The indexing information is stored in a full-text information retrieval system from which clients of the library can search for quick retrieval and browsing of multimedia objects.

The remainder of this paper is organized as follows: Section 2 discusses related work; Section 3 gives an overall architecture of the VISION; Section 4 details the procedures to populate the video library; Section 5 discusses our digital video library system from the point of view of patrons, presenting our multimodal information retrieval over the networks; Section 6 outlines our future plans.

2. RELATED WORK

Digital video libraries distinguish themselves from traditional "video-on-demand" services or other similar projects in that they integrate image and video processing and understanding, speech recognition, distributed data systems, networks, and human-computer interactions in a comprehensive system. A key component of this difference is the use of content-based indexing and retrieval algorithms to enable users to interact with the video library rather than simply playing back entire movies or broadcasts. As a consequence, there has been considerable activity developing improved tools for video processing and content analysis. There has also been important progress made developing real-time multimedia systems capable of displaying video to users on different platforms and over a variety of networks. Systems which share features and goals with the VISION system are described below.

Several approaches have been proposed to decompose raw video into *shots* (a continuous roll of a camera) and *scenes* (collections of shots which occur in a single location or are temporally unified). The problem of identifying *cuts* (sharp transitions between shots) has been typically approached from a bottom up perspective, looking for rapid changes in color histogram or image intensity (Arman, 1993; Nagasaka, 1992; Zhang, 1993). Model-based algorithms have also been developed to successfully detect fades, dissolves, and page translate edits (Hampapur, 1994). Once shots have been identified, keys frames which characterize the shot can be selected by considering the motion of objects within the shot. Here, we can either select frames which are as still as possible (Wolf, 1996) or identify the background and moving objects explicitly and select an image which focuses on one or the other (Sawheney, 1996). Another related approach is to combine information from multiple frames of an image sequence to create a "salient video still" which characterizes the shot in some way (Teodosio, 1993). These methods vary considerably in their computational complexity and effectiveness for different video sources, but each has its merits.

Although the problem of shot detection is essentially solved, the problem of combining shots to obtain scenes presents significant challenges. One approach used by the Princeton Deployable Video Library (PDVL) (Wolf, 1995) is to use identify key frames in each shot and use image-based clustering to construct a scene transition

graph to visually present the relationships among shots. By browsing through a collection of graphs users can locate scenes of interest (eg. two person interviews). The scene transition graph can then be used to navigate through the video. The Algebraic Video System (Weiss, 1995) uses an alternative technique where shots are organized in a hierarchical structure which allows nested stratification (subtrees may refer to overlapping portions of the raw video). This system uses the VuSystem (Lindblad, 1994) for recording and processing video but hierarchy construction is currently performed manually. A model-based approach has been proposed to parse video by an *a priori* model of the video structure (Zhang, 1995). Such a model represents a strong spatial order within the individual frames of shots and/or strong temporal order across a sequence of shots. For example, it is required that all shots of the news anchorperson conform to a spatial layout. For many tasks it will be difficult or impossible to define models for the video. In their system the text description of the video contents are input by an operator. This yields high accuracy but makes production of large video collections very expensive.

Automatically identifying the content of a video segment is a particularly challenging problem. Three basic approaches have been investigated for this purpose: image understanding, speech recognition, and caption processing. Although the human visual system is very effective, research in computer vision over the past 20 years has had success in only limited domains (Haralick and Shapiro, 1992). For this reason, many approaches for image-based content identification have focused on feature-based classification schemes. For example, images can be indexed using color histograms (Swain, 1991) or combinations of shape and color features (Smoliar, 1994). The QBIC (Query By Image Content) project (Faloutsos, 1994) investigated methods to query large on-line image databases using the image contents, such as color, texture, shape, and size. Although feature-based classification is quite fast, one drawback is that very different objects may have the same features (eg. a red car and a red apple).

Moving away from pure feature-based matching, it has been shown that similarity-based image retrieval can be also accomplished using Hidden Markov Models (HMM) which have been trained with representative images of outdoor scenes (rivers, trees, mountains) (Yu, 1995). Multiresolution wavelet decompositions have also been used for rapid image matching and retrieval (Jacobs, 1995). Here, a low resolution example (either hand drawn or scanned) is used as a query and multiscale matching is used to locate the most similar image in the database. More ambitious indexing based on texture, shape and appearance have been investigated within the Photobook system (Picard, 1994; Pentland, 1996). Although this system has had excellent success within a restricted domain of images (textures and faces) the computational expense associated with computing Eigenimages may limit its use for identifying video content.

