A Practical Video Database Based on Language and Image Analysis

Yiqing Liang, Bede Liu, Wayne Wolf, Thomas Yeh

Abstract

We integrated a practical digital video database system based on language and image analysis with components from digital video processing, still image search, information retrieval, closed captioning processing. The attempt is to utilize the multiple modalities of information in video and implement data fusion among the multiple modalities. Keyframes are extracted to represent shots based on video content. Logical structure of video is discovered and represented as Scene Transition Diagram. Storyboard and STG, and catalog information provide means to browse video content non-linearly. Text information from closed captioning is extracted and associated with shots. Text information search and image similarity search are provided to get direct access to video shots.

1. Multiple Modality from Video Information

One key feature of our research is data fusion (i.e., making use of all types of existing information which coexists in video information). What is needed is a system that can utilize multiple tracks of information embedded in digital video in order to obtain retrieval terms that characterize the situation shown in the video. This should include audio, video, closed captions, caption, sound, music, special effects, catalog information, and transcripts. Audio, text, sound track, and image analysis should work hand-in-hand to disambiguate the terms, with imagery identifying speakers or objects, with language analysis providing keys needed for image analysis, and the sound track (i.e., music and special effects) assisting further interpretation of the contents.

Text from Speech Recognition

Extracting transcripts from audio track through speech recognition was the first task examined. Our intention is to study the application of existing speech recognition algorithms for the extraction of words from the audio track. The development of new speech recognition algorithms, although not the goal of our research, was examined for the purpose of extracting keywords as part of the data fusion effort.

The intent of this research is to discover an appropriate Speech Recognition system capable of performing dictation and transformation of human speech from the audio portion of digitized video data into text for use in text inquiries. A Speech Recognition engine must be capable of the following:

- Speaker-independent: requirement of preparing various individuals in a video must be a bare minimum. The system should be adaptive in order to increase recognition accuracy.
- Continuous speech recognition: system must be able to perform recognition of continuous natural human speech. Discrete speech recognition requires short silences to separate individual words, thereby performing poorly on ordinary human speech. Speaker and vocabulary adaptation is desirous.
- Unrestricted dictation: restriction in term of noise-level, domain-specifics, or vocabulary-size are not acceptable in the application.

Many commercial packages for speech recognition were reviewed including ones in the research effort. Special attention was paid to packages capable of reading from MPEG files in order to recognize speeches in view of our plan to use MPEG as the digital video standard and the fact that MPEG files for video and audio can be generated at digitization time. None of the reviewed packages including IBM VoiceType Dictation, Microsoft Whisper, PowerSecretary from Articulate Systems, Inc., AT&T WATSON Advanced Speech Applications Platform, DragonDictate Dragon System, Philips SpeechMagic can...
meet the requirements defined above, though we did not have the opportunities to review some advanced packages such as CMU' SPHIX III.

Closed Caption

Extracting text information from closed captions in the video is yet another approach to increase indexing and retrieval keys for video frames. We studied the off-the-shelf software/hardware products capable of extracting closed caption information from video thereby deciding to use SoftTouch's equipment for this project. SoftTouch's HUBCAP is an economical stand alone caption and XDS decoder, caption and XDS data recovery, and caption character generator device. The test indicates closed caption text can be extracted and time stamp grabbed. These time stamps assist in building a relationship between video frames and closed caption text.

Complications are anticipated in such a relationship. A closed caption statement may span many frames, even more than one keyframe we detect using our algorithms. Conversely, one video shot that is represented by one keyframe may correspond to multiple closed caption text sentences. The topic upon which the research will be conducted is how to build such a relationship. Currently, we employ a straightforward algorithm, i.e., matching the time stamped sentences with the keyframes that have corresponding time stamps.

Catalog Information

Available cataloging information will also be investigated for additional data fusion. Cataloging information for video clips is usually available, although predominately entered manually. This information is useful to search entire video programs, and necessary in identifying which video programs a specific video keyframe belongs to in our system when video program storyboards are available.

Music and Special Effect

Plans are under way to employ music and special effects on the sound tracks to capture more information associated with video frames and speech.

Text Search Engine

Another key technology to make the best use of all text information extracted from video is a text search engine. The text search engine provides both indexing and retrieval points for text information. Once a relationship is established between text information extracted from audio, closed captions, captions, and cataloging information with video frames, especially keyframes, it will provide the use of a text search engine for search and retrieval of video frames.

When the problem of associating text information from speech, closed caption, and cataloging information with video frames, key frames in particular, is resolved, text associated with a keyframe can be treated as a document. Video programs are comprised of shots. Each shot is represented by one or more keyframes that we extract using image processing techniques. We suppose that speeches or text information associated with a shot constitutes an independent unit and can be treated as an independent document. The problem of retrieving a shot meeting text retrieval criteria becomes a problem of retrieving the document associated with the keyframe that represents the shot. Thus, video shot/keyframe retrieval through text becomes an information retrieval problem.

