2005

Segmenting Lecture Videos by Topic: From Manual to Automated Methods

Ming Lin
University of Arizona, mlin@cmi.arizona.edu

Christopher B.R. Diller
University of Arizona, cdiller@cmi.arizona.edu

Nicole Forsgren
University of Arizona, nforsgren@cmi.arizona.edu

Yunchu Huang
University of Arizona, yunchu@eller.arizona.edu

Jay F. Nunamaker, Jr.
University of Arizona, jnunamaker@cmi.arizona.edu

Segmenting Lecture Videos by Topic: From Manual to Automated Methods

Ming Lin
University of Arizona
mlin@cmi.arizona.edu

Christopher B.R. Diller
University of Arizona
cdiller@cmi.arizona.edu

Nicole Forsgren
University of Arizona
nforsgren@cmi.arizona.edu

Yunchu Huang
University of Arizona
yunchu@eller.arizona.edu

Jay F. Nunamaker, Jr.
University of Arizona
jnunamaker@cmi.arizona.edu

ABSTRACT

More and more universities and corporations are starting to provide videotaped lectures online for knowledge sharing and learning. Segmenting lecture videos into short clips by topic can extract the hidden information structure of the videos and facilitate information searching and learning. Manual segmentation has high accuracy rates but is very labor intensive. In order to develop a high performance automated segmentation method for lecture videos, we conducted a case study to learn the segmentation process of humans and the effective segmentation features used in the process. Based on the findings from the case study, we designed an automated segmentation approach with two phases: initial segmentation and segmentation refinement. The approach combines segmentation features from three information sources of video (speech text transcript, audio and video) and makes use of various knowledge sources such as world knowledge and domain knowledge. Our preliminary results show that the proposed two-phase approach is promising.

Keywords

Lecture video, video segmentation, segmentation features, knowledge bases

INTRODUCTION

Video is very useful for knowledge sharing and learning because of its ability to carry and transmit “rich” information (Daft and Lengel, 1986). Nowadays more and more universities and corporations are starting to provide videotaped lectures online (Stanford online). However, people often have difficulties finding specific information in lecture videos because of the unstructured and linear nature of video. The nature of video hides the information structure of its contents and provides an unfriendly interface for learning. For instance, if a student wants to learn about a specific topic (e.g., “what is a constructor?” in a Java programming class), he may have to spend hours searching for the relevant video and then actually watch the whole thing.

Segmenting long videos into small clips (e.g., 5-10 minutes) by topic and listing the topics makes information searching much easier. From a learning perspective, the “Advanced organizer” approach (Ausubel, 1960) indicates that presenting outlines of information can help students “see the big picture” and therefore enhance learning. Furthermore, previous research (Cao et al., 2003) shows that segmenting lecture videos into topics and providing an outline enables more effective learning because it allows learners to jump from one topic to another easily and thus control their own learning pace. In both the Stanford online system and the LBA system (a multimedia-based learning systems developed at the University of Arizona) (Zhang, 2002), each lecture video is segmented into short clips by topic and a list of topic titles is provided as an outline.

Although segmenting lecture videos into topics is beneficial, the actual segmentation process itself is not an easy task. While manual segmentation provides the most accurate results and is usually used as a benchmark for segmentation evaluation

(Allan et al., 1998), it is very time consuming. Automated video segmentation can save labor time, but its accuracy is still far from optimal. In addition, existing video segmentation methods are not suitable for lecture videos. The most commonly used video segmentation methods focus on scene and shot changes (Wactlar, 2000; Zhang and Smoliar, 1994). Lecture videos, however, usually have very few scene and shot changes. For instance, in many situations there is only a "talking instructor" in the video. Furthermore, the scene and shot changes usually do not match the topic transition boundaries. Works in the broadcast and news domain (Allan et al., 1998) address the story segmentation problem, but its larger granularity makes it a relatively easy task compared with topic segmentation. Furthermore, topic boundaries in lecture videos are much more subtle because of the spontaneous speech of instructors and their various instructional styles.

