RC24319 (W0707-150) July 25, 2007
Computer Science

IBM Research Report

Semantic Analysis for Topical Segmentation of Videos

Youngja Park, Ying Li
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598

Semantic Analysis for Topical Segmentation of Videos

Youngja Park and Ying Li
IBM T.J. Watson Research Center, NY
{young_park, yingli}@us.ibm.com

Abstract

Topic segmentation of videos enables topic-based categorization, retrieval and browsing and also facilitates efficient video authoring. Existing video topic segmentation techniques, however, are domain specific to news or narrative videos while generic approaches based on video shot analysis generate too fine-grained micro-segments. This paper addresses this challenge through a multi-modal semantic analysis technique for recognizing topical segments.

We analyze the content of a video by using textual and audio features such as keyword synonym sets, sentence boundary information, silence/music breaks and speech similarity. Specifically, we propose a new natural language processing (NLP) technique for constructing synonym sets from video transcripts. A synonym set is a list of domain-specific keywords that are semantically related and represent a topic. We align the synonym sets with audio cues to identify the topical segments.

Our experiments with six instructional videos show that the system produced very small number of false positives, and the topical segments generated by our system are 5.5 times longer on average compared to those generated by a state-of-the-art micro-segmentation system. The system has been embedded in an e-Learning project, and the user feedback on using the generated topical segments is very encouraging. The experiments were conducted with instructional videos, but our approach is domain-general and is not restricted to instructional videos.

1 Introduction

The demand on automatic topic segmentation of videos is huge as the amount of videos available to the public is fast growing. Much work on video segmentation has focused on video shot detection or micro-segmentation [5, 1, 2, 6, 7]. While video shots are useful for some applications, they are not adequate units for video content analysis applications such as video search, classification or new video authoring. These high-level applications can be enhanced greatly with high quality topic segmentation tools.

Recently research on detecting high-level content semantics moving beyond individual shots has gained much attention. Phung *et al.* applied several probabilistic models to detect topically correlated units in educational and training videos [15]. Specifically, they built models based on pre-labeled shots with the help of heuristic knowledge rules. For instance, they assume that a topic usually starts with either direct-narration, assisted-narration or functional linkage, while the main body contains assisted-narration, voice-over or expressive linkage. These rules are, however, very specific to certain kinds of videos and are not generally applicable to other videos. Moreover, the two proposed states (*intro* and *main body*) are insufficient to capture the hierarchical storyline with nested topics and subtopics structure. Another effort was made by Chaisorn *et al.* which proposed to segment broadcast news videos into stories using a Hidden Markov Model (HMM) framework [3]. Similarly, the two-state HMM model was built upon shots which were pre-classified into 17 news categories including anchor, sports, finance, weather and commercial. While encouraging results were obtained on the TRECVID 2003 evaluation set, the proposed approach is highly adapted to news videos, and may not be effective for generic videos.

This paper presents semantic analysis techniques for recognizing topical segments in instructional videos. While most existing work along this direction attempts to detect high-level video units that are semantically as complete as possible, our work has a slightly different goal. The primary goal of our system is to provide topic-based video browsing and efficient authoring of new videos from existing segments. Therefore, the focus of this work is more on identifying topical segments where each of them solely concentrates on one single subject than on detecting a video unit that covers a complete thematic topic. In this paper, we call these topical segments atomic topical segments (ATS).

Automatic topic segmentation of video data requires understanding of semantic information in the given videos. The semantic information can be extracted from multimodal sources. Especially, text from the video sequence contains abundant semantic information such as keywords, named entities and events. However, most existing work in this area relies mostly on audiovisual features and only uses superficial textual cues such as texts shown in the video's text-overlay or several predefined cue words. Our approach, on the other hand, is based on semantic-level NLP techniques as well as audio-visual content analysis. Furthermore, the techniques we use are domain general, and are not dependent on the characteristics of particular videos.

Specifically, our system consists of three components; audiovisual content analysis component, text content analysis component and finally atomic segment detection component. The audiovisual content analysis component identifies the shots (or micro-segments) in a video by using state-of-the-art audio and video classification tools. It also annotates the shots with fifteen pre-defined content types such as "narrator presentation" and "informative text with voice-over". The text content analysis component first extracts domain-specific keywords from a video transcript and constructs synonym sets from the keyword set. It also generates sentence boundary information. Finally, the atomic topical segment detection component integrates all these cues together to identify topical segments. The overall framework of our system is depicted in Figure 1. In this paper, we focus on the text content analysis and briefly describe atomic topical segment components. More details on atomic topical segmentation can be found in [11], and on the audiovisual content analysis component in [8, 9] respectively.

