Intelligent Multimedia Information Retrieval, Mark T. Maybury, Ed.. AAAI Press, pps. 213-239, 1997.

Informedia: News-on-Demand Multimedia Information Acquisition and Retrieval

Alexander G. Hauptmann and Michael J. Witbrock
School of Computer Science, Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA, 15213
United States of America
+1 412 268-5576
alex@cs.cmu.edu witbrock@cs.cmu.edu

Abstract

In theory, speech recognition technology can make any spoken words in video or audio media subject to text indexing, search and retrieval. This article describes the News-on-Demand application created within the Informedia™ Digital Video Library project and discusses how speech recognition is used for transcript creation from video, time alignment of closed-captioned transcripts, a speech query interface, and audio paragraph segmentation. Our results show that speech recognition accuracy varies dramatically depending on the quality and type of data used, but the system is quite useable with only moderate speech recognition accuracy.

1. What is Informedia: News-on-Demand

The Informedia™ digital video library project [Informedia95, Wactlar96] at Carnegie Mellon University is creating a digital library in which text, image, video and audio data are available for full content retrieval. News-on-Demand is an application within Informedia which monitors news from TV, radio and text sources and allows the user to retrieve news stories of interest.

This paper gives a brief overview of the Informedia digital video library project [Christel94a, Stevens94, Christel94b, Informedia95] followed by a detailed description of the News-on-Demand application [Hauptmann95]. Both the automated library creation process for News-on-Demand and the news library exploration process will be explained. We show how speech recognition fits into the various digital news library processing steps. Results are presented for speech recognition on actual broadcast news data. Finally we discuss some active areas of research relevant to the multimedia information acquisition and retrieval problem.

1.1 An Overview of the Informedia Digital Video Library Project

Vast digital libraries of information will soon become available on the World Wide Web as a result of emerging multimedia computing technologies. However, it is not enough simply to store and play back information as many commercial video-on-demand services apparently intend to do. New technology is needed to organize and search these vast data collections, retrieve the most relevant selections, and permit the to be effectively reused.

Through the integration of technologies from the fields of natural language understanding, image processing, speech recognition and video compression, the Informedia project [Christel-94a] allows a user to explore multimedia data in depth as well as in breadth. The Informedia digital video library project goes far beyond the current paradigm of video-on-demand, where a user can

select one video from a limited set and view that video after a delay of a perhaps a few minutes. The computer adds no substantial benefit to this video-on-demand model over a VCR with each video on a tape; the user remains a passive observer of someone else's produced material. By contrast, the Informedia Project segments hours of video into logical pieces and indexes these pieces according to their raw content (dialog, images, narration). The users can actively explore the information by finding sections of content relevant to their search, rather than by following someone else's path through the material (as one does when using the current generation of educational CD-ROMs) or by viewing a large chunk of pre-produced material (as with video on demand). Through the active, dynamic exploration supported by a deep, rich library and the indexing and retrieval capabilities of the computer, the user is more motivated and may learn more from the data set. Using such a library, a large body of video material can be searched with very little effort.

Users are able to explore Informedia libraries through an interface that allows them to search using typed or spoken natural language queries, to select relevant documents retrieved from the library and to play or display the material on their PC workstations. The library retrieval system can effectively process natural spoken queries and deliver relevant video data in small video paragraphs, based on information associated with the video during library creation. Video and other data may be explored in depth for related content. During retrieval based on keyword searches by a user, only the query-relevant video segments are displayed.

The Informedia project is developing new technologies and embedding them in a video library system primarily for use in education and training. The Informedia project will establish an on-line digital video library consisting of over 1000 hours of video material. In order to be able to process this volume of data, practical, effective and efficient tools are essential.

In the United States, schools and industry together spend between $400 and $600 billion per year on education and training, an activity that is 93% labor-intensive, with little change in teacher productivity ratios since the 1800s. The new digital video library technology will bring about a revolutionary improvement in the way education and training are delivered and received.

The initial Informedia test-bed system has been installed in a K-12 school, where students use the Informedia System to explore multimedia data for educational purposes. We plan to extend this test-bed to other Pittsburgh schools. During library creation for the test-bed, video material obtained from our Informedia Project Partners such as WQED/Pittsburgh and the British Open University is used. Our project plan calls for four test-bed installations with users ranging from grade school children to university faculty. In addition, we will provide networked access to the primary test bed, and export portions of the system and data to other sites for their local exploration and experimentation.

The user tests will be conducted at Carnegie Mellon University, the Winchester Thurston School in Pittsburgh, the Fairfax County (VA.) public school system, and with the Open University in the UK. Users will be of many different types, as we test the practicality of the concept of multimedia library search and the usability of the user interface for various age and interest groups.

Universal access to large amounts of low-cost digital information and entertainment will significantly affect the conduct of business, professional, and personal activity. The initial impact of the Informedia project's activity will be by enabling broad accessibility and reuse of existing

video materials (e.g., documentaries, news, vocational, training) previously generated for public broadcast, public and professional education, and vocational, military and business training.

1.2 The Informedia: News-on-Demand Application

One compelling application branch of the Informedia project is the indexing and retrieval of television, radio and text news. The Informedia: News-on-Demand application [Hauptmann95] is an innovative example of indexing and searching broadcast news video and news radio material by its text content. News-on-Demand is a fully-automatic system that monitors TV, radio and text news and allows selective retrieval of news stories based on spoken queries. The user may choose among the retrieved stories and play back the news stories of interest. The system runs on a Pentium PC using MPEG-I video compression. Speech recognition is currently done on a separate platform using the Sphinx-II continuous speech recognition system [CMU-Speech95].

The News-on-Demand application forces us to consider the limits of what can be done automatically and in limited time. Since news events happen daily, it is not feasible to process, segment and label news through manual or "human-assisted" methods. Immediate availability of the library information is important, as is continuous updating of the contents.