The objective of our research study, therefore, is to design an automated segmentation method suitable for lecture videos with high accuracy. To achieve accuracy rates as high as manual segmentation, it is critical and beneficial to study how humans, especially experts, perform the video segmentation manually. The rules and heuristics that humans use in their segmentation process could be used as foundations for the design of the automated method. This paper investigates how humans perform manual segmentation and applies the collected rules to the design of an automated segmentation method for lecture videos. The rest of this paper is organized as follows: The next section reviews existing automated video segmentation techniques. We then describe a case study about manual segmentation. Based on the rules and heuristics found from the case study, we propose the design of an automated lecture video segmentation approach. We also show a preliminary evaluation of the proposed method. Finally we conclude our research and outline some future research directions.

RELATED RESEARCH

In this paper, we classify the segmentation methods based upon the input sources (video, audio or text) from which the segmentation features were extracted. We will also review domain-independent text segmentation literature because it provides possible methods for video segmentation when the input source is text.

Segmentation Using Video Input

The most commonly used video segmentation methods based on video input focus on detecting shot and scene changes. Wactlar (2000) used color histogram distance computation between successive images to detect scene changes. Zhang and Smoliar (1994) proposed a method for progressive transition detection by combining both motion and statistical analysis. As mentioned in the previous section, the methods that detect shot/scene changes are inapplicable for lecture-type videos because there are few shot/scene changes and there is often no mapping between shot/scene changes and topic changes. Research focusing on lecture video segmentation uses a method of finding topic changes by detecting lecture slide changes in the video images (Mukhopadhyay and Smith, 1999; Ngo et al., 2003). However, these methods assume that slides are present in all frames of the video, which is not true for most lecture videos. Furthermore, slides may not exist in a lecture video because many instructors do not use them.

Besides image cues, researchers have also explored the relationship between human gestures and discourse topic structure (Quek et al., 2000; McNeil et al., 2001). They look at many different kinesics such as hand gestures, gaze and posture shift. They found that, for instance, posture shifts occur more frequently at segment boundaries. However the research has never been applied to lecture videos.

Segmentation Using Audio Input

Audio data is often a rich source of information in a lecture video because the speech in this channel is the major source of content. Research utilized prosodic features (the information gleaned from the timing and melody of speech) such as pausing, pitch change or rhyme duration (Shriberg et al., 2000; Tur et al., 2001). Research shows that pause and pitch features are highly informative for segmenting speech. However the methods in the research have never been applied to lecture videos.

Segmentation Using Text Input

Segmentation methods utilizing text input usually make use of transcribed text or closed captions. With the time stamps that synchronize the video stream and transcribed text (Blei and Moreno, 2001), the output of transcribed text segmentation can be mapped back to video segmentation. Work in this area has been largely motivated by the topic detection and tracking (TDT) initiative (Allan et al., 1998). The story segmentation task in TDT is defined as the task of segmenting the stream of data (transcribed speech) into topically cohesive stories. They usually focus on the broadcast and news domain in which the formal presentation format and cue phrases can be explored to improve segmentation accuracy. For instance, in CNN news stories, the phrase "This is Larry King…" normally implies the beginning or the ending of a story or topic. Machine learning methods (Yamron et al., 1997) were commonly used and large sets of training data were provided in TDT. However, the presentation format of lectures is often more informal than news reports and large sets of training data are not available for lecture videos. Furthermore, the large variety of instructional styles makes the problem even more challenging.

Domain-independent Text Segmentation

Alternatively, research in the area of domain-independent text segmentation provides possible methodologies without requiring formal presentation format and training. Most existing work in this area has been derived from the lexical cohesion theory suggested by Halliday and Hasan (1976). They proposed that text segments with similar vocabularies are likely to be in one coherent topic segment. Thus, finding topic boundaries could be achieved by detecting transitions in vocabulary change. Researchers used different segmentation features to detect cohesion such as word stem repetition (Youmans, 1991; Hearst, 1994), word n-grams, and phrases (Reynar, 1998; Kan et al., 1998). The first use of a word has also been used by researchers because a large percentage of first-used words often accompany topic shifts (Youmans, 1991; Reynar, 1998). Reynar (1998) proposed a method which incorporates all these features into a maximum entropy model. A problem for methods in this category is that they are designed for "written text" with typographic cues such as paragraphs and punctuation, which do not exist in speech text extracted from lecture videos.