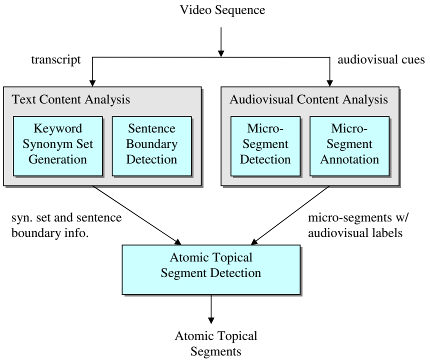

Video Sequence
transcript
audiovisual cues
Text Content Analysis
Keyword Synonym Set Generation
Sentence Boundary Detection
Audiovisual Content Analysis
Micro-Segment Detection
Micro-Segment Annotation
syn. set and sentence boundary info.
micro-segments w/ audiovisual labels
Atomic Topical Segment Detection
Atomic Topical Segments

Figure 1. Overall system framework for detecting topical segments in videos. The system takes a video sequence as the input and applies the audio-visual content analysis and the text content analysis units to produce micro-segments and keyword synonym sets respectively. As a final step, the atomic topical segment detection units combines the extracted information and produces topical segments.

For system evaluation, we used six instructional videos downloadable from various DHS (Department of Homeland Security) web sites. The videos are learning materials on terrorism, and the topics vary from bioterrorism history to school terror prevention. Our initial experiments with the six videos show promising results. The system merged, on average, five micro-segments into one topical segment, and produced very small number of false positives. Also, the user feedback on using the generated topical segments is very encouraging. The experiments were conducted with instructional videos, but our approach is domain-general and is not restricted to instructional videos.

2 Text Content Analysis for Topic Segmentation

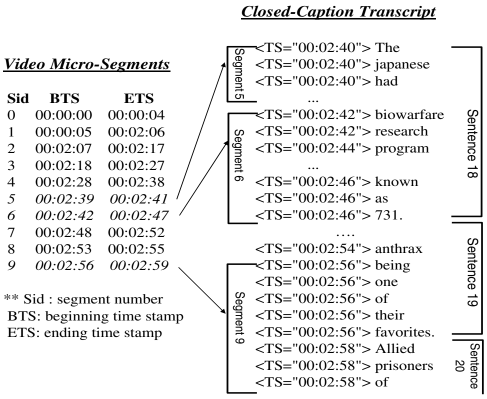
While audio and visual information is valuable source for topic analysis in a video, text from the video provides more abundant semantic information regarding to the topics mentioned in the video. We therefore emphasize semantic analysis of natural language text (i.e., video transcript). Our approach is based on the assumption that different topics are represented by different words. In other words, people use different set of words to discuss different topics, and we can detect the topic change by observing the word change.

The text content analysis comprises the following two steps. First, we identify sentence boundaries in the transcript and align video micro-segments with the sentences (section 2.1). Second, we extract domain-specific keywords from the transcript and build synonym sets from the keyword set. We also extract the occurrences of each synonym set across the micro-segments. The topic changes are identified by observing the distribution of the synonym sets over video segments (section 2.2 and section 2.3 respectively).

2.1 Sentence Boundary Detection and Video-Text Alignment

The closed-caption transcript is first extracted from a video, in which each word is associated with a time stamp. The text content analysis component uses a tokenization system to identify sentence boundaries in the transcript. The tokenization system uses the rules of punctuation to identify sentence boundaries. The system applies special disambiguation rules to distinguish a period used as a sentence end mark or as an abbreviation. We assign two time stamps to the sentences based on the word time stamps; the beginning time stamp and the ending time stamp. The beginning time stamp of a sentence is the time stamp of the first word in the sentence. Similarly, the ending time stamp of a sentence is the time stamp of the last word in the sentence.

As a separate step, the audiovisual content analysis component produces a file containing micro-segments along with the beginning time-stamp and the ending time-stamp [9]. The text content analysis component reads in the file and aligns the video micro-segments with words in the text based on the time-stamps in the following way. A video micro-segment sg is regarded ranging from word w_b to word w_e, if the time stamp of w_b is equal to or greater than the beginning time-stamp of sg and the time stamp of w_e is equal to or less than the end time-stamp of sg. Note that some micro-segments don't have corresponding text because these segments contain only non-speech sounds or transitional video shots. The alignment process is depicted in Figure 2.