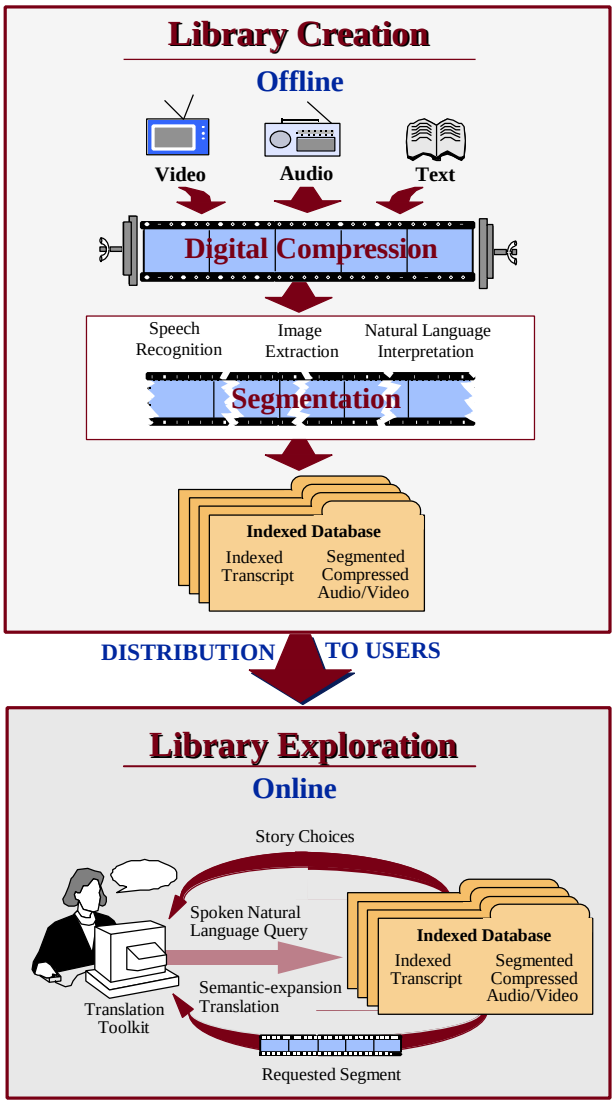

Figure 1: Overview of the News-On-Demand System

Unlike other Informedia prototypes which are designed to be educational test beds, the News-on-Demand system is fully automated. While the educational test bed prototypes' library is developed using computer-assisted methods that require human post-processing, the News-on-Demand system is fully automatic. We are forced, therefore, to fully exploit the potential of computer speech recognition without the benefit of human corrections and editing.

While our work is centered around processing news stories from TV broadcasts, the system exemplifies an approach that can make any video, audio or text data accessible. Similar methods can be used to index and search other streamed multimedia data by content for other Informedia applications.

Currently, TV and radio news is broadcast at particular times, and if a person is not in front of a TV or radio at that time, the information becomes virtually inaccessible. There is simply not enough time to scan through tapes of yesterday's news for relevant stories. Even when it is

possible to watch when the news if being broadcast, the viewer must spend the time required to view all the stories in a news show, since broadcasts do not provide the ability to select which stories to skip and which to pursue in more detail.

Furthermore, a person can only attend to one news channel at a time. Similar or related information broadcast on another news channel at the same time cannot easily be viewed. In contrast, text news on a topic is generally available in overwhelming quantities but cannot provide the comprehensive visual and audio information available in radio and video material.

The solution is to compress, digitally store and analyze news broadcasts on the computer. All information is made accessible through interactive queries. These queries allow the user to retrieve relevant segments from all the news programs that carried stories on the topic of interest. Each individual news story is indexed based on a text transcript. All news information becomes accessible through interactive queries at the user's convenience, permitting the retrieval of relevant news stories from all the networks and news sources that covered the topic of interest. An outline of the system is given in Figure 1.

Most other attempts at solving the news retrieval problem by providing news databases have restricted the data to text material only. Video-on-demand allows a user to select, and pay for, a complete program, but does not allow for selective retrieval. The closest approximation to News-on-Demand can be found in the "CNN-AT-WORK" system offered to businesses by a CNN/Intel cooperation. At the heart of the CNN-AT-WORK solution is a digitizer that encodes the video into INDEO format compression format and transmits it to workstations over a local area network. Users can store headlines together with video clips and retrieve them at a later date. However, this service depends entirely on the separately transmitted "headlines" and does not include other news sources than CNN. In addition, CNN-AT-WORK does not feature an integrated multimodal query interface [CNN-AT-WORK95].

Preliminary investigation into the use of speech recognition for analysis of a news story was carried out by [Schauble95]. Without a powerful speech recognizer, their approach used a phonetic engine that transformed the spoken content of the news stories into a, possibly errorful, phoneme string. The query was also transformed into a phoneme string and the database searched for the best approximate match. Errors in recognition and word prefix and suffix differences did not prevent the operation of the system since these errors scattered evenly over all documents allowing the well-matching search scores to dominate the retrieval.

Another news processing system that includes video materials is the MEDUSA system [Brown95]. The MEDUSA news broadcast application can digitize and record news video and teletext transcriptions, which are equivalent to closed-captions. Instead of segmenting the news into stories, the system uses overlapping windows of adjacent text lines for indexing and retrieval. During retrieval the system responds to typed requests by returning an ordered list of the most relevant news broadcasts. Query words are stripped of suffixes before search and the relevance ranking takes word frequency in the segment and over all the corpus into account, as well as the ability of words to discriminate between stories. Within a news broadcast, it is up to the user to select and play a region, using information provided by the system about the position of the matched keywords. The focus of MEDUSA is in the system architecture and the information retrieval component. No image processing and no speech recognition is performed.

Other projects that seek to index and retrieve from video news sources include the Conceptually Indexed Video project at Sun [Woods96], which is attempting to build conceptual taxonomies of query terms to improve the quality of returned stories, and the VISION system at the University of Kansas which, while similar in aim to Informedia, is concentrating on the problems of compressing video data and delivering it over the Internet. It is also distinguished by its stated concentration on the use of pre-existing, mature, domain independent indexing technologies [Li96].

The Broadcast News Navigator (BNN) system [Maybury96][Mani96] has concentrated on the automatic segmentation of stories from news broadcasts using discourse structure. While a great deal of success has been achieved so far using heuristics based on stereotypical features of particular shows (e.g. "still to come on the NewsHour tonight…"), the longer term objective is to use multi-stream analysis of such features as speaker change detection, scene changes, appearance of music and so forth to achieve reliable and robust story segmentation. The system also aims to provide a deeper level of understanding of story content than is provided by simple full text search, by extracting and identifying, for example, all the named entities in a story.

