Efficient Archiving and Content-Based Retrieval of Video Information on the Web

Behzad Shahraray & David C. Gibbon

Multimedia Processing Research Department
AT&T Labs - Research
101 Crawfords Corner Road
Holmdel, NJ 07733-3030
{behzad,dcg}@research.att.com

Abstract

This paper summarizes an ongoing work in multimedia processing aimed at the automated archiving and selective retrieval of textual, pictorial and auditory information contained in video programs. Video processing performs the task representing the visual information using a small subset of the video frames. Linguistic processing refines the closed caption text, generates table of contents, and creates links to relevant multimedia documents. Audio and video information are compressed, and indexed based on their temporal association with the selected video frames and processed text. The derived information is used to automatically generate a hypermedia rendition of the program contents. This provides a compact representation of the information contained in the video program. It also serves as a textual and pictorial index for selective retrieval of the full-motion video program. A fully automatic system has been set up that generates HyperText Markup Language (HTML) renditions of television programs, and makes them available for access over the Internet within seconds of their broadcast.

Introduction

The availability of low-cost and fast mass storage devices, coupled with effective media compression algorithms have made it feasible to store large volumes of multimedia information in digital form. Advances in network technology have made it possible to access this information remotely. Internet technology and the user interfaces provided by client software have enabled the effective presentation and browsing of multimedia and hypermedia documents.

The majority of the documents that currently exist on the World Wide Web have been created, or converted from other representations to the proper format, manually. This is also true for the indexing that is needed to allow for efficient navigation through this data. Such manual creation, conversion and indexing involves spending large amounts of labor and prevents the presentation of important information in a timely manner. Effective creation and utilization of these large on-line information repositories requires fast automated techniques to organize, condense and index multimedia information for searching, browsing, and selective retrieval.

We describe a number of media processing algorithms and a fully automated system based on these algorithms that generates hypermedia renditions of closed captioned television programs.

Video Processing

An important step in the creation of a compact representation of a video program is to condense and index the video information. Image processing is used to analyze every frame of the video stream in real-time to segment it into the constituent shots. This is achieved by a content-based sampling algorithm (Shahraray 1995). Block matching between consecutive video frames generates motion vectors and match values between corresponding blocks in these frames. The block match values are combined to generate a frame match signal measuring the global match between consecutive frames (figure 1).

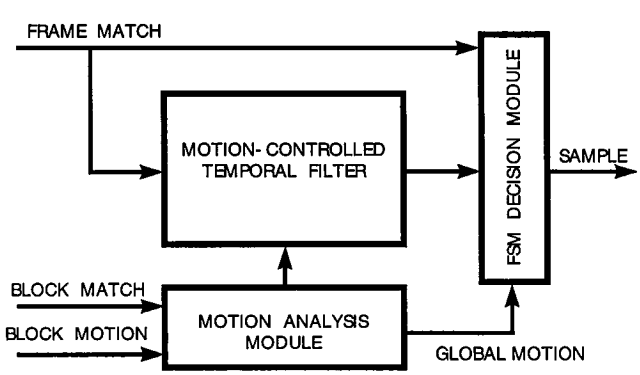

Figure 1. Content-Based Sampling of Video

A motion analysis module performs the task of estimating the global motion parameters pertaining to the camera pan and tilt operations. A motion-controlled filter operating on the frame match signal working in conjunction with a finite state decision module enables the detection of cuts and gradual transitions between shots. Quantitative information about camera operations is used to detect camera-induced scene changes and segment individual shots. A single frame from each segment is retained to represent the visual information within the segment. These representative frames collectively provide a compact representation of the visual contents of the program. They also serve as indices into the audio portion of the program, as well as the full-motion video.

Further processing of the video includes measurement of the degree of similarity between the representative images to detect matching scenes. This information helps improve the final representation of the information. We are also working on incorporating algorithms for finding heads and faces in the video processing (Graf et al. 1996) to improve the characterization of the video.

The video processing stage uses only the image brightness values and does not rely on color information. This is an important issue when dealing with existing archival data that contain large volumes of black and white video material. In these cases, the reliance of the processing algorithms on color information would preclude their usefulness. Efficient computational methods have been devised that enable the real-time processing of video using a fraction of the cycles of a common PC platform.

Text Processing

Currently, the textual component of the Pictorial Transcripts is derived from the closed caption text that accompanies an increasing number of television broadcasts. The closed caption information is recovered from the video signal during the real-time acquisition (Figure 2).