The text search engine has been the main research topic in the area of information retrieval. Traditional information retrieval approaches include: a) full text scanning; b) use of signature files; c) inversion files, the approach most commercial text retrieval engines adopt, and; d) vector model and clustering, of which Salton, Cornell University's SMART system, is a representative. Recent research has started to use natural language processing techniques, neural network technology, and Latent Semantic Indexing.

Latent Semantic Indexing (LSI) (Deerwester et al. 1990) is a vector space information retrieval method demonstrating improved performance over traditional vector space techniques utilized in Salton's SMART system. LSI uses singular-value decomposition allowing re-arrangement of space to reflect major associative patterns in the data, thereby ignoring smaller, less important influences. Position in space then serves as a new kind of semantic indexing. Retrieval proceeds through use of the query terms to identify a point in space, and return documents in its neighborhood. LSI is suitable for document retrieval with sparse indexing information as is the case in video where text information is sparse with respect to image information. We incorporated LSI into our video database at Princeton University as text search engine to help index and retrieve video shots. In addition, we made effort to integrate LSI with IBM's Query By Image Contents (QBIC) still image database to make it a complete system.

2. Image Understanding

Image understanding is another important key to research in digital video database. Indexing and retrieval of digital video at the original video program level can be achieved through traditional relational database and cataloging information. Nonetheless, this traditional and mostly commonly practiced industry approach is not only time consuming, thereby expensive to index video clips, but also can not satisfy often raised questions about finding
one specific video shot. Indexing and retrieval of digital video can not cease at the original video program level. Improved technology is greatly needed.

Scene Change Detection and Key Frames
One approach in achieving this goal is to use scene change detection or video segmentation techniques to divide video stories into video shots. We have employed algorithms invented at Princeton University to meet this goal. Two types of algorithms have been used. One algorithm operates directly on compressed MPEG files, achieving excellent performance. We also used an algorithm that operate on uncompressed video file to identify scene change first and then compares adjacent keyframes using color histogram and rejects duplicates (Philips & Wolf 1996).

Storyboard and Scene Transition Graph (STG)
Once keyframes are extracted, a storyboard can be generated using these key frames. These storyboards give people an overview of the key contents in the video program without viewing the entire video, thereby achieving non-linear access. In addition, algorithms have been invented to generate a hierarchical storyboard, referred to as Scene Transition Graph (STG) (Yeung, Yeo, & Liu 1996). STG not only lays out all key frames, but also provides time stamps for each keyframe and organizes them into a Directed Arc Graph with time sequence represented by arcs among key frames. In addition, similar key frames are clustered into one frame which will appear in the STG allowing each key frame to be decomposed into a set of similar frames. The algorithm has been improved to generate the STG to capture story units in the video program. The figure below on this page is an example of the STG output from our recent algorithm.

Video Browsing
Based on the key frame storyboard and the Scene Transition Graph, the building of a browse mechanism allowing people to browse and navigate through our digital video library is possible.

In the video browsing system, users can enter a digital video library and view the table of contents listing of video programs in the database. They can select one of the programs by clicking on the corresponding item in the table of contents. Also, users can apply the text search to retrieve corresponding program items with key words when the list gets large.

Once they get into a particular program, they will see all cataloging information including: author, date, title, provenance, length, key words, and description. Users have several options: playing back the entire video program; playing back only audio part; viewing the video storyboard; or viewing the keyframe with associated text information and the position of the shot in the video program.

Upon selecting storyboard, the storyboard might be either the regular storyboard laying out of all key frames for that video program or a Scene Transition Graph for that program.

3. Retrieving Video Key Frames and Video Shots from the Database
The video browsing mechanism provides a fast, non-linear access to the contents of a particular video program one at a time. Often users prefer to search for a particular scene from all the video programs in the database rather than from one program at a time. Thereby, a new mechanism is needed to meet this requirement.
The Concept

We have designed a database indexing and retrieval mechanism for digital video contents. This mechanism is based on the keyframes extracted from video programs. Since keyframes capture the most important contents in video programs, consideration for retrieving video shots should be conducted by retrieving key frames representing that shot. If we organize all the video key frames from all video programs in one single database, a retrieval of certain keyframes should allow us to access all video contents in a non-linear manner to the level of video shots. Consequently the problem of retrieving certain video shots is reduced through retrieving certain video key frames. Each video keyframe can be a still image, possibly a JPG file or a DC image. Thereby retrieving a desired still image is the approach capable of solving the problem of accessing video information.

Similarity Measurement for Video Keyframe Access

Still image indexing and retrieval has been a research topic in image processing for many years. Numerous approaches have been proposed to resolve this problem. One commonly accepted conclusion is that the traditional and fundamental operation in a traditional database (i.e., matching) will not work for image search within a database. Alternately, a similarity search is the correct approach in image database query. A similarity measurement that behaves well for any pair of images must be discovered. There are many different kinds of similarity measurement algorithms: Vector Space model; Metric Space model; and More General models or the algorithms can be classified into: Metric measure; Set-Theoretic based measures; and Signal Detection Theory based measures. Each group is further subdivided into; a) measures based on crisp logic, and b) measures based fuzzy logic (White & Jain 1996).

