Automated Generation of News Content Hierarchy By Integrating Audio, Video, and Text Information

Qian Huang Zhu Liu* Aaron Rosenberg David Gibbon Behzad Shahraray
AT&T Labs - Research
100 Shulz Drive
Red Bank, NJ 07701
{huang,zliu,aer,dcg,behzad}@research.att.com

Abstract

This paper addresses the problem of generating semantically meaningful content by integrating information from different media. The goal is to automatically construct a compact yet meaningful abstraction of the multimedia data that can serve as an effective index table, allowing users to browse through large amounts of data in a non-linear fashion with flexibility, efficiency, and confidence. We propose an integrated solution in the context of broadcast news that simultaneously utilizes cues from video, audio, and text to achieve the goal. Some experimental results are presented and discussed in the paper.

1 Introduction

The amount of information generated in today's society is growing exponentially. Moreover, the data is made available in more than one dimension across different media such as video, audio, and text. This mass of multimedia information poses serious technological challenges in terms of (1) how large amounts of multimedia data can be efficiently stored and transmitted with acceptable quality and (2) how multimedia data can be integrated, processed, organized, and indexed in a semantically meaningful manner to facilitate effective retrieval. This paper addresses the second challenge.

When the amount of data is small, a user can retrieve desired content in a linear fashion by simply browsing the data sequentially. With the large amounts of data now available, and expected to still grow massively in the future, such linear searching is no longer feasible. What is needed is the capability of abstracting the essential content from the multimedia data and then forming an extremely compact yet meaningful representation of the data as a roadmap (or index table) for effective retrieval. The larger is the amount of data, the more abstraction is needed. In most cases, multiple layers or hierarchy of abstractions are needed. In order to automatically abstract multimedia data, the following issues must be addressed simultaneously: (1) integrating the data from different media and (2) organizing the data into retrievable units at semantically meaningful levels of abstraction.

This paper addresses these two issues in the context of broadcast news programs. Specifically, we are interested in automatically organizing broadcast news programs into a content hierarchy at various levels of abstraction via effective integration of video, audio, and text data available from the programs. The levels of abstraction hierarchy must be consistent with what users perceive as useful and meaningful (e.g., news stories, news summaries, news introduction, or commercials).

Studies have been reported in the literature addressing both media integration and automatic segmentation of semantically meaningful events [5, 4]. For media integration, studies have been aimed at extracting predefined sets of events, such as "touch-down" in a televised football game, by combining visual and audio data [5, 4]. In segmenting multimedia data into semantically meaningful units, some studies make use visual cues but most use textual information [2]. Multimedia information retrieval systems have been developed that archive national news programs in such a way that a user can later select specific segments to review. The retrievable units in these systems are often "scene cuts" segmented based on visual information. Although browsing based on such units can obviously save some time, much better effectiveness can be achieved if data is segmented into semantically meaningful units such as stories that are consistent with what users desire.

Users have different needs when searching for information. For instance, some users may want to directly retrieve a story; some may like to listen to the news summary of the day in order to decide which story sounds interesting before choosing what to listen to further; a user in advertising may have a totally different need: to track the commercials from competitors so that the company can quickly come up with a competing commercial. To satisfy different requirements, a segmentation mechanism is needed that partitions broadcast data in different

*The author works at AT&T as an intern student.

ways so that direct indices to the events at different levels of abstraction can be automatically established.

2 Integrated Solution to Automated Generation of Content Hierarchy

Our intention is to provide users the ability to retrieve broadcast news programs in a semantically meaningful way at different levels of abstraction. Segmentation algorithms are developed, aiming at automatically generating a content hierarchy illustrated in Figure 1. The

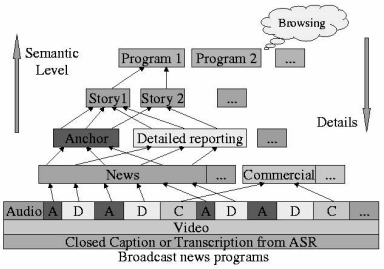

Figure 1: Content hierarchy of broadcast news programs.

lowest level contains the original multimedia data (audio, video, text). The next level separates news from commercials. Then the news is segmented into the anchorperson's speech and the speech from others (reporters, interviewees, etc.). Then, based on this information higher levels of semantics can be invoked to further segment the data into news stories and news summaries. In turn, each news story can be segmented into an introduction by the anchorperson followed by detailed reporting.

