ADL-98 Advances in Digital Libraries Conference, Santa Barbara, CA, April 22-24, 1998.

Story Segmentation and Detection of Commercials In Broadcast News Video

Alexander G. Hauptmann
Department of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213-3890, USA
Tel: 1-412-348-8848
E-mail: alex+@cs.cmu.edu

Michael J. Witbrock
Justsystem Pittsburgh Research Center
4616 Henry St.
Pittsburgh, PA 15213, USA
Tel: 1-412-683-9486
E-mail: witbrock@justresearch.com

ABSTRACT

The Informedia Digital Library Project [Wactlar96] allows full content indexing and retrieval of text, audio and video material. Segmentation is an integral process in the Informedia digital video library. The success of the Informedia project hinges on two critical assumptions: that we can extract sufficiently accurate speech recognition transcripts from the broadcast audio and that we can segment the broadcast into video paragraphs, or stories, that are useful for information retrieval. In previous papers [Hauptmann97, Witbrock97, Witbrock98], we have shown that speech recognition is sufficient for information retrieval of pre-segmented video news stories. In this paper we address the issue of segmentation and demonstrate that a fully automatic system can extract story boundaries using available audio, video and closed-captioning cues.

The story segmentation step for the Informedia Digital Video Library splits full-length news broadcasts into individual news stories. During this phase the system also labels commercials as separate “stories”. We explain how the Informedia system takes advantage of the closed captioning frequently broadcast with the news, how it extracts timing information by aligning the closed-captions with the result of the speech recognition, and how the system integrates closed-caption cues with the results of image and audio processing.

KEYWORDS: Segmentation, video processing, broadcast news story analysis, closed captioning, digital library, video library creation, speech recognition.

INTRODUCTION

By integrating technologies from the fields of natural language understanding, image processing, speech recognition and video compression, the Informedia digital video library system [Wactlar96, Witbrock98] allows comprehensive access to multimedia data. News-on-Demand [Hauptmann97] is a particular collection in the Informedia Digital Library that has served as a test-bed for automatic library creation techniques. We have applied speech recognition to the creation of a fully content-indexed library and to interactive querying.

The Informedia digital video library system has two distinct subsystems: the Library Creation System and the Library Exploration Client. The library creation system runs every night, automatically capturing, processing and adding current news shows to the library. During the library creation phase, the following major steps are performed:

1. Initially a news show is digitized into MPEG-I format. The audio and video tracks from the MPEG are split out and processed separately, with their relative timing information preserved, so that derived data in the two streams can be resynchronized to the correct frame numbers.

2. Speech contained in the audio track is transcribed into text by the Sphinx-II Speech Recognition System [Hwang94]. The resulting text transcript also contains timing information for each recognized word, recording to within 10 milliseconds when it began and when it ended.

3. Images from the video are searched for shot boundaries and representative frames within a shot. Other video processing searches for and identifies faces and text areas within the image, and the black frames frequently associated with commercials.

4. If closed-captioning is available, the captions are aligned to the words recognized by the speech recognition step. This enables the timing information provided by the speech recognition system to be imposed on the closed captions, which are usually a more accurate reflection of the spoken words.

5. **The news show is segmented into individual news stories or paragraphs, allowing for information retrieval and playback of coherent blocks.**

6. Meta-data abstractions of the stories including titles, skims, film-strips, key frames for shots, topic identifications and summaries are created [Hauptmann97b].

7. The news show and all its meta-data are combined with previous data and everything is indexed into a library catalog, which is then made available to the users via

the Informedia client programs for search and exploration [Hauptmann97, Witbrock98, Witbrock97, Hauptmann95b].

This paper will describe, in detail, *step 5* of the library creation process described above. This is the procedure that splits the whole news show into story segments.

THE PROBLEM OF SEGMENTATION

The broadcast news is digitized as a continuous stream. While the system can separate shows by using a timer to start daily recording at the beginning of the evening news and to stop recording after the news, the portion in between is initially one continuous block of video. This block may be up to one hour long. When a user asks the system for information relevant to a query, it is insufficient for it to respond by simply pointing the user at an entire hour of video. One would expect the system to return a reasonably short section, preferably a section only as long as necessary to provide the requested information

Grice's maxims of conversation [Grice75] state that a contribution to a conversation should be as informative as required, as correct as possible and relevant to the aims of the ongoing conversation. A contribution should be clear, unambiguous and concise. The same requirements hold for the results returned in response to a user query. The returned segment should be as long as is necessary to be informative, yet as short as possible in order to avoid irrelevant information content.

In the current implementation of the News-on-Demand application, a "news story" has been chosen as the appropriate unit of segmentation. Entire segmented news stories are the video paragraphs or "document" units indexed and returned by the information retrieval engine in response to a query. When browsing the library, these news stories are the units of video presented in the results list and played back when selected by the user.

This differs from other work that treats video segmentation as a problem of finding scene cuts [Zhang95, Hampapur94] in video. Generally there are multiple scene cuts that comprise a news story, and these cuts do not correspond in a direct way to topic boundaries.

Work by Brown *et al* [Brown95] clearly demonstrates the necessity of good segmentation. Using a British/European version of closed-captioning called "teletext", various news text segmentation strategies for closed-captioned data were compared. Their straightforward approach looked at fixed width text windows of 12, 24, 36, and 48 lines in length[1], overlapping by half the window size. The fixed partitioning into 36 lines per story worked best, and the 48-line segmentation was almost as good. To measure information retrieval effectiveness, the authors modified the standard IR metrics of precision and recall to measure the degree of overlap between the resulting documents and the objectively relevant documents. In experiments using a very small corpus of 319 news stories, with 59 news headlines functioning as queries into the text, the average precision dropped to 0.538, compared with 0.821 for information retrieval from perfectly segmented text documents. This preliminary research indicated that one may expect a 34.5 % drop in retrieval effectiveness if one uses a fixed window of text for segmentation. Another metric showed that precision dipped to 0.407 with fixed window segmentation, from 0.821 for perfect text segmentation, a 50.5% decrease.

[Merlino97] presented empirical evidence that the speed with which a user could retrieve relevant stories that were well segmented was orders of magnitude greater than the speed of linear search or even a flat keyword-based search.

Achieving high segmentation accuracy remains, therefore, an important problem in the automatic creation of digital video and audio libraries, where the content stream is not manually segmented into appropriate stories. Without good segmentation, all other components of a digital video library will be significantly less useful, because the user will not be conveniently able to find desired material in the index.

RELEVANT RESEARCH

[Beeferman97] introduced a new statistical approach to automatically partitioning text into coherent segments. The proposed model enlists both long-range and short-range language models to help it sniff out likely sites of topic changes in text. To aid its search, the model consults a set of simple lexical hints it has learned to associate with the presence of boundaries through inspection of a large corpus of annotated text data. To date, this approach has not been extended to cover non-textual information such as video or audio cues. Beeferman also proposed a new probabilistically motivated error metric, intended to replace precision and recall for appraising segmentation algorithms. We use a modified version of this metric later in this paper.