Segmentation of Lecture Videos

Compared with other video types such as movies and broadcast news, lecture video is special because it has few scene and shot changes, rich (but informal and spontaneous) speech, and usually a talking instructor as the focus. As illustrated above, there is very little research in the segmentation of lecture videos. Most existing video segmentation approaches are not suitable or simply cannot be applied to lecture videos. Since most segmentation features proposed in the research literature have not been applied to lecture videos, it is critical to study those features in a lecture video domain. Questions such as, "What are the effective segmentation methods?" and "What are the reliable segmentation features for lecture videos?" need to be answered in order to design and develop a highly accurate automated segmentation method. Previous research (Lin et al., 2005) utilized the linguistic features extracted from speech text and improved the segmentation performance. It achieves 70% accuracy in terms of F-Measure if one sentence away from the actual topic boundary is allowed and 40% accuracy if an exact match is required. Unfortunately the performance is still far from optimal and cannot be compared with human performance. Therefore, in order to answer these questions and gain some insight into the lecture video segmentation problem, we conducted a case study on how humans perform the segmentation of lecture videos before designing an automated segmentation method.

MANUAL SEGMENTATION – A CASE STUDY

The objective of our case study is to understand the segmentation process of humans and determine the rules and heuristics that they use. More specifically, we are interested in answering the following research questions:

- What methods or processes do humans use in manual segmentation?
- What features do humans use in segmentation, and which ones do they consider to be the best?

Study Design

We conducted a case study with undergraduate students enrolled in an MIS course at a southwestern university during the fall semester of 2003. Each participant selected had earned a high mid-term score in the class. They were asked to segment lecture videos from the same class. We assume that the students' superior class performance justifies their identification as experts in the topic (besides, they had already attended the videotaped class). We also believe that their overall segmentation behavior is a good reflection of domain experts' view on segmentation.

The study consisted of two parts: a segmentation task and a questionnaire. In the segmentation task, every participant was asked to segment three lecture videos. Each lecture video was segmented by three different participants in order to check the consistency among different people. In addition, human-corrected text transcripts of the lecture video were provided to facilitate segmentation. The students were also required to record certain statistical information such as the time they spent on the segmentation task and how many times they reviewed the video or/and transcript. After the segmentation task, they were asked to fill out a questionnaire. The questionnaire included three parts: segmentation, video/audio/transcript qualities, and general questions.

The segmentation part included open-ended questions asking participants to identify the segmentation features they used to segment the videos. The questions were classified into five categories (video, audio, text, content and others) according to the type of input source from which the segmentation features were extracted. The "content" category was used to describe the situations when humans may only concentrate on the overall content understanding and are not aware of any specific features or cues. A list of sample segmentation features with explanations (extracted from automated segmentation research) was also provided for each category.

The video/audio/transcript qualities part included both closed- and open-ended questions. The closed-ended questions asked participants to provide various ratings on a 7-point Likert scale. The open-ended questions asked for an explanation of their ratings of the qualities. We were interested in the impact of video, audio and text transcript qualities on segmentation because these qualities bias the results of what segmentation features are utilized by humans in the study, and thus affect our judgment on potential useful segmentation features. For instance, bad transcript quality may make it more difficult for people to extract text-based features. But it does not imply that text-based segmentation features are not good segmentation features.

The general questions part included five open-ended questions. The participants were asked to describe their general processes in segmentation task; how they thought the video/audio/transcript qualities affected their segmentation process; and general problems and suggestions on the segmentation task.

Results Analysis and Findings

Thirteen participants submitted the segmentation results of eleven lecture videos and their questionnaires. Each video in the eleven videos was segmented by three out of the thirteen participants. All videos were approximately one hour (mean = 66 minutes). Average time spent on the segmentation task was 1 hour 53 minutes. After analyzing the participants' responses in the questionnaires and removing the influences of video/audio/transcript qualities, we summarized the findings as follows.