Video Micro-Segments

Sid	BTS	ETS
0	00:00:00	00:00:04
1	00:00:05	00:02:06
2	00:02:07	00:02:17
3	00:02:18	00:02:27
4	00:02:28	00:02:38
5	*00:02:39*	*00:02:41*
6	*00:02:42*	*00:02:47*
7	00:02:48	00:02:52
8	00:02:53	00:02:55
9	*00:02:56*	*00:02:59*

** Sid : segment number
BTS: beginning time stamp
ETS: ending time stamp

Closed-Caption Transcript

<TS="00:02:40"> The
<TS="00:02:40"> japanese
<TS="00:02:40"> had
...
<TS="00:02:42"> biowarfare
<TS="00:02:42"> research
<TS="00:02:44"> program
...
<TS="00:02:46"> known
<TS="00:02:46"> as
<TS="00:02:46"> 731.
....
<TS="00:02:54"> anthrax
<TS="00:02:56"> being
<TS="00:02:56"> one
<TS="00:02:56"> of
<TS="00:02:56"> their
<TS="00:02:56"> favorites.
<TS="00:02:58"> Allied
<TS="00:02:58"> prisoners
<TS="00:02:58"> of

Figure 2. An example of the alignment of micro-segments, words and sentence boundaries in a video. We align micro-segments and words based on time stamps. As a separate step, we identify sentence boundaries and assign the beginning time stamp and the ending time stamp for each sentence. We then obtain the alignment of micro-segments and sentences.

2.2 Domain-specific Keyword Extraction

As we pointed out, our basic assumption for finding topical segments is that we can detect the topic change by observing how word usage change throughout the text or speech. Therefore, a critical process for understanding the topic is identifying keywords that are relevant to the topic (a.k.a., domain-specific keywords). Furthermore, people often use many synonymous words alternatively for referring to a same object or concept. Therefore, the recognition and aggregation of the synonymous words (i.e., synonym set construction) is also an important step.

We use a text-based keyword extraction system to extract domain-specific keywords in a given transcript. A brief description on the algorithm is given below (see [13] for details about the keyword extraction system). First, the algorithm identifies candidate keywords by applying syntactic grammars for recognizing noun phrases. It also recognizes different expressions for a term such as abbreviations, misspellings or alternative spellings, and orthographic variants. Orthographic variants include compound words (e.g., ”air bag” vs. ”airbag”) and hyphenated words (e.g., ”bioterrorism” vs. ”bio-terrorism”). All the different expressions are normalized into one canonical form. Next, the candidate keywords are ranked based on two pieces of statistical information — (1) the frequency of a keyword in the given text and (2) the relative probability of occurrences of a keyword in the given text against a general-purpose corpus (a.k.a. domain-specificity). Note that the statistical information is computed for normalized forms. The frequency of a term is the aggregated frequency of the canonical form and its variants. Finally, top-ranked keywords are selected as domain-specific keywords.

Having identified domain-specific keywords in the transcript, we then locate all the occurrences of the selected keywords in micro-sgements and generate keywords for each micro-segment. For instance, the system generates domain-specific keywords “offensive biowarfare research program” and “anthrax” from the transcript snippet shown in Figure 2, and their occurrences being segment 6 and segment 8 respectively.

2.3 Keyword Synonym Set Generation

In this work, we define two keywords w_i and w_j be synonymous if the two keywords satisfy one of the following conditions:

1. w_i is an abbreviation of w_j. For example, keywords “WMD’ and “Weapons of mass destruction” are synonymous.
2. w_i and w_j differ only orthographically. This includes cases of w_i having a special character such as a hyphen or a dash, and w_i being an alternative spelling or a misspelling of w_j. Examples include “antiterrorism” and ”anti-terrorism”.
3. w_i and w_j are two different names for an object or concept. For instance, “tularemia” and “rabbit fever” belong to this case.
4. w_i is a higher-level concept of w_j. Examples are “disease” and “anthrax”, and “country” and “congo”.

Please refer to [12] and [13] for detailed description on how to find the first two types of synonyms respectively. In this paper, we focus only on the last two types (3. and 4.) of synonyms.