[Hearst93] introduced the use of *text tiles* for segmentation of paragraphs by topic. Text tiles are coherent regions that are separated through automatically detected topic shifts. These topic shifts are discovered by comparing adjacent blocks of text, several paragraphs long, for similarity, and applying a threshold. The text tiling approach was initially used to discover dialog topic structure in text, and later modified for use in information retrieval. Unlike the segmentation approach presented here, it was based entirely on the words in the text. Our own preliminary experiments showed that the text tiling approach was not easily adaptable to the problem of combining multiple sources of segmentation information.

Yeung and her colleagues [Yeung96, Yeung96b] used an entirely image based approach to segmentation. Their video storyboard lays out a two dimensional flow of the scenes, indicating where the broadcast returns to the same scene. Video storyboards rely heavily on the detection of similar images in scenes. This is similar to the Video Trails idea presented by [Kobla97]. Video trails use the MPEG encoded features and map them into a three dimensional space.

[1] With the line lengths defined by the output of the teletext decoder.

Clusters of scenes in the same region of this space indicate that there is a return to the same scene after a digression to other scenes. This technique can easily identify the anchor narrating a story, a segment of reporting from the field, and the anchorperson returning in a later scene. The approach is, however, unable to distinguish the fact that the anchorperson is merely reading two consecutive stories, without intervening video. While the story board and video trails ideas are useful, work with large collections of broadcast news shows that text information provides important additional information which should not be ignored. None of the available papers on Video Trails or video storyboards have reported segmentation effectiveness in a quantitative way.

Perhaps the most similar research to that presented here is MITRE's Broadcast News Navigator. The Broadcast News Navigator (BNN) system [Merlino97, Maybury96, Mani96] has concentrated on automatic segmentation of stories from news broadcasts using phrase templates. The BNN system uses a finite state network with 22 states to identify the segmentation structure. In contrast to the approach presented here, the BNN system is heavily tailored towards a specific news show format, namely CNN Prime News. As a result, the system exploits the temporal structure of the show's format, using knowledge about time, such as the fact that a particular CNN logo is displayed at the start of the broadcast. The system also makes extensive use of phrase templates to detect segments. For example, using the knowledge that, in this show, news anchors introduce themselves at the beginning and the end of the broadcast, the system tries to detect phrase templates such as "Hello and welcome", "I'm <person-name>", which signal the introduction to the show. 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 of BNN 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 BNN system also aims to provide a deeper level of understanding of story content than is provided through simple full text search, by extracting and identifying, for example, all the named entities (persons, places and locations) in a story.

Our work differs from especially the latter in that we have chosen not to use linguistic cues, such as key phrases for the analysis of the segmentation. This allows our system to be relatively language independent. We also don't exploit the format of particular shows by looking for known timing of stories, special logos or jingles. However, like the BNN system we do exploit the format provided through the closed-captioned transcript. Unlike the BNN, we use a separately generated speech recognition transcript to provide timing information for each spoken word and to restore missing sections of the transcript that had not initially been captioned. In addition, we make extensive use of timing information to provide a more accurate segmentation. As a result, our segmentation results are relatively accurate at the frame level, which is not possible to achieve without accurate alignment of the transcript words to the video frames.

SOURCES OF INFORMATION FOR SEGMENTATION

The Informedia Digital Video Library System's News-on-Demand Application seeks to exploit information from image processing, closed-captioning, speech recognition and general audio processing. The following describes the features we derive from each of these sources. For each of the sources, there are features the Informedia system actually uses, and features the system could use but does not yet use. These latter features are being investigated, but are not yet sufficiently well understood to be reliably integrated into the production version of the Informedia News-on-Demand library.

We are not exploiting specific timing information or logos, or idiosyncratic structures for particular shows at particular times. Such cues could include detection of the CNN logo, the WorldView logo, the theme music of a show, the face of a particular anchorperson, and similar features specific to particular news show formats. Despite omitting these potentially useful markers, we have successfully applied our segmentation process to shows such as CNN World View, CNN Prime News, CNN Impact, CNN Science and Technology, Nightline, the McNeil-Lehrer News Hour, Earth Matters, and the ABC, CBS and NBC Evening News.

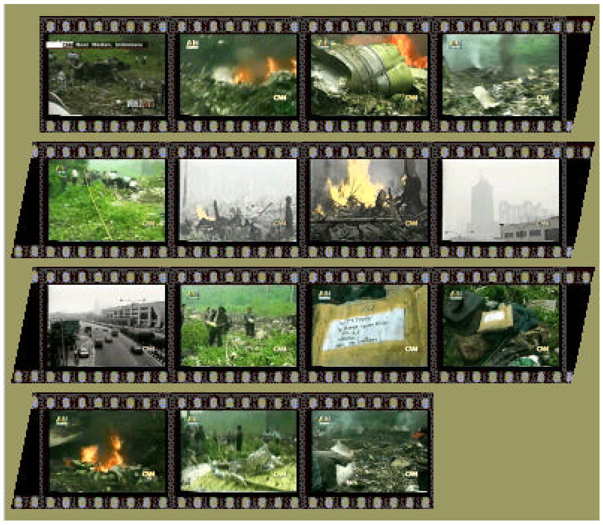

Figure 1: Scene breaks in the story whose transcript appears in Figure 2.

Image Information

The video images contain a great deal of information that can be exploited for segmentation.

Scene breaks. We define a scene break as the editing cut between individual continuous camera shots. Others have referred to these as shot breaks or cuts. Fade and dissolve effects at the boundaries between scenes are also included as

```
 1 >>> I'M HILARY BOWKER IN
 1 LONDON.
 2 CONFRONTING DISASTERS
 2 AROUND THE GLOBE.
 3 ASIAN SKIES WEIGHED DOWN
 3 UNDER A TOXIC BLANKET OF
 4 SMOKE.
 4 IS IT THE CAUSE OF A PLANE
 5 CRASH IN INDONESIA?
 6 >> I'M SONIA RUSELER IN
 7 WASHINGTON.
13 RARE DESERT FLOODING IN
14 ARIZONA THREATENS FAMILIES
15 AND THEIR HOMES.
20 [CLOSED CAPTIONING
21 PROVIDED BY BELL ATLANTIC,
21 THE HEART OF COMMUNICATION]
22
27 [CLOSED CAPTIONING PERFORMED
28 BY MEDIA CAPTIONING SERVICES
29 CARLSBAD, CA.]
44
44
45
46 >>> AN AIRLINE OFFICIAL IN
47 INDONESIA SAYS HAZE
47 PRODUCED BY RAMPANT FOREST
48 FIRES PROBABLY HAD A ROLE
48 IN AN AIRLINE CRASH THAT
49 KILLED 234 PEOPLE.
52 THE AIRBUS A-300, ON A
53 FLIGHT FROM JAKARTA,
55 CRASHED ON APPROACH TO
55 MEDON.
58 MEDON IS ON THE INDONESIAN
59 ISLAND, SUMATRA, CITE OF
60 MANY OF THE FIRES THAT HAVE
61 SENT A SMOKE CLOUD OVER SIX
62 COUNTRIES.
63 CNN'S MARIA RESSA IS IN THE
64 MALAYSIAN CAPITAL, KUALA
65 LUMPUR, AND FILED THIS
66 REPORT.
71 >>> RESCUERS FOUND ONLY
72 SCATTERED PIECES OF THE
73 AIRBUS 300.
```