Given the difficulty of image-based content analysis, processing the audio track and closed caption information is an attractive alternative. The goal of the Informedia Digital Video Library project is to establish a large, on-line library featuring full

content and knowledge-based search and retrieval of digital video (Christel, 1994; Christel, 1995; Wactlar, 1996). To automatically transcribe narratives and dialogues into text files, Informedia uses the SPHINX-II speech recognition system, which is a large-vocabulary, speaker-independent, continuous speech recognizer developed at CMU. Closed caption transcripts are also recorded whenever it is available. To process queries, Informedia will employ natural language processing technologies to extract users' subject or content of interest without forcing them to specialized syntax or complicated command forms. Users can input their queries into Informedia by either keyboards or microphones. The extensive use of AI techniques, particularly natural language processing and speech recognition in the presence of noise and background music, makes this an ambitious, high-risk project.

The VISION system (Gauch, 1994) shares many of the goals of the Informedia project, but much more limited processing of the audio track is used to perform automated scene segmentation and content analysis. In particular, audio information is used to combine shots and for limited word spotting. Closed captions are recorded in the library when available and used in a full-text retrieval engine to search the video library for material of interest.

3. ARCHITECTURE OF VISION

VISION is constructed in a client-server architecture in which each subsystem provides a major function to other subsystems through network protocols. It is logically divided into the following three subsystems:

- *Video Processing System (VPS)* which captures video/audio and closed-captions if available, and produces compressed and segmented video clips.

- *Library Server* which builds search indices on video clips based on transcripts or keywords, and provides query services to library clients.

- *Library Client Browser*, the graphical user interface of the system, which transmits queries to the library server and plays back video/audio.

The architecture of VISION is summarized in Figure 1. The heart of the library is an object-oriented database management system containing segmented video/audio clips and transcripts or keywords files. The DBMS defines attributes and relations among these entities in an object-oriented approach. Due to the data abstraction and information encapsulation inherent in the object-oriented approach, VISION is designed to allow for fast incorporation of emerging standards for video compression, evolving high-speed communications services, and other new technologies and products.

4. VIDEO PROCESSING SYSTEM (VPS)

The VPS has the following three main functions: audio/video capture and

compression; video segmentation; and content-based indexing.

4.1 Audio/Video Capture and Compression

The audio/video digitization subsystem in VISION employs a DEC Sound and Motion J300 board installed in the TURBOchannel of a DEC 3000 Alpha AXP workstation. Video can be captured in real-time (NTSC 640x480 pixels, PAL or SECAM 768x576 pixels in either composite or S-video format, JPEG compressed or 4:2:2 YUV format); audio can be digitized with sample rates 8-bit (8 Khz) to 16-bit (up to 48 Khz). The system is currently being ported to Windows '95 environments.

Compression techniques clearly play a crucial role in digital video library systems. Audio, image, and video signals produce a vast amount of data. For example, a single frame of color video, with 640x480 pixel frames at 24 bits per pixel, would take up about 0.92 Mbyte. At a real-time rate of 30 frames per second, that equals 27.7 Mbytes for one second of video. A modest digital video library might contain 100 hours and would require about 10,000 Gbytes of storage for video! Even if we had enough storage available, we wouldn't be able to play back the video in real time due to the insufficient bit rate of the storage device. At the present state of technology, to make effective use of video storage facilities, the only solution is to compress the video data before storage and decompress it during playback. Compressing the video at by factor of 150:1 makes the total storage requirement for a 100 hour digital video library to be about 70 Gbytes, which is much more feasible.

There are currently two major compression algorithms for motion videos: JPEG and MPEG. JPEG (Joint Photographic Experts Group) is based on the discrete cosine transform (DCT). It compresses single frames by discarding and quantizing DCT coefficients for each 8x8 (or 16x16) block of pixels in an image. JPEG was designed for still images and can achieve compression rates between 2:1 and 200:1. It does not capture the inter-frame coherence of video sequences, resulting in a relatively low compression rate for motion videos. In contrast, MPEG (Motion Picture Experts Group) compresses video by encoding the first frame of a video sequence and using motion estimation to predict subsequent frames. The differences between the predictions and the original images are then encoded using block-based DCT coding. By utilizing the inter-frame coherence in video sequences, MPEG achieves higher compression rates than JPEG at a fixed level of image quality.

Although MPEG is more space efficient than JPEG, there are two disadvantages of MPEG compression, from the point view of VISION design and implementation. First, JPEG is better suited for random access of individual video frames than MPEG. Random access of JPEG frames is trivial since there no inter-frame information is used. Conversely, randomly selected individual video frames cannot be retrieved and decompressed from an MPEG stream without surrounding I, P, and B frames. A second more serious problem occurs if data is lost or delayed during transmission over networks. If selected JPEG frames are lost, the client and server can easily resynchronize video transmission and playback. Adjusting the playback rate for

MPEG in this situation is much more complex. For these pragmatic reasons, VISION uses JPEG compression.