There are two widely used approaches in building synonym sets in the area of NLP. One approach is based on statistics of word occurrences. In this approach, two words are regarded *similar* if the two words co-occur many times in a given document set, and are grouped together in a word cluster [14]. By repeating the clustering process, this approach can produce a hierarchy of word clusters. Statistical approaches, however, have the following restrictions:

(1) they require a lot of documents to obtain reliable statistical information; (2) they produce a set of related words but not always synonymous each other (e.g., "school", "student" and "class" in a cluster); (3) they create a hierarchy of word clusters but does not provide semantic relationships between two clusters.

The other approach is based on a lexical knowledge source such as a thesaurus. These lexical knowledge sources provide word meanings (or senses) and semantic relationships between words such as hypernym (or ISA) relationship. The major challenge in using a lexical knowledge base is that we need to know which sense of a word is used in a given context (a.k.a., word sense disambiguation) in order to get the definition or semantic relationships. Another problem is out-of-vocabulary (i.e., a word does not exist in the lexical knowledge source). The problem of out-of-vocabulary is more common when we analyze domain-specific documents.

We choose the second approach for identifying topical segments for the following reasons. First, it is unrealistic to expect to have a large collection of videos for a certain domain because of high production cost. Therefore, we can not expect reliable statistical information. Second, a learning video usually addresses one topic and consists of several sub-topics which all together explain the main topic. Therefore, fine-grained synonym sets are more desirable than coarse synonym sets for our task. In this work, we use *WordNet* as the lexical database [4]. *WordNet* is online lexical database that defines senses of words, groups senses into sets of synonyms called *synsets*, and connects the *synsets* with various semantic relationships such as *hypernym* and *part-holonym*. We address the two issues in using a lexical resource as follows.

We resolve the out-of-vocabulary problem in two different ways. First, if an out-of-vocabulary word is a multi-word term, and its substrings exist in the lexical source, we take the longest match strategy (i.e., substituting a multi-word keyword with the longest subset of the keyword found in the *WordNet*). For instance, *anthrax bomb* does not exist in the *WordNet*, but *bomb* exists in the database. In this case, we consider *anthrax bomb* is *bomb* and extract the senses of *bomb*. Second, if an out-of-vocabulary is a single word, then we create an artificial *synset* for the word with no semantic relationship with other *synsets*.

The second and more profound issue in using *WordNet* is sense disambiguation. In order to extract a sense of a word and its corresponding semantic relationships with other senses, we need to know which sense among all the senses of the word is used in the given context. In this work, we propose a new sense disambiguation method based on the "One Sense Per Discourse" hypothesis; it is very unusual to find two or more senses of a polysemous word in the same discourse [17]. The basic idea of our method is that the sense of a polysemous keyword is decided to be the most frequently used sense in the given transcript. Algorithm 1 describes our method for resolving word sense ambiguity.

```
Algorithm: Word Sense Disambiguation
1. Let W be the set of ambiguous keywords
      $\mathbf{W} = \{w_1, w_2, \cdots, w_m\}$;
2. Let S be the set of all senses for words in W
      $\mathbf{S} = \{s_1, s_2, \cdots, s_n\}, n > m$;
3. Compute the initial weight for each sense $s_i$ in S,
   $weight(s_i)$, based on the number of keywords which
   have $s_i$ as a sense and on the number of other senses
   with which $s_i$ has semantic relationships in the
   WordNet ;

      $weight(s_i) = \sum_{1 \le j \le m} f(w_j) + \sum_{1 \le k \le n} g(s_k)$

      $f(w_j) = \begin{cases} 1.5 & \text{if } s_i \text{ is the first sense of } w_j \\ 1 & \text{if } s_i \text{ is one of other senses of } w_j \\ 0 & \textit{otherwise} \end{cases}$

      $g(s_k) = \begin{cases} 1 & \text{if } s_k \text{ is a hypernym (hyponym) or} \\ & \text{a part-holonym (-meronym) of } s_i \\ 0 & \textit{otherwise} \end{cases}$

4. repeat
     4.1 Select a sense with the highest weight, $s_h$;
     4.2 Let $W$ be the set of keywords which have $s_h$
     as a sense, $W = \{w_1, w_2, \cdots, w_l\}$;
     4.3 $k = 1$;
     4.4 repeat
         4.4.1 Let $S$ be the set of senses of $w_k$;
         4.4.2 Discard all senses of $w_k$ except $s_h$;
         4.4.3 Recalculate the weights of the discarded
         senses, $weight(s)$;
         4.4.4 $k = k + 1$;
     until $k \le l$ ;
until there is no more ambiguous keyword ;
```

Algorithm 1: **Algorithm for word sense disambiguation. The correct sense of a polysemous word is decided based on the number of words which contain the sense and the number of other senses with which the sense has hyponym/hypernym or part-holonym/part-meronym semantic relationships**

Having decided the correct senses of all keywords in the transcript, we build synonym sets from the keyword set by using the two methods described below.