We observe that a typical national news program consists of news and commercials. News is composed of several headline stories, each of which is usually introduced and summarized by the anchor prior to and following the detailed reporting conducted by correspondents and others. Commercials are usually between different news stories. With this structure of the data, we propose an integrated solution to achieve automatic segmentation of news data into the content hierarchy shown in Figure 1 by utilizing cues from different media.

To separate news from commercials, audio and video information is combined (Section 3). Within each news segment, we further identify the anchorperson's speech based on speaker detection techniques (Section 4). Each segment of the anchor's speech is a hypothesized starting point for a new story. The audio-based processing results are then integrated with text-based information processing to obtain higher levels of semantically meaningful abstraction such as stories, story summaries, summary of the day, etc. (Section 5).

Our solution differs from existing methods in the following: (1) it combines visual/auditory/textual cues simultaneously; (2) instead of extracting a predefined set of isolated events, it generates different partitions over the entire news program with each partition corresponding to one level of abstraction in the hierarchy; and (3) the segmented content at different levels are semantically meaningful to users.

3 News/Commercials Separation Using Audio

News and commercials are separated based on audio measurements [3]. Nine acoustic features are extracted from audio clips: Non Silence Ratio (NSR), Standard Deviation of Zero crossing rate (ZSTD), Volume Standard Deviation (VSTD), Volume Dynamic Range (VDR), Volume Undulation (VU), 4 Hz Modulation Energy (4ME), Smooth Pitch Ratio (SPR), Non-Pitch Ratio (NPR), and Energy Ratio in Subband (ERSB). These features are chosen so that the underlying audio events (news vs. commercials) can be reasonably separated in the acoustic feature space. Clip level features are computed from frame level features, where each frame consists of 512 samples and adjacent frames are overlapped by 256 samples. Each clip is composed of a set of frames [3]. Three different classification methods were tested in separating news from commercials: linear classifier, fuzzy classifier, and GMM model based classification. Even though the classification is performed on each clip, the precise boundary between news and commercials (which can be in the middle of a clip) is determined by also considering the video processing results: the boundary cannot be in the middle of a scene cut. Experimental results of news/commercial separation are reported in Section 6.

Separating news from commercials is valuable in several aspects. Not only does it benefit the users who want to browse commercials, but it can also be used as a preprocessing step in Automatic Speech Recognition (ASR) to enhance the performance. We are currently using this tool to exclude commercials from ASR processing.

4 Anchor Identification

We use Gaussian mixure models (GMM) to perform text independent anchorperson recognition. The segmentation at this level partitions each news segment into anchor segments and everything else. The target speaker, background speakers, and other background audio categories are represented by 64 mixture component, diagonal covariance matrices. The models are constructed using 12 mel cepstral coefficients augmented by 12 delta cepstral features. A target speaker detection method based on likelihood ratio values evaluated from the models is developed with appropriate normalization and smoothing mechanisms [1].

Different training strategies were tested and compared. Benchmarking experiments against different thresholds were also conducted in order to choose the most effective system setting. Performance is measured at two different levels: segment level (the hit ratio of correct segment) and frame level measured (the hit rate of correct frames). Some of the experimental results are presented in Section 6.

Integrating multiple cues from different media can improve performance. For example, with audio information, locating the precise position where the anchor starts can be difficult because anchor often speaks with strong theme music in the background. By fusing visual information from the studio setting with audio information, the precise moment where the anchorperson's speech segment starts may be determined.

Segmentation at the anchor level provides a set of hypothesized story boundaries. (Typically each half-hour news program yields 13-15 segments of anchor speech of which 5-6 correspond to the beginning of a new story). Since not every anchor speech segment starts a new story, further analysis is needed to detect story boundaries. The set of starting points of anchor's speech correspondingly partitions the synchronized text data (either from closed caption or from an ASR) into blocks of text. Due to the structure of the broadcast data, even though each block does not necessarily represent one story, a new story cannont start in the middle of any block. Therefore, generating higher levels of abstraction (stories, summaries, etc.) helps to determine how these blocks of text can be merged to form semantically consistent content based on appropriate criteria.