The VISION system is designed to support three classes of users with dramatically different network bandwidth between the client and server. For K-12 schools with low bandwidth access (28.8 Kb/s), very small images (160x120 pixels) and a reduced frame rate are necessary when delivering video. For users connected via the Internet with 1 Mb/s bandwidth, an intermediate quality image (320x240 pixels) can be transmitted at nearly full frame rates. Users with access to our ATM testbed have access to our highest quality video (640x480 pixels, 30 frames/sec).

To accommodate the requirements of the users at the different levels, the VPS supports flexible selection of the digitization parameters (scaling of the video signal before compression and selecting the video compression ratio) and selection of display parameters (scaling the video after decompression). We conducted a series of informal experiments varying image sizes between 640x480, 320x240, and 160x120 pixels, and JPEG compression ratios from 4:1, 16:1, 50:1, to 150:1. The following tables summarize the capture size, compression rate, playback size, and bandwidth requirements we feel are suitable for the three classes of users VISION supports.

Capture Size	Compression Ratio	Playback Size	Bandwidth (KB/s)
640x480	4	640x480	2304
640x480	16	640x480	576
640x480	50	640x480	184

Table 1: High Quality Video

Capture Size	Compression Ratio	Playback Size	Bandwidth (KB/s)
320x240	16	320x240	144
320x240	50	320x240	46
160x120	16	320x240	36

Table 2: Intermediate Quality Video

Capture Size	Compression Ratio	Playback Size	Bandwidth (KB/S)
320x240	150	320x240	15.3
160x120	50	320x240	11.5
160x120	150	320x240	3.8

Table 3: Low Quality Video

When populating the video library with a video source, the librarian should first decide which potential group of users would access the video most. For example, if the video tape is to teach pupils elementary mathematics, users from K-12 schools will probably patronize it more than any others. Another example is CNN Headline Sports, which could be accessed by student fans at universities more than by other communities. Based on the possible user groups of a video tape, we can choose the

digitization parameters. For the video tapes on elementary mathematics, scaling the original size to 160x120 or even smaller and compressing the video with greatest possible compression ratio is necessary to fit the low-bandwidth networks at most K-12 schools. For video tapes for a general audience, we can digitize them with an intermediate compression ratio, at 320x240 or 160x120 if the disk space and transmission bandwidth are limited. For the archival video tapes such as surgical operations video for educational purposes which do not allow image distortion and loss, we should digitize as clearly as possible. For instance, we shall use a capture size of 640x480 and less compression. Video digitization parameter selection is a tradeoff between storage space, transmission time, and the quality of video in the digital library.

Our three quality levels strategy does not prevent the users connected via low-bandwidth networks from accessing the high quality video stored in the library. When such a user wants to access the better quality video, our current system will rescale and recompress the video clips to decrease the transmission time and bandwidth requirements. In our pilot system, there was little contention for the JPEG compression board in our video server, so this approach delivered acceptable performance. Given the rapidly decreasing cost of disk storage, a more effective alternative would be to store multiple copies of the video at different compression rates and transmit the appropriate video stream for each class of user.

4.2 Video Segmentation

The data volume in a digital video library is dominated by its digital video component. The effective management of this spatial, temporal, and unstructured component is essential to the development of successful digital video libraries. From a logical point of view, production video footage can be thought of as a collection of scenes which illustrate different subtopics. Typically each of these scenes is comprised of one or more camera shots which have been spliced together in some manner. The two goals of video segmentation are: 1) to locate the start and end of each camera shot, and 2) to combine camera shots to identify the starting point and endpoint of each scene in the video. Once scenes are located, a representative frame can be extracted and the group of frames comprising the scene can be treated as a unit for indexing and searching. Thus, the library will be capable of providing the search and retrieval of randomly accessed video clips appropriate to a user's needs and desires.

Several approaches to the problem of automatic location of camera motion breaks in video sequences have been investigated. Nagasaka and Tanaka (Nagasaka, 1992) have evaluated a number of image processing measures for detecting cut edits in video sequences by detecting shot boundaries in digital video. Their conclusion is that the best measurement is the sub-window-based histogram comparison. Zhang et al (Zhang, 1993) have also presented the evaluation of different image processing routines for detection of cut edits. They tried to detect special effects having gradual transitions like fades and dissolves by using a dual threshold. Hampapur et al

(Hampapur, 1994) approached the problem of digital video segmentation by proposing a model for video based on the production process and classifying video edit effects based on these models. The edit effect models are used to design feature detectors, which are used in a feature-based classification approach to segment the video. Arman, Hsu, and Chui (Arman, 1993) presented a technique operating directly on compressed video detect shot boundaries. Their technique relies on the properties of the coefficients of the discrete cosine transform used in encoding the video to detect the transitions.

All of these solutions address the first problem, locating the start and end of each camera shot. In order to address the second more difficult problem of combining shots to obtain scenes, we have chosen to look beyond image information and consider the associated audio, which is a very rich source of information. For example, a change in speaker may indicate different topics or different aspects of a topic are being discussed. Conversely, when a single speaker is active over multiple camera shots, there is evidence that the shots should be considered part of one scene. Therefore, integration of effective analysis of audio signals with results obtained from image analysis should reduce the misdetection rate and enhance the robustness of the traditional video segmentation algorithms.