breaks. However, we want to avoid inserting scene breaks when there is merely object motion or some other dramatic visual effect such an explosion within a single scene. Some of the more image oriented vision research has actually referred to the detection of scene breaks as video segmentation. This view differs dramatically from our interpretation of segmentation as the detection of story boundaries. In general, news stories contain multiple scene breaks, and can range in duration anywhere from 15 seconds to 10 minutes. Scene breaks, on the other hand, appear, on average, at intervals of less than 15 seconds. To detect scene breaks in the Informedia Digital Video Library System , color histogram analysis and Lucas- Kanade optical flow analysis are applied to the MPEG-encoded video [Hauptmann95]. This also enables the software to identify editing effects such as cuts and pans that mark shot changes. An example of the result of this process is shown in Figure 1. A variation of this approach [Taniguchi95] uses a high rate of scene breaks to detect commercials.

Black Frames. For technical reasons, commercials are usually preceded and followed by one or more frames that are completely black. When we can detect these blank frames, then we have additional information to aid in the segmentation of the news text. However, blank, or black frames also occur at other points during regular broadcasts. Because of the quality of the MPEG encoded analog signal, it may also not be possible to distinguish a very dark frame from a black frame. [Merlino97] also found black frames to be useful for story segmentation. Like so many of these cues, black frames are not by themselves a reliable indicator of segment boundaries. However, they provide added information to improve the segmentation process.

Frame Similarity. Another source of information is the similarity between different scenes. The anchor, especially, will reappear at intervals throughout a news program, and each appearance is likely to denote a segmentation boundary of some type. The notion of frame similarity across scenes is fundamental to both the Video Trails work [Kobla97] and to video storyboards [Yeung96, Yeung96b]. In the Informedia system we have two different measures of similarity available, each of the measures looks at key frames, a single chosen representative frame for each scene, and compares them for similarity throughout the news broadcast.

1. Color histogram similarity is computed based on the relative distribution of colors across sub-windows in the current keyframe. This color similarity is computed between all pairs of key frames in the show. The key frame that occurs most frequently in the top 25 similarity candidates and has the highest overall similarity score to others is used as the root frame. It and its closest matches will be used as the candidates for segmentation based on image similarity.

2. Face similarity is computed by first using CMU's face detection algorithm [Rowley95]. Any faces in all key frames are detected, and then these faces are compared using the eigenface technique developed at MIT [Pentland94]. Once a matrix of pair-wise similarity coefficients has been computed, we again select the most popular face and its closest matches as candidates for segmentation boundaries.

While each of these two methods is somewhat unreliable by itself, combining evidence from both color histogram and face similarity estimates gives a more reliable indication that we have detected a frame containing an anchor person.

MPEG optical flow for motion estimation: In the MPEG video stream there is a direct encoding of the optical flow within the current images [MPEG-ISO]. This encoded value can be extracted and used to provide information about whether there is camera motion or motion of objects within the scene. Scenes containing movement, for example, may be less likely to occur at story boundaries.

While we have experimented with all these cues for segmentation, in the Informedia News-On-Demand production system, we currently only exploit black frames and scene breaks. Figure 5 shows the various image features as well as corresponding manual segmentation. Not all features correlate well to the segments.

Closed-Captioned Transcripts

The closed caption transcript that is broadcast together with the video includes some useful time information. The transcript, while almost always in upper case, also contains syntactic markers, format changes between advertisements and the continuous news story flow. Finally useful information can be derived from both the presence and absence of closed captioning text at certain times during the video. Note that this information almost always contains some errors.

For the production Informedia Digital Video Library system, we currently exploit all of these cues in the closed captioned transcript. A sample closed-caption transcript is given below in Figure 2.

This transcript shows the actual transmitted caption text, as well as the time in seconds from the start of the show when each caption line was received. The particular format is specific to CNN, but similar styles can be found for captioning from ABC, CBS and NBC as well.

There are several things to notice in this closed-captioned transcript. First of all, there are ">>>" markers that indicate topic changes. Speaker turns are marked with ">>" at the beginning of a line. These markers are relatively reliable, but like anything done by people, are subject to errors. Most of these are errors of omission, but occasionally there are also spurious insertions of these discourse markers.

Secondly, notice that the text transcript is incomplete and contains typing errors, although these are relatively few in this example. The transcript also contains text that was not actually spoken (e.g. "[CLOSED CAPTIONING PROVIDED BY BELL ATLANTIC, THE HEART OF COMMUNICATION]").

```
3914    AN                 5340    KNOWN
3929    AIRLINE            5416    AND
3971    OFFICIALLY         5436    ON
4027    ENDED              5455    HIS
4056    ASIA'S             5469    ON
4093    SAYS               5487    HIS
4122    HEY                5540    SHOULD
4150    IS                 5554    ISLAND
4164    PRODUCED           5593    OF
4209    BY                 5600    SUMATRA
4230    RAMPANT            5658    CITE
4278    FOREST             5695    AN
4319    FIRE               5724    IDEA
4345    IS                 5753    THAT
4356    PROBABLY           5768    FIRES
4421    OUR                5822    THAT
4443    ROLE               5836    HAVE
4474    IN                 5850    SENT
4488    ANY                5881    US
4510    AIRLINE            5905    HOW
4550    CRASH              5929    CLOUD
4632    THAT               5964    OVER
4652    KILLED             5992    SIX
4700    TWO                6031    COUNTRIES
4727    HUNDRED            6133    C.
4768    AND                6152    N.
4793    THIRTY             6161    N.'S
4824    FOUR               6180    MARIA
4850    PEOPLE             6213    RESSA
4934    THE                6258    IS
4950    AIR                6273    IN
4966    WAS                6286    THE
5003    A                  6293    LOW
5023    THREE              6310    WAGE
5049    HUNDRED            6338    AND
5085    ON                 6351    CAPITAL
5099    A                  6392    QUELL
5107    FLIGHT             6435    INFORMED
5139    FROM               6506    AND
5154    JAKARTA            6518    HAS
5213    CRASHED            6539    FILED
5254    ON                 6575    THIS
5268    APPROACH           6595    RIFT
5307    TO
5324    THE
```

Figure 3. Sample of a speech recognizer generated transcript for the captions between 46 and 66 seconds in Figure 2. Times are in 10 millisecond frames.