1. Global Merge: If two keywords w_i and w_j have a same sense (i.e., belong to a same *WordNet synset*), the two keywords are synonymous and are merged into a synonym set regardless of the distance of the two keywords in the transcript.

2. Local Merge: If the senses of two keywords w_i and w_j are linked through HYPERNYM (HYPONYM) or PART-HOLONYM (MERONYM) relationships in *WordNet* and the length of the path in the hierarchy is shorter than a predetermined threshold, ρ, the two keywords are considered as synonymous. In this case, the hypernym or part-holonym (i.e., more general sense) is merged into the closest hyponym or part-meronym (i.e., more specific sense) respectively. ρ is a threshold, and the value is empirically set to 5.

Table 1 shows several examples of the synonym sets. A synonym set consists of a list of domain-specific keywords and the micro-segments in which the keywords appear. For the rest of the paper, if synonym set S contains keyword w which is extracted from segment sg, we say that "segment sg belongs to synonym set S". Moreover, all segments in S will be sorted in temporal order and denoted by $\{sg_1, sg_2, ..., sg_n\}$. Note that the subscript i $(1 \leq i \leq n)$ here denotes a sequence, and does not mean the segment's actual index within the video. For instance, sg_1 could refer to segment 10, while sg_2 is actually segment 15.

Keywords in synonym sets and occurrences in micro-segments

disease {1;4;69;72}
illness {65}

anthrax {1;1;8;17;23;28;37;41;45;47;48;55;114}
dangerous anthrax {56}
liquid anthrax {116}
disease {8}

plague {1;66;70;77;79;81;84;87;88;136;142}
mysterious plague {128}
disease {69;72;128;129}
dangerous disease {65}
illness {65}
mass illness {129}

smallpox {1;90;94;96;96;99}
endemic smallpox {93}
disease { 92}

tularemia {1;114;132;132;133;136;141;142;143;148}
rabbit fever {132}
disease {132;133;147}
illness {141}
plague-like illness {132}

Table 1. This table shows five synonym sets constructed from a video on bioterrorism history. Each synonym set consists of more than one keywords and their occurrences in micro-segments in the video. The numbers denote micro-segment indices.

3 Atomic Topical Segment Detection

As defined in Section 1, an atomic topical segment (ATS) is both visually and aurally complete which implies that it shall not start or end in the middle of a sentence or a continuous visual scene. Moreover, it should strictly concentrate on one single subject without including any non-related materials. In align with these requirements, we apply an agglomerative approach to merge neighboring micro-segments into one topical segment under the condition that: 1) these micro-segments are semantically related to the same subject topic, and 2) they contain complete sentences. Keeping these requirements in mind, we perform the merging process by using the the following two steps.

3.1 Synonym Set-based ATS Detection

For each synonym set S, we perform a successive temporal clustering to conditionally group micro-segments that belong to this synonym set into one cluster. Specifically, assuming that S contains micro-segments $\{sg_1, sg_2, ..., sg_n\}$, we propose to merge $\{sg_i, ..., sg_{i+m}\}, 1 < i < i+m < n$, into one topical segment if the following three conditions are satisfied:

1. Any two neighboring segments within this group are temporally close to each other. Particularly, we require them to be either less than T_{sg} segments apart, or less than T_{sd} seconds apart. Here, T_{sg} and T_{sd} are two thresholds whose values are empirically set to be 5 and 90 based on experimental results. A better solution would be to determine them adaptively.
2. It does not contain a silent segment. This is mainly used to avoid merging segments that could possibly belong to different subjects. Based on our study, we notice that there is usually a silent transitional shot between two topical sections in instructional videos.
3. m is sufficiently large. Specifically, we require that m be larger than 4 so as to ensure that this synonym set indeed corresponds to a serious topical subject.

In a word, if similar keywords are frequently mentioned among a series of temporally adjacent segments, then very likely they are semantically related to the same subject topic and consequently should be merged together. One such example is shown here. Assume that S contains segments {6, 8, 9, 13, 30, 60, 64, 66, 69, 70, 80, 84}, then two topical segments could be derived from this set: one containing from segment 6 to 13, and the other from segment 60 to 70. Note that segments such as 7 and 10 that are not contained in set S will also be included into the corresponding topical segments accordingly. This step thus leads to a list of atomic topical segments where overlapped ones will be merged.