In VISION, we developed a novel algorithm that combines audio analysis with traditional image-based segmentation methods. It performs the segmentation task in two steps: initial segmentation based on scene changes in video; and merger based on features in the audio track. Work is also underway using the contents of the closed-captions as a third source of scene change information.

Video-based Segmentation: In the first step of video segmentation procedure, we employ two basic measurements: pixel-by-pixel or color histogram differences between successive frames in a video sequence.

Let $I_n(t)$ be the intensity of pixel n at time t, $\Delta I_n(t)$ be pixel's absolute difference defined as follows:

$$\Delta I_n(t) = \left| I_n(t+\Delta t) - I_n(t) \right|$$

Then the absolute sum of the inter-frame difference will be:

$$D(t) = \frac{1}{N} \sum_{n=0}^{N} \Delta I_n(t)$$

where N is the number of pixels in a frame.

Large values of $\Delta I_n(t)$ represent a cut. Intermediate values are yielded by object motion and camera operations. Unfortunately, the principal problem with this metric is the fact that it does not reflect the distribution of differences between the

two frames. For example, it is difficult to distinguish between a large change in a small area and a small change in a large area. It tends to falsely detect scene changes when a small part of a frame undergoes a large, rapid change.

With this method, an 8 minute clip of CNN Headline News which contains 10 news items was segmented into 42 camera shots. Several shots were incorrectly split into multiple pieces because there are many moving objects in a frame combined with camera movements (such as tilt and zoom) across frames. To avoid this problem, we make use of the absolute difference of the color histogram between successive frames, which is less sensitive to object motion and camera movements. The absolute difference of the color histogram between two successive frames, ΔH , is defined as follows:

$$\Delta H = \sum_{b=0}^{B} \left| h(t+\Delta t, b) - h(t,b) \right|$$

where B is the total number of colors in a frame, and $h(t,b)$ is the height of the color b's bin at time t. Again, values of ΔH above a threshold indicate a camera cut. By using this algorithm on the same CNN footage, we correctly identified the 31 camera shots which comprised the 10 news items we recorded. In order to obtain the full scenes corresponding to each news item, these camera shots must be combined.

Audio-based Merger: A video sequence for one logical scene may contain many edit cuts. In order to produce meaningful segments, the results of the video segmentation should be post-processed to merge some contiguous segments back together. We do this by analyzing audio features extracted from speech signals such as endpoints detection and speaker identification. For example, by using a speaker identification algorithm, the segmentation procedure would identify a talk given by a lecturer in a conference as a complete video clip instead of cutting it into several clips based purely on scene changes. In TV news, it is common for editors to use several clips taken from different view points, places, or even times to compose one news item while letting the anchorperson narrate the story. A video-based segmentation algorithm would ruin the editor's product by decomposing the video into its components, which are likely to be difficult to understand without suitable context. This is a fundamental problem with CNN Headline News video that any video-based segmentation algorithm alone cannot address.

We are currently using endpoints algorithms to detect changes in both time and frequency domain representations as supplemental sources of criteria for segment boundaries. Endpoints detection is one of the most basic aspects of speech signal processing. Unlike the hardware neural network solution proposed by Newman (Newman, 1990), our endpoints detection algorithm is entirely implemented in software. We base the measurement on the audio signal short-time energy and zero-crossing rate, and attempt to detect the changes that these quantities undergo at the beginning and end of an utterance. The short-time energy function of speech

may be computed by splitting the speech signal into N samples and computing the total squared value of the signal in each sample. Splitting the signal into samples can be achieved by multiplying the signal by a suitable window $W[n], n = 0,1,2,...,N-1$, which is zero for n outside the range $(0, N-1)$. A simple rectangular window of duration 10-20 ms is suitable for this purpose. For a window starting at sample m, the short-time energy function E_m may be formulated as:

$$E_m = \sum_m \{x[n] \cdot W[n-m]\}^2$$

The energy of speech is generally greater than that of silence or background noise. A speech threshold can be determined that takes into account the silence energy and the peak energy. Initially, the endpoints are assumed to occur where the signal energy crosses the threshold. Corrections to these initial estimates are then made by computing the zero-crossing-rate in the vicinity of the endpoints. Zero-crossing-rate is a measure of the number of times in a given time interval (frame) that the speech signal amplitude passes through a value of zero. Normally, the rate for silence is greater than that of voiced speech. If detectable changes in zero-crossing-rate occur outside the initial points, the endpoints will be changed to the points at which the changes take place.