The Informedia project, in as much as it involves the indexing of non-textual data, also bears similarities to projects such as QBIC [Flickner95], which applies both automatic image characterization and hand-annotation to images, and supports their retrieval using image similarity. One of the more interesting features of the QBIC system is that it allows query by demonstration, with the user sketching the features desired in the retrieved image. A similar effort, which also encompasses some video material, is the Photobook system [Pentland94]. Photobook employs relatively sophisticated statistical characterizations of selected image features, such as faces, shapes and textures, to support accurate retrieval by image similarity. A final example of an image retrieval system is Chabot [Ogle95], a part of the Berkeley digital library project. This system includes an element of cross-modal operation, allowing users to search simultaneously in pre-existing annotations and color content characterizations of a large set of landscape images. This allows searches for objects such as "yellow flowers", that might not have been easily identified from the annotations or image qualities alone.

1.3 Component Technologies

There are three broad categories of technologies we can bring to bear to create and search a digital video library built from broadcast video and audio materials [HauptmannSmith95]:

Text processing looks at the textual (ASCII) representation of the words that were spoken, and at other text annotations that may be derived from the transcript, from the production notes, or from closed-captioning that is sometimes broadcast with the news stories. Text analysis can work on an existing transcript to help segment the text into paragraphs [Mauldin89]. An analysis of keyword prominence allows us to identify important sections in the transcript. Other more sophisticated language based criteria are under investigation. We currently use two main techniques for text analysis:

1. If we have a complete time-aligned transcript available from the closed-captioning or from a human-generated transcription, we can exploit "structural" text markers such as caption punctuation to identify news story boundaries.

2. To rank the contents of news segments, we use the well-known TFIDF (term frequency, inverse document frequency) weighting scheme to identify critical keywords and their relative importance for the video document [Salton83].

Image analysis looks at the images in the video stream. Image analysis is primarily used for the identification of scene breaks and to select static frame icons that are representative of a scene. Primitive image features based on image statistics, such as color histograms, are used for indexing, matching and segmenting images [Zhang95]. The following two techniques are currently implemented in our News-on-Demand production system:

1. Using *color histogram analysis*, video is segmented into scenes through the use of comparative difference measures. Images with small histogram disparity are considered to be relatively similar. By detecting significant changes in the weighted color histogram of successive frames, image sequences can be separated into individual scenes. A comparison between cumulative distributions is used as a difference measure. This result is passed through a high pass filter to further isolate peaks and an empirical threshold is used to select only those regions where scene breaks occur.

2. *Optical flow analysis* is an important method of visual segmentation and description based on interpreting camera motion. We can identify camera motion as a pan or zoom by examining the geometric properties of the optical flow vectors. Using the Lucas-Kanade [Lucas81] gradient descent method for measuring optical flow, we can track individual regions from one frame to the next. By measuring the velocity of individual regions over time, a motion representation of the scene is created. Sudden, widespread changes in this flow suggest random motion, and therefore, new scenes. Optical flow changes also occur during gradual transitions between images such as fades or special effects.

Only regions of low ambiguity are selected for tracking. Trackable regions are found by searching the entire image for sub-windows whose gradient derivatives exhibit relatively similar eigenvalues. In order to accurately track a region over large areas, a multi-resolution structure is used. With this structure we can track regions across many pixels and reduce the time needed for computation. When optical flow is minimal the frames are suitable for use in an iconic frame representation.

These techniques work well when scene changes are abrupt, however, camera motion and gradual transitions can severely affect the scene segmentation accuracy of the system. When changes are gradual, we combine the optical flow results with histogram analysis. This allows for segmentation under conditions that do not involve sudden changes in image content.

Speech analysis provides the basis for analyzing the audio component of the news broadcast.. To transcribe the content of the video material, we use the Sphinx-II speech recognition engine, a large-vocabulary, speaker-independent, continuous speech recognizer created at Carnegie Mellon [CMU-Speech95, Hwang94]. Sphinx-II uses senonic semi-continuous hidden Markov models (HMMs) to model between-word context-dependent phones. The system uses four types of codebooks: mel-frequency cepstral coefficients, first cepstral differences, second cepstral differences, and power and its first and second differences. Twenty-seven phone classes are

identified, and a set of four VQ codebooks is trained for each phone class. Cepstral vectors are normalized using an utterance-based cepstral mean value. The semi-continuous observation probability is computed using a variable-sized mixture of the top Gaussian distributions from each phone-dependent codebook.

The recognizer processes an utterance in four steps:

1) A forward time-synchronous pass using between-word senonic semi-continuous acoustic models with phone-dependent codebooks and a bigram language model is performed. This produces a set of possible word occurrences, with each word occurrence having one start time and multiple possible end times.

2) A backward pass using the same system configuration is then performed, resulting in multiple possible begin times for each end time predicted in the first pass.

3) An A* algorithm is used to generate the set of N-best hypotheses for the utterance from the results of the forward and backward passes. Any language model can be applied in this pass — the default is a trigram language model. This approximate A* algorithm is not guaranteed to produce the best-scoring hypothesis first.

4) The best-scoring hypothesis is selected from among the N-best list produced. This hypothesis is output as the recognizer's result.

The language model consists of words with probabilities, bigrams or trigrams which are word pairs or triplets respectively with conditional probabilities for the last word given the previous word or word-pair. Normally, a word trigram is used to predict the next words for the recognizer. However, a back-off procedure allows the next word predicted from only the current word, using bigram probabilities, at a penalty. A word may also occur independently of context, based only on its individual probability, with another larger back-off penalty. Our current largest and most accurate language model was constructed from a corpus of news stories from the Wall Street Journal collected from 1989 to 1994, and from the Associated Press news service stories from 1988 to 1990. Only trigrams that were encountered more than once were included in the model, but the most frequent 58800 words in the corpus, and their bigrams, were all included [Rudnicky95].

Acoustic signal analysis is used to identify segment boundaries of paragraph size. We can detect transitions between speakers and topics which are marked by silence or low energy areas in the acoustic signal. To detect breaks between utterances we use a Signal to Noise ratio (SNR) computation. This algorithm computes the power of digitized speech samples where each s_i is a pre-emphasized sample of speech gathered over a twenty millisecond frame. A low power level indicates that there is little active speech occurring in the frame. Segmentation breaks between utterances are set at the point of minimum power after smoothing over a one second window. To prevent unusually long segments, we force the system to place at least one break within each thirty seconds. This algorithm seems to be fairly robust in segmenting speech at silences or speaker changes. An empirical evaluation of the algorithm is in progress.

