

# Augmented Segmentation and Visualization for Presentation Videos

Alexander Haubold  
Department of Computer Science  
Columbia University  
New York, NY 10027  
+1 (212) 939-7152  
ahaubold@cs.columbia.edu

John R. Kender  
Department of Computer Science  
Columbia University  
New York, NY 10027  
+1 (212) 939-7115  
jrk@cs.columbia.edu

## ABSTRACT

We investigate methods of segmenting, visualizing, and indexing presentation videos by both audio and visual data. The audio track is segmented by speaker, and augmented with key phrases which are extracted using an Automatic Speech Recognizer (ASR). The video track is segmented by visual dissimilarities and changes in speaker gesturing, and augmented by representative key frames. An interactive user interface combines a visual representation of audio, video, text, key frames, and allows the user to navigate presentation videos. User studies with 176 students of varying knowledge were conducted on 7.5 hours of student presentation video (32 presentations). Tasks included searching for various portions of presentations, both known and unknown to students, and summarizing presentations given the annotations. The results are favorable towards the video summaries and the interface, suggesting faster responses by a factor of 20% compared to having access to the actual video. Accuracy of responses remained the same on average. Follow-up surveys present a number of suggestions towards improving the interface, such as the incorporation of automatic speaker clustering and identification, and the display of an abstract topological view of the presentation. Surveys also show alternative contexts in which students would like to use the tool in the classroom environment.

## Categories and Subject Descriptors

H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing – *Indexing methods*; H.5.1 [Information Interfaces and Presentation]: Multimedia Information Systems – *Evaluation/methodology, Video*.

## General Terms

Algorithms, Management, Design, Human Factors.

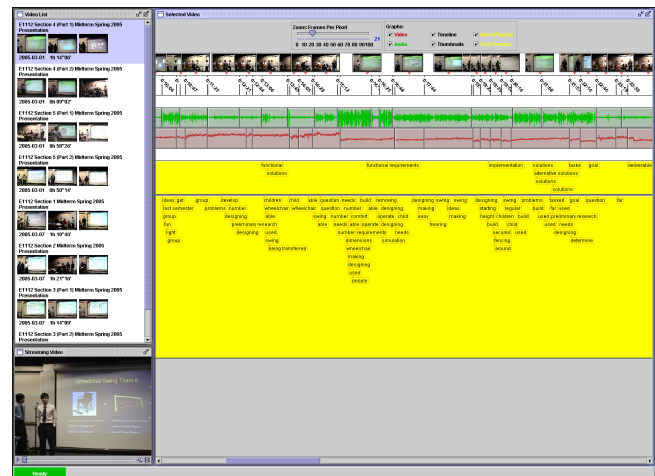
## Keywords

Presentation video, audio, text, multimedia, segmentation, summarization, browsing, visualization, user interface, gesture, speaker, presenter, cross-reference, video library, user studies.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

MM'05, November 6–11, 2005, Singapore.

Copyright 2005 ACM 1-59593-044-2/05/0011...\$5.00.



**Figure 1. User Interface for video/audio segmentation and text augmentation.** Videos are selected from the list on the left side. Streaming video contents appears in the lower left frame. Video summaries in the right frame appear on a segmented linear timeline, where one hour of video is compressed between 1 and 100 frames/pixel. The row of thumbnails provides access to high-quality keyframes. The timeline presents a combined segmentation of audio and video. The green row shows audio activity, the red row video activity, and the yellow rows displays index and content phrases.

## 1. INTRODUCTION AND MOTIVATION

Video segmentation, visualization, and indexing have received much attention for providing means to organize and access video libraries. With the growing use of videos in classrooms other than for recording lectures, we investigate a novel application of video libraries for classroom presentation videos. Characteristically, classroom presentations are carried out by several students and follow a known structure. The recorded videos differ from lecture videos in several critical respects: their shots are longer without distinct visual cuts, presentations are carried out by multiple speakers, audio quality varies significantly, and a repetitive formal structure of keywords exists across presentations. Analysis of such videos should take advantage of this a priori knowledge.