The Two-phase Process

We found that most participants used a two-phase procedure in segmenting the videos: rough segmentation first and then refinement. Participants usually watched the video, read the associated transcript, and tried to understand the content first without marking any exact topic boundaries. Or they might simply write down some rough time stamps. Later when they reviewed the video again, they started to refine the topic boundaries and narrow down to the exact time points. Participants also claimed that they only started to notice specific indicators or cues for segmentation during refinement. For instance, some participants indicated that the instructor will say something explicitly like "ok, let's go to next topic…" The quantitative data also supported the idea that manual segmentation is multi-step process: most students watched the videos two or three times (with 2.66 mean review times).

Potential Segmentation Features: Combining All Input Sources

Various features from each of the sources (video, audio and text) were reported. For instance, many participants indicated that the instructor usually said some cue phrase such as "all right" or "ok" when switching to the next topic (see Figure 1). Shot change occurred when the instructor switched the screen from a homework review demo (topic 2) to a PowerPoint slide (topic 3). The most commonly used features reported by participants are listed in Table 1.

Topic 1: Introduction to Lecture Content

"All right, today we get to do the funner stuff. This is where we get into the whole purpose behind java, so today is gonna be good. Object oriented programming, that is what we are all here for …"

Topic 2: Homework review-two-dimensional arrays

"Ok but first of course we need to go through the homework assignment that I assigned for this time, which was to write an assignment called print2darray that accepts a two dimensional string array as a parameter …"

Topic 3: Object Oriented Programming Lecture

"All right, so that was that. Alright, object oriented programming. We are finally in the real meat. This is the real whole purpose behind java the fact even exist followed up the object oriented programming …"

Figure 1. Part of the transcript for a lecture video about Java Programming.

NOTE: The "Topic" headers indicate the boundaries identified by study participants. Certain words or phrases are bolded to show the segmentation features used.

However, we found that no single source or universal feature was used by all of the participants across the segmentation process. Instead, participants made use of features from all three sources: video, audio and text. For instance, in Figure 1 while judging the topic transition from topic 2 to topic 3, people made use of features from all three sources: the shot change from the video, the cue phrase "all right" from the text transcript and the high pitch from the audio when speaking the "all right." This implies that humans use complementary features from all three sources in order to make a decision on a topic boundary. Furthermore, several students also reported that they found the lecture slides to be very helpful in the segmentation process.

Category (Input Source)	Potential Segmentation Features
Video	*Scene/shot changes:* "When the instructor is explaining about java theory, the camera also moves to the white board - when the instructor is doing the homework problem, the camera will be on the slides" (problem: course and instructor style specific) *Gesture/posture:* "Turn around to whiteboard or erase the whiteboard"
Audio	*Long pause or silence:* "Instructor usually pauses before changing topic" *High pitch:* "Alright!", "Yes!" with higher pitch
Text	*Cue phrases:* "Ok so let's go on to…", "Now we are going to talk about….", "Instructor usually said explicitly that he wanted to move to next topic" *Introduction of new vocabulary:* "the instructor stated the topic prior to talking about the new topic"; new topic words were discussed (e.g. "composition", "inheritance", etc) before going to that topic.
Content	*Overall content changes:* Comments such as "where the instructor talks about another topic"; "When he started to talk about something other than what he was talking about earlier" are very common
Others	*Questions:* Topics were changed by students' questions when the instructors asked students if they had further questions on the current topic.

Table 1. Potential Segmentation Features Identified in Manual Segmentation

PROPOSED AUTOMATED SEGMENTATION METHOD

Based on the findings from the manual segmentation study and related research of automated segmentation, we propose an automated video segmentation method for lecture videos. We believe that this method is novel because it 1) simulates the two-phase manual segmentation process with initial segmentation and segmentation refinement, 2) combines segmentation features from all three input sources, and 3) utilizes various knowledge sources including world knowledge, domain knowledge and extra knowledge. As shown in Figure 2, our proposed automated segmentation method has three steps: *preprocessing, feature extraction,* and *features fusion and segmentation*.