We can distinguish two distinct phases during News-On-Demand processing: library creation and library exploration. Library creation deals with the accumulation of information, transcription, segmentation and indexing. Library exploration concerns the interaction between the system and the user trying to retrieve selections in the database. The following section illustrates how the different technologies interact in the creation of a multimedia digital news library.

2. News-on-Demand Library Creation

Unlike other Informedia prototypes [Christel-94b,Christel-94c] which are designed to be test-beds for educational uses of the system, News-on-Demand focuses on the research goal of finding rapid and fully automatic methods for the creation of an indexed digital video library. While the educational content of the other Informedia prototypes allows time for careful computer-assisted human editing, the short-lived nature of news requires library creation and update for News-on-Demand to be completely automatic. In our early work on the Informedia digital video library, all segmentation was done by hand. Due to the time constraints of continuous daily news coverage and the volume of data, we are now using fully automatic news library creation methods.

The following steps are performed during library creation by a set of cooperating scripts and programs:

1. Digitize the video and audio data using MPEG-I compression format.
2. Create a time-aligned transcript from closed-captioning or from speech recognition output.
3. Segment shows at story boundaries.
4. Segment images by finding scene breaks and select key frames for each scene
5. Index all stories for access using the Informedia client program.

In more detail, these steps proceed as follows:

1. Our current library creation process starts with a raw digitized video tape. Generally, the videos in the *Informedia: News-on-Demand* Library are half-hour news shows selected from amongst the evening news broadcasts. Radio news shows such as *"All Things Considered"* or the news broadcasts from the Voice of America are also compressed and incorporated into the library. Using inexpensive off-the-shelf PC-based hardware we can compress video and audio to about 520 Mbytes/hour of video in MPEG-I format. The audio-only data is compressed to about 80 Mbytes/hour.
2. The audio portion of the signal is extracted and fed through the speech analysis routines, which produce a transcript of the spoken text. A subset of the news stories also have closed-captioning text available. However, the closed-captioned data, if available, may lag up to twenty-five seconds behind the actual words spoken. This problem, and general inaccuracies in transcription, are especially glaring when the broadcast is ``live''. In News-on-Demand we use speech recognition in conjunction with closed-captioning, when available, to improve the time-alignment of the transcripts. To create a time-aligned transcript from closed-captioned text, the speech recognizer output is aligned against the closed-caption transcript words using a orthographic distance measure with a standard dynamic programming algorithm. Through this

alignment each word in the closed-captioned transcript is assigned an accurate time marker derived from the corresponding word in the speech recognition output.

For the news broadcasts that are not closed-captioned, we use a transcript generated exclusively by the speech recognition system. The vocabulary and language model used here approximate a ``general American news'' language model. It was based on a large corpus of North American business news from 1987 to 1994 [Rudnicky95]. In addition, transcripts or closed-captioning text maybe obtainable for the video data. If there is a text transcript available during library creation, speech recognition helps create a time-aligned transcript of the spoken words as well as segmenting the broadcast into paragraphs. The library index needs very specific information about the start and end of each spoken word, in order to select the relevant video "paragraph" to retrieve and to find individual words within the story.

3. To allow efficient access to the relevant content of the news data, we need to break up the broadcasts into small pieces or news stories. To answer a user query by showing an entire half-hour long news show is rarely a reasonable response. Initially, the speech signal is analyzed for low energy sections that indicate acoustic "paragraphs" by the presence of silence. This is the first pass at segmentation; if a closed-captioned text transcript is available, we also use structural markers such as punctuation and paragraph boundaries to identify news stories. If only a speech recognition generated transcript is available, the acoustically determined paragraph boundaries identified using silence detection are used.

4. Image analysis is primarily used for the identification of breaks between scenes and for the identification of a single static frame icon that is representative of a scene. Primitive image features based on image statistics, such as color histograms, and their time functions, are computed and used for indexing, matching and segmenting images.

The Informedia library creation phase uses three different levels of segmentation for a video. The largest segment type consists of a "**news story**" — a series of related scenes with a common content. The system needs to determine the beginning and ending of an individual news story. In the ideal case, a news story starts at the natural boundary of the relevant content and ends wherever the video moves to a different context.

The second level of segmentation identifies an **individual scene** of video within the news story. Segment breaks produced by image processing are examined along with the boundaries identified by the speech and natural language processing of the transcript, and an improved set of segment boundaries are heuristically derived to partition the news story into scenes.

Finally, within a single scene we also need to be able to select a representative, characteristic **frame icon** for static display. Such a single frame is displayed as the representative for the whole video segment. This is used in a static display showing the results of a user query. Showing frame icons allows the user to look simultaneously at a static representation of multiple video paragraphs and to obtain some information about their content and possible relevance to the user's query, before selecting any one paragraph to be played. In choosing a static icon representative of a video clip, we rely exclusively on the image data. The paragraph bounds are determined by the transcript and keywords. Within the paragraph the most prominent keywords identify the most prominent scene. The scene boundaries are determined by color histogram differences and optical flow analysis. Within the scene we select the key frame icon using optical flow analysis.

5. The keywords and their corresponding paragraph locations in the video are indexed for inclusion in the Informedia library catalogue. An inverted index is created using the Pursuit search engine [Mauldin89]. To obtain video clips that are candidate matches to a user's query, the system searches for keywords from the query in the recognition transcript. When a match is found and selected by the user, the surrounding video paragraph is returned.

3. Library Exploration

Library exploration concerns the interaction between the system and the user trying to retrieve news material in the database.