There is a large gap in the text, starting after about 29 seconds and lasting until the 44-second mark . During this gap, the captioner transcribed no speech, and a short commercial was aired advertising CNN and highlights of the rest of the evening's shows. Although it is not visible in the transcript, closed-captioned transcripts lag behind the actually spoken words by an average of 8 seconds. This delay varies anywhere between 1 and 20 seconds across shows. Surprisingly, rebroadcasts of shows with previously transcribed closed-captioning have been observed at times to have the closed-captioning occur *before* the words were actually spoken on the broadcast video. Exceptional delays

```
4600 an 3914 AN             5500 medon 5324 THE
4600 an 3929 AIRLINE        5500 medon 5340 KNOWN
4600 an 3971 OFFICIALLY     5500 medon 5416 AND
4600 airline 3971 OFFICIALLY5800 medon 5436 ON
4600 official 4027 ENDED    5800 is 5455 HIS
4600 in 4056 ASIA'S         5800 on 5469 ON
4700 indonesia 4093 SAYS    5800 the 5487 HIS
4700 says 4122 HEY          5800 indonesian 5540 SHOULD
4700 haze 4150 IS           5900 island 5554 ISLAND
4700 produced 4164 PRODUCED 5900 sumatra 5593 OF
4700 by 4209 BY             5900 sumatra 5600 SUMATRA
4700 rampant 4230 RAMPANT   5900 cite 5658 CITE
4700 forest 4278 FOREST     5900 of 5695 AN
4800 fires 4319 FIRE        6000 many 5695 AN
4800 probably 4345 IS       6000 of 5724 IDEA
4800 had 4356 PROBABLY      6000 the 5753 THAT
4800 a 4421 OUR             6000 fires 5768 FIRES
4800 role 4443 ROLE         6000 that 5822 THAT
4800 in 4474 IN             6000 have 5836 HAVE
4800 an 4488 ANY            6100 sent 5850 SENT
4800 airline 4510 AIRLINE   6100 a 5881 US
4800 crash 4550 CRASH       6100 smoke 5905 HOW
4800 that 4632 THAT         6100 cloud 5929 CLOUD
4900 killed 4652 KILLED     6100 over 5964 OVER
4900 two 4700 TWO           6100 six 5992 SIX
4900 thirty 4727 HUNDRED    6200 countries 6031 COUNTRIES
4900 thirty 4768 AND        6300 cnn's 6133 C
4900 thirty 4793 THIRTY     6300 cnn's 6152 N
4900 four 4824 FOUR         6300 cnn's 6161 N.'S
4900 people 4850 PEOPLE     6300 maria 6180 MARIA
5200 the 4934 THE           6300 ressa 6213 RESSA
5200 airbus 4950 AIR        6300 is 6258 IS
5200 airbus 4966 WAS        6300 in 6273 IN
5200 a 5003 A               6300 the 6286 THE
5200 three 5023 THREE       6400 malaysian 6293 LOW
5200 hundred 5049 HUNDRED   6400 malaysian 6310 WAGE
5200 on 5085 ON             6400 malaysian 6338 AND
5200 a 5099 A               6400 capital 6351 CAPITAL
5300 flight 5107 FLIGHT     6400 kuala 6392 QUELL
5300 from 5139 FROM         6500 lumpur 6435 INFORMED
5300 jakarta 5154 JAKARTA   6500 and 6506 AND
5500 crashed 5213 CRASHED   6500 filed 6518 HAS
5500 on 5254 ON             6500 filed 6539 FILED
5500 approach 5268 APPROACH 6500 this 6575 THIS
5500 to 5307 TO             6600 report 6595 RIFT
```

Figure 4 Sample alignment of the closed captions and the speech recognition in the previous examples

can also occur when there are unflushed words in the transcription buffer, which may only be displayed after a commercial, even though the words were spoken before the commercial.

Finally, in the captions shown in Figure 2, there are formatting cues in the form of blank lines after the break at 44 and 45 seconds into the program, before captioning resumes at 46 seconds.

AUDIO INFORMATION

Speech Recognition for Alignment

To make video documents retrievable with the type of search available for text documents, where one can jump directly to a particular word, one needs to know exactly when each word in the transcript is spoken. This information is not available from the closed captioning and must be derived by other means. A large vocabulary speech recognition system such as Sphinx-II can provide this information [Hwang94]. Given exact timings for a partially erroneous transcription output by the speech recognition system one can align the transcript words to the precise location where each word was spoken, within 10 milliseconds.

To keep the processing speeds near real time for Informedia, the speech recognition is done with a narrow beam version of the recognizer which only considers a subset of the possible recognition hypotheses at any point in an utterance , resulting in less than optimal performance. This performance is, however, still sufficient for alignment with a closed-captioned transcript. Methods for improving the raw speech recognition accuracy when captioned transcripts are available before recognition are outlined in [Placeway96].

The basic problem for alignment is to take two strings (or streams) or data, where sections of the data match in both strings and other sections do not. The alignment process tries to find the best way of matching up the two strings, allowing for pieces of data to be inserted, deleted or substituted, such that the resulting paired string gives the best possible match between the two streams. The well-known Dynamic Time Warping procedure (DTW) [Nye84] will accomplish with a guaranteed least cost distance for two text strings. Usually the cost is simply measured as the total number of insertions, deletions and substitutions required to make the strings identical.

In Informedia, using a good quality transcript and a speech recognition transcript, the words in both transcripts are aligned using this dynamic time warping procedure. The time stamps for the words in the speech recognition output are simply copied onto the clean transcript words with which they align. Since misrecognized word suffixes are a common source of recognition errors, the distance metric between words used in the alignment process is based on the degree of initial sub-string match. Even for very low recognition accuracy, this alignment with an existing transcript provides sufficiently accurate timing information.

The Informedia system uses this information to aid in segmentation, allowing more accurate segment boundary detection than would be possible merely by relying on the closed captioning text and timings.

The actual speech recognition output for a portion of the story in Figure 2 is shown in Figure 3. An example alignment of the closed-captions to the speech recognition transcript is shown in Figure 4.

OTHER AUDIO FEATURES

Amplitude. Looking at the maximal amplitude in the audio signal within a window of one second is a good predictor of changes in stories. In particular, quiet parts of the signal are correlated with new story segments.

Silences. There are several ways to detect silences in the audio track. In the amplitude signal described above, very small values of the maximum segment amplitude suggest a silence. Low values of the total power over the same one second window also indicate a silence. Alternatively, one can use the silences detected by the speech recognizer, which explicitly models and detects pauses in speech by using an acoustic model for a “silence phone” [Hwang94]. Silences are also conveniently detected by the CMUseg Audio Segmentation package [CMUseg97].