Figure 2a. Presentation

Figure 2b. Discussion and Q/A

Figure 2c. Film clip screening

**Figure 2. Examples of the many kinds of video imagery from typical presentation video. Segmentation of such videos should take advantage of cues from audio, video, and text, leading to integrated approaches for summarization and indexing.**

The methods and tools discussed here address the needs of two audiences involved in the presentations: instructors and students. Presentations are used by instructors to evaluate and grade the performance of teams and individual students. Recorded video material is used to revisit the presentations as necessary in some cases to re-evaluate and discuss with students. However, the inherently serial nature of video inhibits the instructor from quickly locating portions of the presentation.

With the introduction of video-taped presentations for archiving purposes, students have gradually become more interested in reviewing this material as well. Since presentations are recorded at the midterm and at the end of the semester, students have found it useful to evaluate their own and their peers' performance towards improving presentation skills.

## 2. RELATED WORK

Work in video segmentation and summarization has focused largely on lecture and news videos; we only make mention of a few related ones here. The Cornell Lecture Browser [11] passively captures and summarizes structured university-style lectures. Stationary and tracking cameras are used to record presentation slides and presenter. Segmentation of a video relies on the changes in presentation slides. Classroom 2000 [1] uses invasive technologies to capture lectures. Instructors provide presentation slides before class for annotation purposes, and are required to use electronic whiteboards during class. IBM CueVideo [3] is intended to serve as a more universal video analyzer and browser, taking cues from video, audio, and speech recognition.

Approaches for segmentation by contents have been presented for structured lecture videos [8], where visual cues from in-class instructor notes are used to determine topics. News videos have been widely explored for shot boundary detection, story unit determination through multimodal fusion, and classification ([5], [2]).

Summarization of videos by extracting the most important contents and producing a new video or a mural-like snapshot has been researched as an alternative to video browsing. In [10], lecture-style audio-video presentations are summarized in segments, which are 20-25% of the original video's length. Cues are taken from pauses in the audio track and slide transitions captured from the presentation software. Video skims [13] have been used to summarize news videos in much more compact representations taking cues from video, audio, and text.

Video segmentation based on textual cues and linguistic features has been explored separately. In [4], noisy ASR transcripts and error-free text streams are successfully segmented by topic using aspect models combined with HMMs. In [12], word n-grams are detected in imperfect lecture transcripts, consisting of specific word classes. Segmentation is performed by first identifying differences in feature frequencies, and then merging segments with similarities. Previous work in audio segmentation includes methods of reliably detecting speaker changes via the Bayesian Information Criterion introduced in [6].

In our analysis of student presentation video, we cannot make assumptions on exclusive use of presentation slides or good (if any) camera tracking. There are practically no cues available from cinematography – lighting remains fixed, cuts between shots are not present, and audio cues can be extracted only from speech. Compared to lectures, student presentations may not be as well prepared and they vary greatly in terms of presentation slide utilization. We thus focus on methods that are neither invasive nor structure-driven, while centering attention on the presenting students.

Our approach to segmenting and building an interactive user interface is based on considering both audio and video data, and combining the resulting segmentations in a user interface. In addition, we generate highly imperfect transcripts using the IBM ViaVoice ASR, from which we filter a small number of meaningful phrases using text analysis presented in our previous work [9]. These phrases are used to index the video, and allow for quick visual scanning of a video's contents.

## 3. DEFINITION OF PRESENTATION VIDEOS

A presentation video contains one or more distinct presentations carried out by students or teams of students. Typically, an electronic medium like PowerPoint is used to accompany the speaker; however, for the purpose of segmentation and visualization, it is not required here. However, intrinsic but unpredictable aspects of these video records are the inclusion of short video sessions, speaker transitions, question and answer sessions (of varying audio quality), interruptions, cameraperson errors, etc. (see Figure 2).

A single camera captures the speaker standing next to the projected image, and a handheld or stationary wireless microphone is used by the speaker to better pick up the audio

signal. The video is neither shot professionally, nor is the presentation space set up specifically for the purpose of video productions. It is not uncommon for students to forget to speak into the microphone, or to appear in the filmed frame. The time period during which a speaker presents does not necessarily overlap with the projected slides; two presenters may share a slide. Separating segmentation and visualization of audio and video is especially useful for these conditions. Although good practice would suggest that the video could be easily segmented by speaker, these and other frequent violations of presentation rules make a more robust approach necessary.

The corpus of presentation videos considered here originates in a collection of videos for the past three years. Since video has become a popular medium of self-evaluation, especially in the educational environment, we foresee the continued production and persisting problems of indexing, browsing, searching, and disseminating such presentation videos.