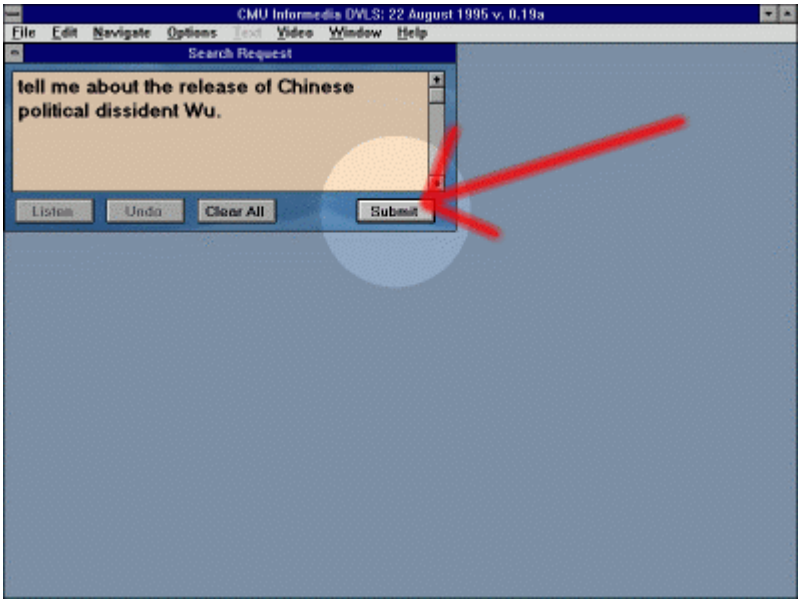

Figure 2. The user has spoken a query and is submitting it to the database search engine.

During library exploration the user generally goes through the following procedure, where steps may be repeated or skipped:

1) The user speaks a query to the system. If the query was misrecognized, the user may repeat or edit it by hand (see Figure 2).

2) The query is submitted and the database is searched for matches.

3) The user selects among the news stories returned as results, possibly refining the query first (see Figure 3).

4) A result story is selected for display (see Figures 4 and 5).

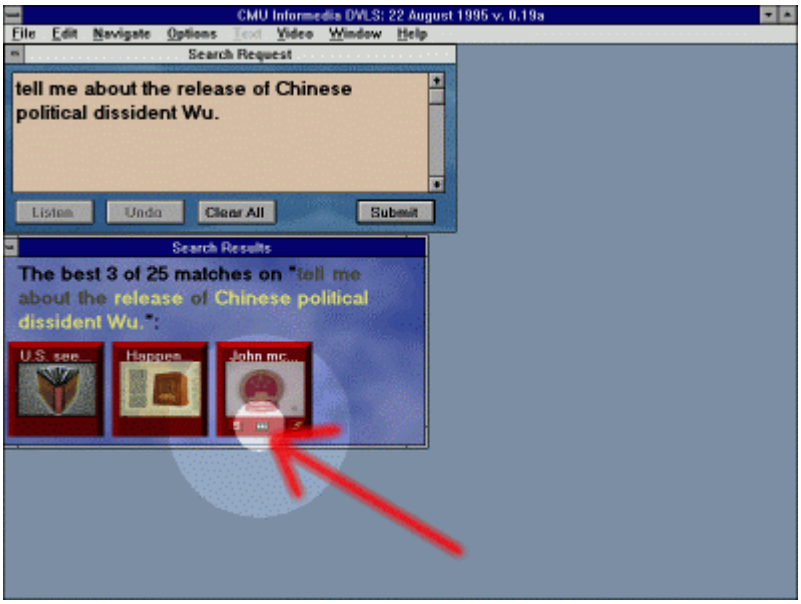

Figure 3.The best three icons matching the query have been displayed. The left result icon is a book, indicating a text source. The middle icon is a radio, indicating and audio-only source. The right, highlighted, icon is a key frame from a video news story. The user selects the "filmstrip" button at the bottom of the right icon.

1. **Speaking a query.** Users are able to explore the Informedia library through an interface that allows them to search the library using typed or spoken natural language queries, select relevant documents retrieved from the library and play or display the material on their PC workstations. The current interface for the Informedia system is very simple. There are four buttons and a query text window. The text window shows the result of the speech recognition and can be edited by selecting portions of the text and typing.

The LISTEN button must be held pressed while the user is speaking. Pushing the LISTEN button starts the recognition process and releasing the button signals the end of the query. In the future we hope to include a continuous listening mode, with the computer deciding when it is being spoken to, enabling us to eliminate this button.

The SUBMIT button sends the text that is currently in the query text window to the search retrieval component, which then returns the relevant matches.

The CLEAR button erases all text from the text query window.

The UNDO button erases just the last recognized utterance from the query text window. All earlier utterances are still part of the current query until they are erased with a CLEAR.

This simple interface seems sufficiently simple and intuitive to learn. However we are currently experimenting with even fewer buttons, where UNDO becomes unnecessary, or is replaced by a voice command, and every query is immediately submitted, eliminating the need for the separate SUBMIT button.

This library retrieval system can effectively process natural spoken queries. During library exploration, the Sphinx-II speech recognition allows a user to query the system by voice, simplifying the interface by making the interaction more direct. The integration of speech with the interface enhances access to the stored video data by allowing more immediate and direct entry of queries.

During library exploration, the Sphinx-II [Hwang94] speech recognition system allows a user to query the system by voice, simplifying the interface by making the interaction more direct. The integration of speech with the interface enhances access to the stored video data by allowing more immediate and direct entry of queries. The language model used during exploration is similar to that in [Rudnicky95] but emphasizes key phrases frequently used in queries such as "How about", "Tell me about", "Is there anything about", etc. In the future we plan to limit the language model to only those words found in the actual library data, making the language model more efficient and smaller

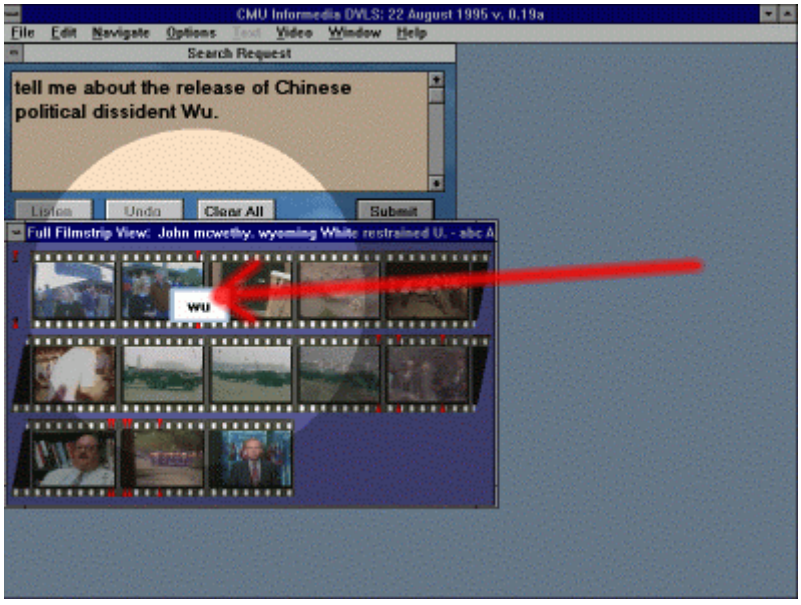

Figure 4. The filmstrip view displays the full news story with one key icon for each scene. Moving the cursor over the red marks the system places on the scene icon shows where the search query words were spoken in the scene.