5 News Story Extraction

Text-based discourse segmentation involves tokenization (the division of the input text into individual lexical units), grouping of processing units (granularity), similarity determination (lexical similarity between two blocks of text), and boundary identification (detection of significant lexical difference based on similarity scores). Both similarity criteria and grouping criteria affect the performance and the precision in discourse segmentation. Most work in the literature uses windows of pre-defined, fixed size for the grouping. The dilemma is that too small window size will make similarity comparison less effective and that too large window size can dramatically reduce the accuracy of identified boundaries.

We propose a grouping criterion based on audio cues. Since anchor-based segmentation has grouped our text input into blocks, in effect, (1) adaptive granularity can be achieved that is directly related to the content, (2) the hypothesized boundaries are more natural than those obtained using a fixed window, (3) blocks formed in this way not only contain enough information for similarity comparison but also have natural breaks of chains of repeated words if true boundaries are present, (4) the original task of discourse segmentation is achieved by boundary verification, and (5) once a boundary is verified, its location is precise. This grouping scheme of integrating audio based analysis provides an excellent starting point for the similarity analysis and boundary detection. Our experimental results show that, with this scheme, we can achieve much improved performance with a simplified solution.

With blocks of text, our task is to organize them into four classes: news stories, story introduction, augmented news stories, and news summary of the day. Our input data for text analysis is two sets of blocks of text: $T_1 = \{T_1^1, .., T_1^i, .., T_1^m\}$ where each T_1^k, $1 \leq k \leq m$, begins with the anchor person's speech; and $T_2 = \{T_2^1, .., T_2^j, .., T_2^n\}$ where each T_2^k, $1 \leq k \leq n$, contains only the anchor's speech. The blocks in both sets are all time stamped so that $T_2^k \subseteq T_1^k$. To find story boundaries, we evaluate the similarity $sim()$ between every pair (T_{b1}, T_{b2}) of adjacent blocks [2]: $sim(T_{b1}, T_{b2}) = \frac{\sum_w f_{w,b1} \times f_{w,b2}}{\sqrt{\sum_w f_{w,b1}^2 \times \sum_w f_{w,b2}^2}}$. Here, w enumerates all the token words in each text block; $f_{w,bi}$ is the frequency of word w in block bi, $i \in 1, 2$; and $0 \leq sim() \leq 1$. The higher the frequency of identical words in the blocks, the more similar are the blocks. We experimentally set up a threshold to determine the story boundaries.

In contrast to most studies in discourse segmentation where the processing is applied only to adjacent blocks of text, some of the classes we need require us to merge disconnected blocks of text. One example is the news summary of the day because the anchor's introduction to different headline stories may be scattered throughout the half-hour program. Therefore, after stories are segmented, we take set T_2 and the stories as input to further extract other classes. For each story, we extract its introduction by finding a T_2^k that has the highest similarity to that story (T_2^k is not necessarily connected) Merging each story with its introduction, we form an augmented story. The news summary of the day is extracted with the criterion that it has to provide the minimum coverage for all the stories reported on that day. Therefore, it is a set of T_2^k's that together covers all the stories of the day without overlap (i.e., each story has to be introduced but only once). With such a higher level of abstraction, users can browse desired information in a very compact form without losing primary content.

6 Experimental Results

The experimental data consists of 17 half-hour broadcasts of NBC Nightly News recorded off the air from January to February of 1998. 13 broadcasts are used for training and 4 are used for testing. The audio data is dig-

Table 1: Classification error rates using linear and fuzzy classifiers.

Error rate	test1	test2	test3	test4
Linear	6.0%	8.0%	9.6%	14.8%
Fuzzy	3.7%	7.9%	2.2%	6.5%

itized at 16 KHz sampling rate with 16 bits per sample resolution. The acquired data is manually labeled, segmented, and tagged according to target speaker, background speaker, commercials, music, noise, and silence. Additional descriptions are also provided for each segment including the gender and identity of the speaker as well as the assessed quality of the recording. Currently, the text source is closed caption text data accompanying the audio data.