## 4. SEGMENTATION

### 4.1 Audio Segmentation

In general, audio segmentation for presentation videos lends itself to segmentation by speaker. We employ the method of detecting speaker changes via the Bayesian Information Criterion introduced in [6]. The audio track is sampled at regular intervals and vectors of 13 Mel Frequency Cepstral Coefficients are determined for each set of audio samples. Using a two-window approach, the BIC is computed for each partition of this interval.

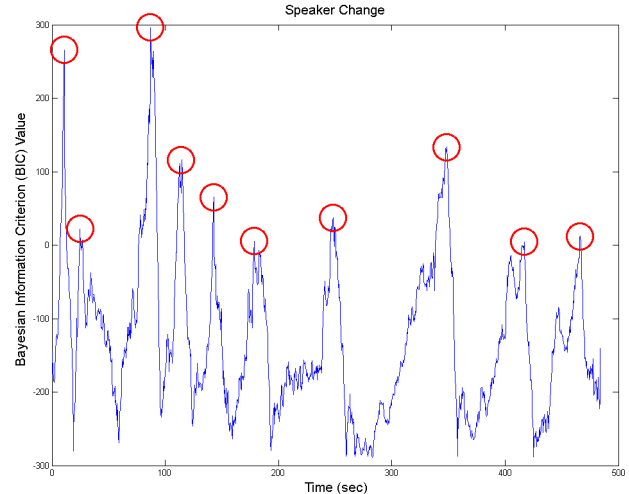
If there exists a clear positive maximum among BIC values, a speaker change has been found, otherwise the interval is extended with additional audio samples. In Figure 3, 10 speaker changes have been identified over 8 minutes of audio data. The maxima between speakers' BIC values can be clearly located.

In evaluating this method on presentation videos, we have found that the best segmentation is achieved with the following settings:

- Number of segments for which MFCC is computed in one second of audio  $\approx 8$ . This corresponds roughly to the number of syllables uttered in such a time frame.
- The portion of each segment used in the Fourier Transform  $\approx \frac{1}{8}$ , and the sample size closest to a power of 2 that matches this time interval is chosen. In terms of common audio sample frequencies, this value is  $\approx \frac{f}{62.5}$  (e.g. 32kHz: 512 samples, 16kHz: 256 samples, 128kHz: 128 samples).

We have experimented with several ranges of values for sampling frequency (8kHz, 16kHz, 32kHz), FT sampling windows (256, 512, 1024), and MFCC vectors per second ( $\frac{1}{256}$ -125). Compared with our chosen sizes, choosing the number of MFCC vectors per second much higher or lower, or selecting much longer or shorter sample sets, results in dramatic over- or under-segmentation.

The results from speaker segmentation are very favorable. From experiments we observed no false negatives, and only few false positives. The latter tend to be introduced by the occurrence of small pauses in the audio track. The final segmentation, as well as the raw audio activity graph (=audio amplitude) are included in the user interface (see Figure 6, row 3).



**Figure 3: Speaker change detection via BIC. Circles mark the points in time at which a speaker change occurs. A speaker change is detected when a maximum BIC value above 0 is measured.**

For our test series of 7.5 hours of presentation videos, we have identified the following measures of precision and recall for speaker segmentation:

$$\text{Precision} = \frac{\# \text{ relevant speaker segments}}{\# \text{ all segments found}}$$

$$\text{Recall} = \frac{\# \text{ relevant speaker segments}}{\# \text{ true speaker segments}}$$