Our initial experiments show that augmenting our video segmentation algorithm with the endpoints detection procedure produces more meaningful clips for later indexing. For example, from the same CNN Headline News footage we got 17 scenes for 10 news items after merging the 31 camera shots generated by the color histogram difference metric; see Figure 2. Obviously, our two-step segmentation algorithm does not work perfectly, but it does improve the segmentation, which is encouraging. Most contiguous clips separated by a special edit such as fade in / fade out are merged back into one clip if the anchorperson narrates the dialogue across the boundary between the two clips. Our procedure fails when there is no sound across the boundary between two successive clips because we cannot detect a change on the short-time energy of the audio signal.

We are trying to remedy this problem in two ways. On the video side, we will incorporate motion analysis techniques to detect gradual transitions. For instance, to detect the chromatic edit such as fade and dissolve, which is usually achieved by manipulating the color or intensity space of the two shots being edited, one must discriminate between intensity changes in the video due to scene activity as opposed to intensity changes in the video due to chromatic editing. The key difference between them is that the changes due to editing are more uniform than naturally occurring changes. On the audio side, we will try other audio analysis techniques. For example, we could calculate the similarity of the speech signal between two contiguous clips, where a high value indicates two clips are closely related and should be merged. We could use another useful source of information, the speakers themselves. Speaker change generally implies content change. Although considerable research has addressed the speaker verification / identification problem,

most results are inadequate in the presence of other speakers or noises. Thus, further study is necessary.

4.3 Closed Caption Decoding and Content-based Video Indexing

Well indexed video is crucial to the success of the digital video library. Index information usually consists of the document identifier, a description of its content, and a number of keywords. Manually assigning keywords is very labor intensive, particularly when shots and scenes must be indexed individually. Furthermore, an individual cataloguer's perspective may vary over time and is likely to be different from the eventual users. Fortunately, many broadcast television programs are closed-captioned to provide a transcription of their audio content for the hearing-impaired. This transcript of the dialog in the scene is often a rich source of information for indexing purposes. For example, if the speaker is describing the eating habits of spider monkeys, the words "monkey" and "diet" are likely to occur in the transcript and would be effective keywords for indexing the video. Given the known limitations of automatic speech recognition (large vocabulary, background music, and other noise) the captions provided by the video producer will provide a more reliable transcript in most situations (Roe, 1994). For this reason, the VISION system uses closed caption text as the primary source of information to describe the content of individual video clips.

In the NTSC television system, caption data is transmitted in scan line 21 of field 1 of the video frames. This scan line does not normally appear on a television screen because it is part of the vertical blanking interval (VBI). Unfortunately, most video capture boards available on the market, such as the J300 we are using, digitize only the active region of the signal, ignoring the vertical blanking interval. One approach would be to incorporate an over-the-counter decoder chip into the video processing subsystem. For example, the TNS group at MIT uses a specially designed digitization board, Vidboard, to decode the closed caption (Lindblad, 1994; Bacher, 1995). However, the Vidboard does not capture full-color video at the full-motion rate (30 frames per second for NTSC). To make use of our J300 board and simultaneously digitize video and decode captions, we split our video source and send one stream to a stand-alone caption decoder, called TextGrabber, developed by Unitec Inc. It works with any closed-caption program, supporting three additional caption and language channels, four independent text channels, and the new Extended Data Service (XDS) channel recently established by the FCC. Figure 3 shows that VISION decodes and displays the closed captions while digitizing the video.

When closed-captions do not exist, the audio content is the best remaining alternative for indexing purposes. Rather than attempting speaker independent continuous speech recognition, we have utilized a more limited form of speech recognition called *word spotting* (Vroomen, 1990). The goal of word spotting is to detect the occurrence of selected words in the audio source. For example, if we wish to index sports news, keywords such as "football", "baseball", "hockey", and "golf"

might be appropriate; whereas for political coverage, the names of candidates would be more appropriate. To implement word spotting, we trained the Entropic Speech Recognition System with four speakers on a test suite of 200 sentences (approximately one hour each). A small set of keywords of interest was selected based on video footage in our library, and word spotting was performed. Manually generated transcripts were then used to quantify the recall and precision rates for different sets of keywords. Although our initial recall rates are promising (~50%), our precision is quite low (~20%). We hope to improve the accuracy of word spotting by more extensive training of the Entropic system and preprocessing to remove audio noise where possible.

5. VISION CLIENT BROWSER

Due to the video segmentation and content extraction discussed above, VISION can provide efficient content-based retrieval functionality to library patrons. A library is only as useful as the retrieval facilities it provides. For a digital video library, it must empower clients with a set of efficient and effective exploration tools based on both text and visual interfaces. At present, VISION provides a Boolean full-text query engine and an easy to use interface for viewing video clips.