In separating news and commercials, all the training data is used. The performance using linear and fuzzy classifiers on 4 test data sets is shown in Table 1. The average classification error rate is 9.6% using a linear classifier and 5.0% using a fuzzy classifier[3]. We are currently experimenting with a GMM model-based approach, in which the average classification error rate improves to 2.4% using 4 component mixtures.

In speaker identification, training data consists of 133 seconds from the target speaker, 275 seconds from other speakers, 468 seconds from commercials, and 63 seconds from music categories. Multiple models are constructed from the training data with 64 mixture components. The performance is measured at both frame and segment levels. At frame level, we use two measures: Target Miss Rate (TMR) and Non-target False-alarm Rate (NFR). At the segment level, we also use two measures: Segment Hit Rate (SHR) and Segment False-alarm Rate (SFR). The results are shown in Table 2. Test data set 3 does not contain any target speaker. The performance at segment level reaches average 92% of hit rate and average 1.5% of false alarm rate. Only one false alarm segment was detected over the 2 hour testing news program. The performance at frame level reaches average 8.7% of TMR (91.3% hit rate) and average 1% of NFR.

Table 2: Performance of speaker identification using GMM models.

Test data	TMR	NFR	SHR	SFR
test1	4.8%	0.8%	100.0%	0.0%
test2	13.8%	2.3%	82.4%	5.8%
test3	-	0.0%	-	0.0%
test4	7.6%	1.2%	93.3%	0.0%

The text analysis generates four classes: stories, augmented stories, story introduction, and the news summary of the day. The segmentation results on 2 hour test data so far has not yielded any incorrect segmentations. Currently, we have implemented a Web-Based browser to demonstrate how users can conveniently retrieve news broadcast data in a non-linear and semantically meaningful manner. With this browser, users can first choose a news program such as from NBC, CNN, ABC, CBS, etc. and then the date of the news program that they are interested in. Once a particular program is chosen, our system will tell users how many headline news stories were reported on that day with each story represented by a set of key words that reflects the content of the story. The news summary of the day is presented separately so that users can play back it first before they decide which story to browse further. Currently, we are experimenting with automatic construction of the visual representation of each story which involves choosing appropriate images from the video to compose a representation that is visually most relevant to the content of the story.

7 Concluding Remarks

This paper addresses the problem of automatically generating abstraction of content for multimedia indexing and retrieval in the context of broadcast news. We proposed an integrated approach to the problem. Our experimental results showed that (1) integrating data from different media can achieve what a single medium approach can not achieve, thereby attaining better performance, (2) abstracting multimedia data into semantically meaningful units significantly improves the quality of the extracted content, and (3) deriving well defined semantics from data makes it possible to form a hierarchical representation for the content that subsequently offers users an extremely compact index structure, allowing them to browse through large amounts of multimedia data with convenience, efficiency, and confidence.

References

[1] A.E.Rosenberg, I. Magrin-Chagnolleau, S. Parthasarathy, and Q. Huang. Speaker detection in broadcast speech databases. In *Proc. of International Conference on Spoken Language Processing*, Sydney, November 1998.

[2] Marti A. Hearst. Multi-paragrah segmentation of expository text. In *The 32nd Annual Meeting of the Association For Computational Linguistics*, pages 9–16, New Mexico, USA, June 1994.

[3] Zhu Liu and Qian Huang. Classification of audio events in broadcasr news. In *Proc. of IEEE Workshop in Multimedia Signal Processing*, December 1998.

[4] J. Nam and A. H. Tewfik. Combined audio and visual streams analysis for video sequence segmentation. In *Proc. of ICASSP*, volume 4, pages 2665–2668, 1997.

[5] Y.L.Chang, W. Zeng, I. Kamel, and R. Alonso. Integrated image and speech analysis for content-based video indexing. In *Proc. of Multimedia*, pages 306–313, Sept. 1996.