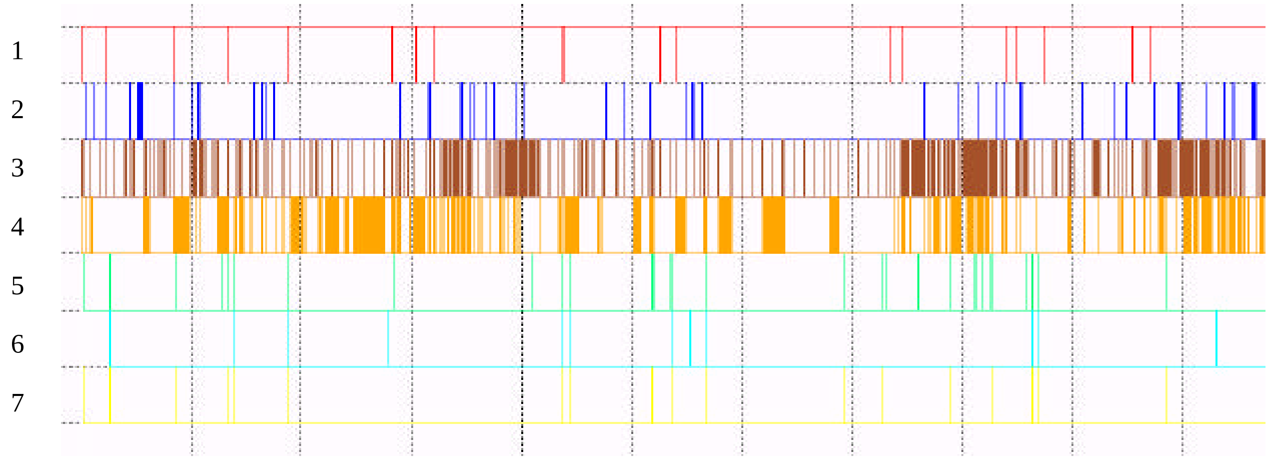

Figure 5: Image Features for Segmentation. The manually discovered segments (1) are at the top, and aligned underneath are MPEG optical flow (2), scene breaks (3), black frames (4), all detected faces (5), similar color image features (6), and similar faces (7) .

Acoustic Environment Change. Changes in background noise, recording channel [2], or speaker changes, for example can cause long term changes in the spectral composition of the signal. By classifying acoustically similar segments into a few basic types, the location of these changes can be identified. This segmentation based on acoustic similarity can also be performed by the CMUseg package.

Signal-to-Noise Ratio (SNR). Signal to noise ratio attempt to capture some of the effects of acoustic environment by measuring the relative power in the two major spectral peaks in a signal. While there are a number of ways to compute the SNR of an acoustic signal, none of them perfect, we have used the approach to SNR computation described in [Hauptmann97] with a window size of .25 seconds.

To date we have only made informal attempts to include this audio signal data in our segmentation heuristics. We will report the results of this effort in a future paper.

Figure 6 shows these audio features as they relate to perfect (human) segmentation.

We fully exploit the speech recognition transcript and the silences identified in the transcript for segmentation. Some of the other features computed by CMUseg are also used for adapting the speech recognition (SR) to its acoustic environment, and for segmenting the raw signal into sections that the recognizer can accommodate. This segmentation is not topic or news story based, but instead simply identifies short "similar" regions as units for SR. These segments are indicative of short "phrase or breath groups", but are not yet used in the story segmentation processing.

METHOD

This section will describe the use of image, acoustic, text and timing features to segment news shows into stories and to prepare those stories for search.

DETECTING COMMERCIALS USING IMAGE FEATURES

Although there is no single image feature that allows one to tell commercials from the news content of a broadcast, we have found that a combination of simple features does a passable job. The two features used in the simplest version of the commercial detector are the presence of black frames, and the rate of scene changes.

Black frames are frequently broadcast for a fraction of a second before, after, and between commercials, and can be detected by looking for a sequence of MPEG frames with low brightness. Of course, black frames can occur for other reasons, including a directorial decision to fade to black or during video shot outdoors at night. For this reason, black frames are not reliably associated with commercials.

Because advertisers try to make commercials seem more interesting by rapidly cutting between different shots, sections of programming with commercials tend to have more scene breaks than are found in news stories. Scene breaks are computed by the Informedia system as a matter of course, to enable generation of the key frames that represent a salient section of a story when results are presented from a search, and for the film strip view [Figure 1] that visually summarizes a story. These scene changes are detected by hypothesizing a break when the color histogram changes rapidly over a few frames, and rejecting that hypothesis if the optical flow in the image does not show a random pattern. Pans, tilts and zooms, which are not shot breaks, cause color histogram changes, but have smooth optical flow fields associated with them.

These two sources of information are combined in the following, rather *ad hoc* heuristic:

1. Probable black frame events and scene change events are identified.

[2] Telephones and high quality microphones, for example, produce signals with distinct spectral qualities.

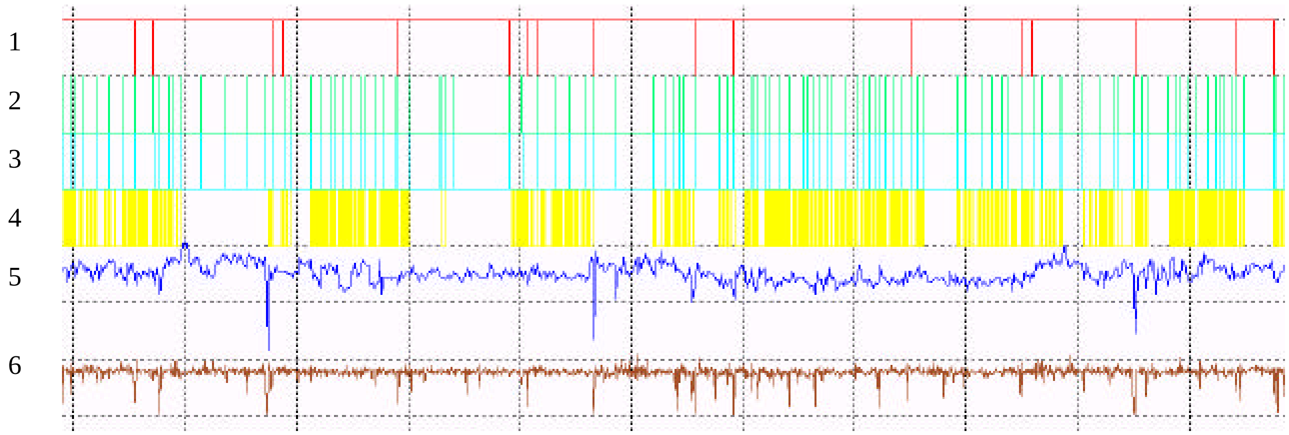

Figure 6: Audio Features for Segmentation. At the top are the manually found segments (1), followed by silences based on spectral analysis (2), speech recognition segments (3) silences in speech (4), signal-to-noise ratio (5), and maximum amplitude (6).

2. Sections of the show that are surrounded by black frames separated by less that 1.7 times the mean distance between black frames in the show, and that have sections meeting that criterion on either side, are marked as probably being commercials on the basis of black frames.

3. Sections of the show that are surrounded by shot changes separated by less than the mean distance between shot changes for the whole show, and are surrounded by two sections meeting the same criterion on either side, are marked as probably occurring in commercials on the basis of shot change rate.

4. Initially, a commercial is hypothesized over a period if either criterion 3 or 4 is met.

5. Short hypothesized stories, defined as non-commercial sections less than 35 seconds long, are merged with their following segment. Then short hypothesized ads, less than 28 seconds long, are merged into the following segment.