5.1 Forming a Query

A Boolean full-text query provides a basic and intuitive access into the library for the library users. Users specify subjects of interest using keywords, which may have optional weights. Queries are sent to the library server, which calls the Illustra text blade to identify video clips from the library collection. The library server then sends summary information back to the user. Figure 4 shows an example query: *(Internet OR Networking) AND (Stocks OR Investment)*. VISION also provides a more advanced text-based query access for experienced patrons. Users can formulate a query by specifying other attributes such as date, source, length, author, and title in addition to keywords. As the library grows, the clips will be assigned to categories which will allow the user to browse clips by subject matter and/or direct their queries to a subset of the database.

5.2 Viewing Results

For query results, we employ thumbnails, mouse sensitive pixelmap icons as shown in Figure 5, which correspond to the key frames for each video clip. As opposed to icons, thumbnails present many of the visual clues of the underlying video clips and can support rapid visual recognition and may directly provide the information sought. At present, only one thumbnail is used per clip, which may not be ideal if the clip consists of several distinct shots. Below each thumbnail is an information button which provides more information about the video clip: its indexing keywords; its caption transcripts if available; its source; its length and other relevant information. The result of any query will be a set of thumbnails listed in sorted order by the quality of the match against the query. All hits are represented by

thumbnail sketches, which are presented in a scrolling window below the query window.

When the user double-clicks a thumbnail, the library server starts sending the underlying video clip to the client machine for playback. Users are provided with a VCR-like interface to control the play of the selected video clip; see Figure 5. Two distinguishing features of this user interface are: fast forward and fast rewind buttons which allows viewers to scan the clip four times faster than real time, and a time slider bar which allows views to manually move to any position and resume playback.

Depending on the configuration of the client machine, the server will send compressed or uncompressed video with or without audio. For example, for clients with a JPEG-decompression board and sound capability, the server will send compressed video with audio; for clients with a sound board but no JPEG-decompression board, the server will send decompressed video with audio. Our experiments show that on DEC Alpha workstations with an average CPU load, we achieve approximately the real-time playback rate of 30 frames per second. This has been realized using 320x240 pixel per frame transferred from a remote server machine linked to client machines via Ethernet in the same building. Our next experiments include transferring video/audio files across campus via ATM switches.

5.3 Supporting Multiple Classes of Users

VISION employs a client-server communications architecture. It is capable of supporting simultaneous access to the library over the Internet at a variety of transmission capacities and with a wide range of client machines. VISION supports three levels of service dependent on the bandwidth available on the path between the server and its clients. The bandwidth demands for VISION are highly asymmetric; queries to the library server require relatively low bandwidth, while delivering video frames may require significantly greater capacity.

Access to VISION will be provided for K-12 schools via the Internet. Because of the low bandwidth available between the schools and the Internet, a high degree of compression is necessary, resulting in minimum service quality. When VISION receives a query from these schools, the server will perform the search and retrieval functions, compress video clips as needed, and transmit the video/audio data to the schools via the network. The low bandwidth prevents students at the schools from playing back videos in real time without storing the video clips locally.

An intermediate level of service is supported over the existing Internet. It includes video retrieval service in the 1 Mb/s to 2 Mb/s per stream range. Our current experiments are done in this level.

The highest quality access will be provided using high-bandwidth ATM/SONET capabilities. This level of service will be limited to sites with these capabilities, which include the University of Kansas campus and the MAGIC Gigabit network testbed sites. This part of the network, with portions operating at 155 Mb/s will allow the full capabilities of our digital video library to be demonstrated. As more sites become connected to MAGIC, direct high-bandwidth access from schools may become possible.

6. CONCLUSIONS AND FUTURE WORK

This paper describes the design and implementation of the VISION digital video library. We have developed automatic mechanisms to populate the library and provide content-based search and retrieval of video over computer networks. The salient feature of our approach is the integrated application of mature image or video processing, information retrieval, speech feature extraction and word-spotting technologies for efficient creation and exploration of the library materials. First, full-motion video is captured in real time with flexible qualities to meet the requirements of library patrons connected via a wide range of network bandwidths. Then, the videos are automatically segmented into a number of logically meaningful video clips by our novel two-step algorithm based on video and audio contents. A closed caption decoder has also been incorporated into the system to extract textual information to index the video clips by their contents. Finally, all information is stored in a full-text information retrieval system for content-based exploration of the library over networks of varying bandwidths.

Video libraries present a number of challenges which remain to be addressed. From a video capture standpoint, there are a number of hardware/software products available to perform real-time video compression/decompression (eg. JPEG, MPEG). Rather than selecting the "best" method for all video libraries, systems which support multiple formats need to be investigated. Video segmentation and content analysis are two problems which are at the heart of library creation. Both areas have room for improvement. Our current work in this area is focusing on getting more information from the audio signal and closed captions. Using an online thesaurus to automatically expand user queries may improve the search performance of our system. Related research in the text domain (Gauch and Smith 1991; Gauch and Futrelle, 1994; Gauch and Chong, 1995) is currently being extended to apply to video indexing. Ongoing work on video servers by a number of groups have made significant progress on the systems side of making digital video libraries available. The problem of automatically adjusting the level of service depending on network capacity is one area we are currently investigating.