In the Video Retrieval Based on Language and Image Analysis project, the intent is to follow both matching and similarity search algorithms. Two text search engines are used in our system: one exact key word match from IBM QBIC, and one of LSI. The text search used in IBM QBIC employs matching algorithms. On the other hand, LSI employs similarity search by concept space and concept mapping to expand the algorithms’ capabilities to resolve the famous synonymy and polysemy problem in information retrieval. Similarity search algorithms will be used to search for query by image content.

Furthermore, when we combine both matching and similarity search to locate desired key frames, many interesting results turned out and more innovative algorithms may be needed. In addition, video consists of frames of images in time sequence. It is unclear whether it is necessary, and if so how, to incorporate time into the search algorithm. However, the research conducted at Princeton University to extract story units with time constraint as represented by STG is worth further pursuing in this direction.

Still Image Search Engine

Currently, we employ IBM’s Query By Image Content (QBIC) (Flicker 1995) as our image search engine. QBIC is a set of software routines providing functions to query collections of images by content, thereby permitting an image collection to be queried for images that have predominantly red colors or striped textures where the color and texture information has been automatically computed. Originally, QBIC was designed for still image search.

It is our belief that QBIC’s still image search engine can be applied to the video key frame search. Although QBIC is restricted by its search capabilities in terms of color, color histogram, and texture, it should be capable of providing a starting base for the video database search. Every effort will be made to add our own features to QBIC, thereby gradually adding more functionality to the search engine.

4. System Architecture

System architecture design is a key to the success of this project. System architecture need not only meet user interface requirements, but also meet all technology requirements that are going to be implemented into the system. The system will incorporate a number of independently developed research systems whose data and functions must be integrated.

We have identified system components including: algorithm to segment video, algorithms to extract keyframes, algorithms to generate STG, algorithms to extract closed captioning information, algorithm to match text with keyframes, image search engine, and text search engine.

The system will have two different types of functions: those that will be executed off-line and time-insensitive, and those that are executed on-line in real-time for the interactive users. In addition, data and network architecture will be part of the system architecture. Network billing function and security will be another subsystem when commercialization starts.

Network (Internet / Intranet)

Network based, in particular Internet based, WWW is the latest evolution in software development. This trend is not only new, but also indicates an important shift in software
development paradigm such as the shift experienced when moving away from mainframe to client/server computing. Hence, building a new system to accommodate the new paradigm and architecture is of great importance, not only because it is new, but because it may take us to a new paradigm. It is our intention to use WWW navigator as the user interface and network computing as part of our fundamental system architecture. It is also desired by some applications such as distribution of digital video on the Internet.

However, the system should be able to be built on Intranet or with other types of user interface, with client/server architecture.

Off-line Processes:
The functions to create database, including extracting key frames from video streams, extracting text and mapping them to key frames, generating indices for key frames, and generating storyboards, are off-line processes. They will be performed without users’ involvement as part of pre-processing.

Real-time Processes:
The processes for users’ interactive functions, including submitting queries, searching the database, returning search results to the users, and playback of video clips, are real time. Network billing will be another real time system.

5. System Tools

Internet Navigator Plug-in for MPEG Video Playback
In the video database query and retrieval, users usually would like to play back a hit video clip immediately when they find it in the retrieval results. This will not only allow the user to instantly find the clips they are looking for, but will give a warm feeling to the end users when they see the live playback of video clips. A very helpful tool in terms of both functionality and user interface.

Many commercially available MPEG player plug-ins can playback MPEG video file on the web browser. However, none can specify the start frame and stop frame for a preset playing section. Therefore, we have successfully developed a 32-bit version which runs in Microsoft Windows NT and can be used for any 32-bit MCI-compliant MPEG decoder, and a 16-bit version of plugin which runs on Window95, both of which allow the start frame and stop frame be specified for playing back only the pre-selected section of the story.

Thumbnail and Full Size Picture; Browsing and Searching
Storyboards and image search employ thumbnails for presentation. We support viewing the full size keyframes by just clicking on a thumbnail to allow user has a good view of the picture.

Connections between Browsing and Searching are built to allow users to easily go from one approach to another.

6. Discussions
We found it is both interesting and hard to evaluate the search results. Image search engine presents similar keyframes as result in terms of similarity on color, texture, color histogram, and color layout. It is rather difficult for user to appreciate some of these similarities. LSI employs a new way of text search, returning most similar documents not based on words. It is also different from what many users are used to. The combination of these two make it even more difficult for users to appreciate. Further research is needed to find the best approach.

Innovation algorithms are needed to segment video not only by syntactical information such as color, histogram, and texture. Algorithms based on semantic information such as object in the video are needed, not only make best segmentation, but also answer the potential concern mentioned above.

Logical structure of video is needed to aid better understanding of the story, thus better retrieving. STG can be used as a start point. Making components of STG indexable and retrievable is an important tasks.

Speech recognition is very important. So are music and special effect, and other sound track information. The association between text, sound, music, special effects and images need more research effort.

Acknowledgment
This research is supported by DARPA Contract No. PAN RTW C3-96 administered through U.S. Army Missile Command. Thanks for IBM Almden Research Lab providing QBIC and Bellcore providing LSI at Princeton University.

References