6. Because scene breaks are somewhat less reliably detected at the boundaries of advertising sections, black

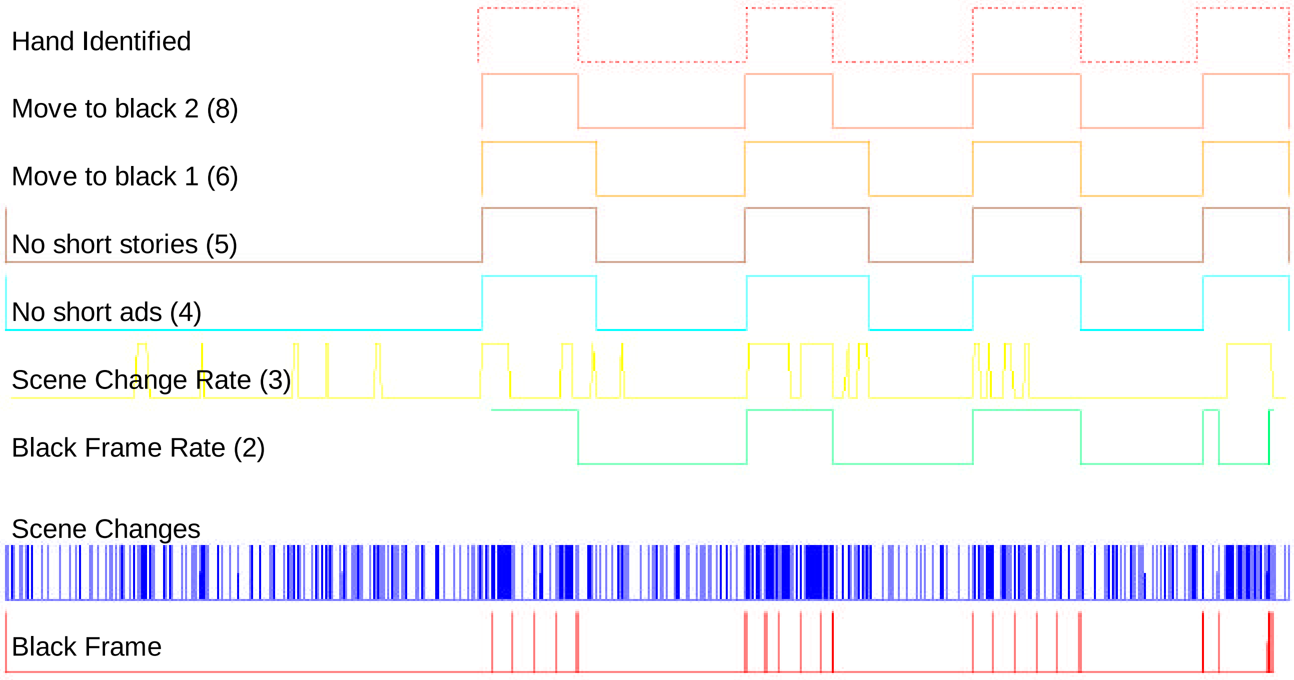

Figure 7: The commercial detection code in Informedia combines image features to hypothesize commercial locations. The bottom two graphs show the raw signals for black frames and scene changes respectively. The graph at the top shows the hand-identified positions of the commercials. The graphs running from bottom to top show successive stages of processing. The numbers in parentheses correspond to the steps outlined in the text.

frame occurrences are used to “clean up” boundaries. Hypothesized starts of advertisements are moved to the time of any black frame occurring up to 4.5 seconds before. Hypothesized ends of advertisements are moved to the time of any black frame appearing up to 4.5 seconds after the hypothesized time.

7. Because scene changes can also occur rapidly in non-commercial content, after the merging steps described above, any “advertising” section containing no black frames at all is relabeled as news.

8. Finally, as a sort of inverse of step 6, because some sections of news reports, and in particular weather reports, have rapid scene changes, transitions into and out of advertisements may only be made on a black frame, if at all possible. If a commercial does not begin on a black frame, its start is adjusted to any black frame within 90 seconds after its start and preceding its end. Commercial ends are moved back to preceding black frames in a similar manner.

Figure 7 shows how this process is used to detect the commercials in an example CNN news broadcast.

DETERMINING STORY BOUNDARIES.

The story boundaries are found in a process of many steps, which include the commercial detection process outlined above:

1. The time-stamped lines of the closed captioned text are first normalized. This normalization removes all digit sequences in the text and maps them into typical number expressions, e.g. 8 becomes eight, 18 becomes eighteen, 1987 becomes nineteen eighty-seven, etc. In order to be able to better match the text to the speech, common abbreviations and acronyms are also transformed into a normalized form (Dr. becomes doctor, IBM becomes I. B. M. *etc*) that corresponds to the output style of the speech recognition transcript. In the example in Figure 2, the number “234” at 49 seconds would be transformed into “two hundred and thirty-four”, “A-300” at 52 seconds would be transformed into “A three hundred”, and other conversions would be done similarly.”

2. The captioned transcript is then examined for obvious pauses that would indicate uncaptioned commercials. If there is a time gap longer than a threshold value of 15 seconds in the closed caption transmission, this gap is labeled as a possible commercial and a definite story segmentation boundary. Similarly, if there are multiple blank lines (three or more in the current implementation), a story boundary is presumed at that location. In Figure 2 above, story segmentation boundaries would be hypothesized at 0, 29, and 44 seconds.

3. Next the image based commercial detection code described in the previous section is used to hypothesize additional commercial boundaries.

4. The words in the transcript are now aligned with the words from the speech recognition output, and timings transferred from the speech recognizer words to the transcript words. After this alignment each word in the transcript is assigned as accurate a start and end time as possible based on the timing found in the speech.

5. Speech recognition output is inserted into all regions where it is available, and where there were no captioned text words received for more than 7 seconds.

6. A story segment boundary is assumed at all previously determined boundaries as well at the times of existing story break markers (“>>>”) inside the caption text.

7. Empty story segments without text are removed from the transcripts. Boundaries within commercials are also removed, creating a single commercial segment from multiple sequential commercials.

8. Each of the resulting segments is associated with frame number range in the MPEG video, using the precise speech recognition time stamps, and the corresponding text, derived from both captions and inserted speech recognition words, is assigned to the segment for indexing.

RESULTS

The actual automatic segmentation results for the data presented above are shown in Figure 8, with the manually generated reference transcript shown at the top.

One metric for segmentation proposed at the recent Topic Detection and Tracking Workshop [Yamron97,

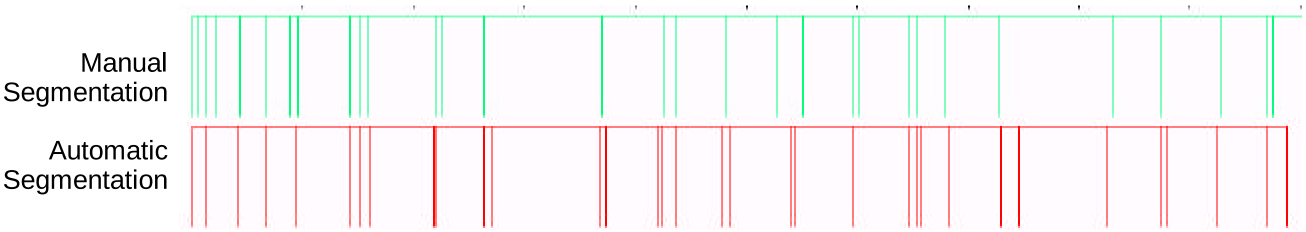

Figure 8: A comparison of manual segmentation with the automatic segmentation described in this paper shows very high correspondence between the segment markers.