ACKNOWLEDGMENTS

This work was supported in part by the University of Kansas Research Development Fund, the Telecommunication and Information Sciences Laboratory (TISL), the Center for Excellence in Computer-Aided Systems Engineering (CECASE), and The

National Science Foundation Award CDA-9401021. Video sequences from CNN Headline News are courtesy of Turner Broadcasting.

REFERENCES

Arman, F., Hsu, A., et al, (1993). Image Processing on Compressed Data for Large Video Databases, *ACM Multimedia '93*, California, USA, 267-272.

Bacher, D.R., Lindblad, C.J. (1995). Content-based Indexing of Captioned Video on the ViewStation, *Technical Report*, MIT.

Christel, M., et al, (1994). Informedia Digital Video Library, *ACM Multimedia '94*,, 480-481.

Christel, M., et al, (1995). Informedia Digital Video Library, *Communications of ACM*, Vol. 38, No. 4, 57-58.

Faloutsos, C., et al., (1994). Efficient and Effective Querying by Image Content, *Journal of Intelligent Information Systems*, Vol. 3, 231-262.

Fox, E.A., et al, (1995). Introduction: Special Issue on Digital Libraries, *Communications of ACM*, Vol. 38, No. 4, 22-28.

Gauch, S., and Smith, J.B., (1991). Search Improvement via Automatic Query Reformulation, *ACM Transactions on Information Systems*, Vol. 9, No. 3, 249-280.

Gauch, S., et al, (1994). The Digital Video Library System: Vision and Design, *Digital Libraries '94* , College Station, Texas, 47-52.

Gauch, S., (1994). Network-based Multimedia Information Services, *NIST Virtual Library Consortium Working Group Report*,.

Gauch, S. and Chong, M.K., (1995). Automatic Word Similarity Detection for TREC-4 Query Expansion, *Proc. of TREC-4: The 4th Annual Text Retrieval Conf.*, Gaithersburg, MD.

Gauch, S. and Futrelle, R.P., (1994). Experiments in Automatic Word Class and Word Sense Identification for Information Retrieval, *Third Annual Symposium on Document Analysis and Information Retrieval*, 425-434.

Hampapur, A., Jain, R., Weymouth, T., (1994). Digital Video Segmentation, *ACM Multimedia '94*, San Francisco, 357-364.

Haralick, R.M., and Shapiro, L.G., (1992). *Computer and Robot Vision*, Addison Wesley.

Jacobs, C.E., Finkelstein, A., Salesin, D.H., (1995). Fast Multiresolution Image Querying, ACM Computer Graphics (SIGGRAPH '95), 277-286.

Lindblad, C.J., et al, (1994). The VuSystem: A Programming System for Visual Processing for Digital Video, *ACM Multimedia '94*, San Francisco, 307-314.

Pentland, A., Picard, R.W., Sclaroff, S., (1996). Photobook: Content-Based Manipulation of Image Databases, *International Journal of Computer Vision*, 18(3), 233-254.

Picard, R.W., Liu, F., (1994). A new Wold ordering for image similarity, *Proc. ICASSP*, Adelaide, Australia.

Nagasaka, A., Tanaka, T., (1992). Automatic Video Indexing and Full-Video Search for Object Appearances, *Visual Database Systems, II*, E. Knuth and L.M.

Wegner, Editors, North-Holland, 119-133.
Newman, W.C., (1990). Detecting Speech With An Adaptive Neural Network, *Electronic Design*, 79-88.
Reinhardt, Andy, (1995). New Ways To Learn, *BYTE*, March, 50-72.
Roe, D.B., Wilpon, J.G., (1994). Voice Communication Between Humans and Machines, *National Academy Press*, Washington D.C.
Sawhney, H.S., and Ayer, S., (1996). Compact Representations of Videos Through Dominant and Multiple Motion Estimation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 18, No. 8, 814-830.
Swain, M., and Ballard, D., (1991). Color Indexing, *International Journal of Computer Vision*, 7(1), 11-32.
Smoliar, S., and Zhang, H., (1994). Content-based video indexing and retrieval, *IEEE Multimedia Magazine*, 1(2), 62-72.
Teodosio, L., Bender, W., (1993). Salient Video Stills: Content and Context Preserved, *ACM Multimedia '93*, California, 39-46.
Yu, H.H., and Wolf, W., (1995). Scenic Classification Methods for Image and Video Databases, *Digital Image Storage and Archiving Systems*, SPIE 2606, 363-371.
Vroomen, L.C., et al, (1990). Robust Speaker-Independent Hidden Markov Model Based Word Spotter, *Speech Recognition and Understanding, Recent Advances*, P. Laface and R. De Mori, Editors, NATO ASI Series, Vol. F75, 95-100.
Wectlar, H.D., et al, (1995). Intelligent Access to Digital Video: Informedia Project, *IEEE Computer*, Vol. 29, No. 5, 46-52.
Weiss, R., Duda, A., Gifford, D.K., (1995). Composition and Search with a Video Algebra, IEEE Multimedia, 12-25.
Wolf, W., Liu, B., Wolf, W., (1995). A Digital Video Library for Classroom Use, *Proceedings of the International Symposium on Digital Libraries*.
Wolf, W., (1996). Key Frame Selection by Motion Analysis, *Proc. ICASSP*.
Zhang, H.J., et al, (1993). Automatic Partitioning of Video, *Multimedia Systems*, Vol. 1, 10-28.
Zhang, H.J., et al, (1995). Automatic Parsing and Indexing of News Video, *Multimedia Systems*, Springer-Verlag, No. 2, 256-266.