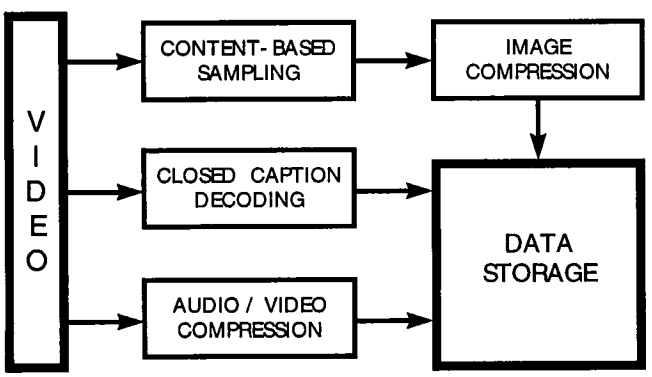

Figure 2. Data Acquisition

This phase also involves the acquisition and processing of audio, video and the extraction of information pertaining to the temporal relationships between different media components.

The raw closed caption text undergoes lexical and linguistic processing to convert it to a form suitable for generating the hypermedia documents. This processing serves several purposes (figure 3). It converts the upper-case text into lower-case while preserving the correct capitalization. The phrase database used for this purpose is generated automatically by off-line analysis of a large corpus of the Associated Press news wire.

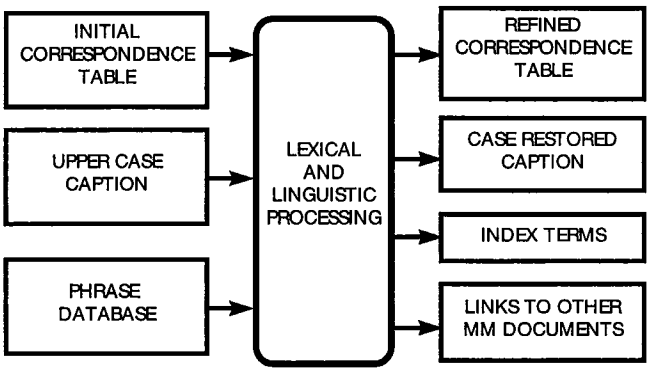

Figure 3. Text Processing

Text processing is also used to refine the temporal relationship between the text and images by adjusting the text segmentation boundaries. Index terms that are used to generate a table of contents are extracted from the closed caption text. Predetermined words and phrases are spotted and used to generate links to relevant material. Currently these tasks rely on phrase databases that have been generated either automatically or manually.

In the absence of closed caption text, speech processing has to be employed to provide the textual component of the hypermedia representation of the video program. Effective application of speech recognition to the automated transcription of video programs has been reported by other researchers (Hauptmann & Witbrock 1977). Such automatically generated transcripts have proved useful for performing text-based queries despite their limited accuracy. We are investigating the application of AT&T's large vocabulary automatic speech recognizer (Riley et al. 1995, Mohri & Riley 1997) to address this problem. This recognizer has an off-line accuracy of 90.5 percent (with a 60,000 word vocabulary size), on the North American Business (NAB) task. The real-time word error rate (with 20,000 word vocabulary size) is about 27 percent. Parallel implementations of the recognition algorithm allows for faster than real-time performance with a 20,000 word vocabulary.

Audio Processing

While the combination of the images and text provides the most compact representation of the information content of the video program, audio is included to convey additional information. When no textual component is present, the inclusion of audio is essential.

A low-complexity, wideband speech coder known as Transform Predictive Coder (TPC) is used to compress the audio stream (Chen & Wang 1996). The TPC coder is used to generate 7KHz bandwidth speech at 16 kilobits / second. This bit-rate is sufficiently low to allow the delivery of the audio, images and text components over a switched telephone networks with commonly used modems. Digitization and compression of audio is performed in real-time during the acquisition phase. Audio is segmented on the basis of its temporal relationship with the video segments.

Further processing of the audio stream for feature extraction and event detection is being investigated. The TPC coder generates the LPC coefficients and pitch estimates that can be utilized to perform additional analysis of the audio stream.

Hypermedia Document Retrieval

In the final phase, the information derived by media processing is used to automatically generate a hypermedia rendition of the program contents in HTML form (Shahraray & Gibbon 1995). A segment of a sample page generated by the system in depicted in figure 4. Audio icons next to each image can be used to initiate the replay of the corresponding audio contents. In this case, the compact representation serves as an index into the audio stream. A similar arrangement is used to initiate selective replay of motion video by selecting the images. This has been achieved through an interface to the FusionNet multimedia information-on-demand system (Civanlar, Cash, & Haskell 1996). Video segments selected using the Pictorial Transcripts as an index, are delivered over the switched telephone network .