Beeferman97] is based on the probability of finding a segment boundary between two randomly chosen words. The error probability is a weighted combination of two parts, the probability of a false alarm and the probability of a missed boundary. The error probabilities are defined as:

$$P_{Miss} = \frac{\sum_{i=1}^{N-k} d_{hyp}(i, i+k) \cdot (1 - d_{ref}(i, i+k))}{\sum_{i=1}^{N-k} (1 - d_{ref}(i, i+k))}$$

$$P_{FalseAlarm} = \frac{\sum_{i=1}^{N-k} (1 - d_{hyp}(i, i+k)) \cdot d_{ref}(i, i+k)}{\sum_{i=1}^{N-k} d_{ref}(i, i+k)}$$

Where the summations are over all the words in the broadcast and where:

$$d(i, j) = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are from the same story} \\ 0 & \text{otherwise} \end{cases}$$

The choice of k is a critical consideration in order to produce a meaningful and sensitive evaluation. Here it is set to half the average length of a story.

We asked three volunteers to manually split 13 TV Broadcast news shows at the appropriate story boundaries according to the following instructions:

> *"For each story segment, write down the frame number of the segmentation break, as well as the type with which you would classify this segment. The types are:*

P *"Beginning of PREVIEW".* The beginning of a news show, in which the news anchors introduce themselves and give the headline of one or more news stories. If each of 3 anchors had introduced 2 stories in sequence, there would be 6 beginning of preview markers.

T *"Beginning of searchable content: TOPIC".* This is the most typical segment boundary that we expect to be useful to an Informedia user. Every actual news story should be marked with such a boundary. Place the marker at the beginning of news stories, together with the frame number at that point. *Do include* directly preceding words, like "Thanks, Jack. And now for something completely different. In Bosnia today …"

S *"Beginning of searchable content: SUBTOPIC".* These subtopics mark the boundaries between different segments within the same news story. As a rule, whenever the anchorperson changes but talks about the same basic issue or topic as in the last segment, this is a subtopic. These subtopics are usually preceded by a phrase like "And now more from Bernard Shaw at the White House". Keep the "And now more from Bernard Shaw…" in the last segment.

C *"Beginning of COMMERCIAL".* This type of segment covers all major commercials with typical duration of 30 or 45-seconds up to one minute. The category also covers smaller promotional clips such as "World View is brought to you in part by Acme Punctuation Products" or "This is CNN!"

For evaluation purposes, we compared the set of 749 manually segmented stories from 13 CNN news broadcasts with the stories segmented from the same broadcasts by the Informedia segmentation process described above. The only modification to the manual segmentation done according to the instructions above was that multiple consecutive commercials were grouped together as one commercial block. On these 13 news broadcasts, the automatic segmentation system averaged 15.16% incorrect segmentation according to this metric. By comparison, human-human agreement between the 3 human segmenters averaged 10.68% error. The respective false alarm and miss rates are also shown in Table 1.

	P(err)	P(FalseAlarm)	P(Miss)
AutoSegment	**15.16%**	**7.96 %**	**26.9 %**
Inter-Human Comparison	**10.68%**	**8.26 %**	**15.35%**
No Segmentation	36.91 %	0.0 %	99.78 %
Segment every second	62.99 %	99.99 %	0.0 %
Segment every 1180 frames (Average Story Length)	60.86 %	91.08 %	9.32 %

Table 1: Performance of the Automatic Segmentation Procedure on evening news shows. Average Human segmentation performance is given for comparison. The results for no segmentation, segmentation every second and segmentation into fixed-width blocks corresponding to the average reference story length are given for reference.

DISCUSSION

Unfortunately, these results cannot be directly compared with either the results in [Brown95] or [Merlino97]. Brown *et al* used a criterion of recall and precision for information retrieval. This was only possible with respect to a set of information retrieval queries, and given the existence of a human relevance judgement for every query against every document. In our study, the segmentation effectiveness was evaluated on its own, and we have yet to evaluate its effect on the effectiveness of information retrieval.

[Merlino95] reported a very limited experiment in story detection with the Broadcast News Navigator. In effect, they only reported whether the main news stories were segmented or not, ignoring all minor news stories, commercials, previews, etc. Thus their metrics reflect the ability of the Broadcast News Navigator System to detect core news stories (corresponding only to segments labeled T and S by our human judges). The metrics for the BNN also ignored the timing of the story. Thus they did not take into account whether the detected story began and ended at the right frame compared to the reference story.

We have achieved very promising results with automatic segmentation that relies on video, audio and closed-captioned transcript sources. The remaining challenges include:

- The full integration of all the available audio and image features in addition to the text features. While we have discussed how various features could be useful, we have not yet been able to fully integrate all of them.

- Gaining the ability to automatically train segmentation algorithms such as the one described here and to learn similar or improved segmentation strategies from a limited set of examples. As different types of broadcast are segmented, we would like the system to automatically determine relevant features and exploit them.

- Completely avoiding the use of the closed-captioned transcripts for segmentation. While the closed-captioned transcripts provide a good source of segmentation information there is much data that is not captioned. We would like to adapt our approach to work without the captioned text, relying entirely on the speech recognizer transcription, the audio signal and the video images.

In the near term we plan to use the EM [Dempster77] algorithm to combining many features into one segmentation strategy, and to learn segmentation from data for which only a fraction has been hand-labeled. Work is also currently underway in the Informedia project to evaluate the effectiveness of the current segmentation approach when closed-captioning information is not available.

CONCLUSION

The current results provide a baseline performance figure, demonstrating what can be done with automatic methods when the full spectrum of information is available from speech, audio, image and closed-caption transcripts. The initial subjective reaction is that the system performs quite well in practice using the current approach. The future challenge lies in dealing with uncaptioned, speech-transcribed data, since the speech recognition generated transcript contains a significant word error rate.

The adequacy of segmentation depends on what you need to do with the segments. We are now in a position to evaluate the effectiveness of our segmentation process with respect to information retrieval, story tracking, or information extraction into semantic frames. Some approaches from the information retrieval literature [Kaszkiel97] claim that overlapping windows within an existing document can improve the accuracy of the information retrieval. It remains for future work to determine if a modification of this technique can circumvent the problem of static segmentation in the broadcast news video domain.

Segmentation is an important, integral part of the Informedia digital video library. The success of the Informedia project hinges on two critical assumptions: That we can extract sufficiently accurate speech recognition transcript from the broadcast audio and that we can segment the broadcast into video paragraphs (stories) that are useful for information retrieval. In previous papers [Hauptmann97, Witbrock97, Witbrock98], we have shown that speech recognition is sufficient for information retrieval of pre-segmented video news stories. In this paper we now have addressed the issue of segmentation and demonstrated that a fully automatic system can successfully extract story boundaries using available audio, video and closed-captioning cues.