3.2 ATS Detection Using Other Cues

This step applies several other cues to detect more ATSs and also to refine existing ones.

Sentence Boundary Information: Since sentence forms the basic unit of thought in language, sentence boundaries are a good indicator for topic changes. However, the video micro-segmentation system often produces multiple micro-segments within a sentence or one micro-segment spanning across parts of two different sentences. We claim that no more than one topic exists in a sentence and propose to merge all micro-segments that are covered by the same sentence into one ATS. One such example is shown in Figure 2 where segments 5 and 6 are merged as they both belong to sentence 18.

Music Information: We notice that segments which share the same background music usually relate to the same topic. This is due to the fact that for most of professionally produced instructional videos, adding in music during the post-production is meant to reflect particular atmosphere under certain circumstance. We thus take advantage of this observation and apply it to the ATS detection.

Speech Information: Based on our study, instructional videos have frequent voice-over segments where a continuous background narration accompanies various visual materials such as historical documents, images and slides shown in the foreground. In this case, although the visual content keeps changing, yet the speech source remains constant, and thus the same topic continues. Consequently, all temporally adjacent micro-segments that share speech from a same narrator are merged together. To identify such scenes, we evaluate the speech similarity of any two neighboring voice-over segments by using their Kullback Leibler 2 (KL2) distance calculated based on Mel-frequency cepstral coefficient (MFCC) feature [16, 10]. A small KL2 distance indicates a continuous speech over the two segments. If the KL2 distance of two micro-segments is less than a preset threshold T_{sp}, we regard the two micro-segments to be acoustically similar, and subsequently merge them together.

Finally, all these three cues are fused together to either extend pre-detected ATS or identify new ones in an iterative manner. When no more change is observed, the ATS detection process ends.

4 Experiments and Performance Evaluation

Six videos downloaded from the web sites of FEMA (Federal Emergency Management Agency), CDC (Centers for Disease Control and Prevention) and other related DHS agencies, were used to evaluate the system. Lengthes of these videos range from 30 minutes to 120 minutes, yet a majority of them is of 1 hour long. All test videos contain various types of sounds, complex video content with frequent gradual content transitions, and various types of visual scenes such as classroom instruction, panel discussions, presentations, and outdoor activities.

Table 2 shows the transcript of an atomic topical segment which was created from a video on bioterrorism history. As we can see, the micro-segments are very short and often break a phrase in the middle, and thus making them difficult to understand the subject in isolation. The topical segment, however, contains a story and is more useful as an unit for search and content re-purposing.

As we pointed out in the introduction, the goal of this work is to facilitate users in quickly browsing video content organized by topical segments as well as conveniently authoring new clips. We are more tolerant to misses than false alarms, since missed topical segments could be easily combined using an editing tool. In this context, the traditional performance measurements such as precision and recall no longer serve as good evaluation criteria for this work, since it is rather difficult to determine the "hits" for a given video. We have thus defined a new measurement, namely, the *compression ratio* between the number of atomic topical segments (ATS) and the number of micro-segments, to evaluate the system performance. In addition, we consider *false alarm* (FA) as the second evaluation index, which flags when a detected ATS contains more than one subjects. Intuitively, a higher compression ratio with fewer false alarms would offer users a faster, easier and more pleasant content navigation and browsing experience.

The performance on ATS detection is shown in Table 3.

Video Data	No. of MS	Avg. Len.	No. of ATS	Avg. Len.	Comp. Ratio	FA
V 1	150	10.7	26	62	5.7	0
V 2	399	8.8	69	51	5.7	0
V 3	330	21.3	93	75.7	3.5	3
V 4	349	10.3	45	79.6	7.7	1
V 5	426	8.4	91	38.4	4.7	1
V 6	341	10.5	64	56.2	5.3	1
Avg.		**11.6**		**60.5**	**5.4**	**1**

Table 3. System performance evaluation on atomic topical segment detection, where MS, ATS and FA stand for micro-segment, atomic topical segment and false alarms, respectively. The average length of each segment is in unit of second.

As we see from the table, the average length of a video unit has now been increased from 11.6 seconds (for a micro-segment) to 60.5 seconds (for an ATS). This leads to a compression ratio of 5.4 on average. However, we did observe that video 3 has a relatively low compression ratio compared to others, which owes to its unusual story structure. Specifically, this video is in the form of three long panel discussions with each focusing on one particular topic. Each discussion starts with a presentation from one individual panelist, and subsequently goes on to another in a sequential

(5) The japanese had a very active offensive biowarfare research

(6) program which included a battalion known as 731.