The system provides several searching and browsing mechanisms to facilitate selective retrieval. These include a calendar interface for selecting a particular program from an archive of television programs accumulated over more than two years. An example of a page used to choose among different programs is shown in figure 5. Each program is organized into pages. Currently, the segmentation of the program into individual pages is based on size, as well as detected commercial boundaries. Work is underway to take advantage of more advanced text segmentation techniques to perform topic segmentation (Salton et al. 1996). An index page enables the selection of the pages using textual and pictorial indices.

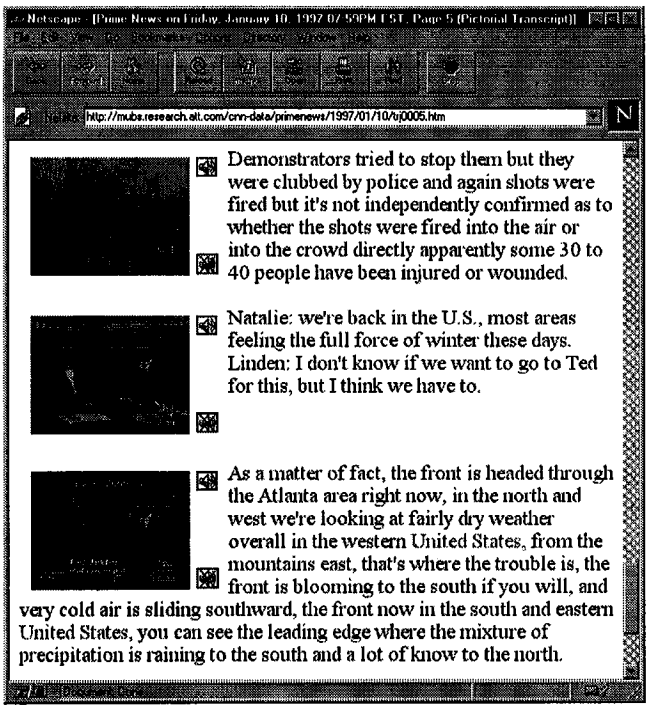

Figure 4. A Sample Page from a Pictorial Transcript

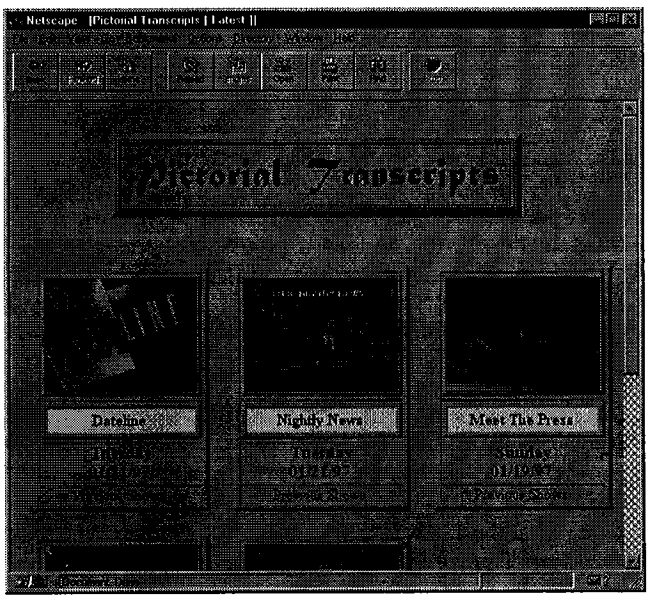

Figure 5. Program Index Page

Figure 6 depicts the pictorial segment of an index page. An archive of video programs collected over the period of several years can be searched using available search engines. The information can then be presented either in HTML form, or using a streaming multimedia player (figure 7) that delivers a synchronized presentation of the

representative images, the text and the audio. The low average bandwidth of the media used enables the delivery of this presentation over telephone lines.