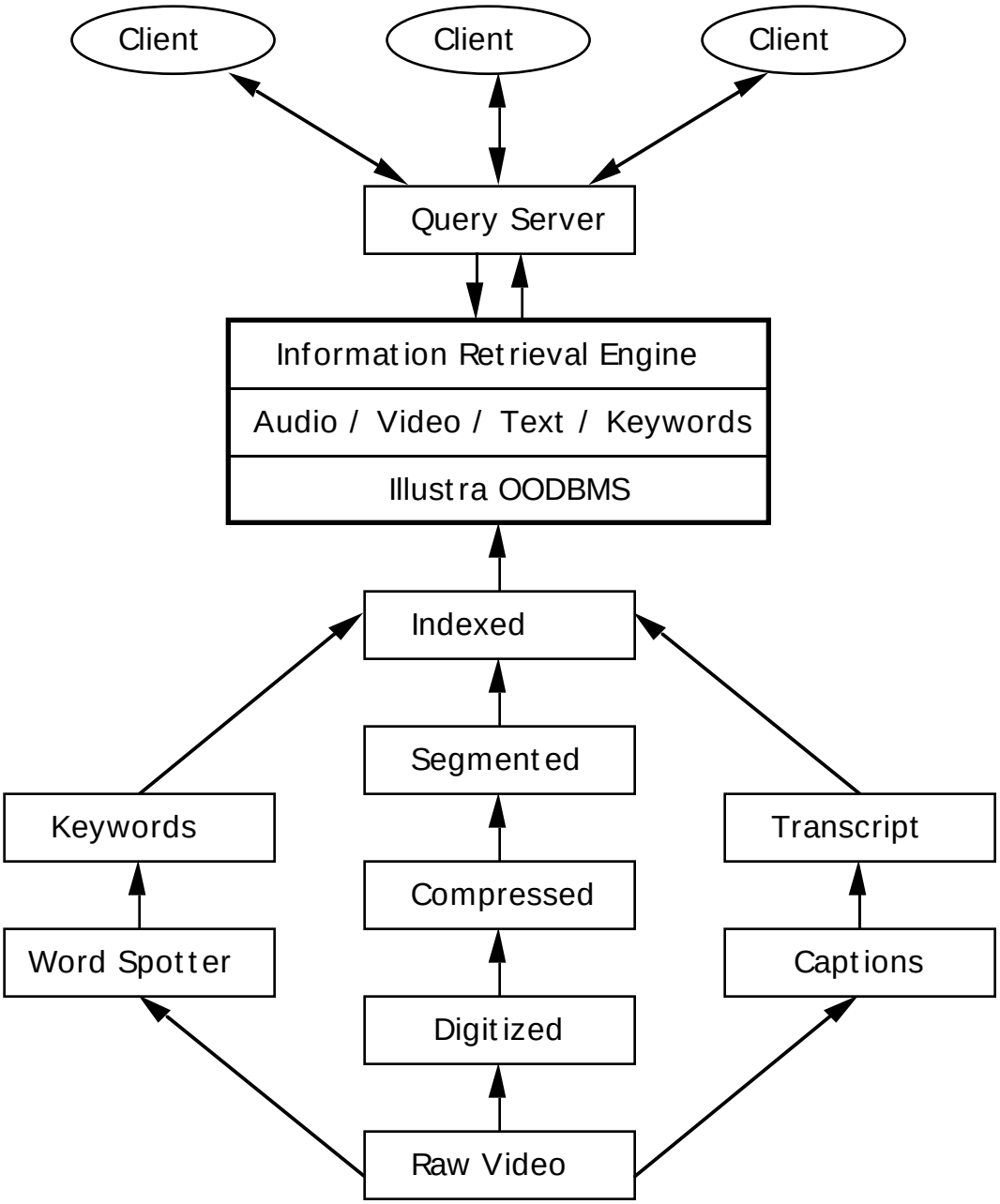

Figure 1. The architecture of the VISION Digital Video Library