(7) In their program, the japanese conducted experiments on humans, using 15 to 20

(8) different disease causing agents, with anthrax

(9) being one of their favorites. Allied prisoners of war and

(10) innocent manchurian civilians in

(11) nearby villages provided an almost endless supply of

(12) experimental subjects.

(13) When word of unit 731 leaked to

(14) the west, allied forces began their own programs

(15) concerned that japan and possibly germany would gain a military advantage in biowarfare research.

(16) Narrator: On the third day after exposure, the casualties begin. Dead sheep can be seen further down the line. It is of course necessary to

(17) confirm that they have died of anthrax.

(18) Cono: In 1942, on gruinard island off the coast of scotland, the british conducted

(19) their first scientifically controlled biowarfare field

(20) trials. Scientists exploded anthrax bombs near immobilized sheep to

(21) determine if the spores would survive an explosion and retain

(22) the ability to infect anyone

(23) nearby. Test results showed that anthrax could in fact be effectively disbursed by explosive devices and could also remain

(24) viable in the soil for decades. This brought home the realization that if an anthrax bomb were dropped on a city like london, the results could have been catastrophic.

(25) Gruinard island was declared off limits until it was decontaminated in the 1980s.

(26) It's now safe for both humans and animals.

(27) Like our allies, the united states responded to the perceived threats from germany and japan. In 1943, we began an offensive biological program with a modest research and development facility at camp dietrich, which

(28) is now fort dietrich, maryland. By the end of the program, we had weaponized a total of seven incapacitating or lethal human agents, including anthrax.

Table 2. An example topical segment from a video on bioterrorism history. Twenty four micro-segments (from segment 5 through segment 28) about Japanese biowarfare program during the second world war were merged into a topical segment.

manner. The discussion finally ends with a Question-and-Answer session which is modulated by the host. In particular, the host directs the questions, each of which focuses on one specific sub-topic, to the panel in a round robin fashion. As a result, each participant will address one particular aspect of the major topic, which naturally results in many topical segments.

Note that the false alarm (FA) rate is really low, which averages to 1 per video. The largest number of FAs is observed from video 3 where two of them are resulted from the failure of separating two sub-topics. Specifically, the first one is due to the mergence of sub-topic “when should a quarantine be ordered” with sub-topic “who makes the decision to order a quarantine”, and the second one with sub-topics “what happened after a quarantine is ordered” and “what are the challenges in quarantine”. The third FA is caused by the wrong mergence of segments that are on two different major topics. By watching the video, we found that these segments are mistakenly “bridged” by the host who introduces the next topic right after summarizing the previous one. Thus officially no segment could be correctly served as the delimiter. The other two false alarms in videos 4 and 5 are caused by the same problem where no transitional period exists between two topics. Finally, the false alarm in video 6 is resulted from the inaccuracy in one of the synonym set which contains both *site* and *hazardous waste site*. These two keywords, however, mean two different things in the video and thus belong to different topics (one on “dump site” and the other on “chemical waste site”).

We also conducted a same experiment by only using textual cues–synonym sets and sentence boundaries–to measure the impact of textual cues on video segmentation. Table 4 shows the experimental results. As we can see, the exclusion of the speech and music cues reduced false alarms. Specifically, this prevented the false mergence of two different sub-topics in video 3. However, the compression ratio is lower than that of shown in Table 3. The average length of ATSs created based only on synonym sets and sentence boundaries are about 17 seconds shorter on average. This fact suggest that the speech and music cues are also important features for topic segmentation.

Finally, we performed a few sessions of usability study

Video Data	No. of MS	Avg. Len.	No. of ATS	Avg. Len.	Comp. Ratio	FA
V 1	150	10.7	28	57.6	5.4	0
V 2	399	8.8	106	33.2	3.8	0
V 3	330	21.3	155	45.4	2.1	1
V 4	349	10.3	65	55.1	5.4	2
V 5	426	8.4	128	28	3.3	1
V 6	341	10.5	84	42.8	4.1	1
Avg.		**11.6**		**43.7**	**4**	**0.8**

Table 4. System performance evaluation on ATS detection without using the speech and music cues. We obtain lower compression ratio but also lower false alarms without the two cues.

for the system. Some examples of tasks that are specific to videos include: "*find a segment on the topic of botulism toxin and list its keywords*", "*locate the point where President Roosevelt is delivering a speech on anthrax*", as well as "*author a video that covers the topic of tularemia fever*". The participants we invited have no prior experience of using any video browsing/navigation/editing tools, and we limit our help to them as little as possible on using the presentation GUI. The study results are very encouraging, all participants finished the tasks within time limit and they answered all questions correctly. Moreover, they were very pleased with the presented video segment information, and gave us encouraging comments such as "cool", "I like it" and "wonderful".