2. **Submitting a query.** Figure 2 shows the user making the query *"tell me about the release of Chinese political dissident Wu."* The words *"tell"*, *"me"*, *"about"*, *"the"*, and *"of"* have been eliminated from the query as stop-words. The stop-words are function words selected from the most frequent query words. The keywords *"release"*, *"Chinese"*, *"political"*, *"dissident"* and *"Wu"* are used for the query. The query also looks for words with the same stems derived from a Houghton-Mifflin electronic dictionary of word roots (e.g. dissident and dissidents). Retrieval is done using an inverted index of the entire transcripts of all news stories, and each indexed story is ranked by relevance based on the frequency of the keywords occurring within the news stories.

3. **Selecting among the returned query results.** Figure 3 shows the results of the query display through 3 types of icons. A book icon indicates that the result is from a text source. The middle

icon indicates a radio only source. The right icon shows the key frame for a video story. The most relevant words from early in each story are displayed at the top of each icon. The user wants to find out more about the video story and clicks on the filmstrip button below the right thumbnail poster frame icon. Figure 4 shows the resulting filmstrip view. In this view, one key-frame thumbnail icon is displayed for each scene in the news story, as determined by the image analysis. Moving the mouse over the scene displays the keywords that were matched within the scene. The location of these words is also marked. These features help the user select among the returned results, and deliver relevant data about the video in the library in a compact format, using information embedded with the video during library creation.

4. **Playing or displaying a story**. Once the user has made a choice, and clicked on a story icon to play, the matching news story is displayed in the video window on the right half of the screen. Figure 5 also shows each word being highlighted as it is spoken. The transcript can be displayed below the video window, and automatically scrolls in sequence with the video.

Figure 5. The user now plays the news story from the chosen scene. A text transcript is displayed below the video on the right side. Each text word is highlighted as it is spoken.

4. News-on-Demand: Speech Recognition Facts and Results

Table 1 shows the results of testing recognition accuracy on a variety of video data. The results on our data show that the type of speech and the environment in which it was created dramatically alter the speech recognition accuracy.

Table 1: Speech Recognition Results

Type of Speech Data	**Word Error Rate** = Insertions + Deletions + Substitutions
1) Speech benchmark evaluation	~ 8% - 12%

2) News text spoken in lab	~ 10%- 17%
3) Narrator recorded in TV studio	~ 20%
4) C-Span	~ 40%
5) Dialog in documentary video	~ 50% - 65%
6) Evening News (30 min)	~ 65%
7) Complete 1-hour documentary	~ 75%
8) Commercials	~ 85%

1) The basic reference point is the standard speech evaluation data which is used to benchmark speech recognition systems with large vocabularies between five and sixty thousand words. The recognition systems are carefully tuned to this evaluation and the results can be considered close to optimal for the current state of speech recognition research. In these evaluations, we typically see word error rates ranging from eight to twelve percent depending on the test set. Note that word error rate is defined as the sum of insertions, substitutions and deletions. This value can be larger than one hundred percent and is regarded as a better measure of recognizer accuracy than the number of words correct. (i.e. words correct = 100% - deletions - substitutions).

2) Taking a transcript of TV broadcast data with an average reader and re-recording it in a speech lab under good acoustic conditions, with a close-talking microphone shows an estimate of word error rate between ten and seventeen percent for speech recognition systems that were not tuned for the specific language and domain in question.

3) Speech that has been recorded by a professional narrator in a TV studio and that does not include any music or other noise gives us an error rate of around twenty percent. Part of the increased error rate is due to poor segmentation of utterances leaving the speech recognizer unable to tell where an utterance started or ended. This problem was not present in the lab recorded data. Different microphones and environmental acoustics also contribute to the higher error rate.

4) Speech recognition on C-Span broadcast data shows a doubling of the word error rate to forty percent. While speakers in this data are mostly constant and always close to the microphone, other noises and verbal interruptions degrade the accuracy of the recognition.

5) The dialog portions of broadcast documentary videos yielded recognition word error rates of fifty to sixty-five percent, depending on the video data. The signal for these sections contains many more environmental noises, speech recorded outdoors.

6) The evening news was recognized with sixty five percent overall error rate. This rate includes recognition accuracy for commercials and introductions as well as the actual news program.

7) A full one-hour documentary video including commercials and music raised the word error rate to seventy-five percent.

8) Worst of all were commercials, which were recognized with an eighty-five percent error rate due to the large amounts of music in the audio channel as well as the unusual speech characteristics, and singing, contained in the spoken portion.

While these recognition results seem dismaying at first glance, they merely represent a first attempt at quantifying the usefulness of speech recognition for broadcast video and audio material.

Fortunately speech recognition does not have to be perfect to be useful in the Informedia digital video library.

The transcript generated by Sphinx-II recognition need not be viewed by users, but can be hidden. However, the words in the transcript are time-aligned with the video for subsequent retrieval. Because, generally, only the timing information from the speech recognition output is used directly, errors in recognition are not directly visible to users and our system can tolerate higher error rates than those that would be required to produce a human-readable transcript.

5. ISSUES FOR FUTURE RESEARCH

The research issues are split into five broad areas: data delivery, user interfaces, image understanding, natural language processing and speech recognition.

There are two main data delivery issues: How can we address the problem of huge storage requirements of the MPEG-I encoded video data accumulated through daily news broadcasts? An hour of video takes up about half a gigabyte of disk space. To populate the news on demand library we need to investigate how much data is necessary for the index to be kept current and when the data can be "forgotten". The data could be degraded to lower quality video at fewer frames per second, and lower resolution. We could also eliminate the video entirely and save only the audio portion. Finally we could retain only the text transcript without audio or video.