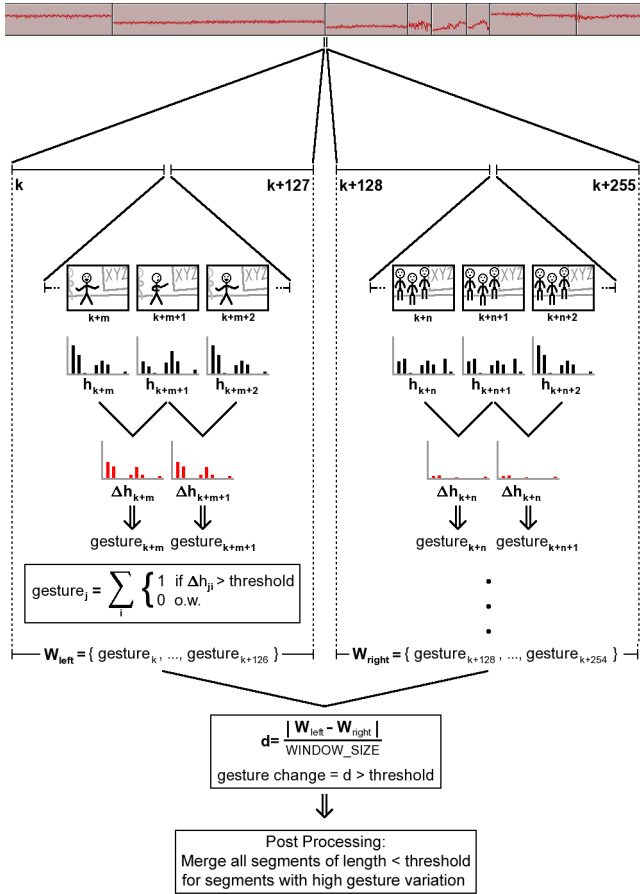
Precision = 88.5%    Recall = 95.7%    True Segments = 395

The minor lack in precision tends to be a result of externally induced environmental changes. This includes scenarios during which some student maintains different distances to the microphone, and is thus identified as two separate speakers.

### 4.2 Visual Segmentation

In this stage of segmentation, visual contents from a video is analyzed for shot boundaries. Here, we use the term “shot” to refer to visual changes during the presentation. This includes changes in electronic slides and changes in a speaker’s tone due to significant changes in speaker motion. Since the closest our data has to “cuts” are infrequent and sometimes visually subtle slide changes, we found methods of histogram comparison to be robust.

We apply methods of computing histogram changes between consecutive frames and detecting long-term changes by comparing the degree of change over time. Comparisons are made between two four-second windows, and a shot boundary is declared if the difference between the windows deviate significantly (see Figure 4). An experimentally derived threshold is used to measure the deviation between the 4-second windows. Figure 5 outlines what degree of visual change is considered a boundary. It is based on observations about presentation slide changes and speaker gesture changes from presentation videos.



**Figure 4: Visual Segmentation based on presentation slide changes and speaker gesture change.**

We have found this method to be robust in detecting changes in presentation slides. More interestingly, this method also detects speaker changes by differentiating the characteristic movement patterns between two speakers. Similar methods of particular gesture detection and classification using HMMs are discussed in [14]. We use a different approach, in which we are not interested in the meaning of gestures, but instead the difference among students and the occurrence of unusual movement. Intuitively, such occurrences tend to describe interesting moments during the otherwise uneventful presentation.

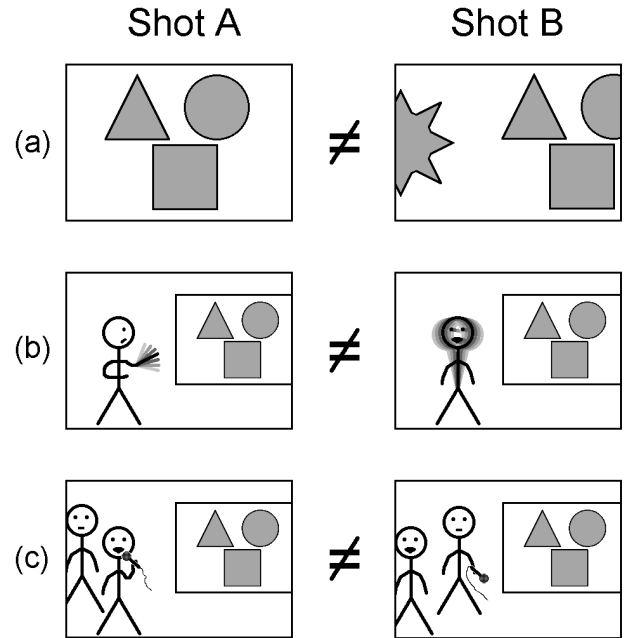
We therefore found it important to include this measure in the user interface as a visual activity graph. (see Figure 6). It is also easy to visually pick out video segments with a high degree of visual change from the raw activity graph.