Automated Segmenter

The *preprocessing step* processes raw video and prepares for feature extraction. Text transcripts are extracted using automated speech recognition (ASR) software. Image sequences and audio are also extracted. In *feature extraction*, various features are extracted, including text-based features (such as noun phrases and cue phrases), and low level video and audio features such as color histogram. All features we identified in the manual segmentation study could be potential candidate features (Table 1). Finally all features are fused and used for segmentation in the *features fusion and segmentation* step. Knowledge sources are used across the segmentation process.

In the *features fusion and segmentation* step, all features are passed to a two-phase segmentation process: *initial segmentation* and *segmentation refinement*. *Initial segmentation* performs a rough segmentation using content-based features. Because the speech in a lecture video contains the majority of the content information, most of the content-based features are extracted from speech-transcribed text. These features are usually linguistic features such as noun phrases, verb classes and word stems, which have already been identified as salient features in our previous study (Lin et al., 2005). Another reason we focus on the text transcript is the low computational cost of text processing compared with video/image and audio processing. After identifying potential topic boundaries in *initial segmentation*, we refine the rough boundaries using more computationally expensive features extracted from video and audio in *segmentation refinement*. Differing from content-based features used in *initial segmentation*, most features used in *segmentation refinement* are discourse-based features. They describe the characteristics of the small body of content immediately adjacent to the potential boundaries proposed in *initial segmentation*. For instance, most of the segmentation features we identified in manual segmentation belong to discourse-based features including video features (e.g., shot changes, gestures and posture shifts), audio features (e.g., pause length and pitch) and text features (e.g., cue phrase and introduction of new words). Furthermore, the computational cost can be significantly decreased when those computationally expensive features are only calculated in the small windows around the potential boundaries.

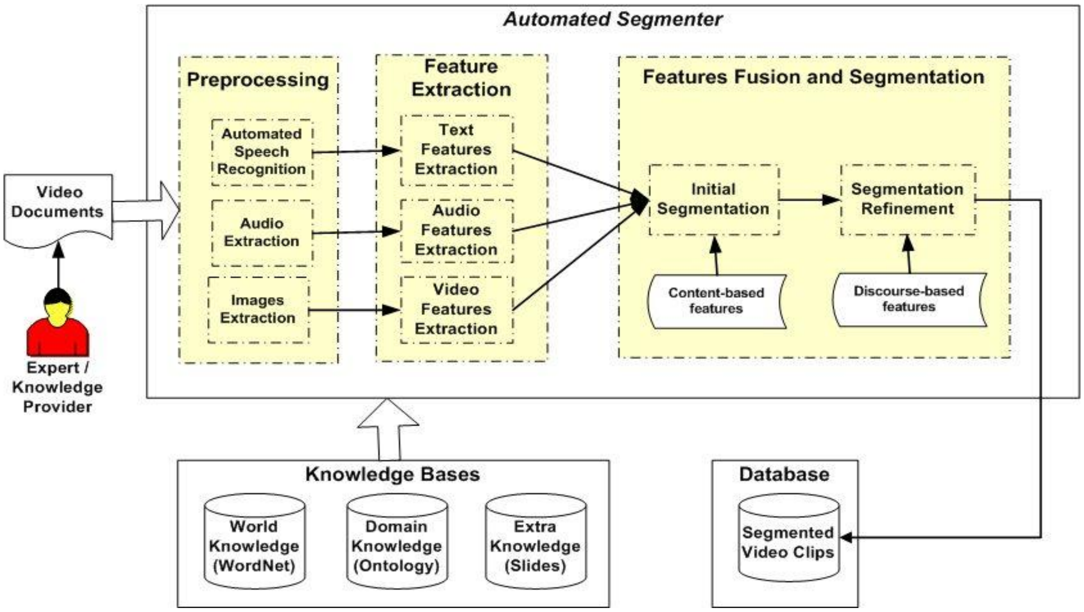

Figure 2. Automated Segmentation Method

Although the combination of segmentation features from the three sources is beneficial, the fusion of these features is not an easy task. One possibility is to adapt the sliding window method used in previous research (Lin et al., 2005). After identifying potential topic boundaries (by finding the points with the most dissimilar neighboring windows) in *initial segmentation*, features applied in refinement are only used to increase or decrease the probability of the potential boundaries as actual boundaries. However, if a training corpus was available, machine learning methods will be more reliable and achieve better performance. One method is to use a decision tree or maximum entropy model (Reynar, 1998). The decision of whether or not the initial segmentation suggests a topic boundary (indicated by a 1 or 0) is one feature that is used (along with all other features) in the refinement. Since segmentation features from different sources have a variety of characteristics, a combination of a decision tree and a Hidden Markov Model (HMM) may be a better strategy. The posterior probability from HMM using text features in *initial segmentation* will be used as a feature in the decision tree with all video and audio features. We can also use HMM as a top-level model (Shriberg et al., 2000). However, we are still testing different machine learning methods at this stage.