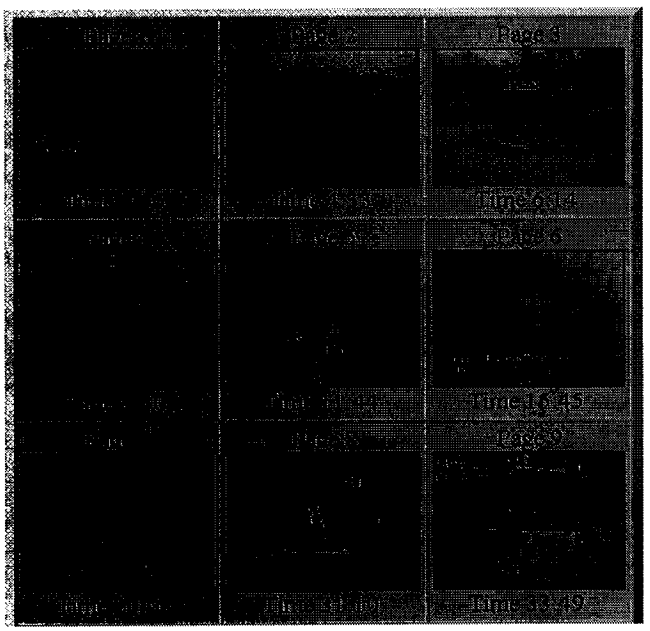

Figure 6. An Example of a Pictorial Index

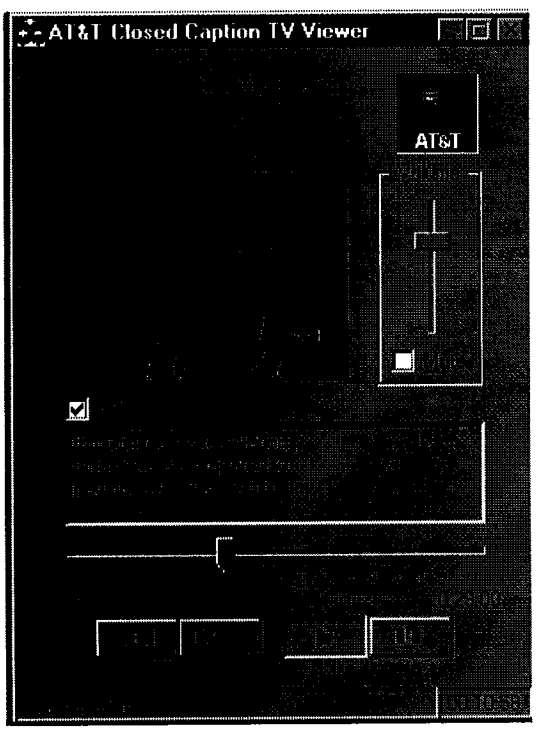

Figure 7. Streaming Image, Text and Audio Player

Conclusions

Several media processing algorithms and a working system based on these algorithms were described. The system is an example of how text, image, audio, and video processing and understanding techniques can be integrated to condense, organize and index information for efficient archiving and intelligent content-based retrieval.

References

Chen J. H., and Wang D. 1996. Transform Predictive Coding of Wideband Speech Signals. In Proceedings of IEEE Int. Conf. Acoustics, Speech and Signal Processing ICASSP'96, 275-278. Atlanta, GA.

Civanlar M. R., Cash G. L., and Haskell B. G. 1996. FusionNet: Joining the Internet and Phone Networks for Multimedia Applications. In Proceedings of the Forth ACM Conference on Multimedia, 431-432. Boston, Mass.

Graf H. P., Cosatto E., Gibbon D. C., Kocheisen M., and Petajan E. 1996. Multi-Modal System for Locating Heads and Faces. In Proceedings of the Second International Conference on Automatic Face and Gesture Recognition, 88-93. IEEE Computer Soc. Press, Los Alamitos, California..

Hauptmann, A.G. and Witbrock, M.J., 1997. Informedia: News-on-Demand - Multimedia Information Acquisition and Retrieval. In Intelligent Multimedia Information Retrieval Maybury, M. (ed), AAAI Press / MIT Press. Forthcoming.

Mohri M., and Riley M. 1997. Weighted Determinzation and Minimization for Large Vocabulary Speech Recognition. In Proceedings of Eurospeech 1997. Forthcoming.

Riley M., Ljolje A., Hindle, D., and Pereira, F. 1995. The AT&T 60,000 Word Speech-to-Text System. In Proceedings of Eurospeech 1995.

Salton G., Singhal, A., Buckley C., and Mitra M. 1996. Automatic Text Decomposition Using Text Segments and Text Themes. In Proceedings of the Seventh ACM Conference on Hypertext, 53-65. Washington DC March 1996.

Shahraray B. 1995. Scene Change Detection and Content-Based Sampling of Video Sequences. In Digital Video Compression: Algorithms and Technologies 1995. Proc. SPIE 2419.

Shahraray B., and Gibbon D. C. 1995. Automated Authoring of Hypermedia Documents of Video Programs. In Proceedings of Third ACM Conference on Multimedia, 401-409. San Francisco, Calif.