The second data delivery issue concerns the transmission of the video news story to a remote user. Essentially, we need to provide fast enough networks to allow MPEG-I bit rates to be transmitted continuously, and servers that can keep up with this demand for many simultaneous users.

The user interface issues deal with the way users explore the library once it is available. Can the user intuitively navigate the space of features and options provided in the Informedia: News-on-Demand interface? What other features should the system provide to allow users to obtain the information they are looking for? We plan to move the system to a test-bed deployment so that we can gain design insights from users and investigate various interface design alternatives.

Natural language processing research for News-on-Demand has to provide acceptable segmentation of the news broadcasts into stories. We also want to generate more meaningful short summaries of the news stories in natural sounding English. Natural language processing also has a role in query matching for optimal retrieval from the story texts. Finally, the system would greatly improve if queries could be parsed to separate out dates, major concepts and types of news sources.

Image processing research [HauptmannSmith95] is continuing to refine the identification of cuts in the video for scene segmentation. Within a scene and within a story, image processing gives us a key frame to represent that scene or story. The choice of a single key frame to best represent a whole scene is a subject of active research. In the longer term, we plan to add text detection and OCR capabilities for reading captions and text off the screen background. In the future, we also hope to include comprehensive similarity-based image matching in the retrieval features available to a user. An initial implementation of image matching is included in a current version of the system.

Speech recognition helps create a time-aligned transcript of the spoken words. While the use of a speech recognizer for transcribing all the audio data into text is a simple concept, we must consider the accuracy of the recognition system. Speech recognition is inherently error prone, and the magnitude of the error rate will determine whether the system is usable or useless. Thus recognizer accuracy is the critical factor in any attempt to use speech recognition for the digital video library. There are a many speech research issues that have been brought to light while studying this data and the uses of speech recognition in the News-on-Demand Library.

When the speech recognizer does not find a trigram (word triple) probability recorded in the language model for the current hypothesized word triplet, it uses bigrams in the language model, although with a probability penalty. Similarly, when an appropriate bigram cannot be found in the language model, individual word probabilities are used, again with a penalty. There were between one and four percent of the spoken words missing from the transcription lexicon. Since each missed word gives rise on average to one and a half to two word errors, this alone accounts for two to eight percent of the error rate. The word pairs (bigrams) in the language model were also inadequate. Depending on the data set, anywhere from eight to fifteen percent of the word pairs were not present in our language model. The trigram coverage gap was quite large. Between seventy and eighty percent of the trigrams in the data were not in the language model —these triples would consequently be recognized with a much lower probability.

By itself, the video library's unlimited vocabulary degrades recognition rate. However, several innovative techniques can be exploited to reduce errors. The use of program-specific information, such as topic-based lexicons and interest-ranked word lists can added to the information sources employed by the recognizer. Word hypotheses can be improved by using adaptive, "long-distance" language models, and we can use a multi-pass recognition approach that considers multi-sentence contexts. Recent research at CMU in long distance language models indicates twenty to thirty percent improvement in accuracy may be realized by dynamically adapting the vocabulary based on words that have recently been observed in prior utterances.

In addition, most broadcast video programs have significant amounts of descriptive text available. These include early the descriptions of the program design called treatments, working scripts, abstracts describing the program, and captions. Words that are likely to appear in the daily news can be obtained from many other sources of news such as the on-line wire services. In combination, these resources can provide valuable additions to dictionaries used by the recognizer. We are exploring the use of these sources of information in our current research.

Speech recognizers are very sensitive to different microphones and to the different environmental conditions in which their acoustic models were trained. Even microphone placement is a factor in recognizer accuracy. Much of the degradation of speech accuracy in the results of Table 1 between the lab and the broadcast data, using identical words, can be attributed to microphone and environment noise. We are actively looking a noise compensation techniques and microphone independence to ameliorate this problem. The use of stereo data from the left and right broadcast channel may also help in reducing the drop in accuracy due to environmental noise.

Perhaps the greatest benefit of speech recognition comes from alignment to existing transcripts or closed-captioning text. The speech recognizer is run independently and the result is matched using a forced text alignment against the transcript. Even though the recognition accuracy may only

provide one correct word in five, this is sufficient to allow the system to accurately find the boundaries of the story paragraphs, and the time at which the words were spoken within those boundaries.

Of course, a fully automated system like News-on-Demand will inevitably make errors. We distinguish five types of errors, all of which are subjects of active research aimed at improving the system:

1. False segmentation of stories. This happens when we incorrectly identify the beginning and end of a video paragraph associated with a single news story. Incorrect segmentation is usually due to inaccurate transcription, either from errors in the closed-captioning itself, errors in processing the closed-captioning text into stories, or errors in the segmentation based on the speech transcription.

2. False words in transcripts. Errors in the transcript are either the result of faulty speech recognition or errors in the closed-captioned text. The result is the appearance of incorrect words in stories and consequent errors in the index and in retrieval.

3. False synchronization. Occasionally words are retrieved that were actually spoken elsewhere in the video. This is generally due to closed-captioning mismatched with the speech recognition.

4. Incorrectly recognized queries. This is the result of an incorrect speech recognition during the library exploration. The user can edit and correct query recognition errors by typing, or simply repeat or rephrase the query.

5. Incorrect set of stories returned for a query. The prevalence of this type of error is measured through information recall and precision. The user might get stories that are not relevant to the query or miss relevant stories. Some of these errors are caused by previously mentioned problems, and others are the result of shortcomings in the processing of the query into retrieval keywords.

6. Conclusions

Despite the drawbacks of a fully automated system, the benefits of News-on-Demand are very dramatic. With News-on-Demand, we can navigate the complex information space of news stories, without the linear access constraint that normally makes this process so time consuming. Thus Informedia News-on-Demand provides a new dimension in information access to video and audio material. In the future, we plan to add OCR capabilities for reading headlines and image processing for visual scene segmentation to the News-on-Demand system.

Speech recognition is not a panacea for retrieval from video libraries. There is no “listening typewriter.” However even speech recognition with reasonable accuracy can be used to great effect in making accessible data that would otherwise be completely unavailable. Especially in conjunction with the use of transcripts or closed-captioning, speech recognition even at high error rates is tremendously useful in the digital video library creation process. For queries, the ability to

quickly correct and edit spoken commands makes the spoken query interface quite usable. Despite the drawbacks of errors in the system, the benefits of speech recognition are very dramatic.