The following measures of precision and recall were determined for visual segmentation:

$$\text{Precision} = \frac{\# \text{ relevant visual segments}}{\# \text{ all segments found}}$$

$$\text{Recall} = \frac{\# \text{ relevant visual segments}}{\# \text{ true visual segments}}$$

Precision = 89.4%    Recall = 82.7%    True Segments = 594



**Figure 5: Definition of a "visual shot". In all three pairs, the difference between A and B is significant enough to be considered separate shots. In (a), the difference is based on changes in the entire frame. In (b), the difference is mainly gesture of a speaker. In (c), the "visual" tone of different speakers deviates.**

Common errors during visual segmentation include accidental movement of the camera, increased movement of students in front of the camera, especially during setup of presentations, and poor lighting (too much or too little) during the presentation.

### 4.3 Combined Audio-Visual Segments

The nature of presentation video leads to the definition of a "presentation unit" for this genre, one in which boundaries are not identified solely by visual or audio scene changes, but by a combination of the two. In some instances, using only one of the two methods would result in an unfavorable series of presentation units:

Speaker segmentation by itself produces no or too few segments for a long presentation lead by only one student.

Visual segmentation on its own may miss dialogue changes, especially when two speakers use the same slide.

The integration of both segmentations takes into account significant audio as well as visual changes, including speaker, gesture, and visual aid changes. We postulate that a "presentation unit" is the smallest unit bounded by either audio or visual change. Intuitively, the sum of both segmentations is more meaningful than its parts, because we do not perceive audio separate from video. In fact, it is unrealistic to have a clear visual source for each audio sample on a given video. And, in many cases, speakers and sound effects are not present in the video frame, suggesting further that separate segmentations must be logically combined.