ACKNOWLEDGMENTS

This paper is based on work supported by the National Science Foundation, DARPA and NASA under NSF Cooperative agreement No. IRI-9411299. We thank Justsystem Corporation for supporting the preparation of the paper. Many thanks to Doug Beeferman, John Lafferty and Dimitris Margaritis for their manual segmentation and the use of their scoring program.

REFERENCES

[Beeferman97] Beeferman, D., Berger, A., and Lafferty. J., *Text segmentation using exponential models*. In Proc. Empirical Methods in Natural Language Processing 2 (AAAI) '97, Providence, RI, 1997.

[Brown95] Brown, M. G., Foote, J. T., Jones, G. J. F., Spärck-Jones, K. and Young, S. J, *Automatic Content-Based Retrieval of Broadcast News,* ACM Multimedia-95, p. 35 - 42, San Francisco, CA 1995.

[CMUseg97] CMUseg, Carnegie Mellon University Audio Segmentation Package, ftp://jaguar.ncsl.nist.gov/pub/CMUseg_0.4a.tar.Z, 1997.

[Dempster77] Dempster, A., Laird, N., Rubin, D., Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society*, **39**, 1, pp. 1 – 38, 1977.

[Grice75] Grice, H. P. *Logic and Conversation*. In P. Cole (ed.) Syntax and Semantics. Vol. 3. New York: Academic Press. 41-58 , 1975.

[Hampapur94] Hampapur, A., Jain, R., and Weymouth, T., *Digital Video Segmentation*, ACM Multimedia 94, pp 357 – 364, ACM Int'l Conf on Multimedia, 15 – 20 Oct. 1994, San Francisco, CA.

[Hauptmann95] Hauptmann, A.G. and Smith, M.A. Text, Speech and Vision for Video Segmentation: the

Informedia Project. *AAAI Fall Symposium on Computational Models for Integrating Language and Vision*, Boston MA, Nov 10-12, 1995.

[Hauptmann95b] Hauptmann, A.G., Witbrock, M.J., Rudnicky, A.I., and Reed, S. *Speech for Multimedia Information Retrieval*, UIST-95 Proceedings of the User Interface Software Technology Conference, Pittsburgh, November 1995.

[Hauptmann97] Hauptmann, A.G. and Witbrock, M.J., Informedia News on Demand: Multimedia Information Acquisition and Retrieval, in Maybury, M.T., Ed, *Intelligent Multimedia Information Retrieval*, AAAI Press/MIT Press, Menlo Park, 1997

[Wactlar96] Wactlar, H.D., Kanade, T., Smith, M.A. and Stevens, S.M., Intelligent Access to Digital Video: Informedia Project. *IEEE Computer*, *29* (**5**), 46-52, May 1996. See also http://www.informedia.cs.cmu.edu/.

[Hauptmann97b] Hauptmann, A.G., Witbrock, M.J. and Christel, M.G. *Artificial Intelligence Techniques in the Interface to a Digital Video Library*, Proceedings of the CHI-97 Computer-Human Interface Conference New Orleans LA, March 1997.

[Hearst93] Hearst, M.A. and Plaunt, C., Subtopic structuring for full-length document access, in Proc ACM SIGIR-93 Int'l Conf. On Research and Development in Information Retrieval, pp. 59 – 68, Pittsburgh PA, 1993.

[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, 1994.

[Kaszkiel97] Kaszkiel, M. and Zobel, J., Passage Retrieval Revisited, pp. 178 – 185, SIGIR-97 Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Philadelphia, PA July 27 – 31, 1997.

[Kobla97] Kobla, V., Doermann, D., and Faloutsos, D., *Video Trails: Representing and Visualizing Structure in Video Sequences,* ACM Multimedia 97, Seattle, WA , Nov. 1997.

[Mani96] Mani, I., House, D., Maybury, M. and Green, M. "Towards Content-Based Browsing of Broadcast News Video", in Maybury, M. T. (editor), Intelligent Multimedia Information Retrieval, 1997.

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

[Merlino97] Merlino, A., Morey, D., and Maybury, M., Broadcast News Navigation using Story Segmentation, ACM Multimedia 1997, November 1997

[MPEG-ISO] International Standard ISO/IEC-CD-11172 *Information Technology – Coding of Moving Pictures & Associated Audio for Digital Storage,* International Standards Organization.

[Nye84] Nye, H. The Use of a One Stage Dynamic Programming Algorithm for Connected Word Recognition, *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. AASP-32, No 2, pp. 262-271, April 1984.

[Pentland94] Pentland A., Moghaddam B., and Starner T. View-Based and Modular Eigenspaces for Face Recognition IEEE Conference on Computer Vision & Pattern Recognition, Seattle, WA, July 1994.

[Placeway96] Placeway, P. and Lafferty, J., "*Cheating with Imperfect Transcripts",* ICSLP-96 Proceedings of the 1996 International Conference on Spoken Language Processing, Philadelphia, PA, October 1996.

[Rowley95] Rowley, H., Baluja, S. and Kanade, T.,, Human Face Detection in Visual Scenes. Carnegie Mellon University, School of Computer Science Technical Report CMU-CS-95-158, Pittsburgh, PA, 1995.

[Taniguchi95] Taniguchi, Y., Akutsu, A., Tonomura, Y., and Hamada, H., *An Intuitive and Efficient Access Interface to Real-time Incoming Video based on automatic indexing*, ACM Multimedia-95, p. 25 - 33, San Francisco, CA 1995.

[Witbrock97] Witbrock, M.J. and Hauptmann, A.G. Using Words and Phonetic Strings for Efficient Information Retrieval from Imperfectly Transcribed Spoken Documents, DL97, The Second ACM International Conference on Digital Libraries, Philadelphia, July 23 - 26, 1997.

[Witbrock98] Witbrock, M.J., and Hauptmann, A.G., Speech Recognition in a Digital Video Library, Journal of the American Society for Information Science (JASIS), 1998, *In press.*

[Yamron97] Yamron, J. "Topic Detection and Tracking: Segmentation Task", Topic Detection and Tracking (TDT) Worksho, 27-28 October 1997, College Park, MD. Also in BNTUW-98, Proceedings of the Broadcast News Transcription and Understanding Workshop, Leesburg, VA, February 1998.

[Yeung96] Yeung, M., and Yeo, B.-L., "*Time-constrained Clustering for Segmentation of Video into Story Units*" in International Conference on Pattern Recognition, August 1996.

[Yeung96b] Yeung, M., Yeo, B.-L., and Liu, B., *"Extracting Story Units from Long Programs for Video Browsing and Navigation*" in International Conference on Multimedia Computing and Systems, June 1996.

[Zhang95] Zhang, H.J., Low, C.Y., Smoliar S.W., and Wu, J.H., *Video Parsing, Retrieval and Browsing: An Integrated and Content-Based Solution*, ACM Multimedia-95, p. 15 – 24, San Francisco, CA 1995.