Universal access to digitally processed news will significantly affect the conduct of business, professional, and personal activity. The initial impact of News-on-Demand will be on the broad accessibility and reuse of all standard news materials, including TV, radio and text, previously generated for broadcast or publication.

The greatest societal impact of the Informedia project is anticipated in K-12 education. The Informedia Digital Video Library represents a critical step toward an educational future that we can hardly recognize today. Ready access to multimedia resources will bring to the paradigm of "books, blackboards, and classrooms" the energy, vitality, and intimacy of "entertainment" television and video games. The persistent and pervasive impact of such capabilities can help to revolutionize education as we've known it, making it as engaging and powerful as the television students have come to love.

7. Acknowledgments

This material is based upon work supported by the National Science Foundation under Cooperative Agreement No. IRI-9411299. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

We would like to thank Michael Smith for his help with News-on-Demand image processing, Mike Christel for the Informedia library interface and Ravi Mosur for the fbs6/fbs8 Sphinx-II implementation. We are also indebted to all the members of the Informedia Project under Howard Wactlar for helping to digitize and record the data as well as the CMU Speech Group under Raj Reddy for all their support.

8. References

[Brown95] Brown, M.G., Foote, J.T., Jones, G.J.F., Sparck Jones, K., and Young, S.J. "Automatic Content-based Retrieval of Broadcast News," *Proceedings of ACM Multimedia*. San Francisco: ACM, November, 1995, pp. 35-43.

[Christel94a] Christel, M., Stevens, S., & Wactlar, H. "Informedia Digital Video Library," *Proceedings of the Second ACM International Conference on Multimedia*, Video Program. New York: ACM, October, 1994, pp. 480-481.

[Christel94b] Christel, M., Kanade, T., Mauldin, M., Reddy, R., Sirbu, M., Stevens, S., and Wactlar, H.","Informedia Digital Video Library", Communications of the ACM", 38 (4), April 1994, pp. 57-58.

[CNN-AT-WORK95] Cable News Network/Intel CNN at Work - Live News on your Networked PC Product Information. http://www.intel.com/comm-net/cnn_work/index.html.

[Flickner95] Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Huang, Q., Dom, B., Gorkani, M., Hafner, J., Lee, D., Petkovic, D., Steele, D., and Yanker, P. "Query by Image and Video Content: The QBIC System". IEEE Computer, September 1995, pp 23-31

[Hauptmann95] Hauptmann, A.G., Witbrock, M.J., Rudnicky, A.I. and Reed, S., *Speech for Multimedia Information Retrieval*, UIST-95, Proceedings of User Interface Software Technology, 1995, in press

[HauptmannSmith95] Hauptmann, A. G., and Smith, M., " Text, Speech, and Vision for Video Segmentation: The Informedia Project," *AAAI Fall 1995 Symposium on Computational Models for Integrating Language and Vision*, Cambridge, MA: MIT, November 1995, pp. 90-95.

[Hwang94] Hwang, M., Rosenfeld, R., Thayer, E., Mosur, R., Chase, L., Weide, R., Huang, X., and Alleva, F., "Improving Speech Recognition Performance via Phone-Dependent VQ Codebooks and Adaptive Language Models in SPHINX-II." *ICASSP-94*, vol. I, pp. 549-552.

[Informedia95] http://www.informedia.cs.cmu.edu/

[Li96] Li, W., Gauch, S., Gauch, J., and Pua, K.M., "VISION: A Digital Video Library", Digital Libraries '96: 1st ACM International Conference on Research and Development in Digital Libraries, Bethesda MD, March 1996.

[Mani96] Mani, I., House, D., Maybury, M. and Green, M. 1996. "Towards Content-Based Browsing of Broadcast News Video", in Maybury, M.T. (editor), Intelligent Multimedia Information Retrieval.

[Mauldin89] Mauldin, M. "Information Retrieval by Text Skimming," Ph.D. Thesis, Carnegie Mellon University. August 1989. Revised edition published as "Conceptual Information Retrieval: A Case Study in Adaptive Partial Parsing, Kluwer Press, September 1991.

[Maybury96] Maybury, M., Merlino, A., and Rayson, J., submitted 1996. "Segmentation, Content Extraction and Visualization of Broadcast News Video using Multistream Analysis", in Proceedings of the ACM International Conference on Multimedia, Boston, MA.

[Ogle95] Ogle, V. and Stonebraker, M. "Chabot: Retrieval from a Relational Database of Images", IEEE Computer, Vol. 28, No 9, September 1995.

[Pentland94] Pentland, A, Picard, R., Sclaroff, S., "Photobook: Tools for Content-Base Manipulation of Image Databases ". SPIE Conf. on Storage and Retrieval of Image and Video Databases II, (SPIE paper 2185-05) Feb 6-10, 1994, San Jose CA, pp34-47

[Rudnicky95] Rudnicky, A., "Language Modeling with Limited Domain Data," Proceeding of the 1995 ARPA Workshop on Spoken Language Technology, in press.

[CMU-Speech95] "http://www.speech.cs.cmu.edu/speech/"

[Salton83] Salton, G. and McGill, M.J. "Introduction to Modern Information Retrieval," McGraw-Hill, New York, McGraw-Hill Computer Science Series, 1983.

[Schauble95] Schauble, P. and Wechsler, M. "First Experiences with a System for Content Based Retrieval of Information from Speech Recordings," IJCAI-95 Workshop on Intelligent Multimedia Information Retrieval, M.Maybury (chair), working notes, pp. 59 - 69, August, 1995.

[Stevens94] Stevens, S., Christel, M. and Wactlar, H. "Informedia: Improving Access to Digital Video". *Interactions* **1** (October 1994), pp. 67-71

[Wactlar96] Wactlar, H.D., Kanade, T., Smith, M.A. and Stevens, S.M. "Intelligent Access to Digital Video: Informedia Project". *IEEE Computer*, **29**(5) May 1996, pp 46-52.

[Woods96] Woods, Bill, "Conceptually Indexed Video: Enhanced Storage and Retrieval" . http://www.sun.com/960201/cover/video.html

[Zhang95] Zhang, H., Low, C., and Smoliar, S. "Video parsing and indexing of compressed data," *Multimedia Tools and Applications* **1** (March 1995), pp. 89-111.