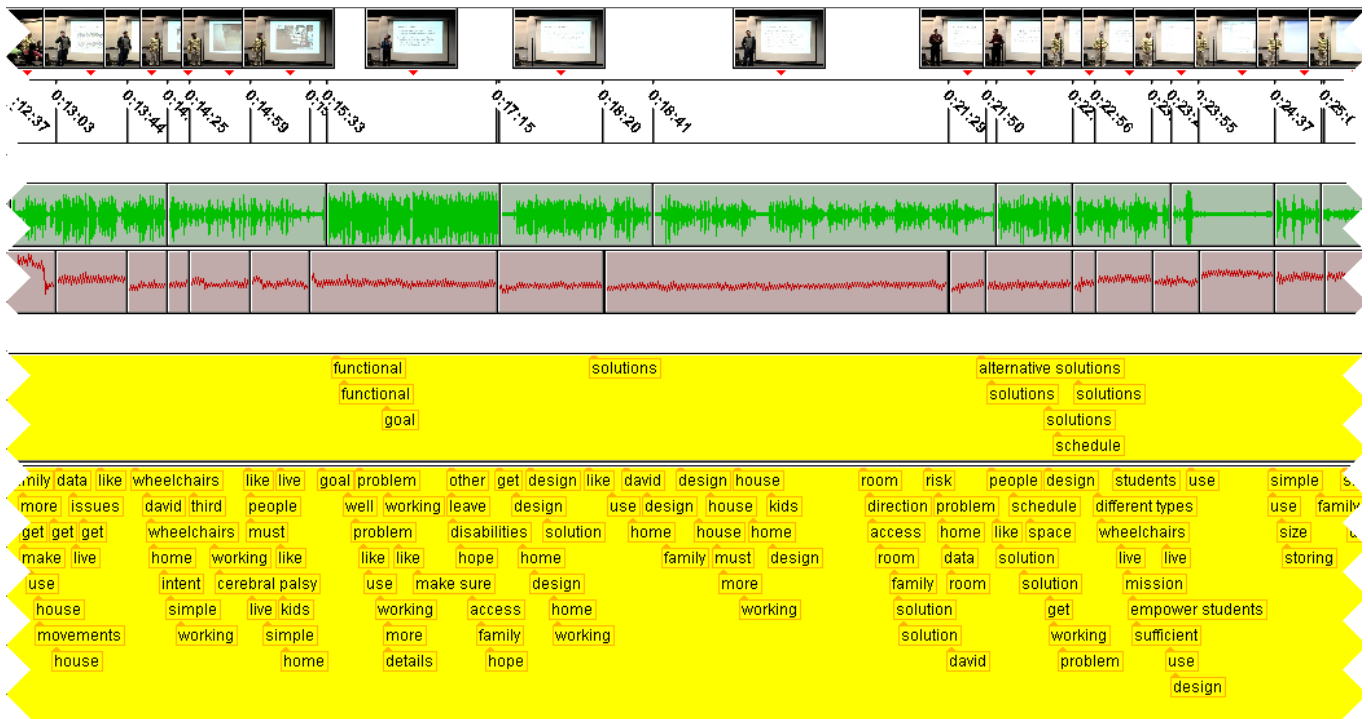


Figure 6. Complete timeline includes thumbnails for sufficiently long video segments (row 1), a timeline with time markers combining video and audio segmentations (row 2), visual video segmentation with activity graph (row 3: red), visual audio segmentation with activity graph (row 4: green), index phrases (row 5: yellow), text phrases (row 6: yellow).



Figure 7: Video segmentation where audio segmentation fails: Gesture and posture hint at where transitions between interesting segments occur. Question and Answer session: (a) Student answers a question, (b) Group listens to comments from the audience, (c) Student answers question.

Combined (ANDed) visual and audio segmentation resulted in the following measures of precision and recall:

$$\text{Precision} = \frac{\# \text{ relevant audio - visual segments}}{\# \text{ all segments found}}$$

$$\text{Recall} = \frac{\# \text{ relevant audio - visual segments}}{\# \text{ true audio - visual segments}}$$

Precision = 89.3%    Recall = 92.7%    True Segments = 710

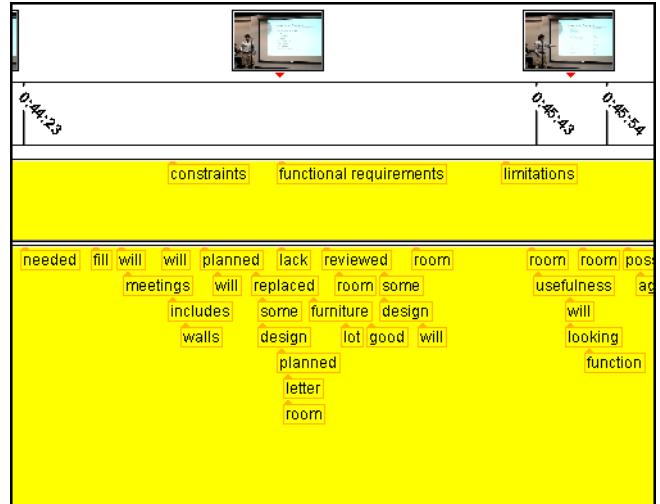
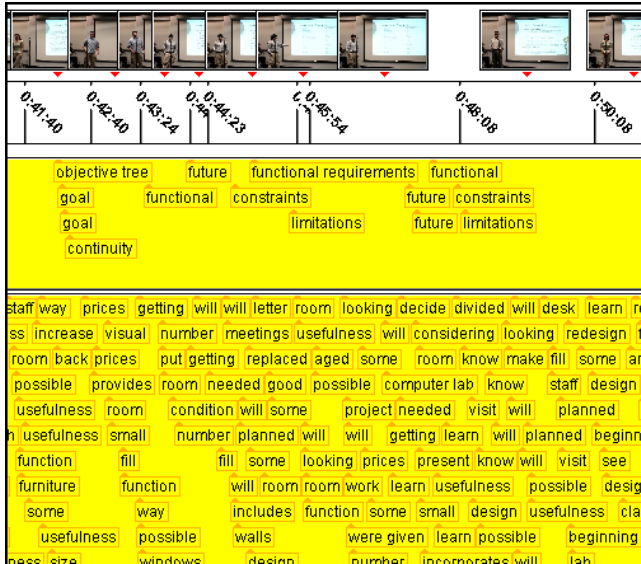
These values are in the same range of audio and video segmentation alone; however, the number of identified audio-visual segments is significantly higher. Precision and recall of

separate segmentations in comparison to audio-visual segmentation is dramatically lower:

Audio                      Precision = 51.3%                      Recall = 53.2%

Video                      Precision = 66.6%                      Recall = 69.2%

Empirically, our method of combining segmentations finds important logical and thematic changes, and suggests that these integrated segments are representative of the presentation units. For example, one of the unobvious presentation units found in presentation videos is the interaction between presenters and audience during the Q&A session. At a question, the camera pans and frames the presenting team, the team temporarily stops