5 Conclusion

This paper presents a new video segmentation scheme which structures videos into units of atomic topical segments. The primary goal of the system is to offer users convenient topic-based video browsing and flexible selection of topical segments to author a new video. Existing video shot detection systems generate too fine-grained micro-segments, while most video topic segmentation techniques are domain specific to certain types of videos. This paper addresses this challenge through a multi-modal semantic analysis technique for recognizing topical segments. Various information cues derived from a video are exploited to accomplish the task, which include keyword synonym set, sentence boundary information, silence/music break and speech similarity.

Experiments carried out on several training videos downloaded from various DHS websites have showed promising results. The system achieved compression ratio of 5.4 compared to a state-of-the-art micro-segmentation system, and showed very small number of false positives. The system has been embedded in an e-Learning project, and the users reported very positive feedback after they tried the new video navigation system based on topical segments. The new topic-based navigation enabled the users to find a video segment of interest in much shorter time, and substantially reduced the effort for merging many segments to create a new video clip.

References

[1] J. S. Boreczky and L. A. Rowe. Comparison of video shot boundary detection techniques. *Proc. of SPIE on Storage and Retrieval for Image and Video Databases*, 1996.

[2] J. S. Boreczky and L. D. Wilcox. A hidden markov model framework for video segmentation using audio and image features. *ICASSP'98*, pages 3741–3744, Seattle, May 1998.

[3] L. Chaisorn, T. Chua, C. Lee, and Q. Tian. A hierarchical approach to story segmentation of large broadcast news video corpus. *ICME*, 2004.

[4] C. Fellbaum. Wordnet: An electronic lexical database. 1998.

[5] B. Gunsel, A. Ferman, and A. Tekalp. Temporal video segmentation using unsupervised clustering and semantic object tracking. *IEEE Workshop on Applications of Computer Vision*, December 1996.

[6] J. Huang, Z. Liu, and Y. Wang. Integration of audio and visual information for content-based video segmentation. *Proc. of IEEE International Conference on Image Processing*, 3:526–529, Chicago, October 1998.

[7] H. Jiang, T. Lin, and H. J. Zhang. Video segmentation with the assistance of audio content analysis. *ICME'00*, New York, 2000.

[8] Y. Li and C. Dorai. SVM-based audio classification for instructional video analysis. *ICASSP'04*, 2004.

[9] Y. Li, C. Dorai, and R. Farrell. Creating MAGIC: System for generating learning object metadata for instructional content. *ACM Multimedia*, 2005.

[10] Y. Li and C.-C. Kuo. A robust video scene extraction approach to movie content abstraction. *International Journal of Imaging Systems and Technology: Special Issue on Multimedia Content Description and Video Compression*, 2004.

[11] Y. Li, Y. Park, and C. Dorai. Atomic topical segments detection for instructional video. *Proceedings of ACM Multimedia Conference*, 2006.

[12] Y. Park and R. Byrd. Hybrid text mining for finding terms and their abbreviations. *EMNLP*, 2001.

[13] Y. Park and R. Byrd. Automatic glossary extraction: Beyond terminology ddentification. *COLING*, 2002.

[14] F. Pereira, N. Tishby, and L. Lee. Distributional clustering of English words. In *Proceedings of the31st Annual Meeting of the ACL*, pages 183–190, 1993.

[15] D. Phung, T. Duong, S. Venkatesh, and H. Bui. Topic transition detection using hierarchical hidden Markov and semi-Markov models. *ACM Multimedia*, 2005.

[16] M. A. Siegler, U. Jain, B. Raj, and R. M. Stern. Automatic segmentation, classification, and clustering of broadcast news. *Proc. of Speech Recognition Workshop*, Chantilly, VA, February 1997.

[17] K. C. W. Gale and D. Yarowsky. One sense per discourse. *Proc. of the 4th DARPA Speech and NaturalLanguage Workshop*, 1992.