Knowledge Bases

During the entire segmentation process, various knowledge bases are utilized to assist in the segmentation. *World knowledge* includes knowledge sources about the general world including sources such as WordNet. WordNet is a lexical knowledge base in which words are organized into synonym sets (Miller et al., 1990). These synonym sets, or synsets, are connected by semantic relationships such as hypernymy or antonymy. WordNet has been used to extract the “verb class” feature in previous research (Lin et al., 2005). *Domain knowledge* includes ontology extracted from sources such as the Internet, electronic textbooks or professional dictionaries. Because many concepts in a lecture are professional terms in a specific domain, they do not exist in a general lexical knowledge base such as WordNet. However, we can extract relationships from, for instance, a professional dictionary (e.g., Webopedia, a dictionary for computer and Internet technology definition) and build a WordNet-like ontology. In many situations, *extra knowledge* such as an instructor-created lecture outline and slides are also available for lecture videos. They provide valuable information for the segmentation. For instance, slides have been used to correct the word errors in the ASR transcripts (Cao, 2004).

Preliminary Results

Although the development of the complete algorithm is still in progress, we conducted a preliminary evaluation using the same method from previous research (Lin et al., 2005). Previous method makes use of content-based text features in the first stage and discourse-based text features in the second stage, which is a good match to the proposed two-phase segmentation process although only text-based features were used. We retested the algorithm using a different dataset by randomly selecting three lecture videos from two different MIS courses and ensuring they were different videos than previously used (in Lin et al., 2005). We focused on evaluating the effectiveness of the two phase segmentation: initial segmentation and segmentation refinement. We achieved similar results to what was previously found (in Lin et al., 2005) (Table 2). We found a small improvement between the algorithm implementing the initial segmentation step only ("Initial") and the algorithm implementing both steps ("Initial + Refinement") (1.6% in F-Measure and 4.0% in Precision). However, the improvement is not significant. One possible reason could be that the cue phrases and pronouns we used (Lin et al., 2005) are too general and may happen rarely in our small test dataset. Another reason could be that all features used in (Lin et al., 2005) come from the text source only. It is expected that the incorporation of video and audio features as proposed in this paper will complement the text features and achieve better performance levels as a result.

Version	Exact Match			Fuzzy (1)		
	Precision	Recall	F-Measure	Precision	Recall	F-Measure
Initial	0.35	0.25	0.29	0.75	0.56	0.64
Initial + Refinement	0.37	0.25	0.30	0.78	0.56	0.65

Table 2. Evaluation of the Effectiveness of Automated Segmentation Method

CONCLUSION AND FUTURE DIRECTIONS

We conducted a case study on manual segmentation to collect common methods and potential segmentation features used by humans. Based on the case study findings, we designed an automated segmentation method which incorporates features from video, audio and speech text, adapts a two-phase segmentation process (initial segmentation plus refinement) from manual segmentation, and utilizes various knowledge bases. However, one limitation of our case study is that only one MIS course was selected for use in our dataset. More studies will be required in order to draw a more generalizable conclusion. Further it will be better if the segmentation results have been cross-verified by lecturer. In future studies, we are interested in issues such as the identification of features used in deciding each specific topic boundary and the segmentation consistency between different individuals. Finally, we are developing and evaluating the proposed automated segmentation method.

REFERENCES

1. Allan, J., Carbonell, J., Doddington, G., Yamron, J. and Yang, Y. (1998) Topic detection and tracking pilot study: Final report. *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*.
2. Ausubel, D.P. (1960) The use of advance organizers in the learning and retention of meaningful verbal material. *Journal of Educational Psychology,* 51, 267-272.
3. Blei, D. M. and Moreno, P. J. (2001) Topic segmentation with an aspect Hidden Markov Model. *Proceedings of the 24th International Conference on Research and Development in Information Retrieval*, New York, NY: ACM Press.
4. Cao, J., Crews, J.M., Lin, M., Burgoon, J.K., and Nunamaker, J.F. (2003) Can people be trained to better detect deception? Instructor-led vs. Web-based training. *Proceedings of the Ninth Americas Conference on Information Systems*, Tampa, Florida.

5. Cao, J. (2004) Question answering on lecture videos: A multifaceted approach. *Proceedings of the Fourth ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2004)*, Tucson, AZ.

6. Daft, R.L., and Lengel, R.H. (1986) Organizational information requirements, media richness and structural design. *Management Science*, 32, 5, 554-571.

7. Halliday, M. and Hasan, R. (1976) Cohesion in English, Longman.

8. Hearst, M. A. (1994) TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23, 1, 33-64.

9. Lin, M., Chau, M., Cao, J., and Nunamaker, J. F. Jr. (2005) Automated video segmentation for lecture videos. *International Journal of Technology and Human Interaction (IJTHI)*, 1, 2, 27-45.

10. Kan, M., Klavans, J.L. and McKeown, K. R. (1998) Linear segmentation and segment significance. *Proceedings of the 6th International Workshop of Very Large Corpora*, 197-205.

11. McNeill, D., Quek, F., McCullough, K-E., Duncan, S., Furuyama, N., Bryll, R., Ma, X-F., and Ansari, R. (2001) Catchments, prosody, and discourse. *Gesture*, 1, 9-33.

12. Miller, G., Beckwith, R., Felbaum, C., Gross, D., and Miller, K. (1990) Introduction to WordNet: An online lexical database. *International Journal of Lexicography (special issue)*, 3, 4, 235-312.

13. Mukhopadhyay, S., and Smith, B. (1999). Passive capture and structuring of lectures. *Proceedings ACM Multimedia* 99, 477-487

14. Ngo, C. W., Wang F., Pong, T. C. (2003) Structuring lecture videos for distance learning applications. *Proceedings of Fifth International Symposium on Multimedia Software Engineering*, 215-222

15. Quek, F., McNeill, D., Bryll, R., Kirbas, C., Arslan, McCullough, K-E., Furuyama, N., and H., Ansari, R. (2000) Gesture, speech, and gaze cues for discourse segmentation. *Proceeding of IEEE Conference on Computer Vision and Pattern Recognition*, 2, 247–254.

16. Reynar, J. C. (1998) Topic segmentation: Algorithms and applications, PhD thesis. Computer and Information Science, University of Pennsylvania, 1998.

17. Shriberg, E., Stolcke, A., Hakkani-Tur, D. and Tur, G. (2000) Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication (Special Issue on Accessing Information in Spoken Audio)*, 32, 1-2, 127-154.

18. Stanford Online, http://scpd.stanford.edu/scpd/students/onlineclass.htm

19. Tur, G., Hakkani-Tur, D., Stolcke, A., Shriberg, E. (2001) Integrating prosodic and lexical cues for automatic topic segmentation. *Computational Linguistics*, 27, 1, 31-57.

20. Wactlar, H. D. (2000) Informedia - search and summarization in the video Medium. *Proceedings of Imagina 2000 Conference*, Monaco.

21. Yamron, J. P., Carp, I., Gillick, L., Lowe, s., and van Mulbregt, P. (1997) Topic tracking in a news stream. *Proceedings of the DARPA Broadcast News Workshop*, 133-136.

22. Youmans, G. (1991) A new tool for discourse analysis: The vocabulary management profile. *Language*, 67, 4, 763-789.

23. Zhang, D. S. (2002) Virtual mentor and media structuralization theory, PhD thesis. University of Arizona, Tucson, AZ.

24. Zhang, H. J. and Smoliar, S. (1994) W. Developing power tools for video indexing and retrieval. *Proceedings of SPIE'94 Storage and Retrieval for Video Databases*, San Jose, CA, USA.