**Figure 8: Text and thumbnail distribution effect upon zooming. (a) The video is zoomed at 34 frames / pixel. The thumbnail row displays overlapping images and the text row displays all corresponding phrases in these 9 minutes, which creates a very busy and difficult to read visual. (b) The video is zoomed at 6 frames / pixel with obvious improvements in the thumbnail and text rows for these 90 seconds. While the words do not form sentences, they can be used to understand the material discussed here.**

talking (or speaks away from the microphone), and visual slides are not advanced. However, segments can be identified by distinct visual breaks between the different portions of the discussion (see Figure 7). This is mainly due to gesture and posture changes of the students when listening to questions versus answering them. Another example occurs when a group of stationary students crowd around the microphone. Turns are made by switching speakers without changing the visual setup of the team. In this case, speaker segmentation provides the best means of distinguishing among segments.

We have included a graphical representation that combines raw audio and video segmentations in the user interface. This timeline marks all relevant breaks in the video and labels them with timestamps (see Figure 6).

## 5. TEXT AUGMENTATION

Parallel to visually summarizing video clips with thumbnails, we use text to summarize audio clips. However, transcripts are not readily available for the presentations, and we cannot make the assumption that every presentation is accompanied by electronic slides.

We thus generate transcripts using the IBM ViaVoice ASR. The resulting transcripts are highly imperfect with large Word Error Rates ( $\approx 75\%$ ) due to several factors. Primarily, the audio quality varies greatly and depends on the individual presenter and the presentation environment. Due to the large number of speakers, it is impractical to apply speech model adaptation. Language model adaptation is also unfeasible, as the contents, style, and fluency of the presentations have high variance as well. We thus apply a generic speech and language model to the audio tracks. Specifically, we use the author's speech and ViaVoice's 10 minute basic language model. Experiments with other speech models yielded few improvements [9].

**Table 1. Topic phrases (slide titles) for presentations in the course "Engineering Design".**

Two word phrases	One word phrases
1. alternative solutions	1. background
2. continuity plan	2. chart
3. design constraints	3. constraints
4. functional requirements	4. continuity
5. future directions	5. deliverables
6. gantt chart	6. demo
7. objective tree	7. functional
8. problem statement	8. future
9. projects goal	9. goal
10. tasks performed	10. implementation
11. team process	11. limitations
12. team development	12. objective
	13. prototype
	14. requirements
	15. schedule
	16. solutions
	17. statement
	18. tasks

Transcripts with high error rates do not lend themselves to known text analyses, in which correlations are found between repetition and uniqueness of words and phrases. In a previous work [9], we have introduced methods by which highly imperfect transcripts from university lecture courses are filtered by using expected significant index terms extracted from external course-related sources such as textbooks, web pages, etc. We apply a similar method to transcripts from presentation videos. While we do not have indicators of the specific contents for a given presentation, we do have some knowledge about the overall structure.

Presentations in the domain of our test video database revolve around Engineering Design projects. We have manually generated a list of 30 frequently used words and phrases from the presentation slide titles, and we use them to filter the transcript (see Table 1). The resulting “theme phrases” are included in the user interface and provide the equivalent of a table of contents for each presentation (see Figure 6, row 5). While the task of extracting title words can be automated, we have chosen to manually do so for this prototype. The list of theme phrases in Table 1 is considered static with respect to the course in which the tool is used. For domains outside of the videos we have used, this list of theme phrases could be compiled by cross-referencing frequently used headers or listing all titles from presentation slides. We address a dynamic and scalable search method in the topical text interface.

Besides identifying theme phrases, we also apply text filtering of all of the phrases found in the source data of the electronic presentation slides, if available. To this end, each line of text in the slides is used as a phrase. The resulting “topic phrases” are included as an additional index in the user interface and give clues about specific items discussed in the presentation, including names, locations, numbers, etc.

Methods of stemming and specific stop word removal are applied in the process of filtering. Among the 109 common stop words are, for example, “a”, “all”, “also”, “an”, “and”, “any”, “are”, “as”, “at”, etc. This list was compiled from empirical observations made during analysis of ASR text corpora.

The ASR software does not create a mapping of words to time during the speech recognition process, but rather provides an uninterrupted flow of text. For the mapping of words to time in the user interface, we make use of a linear fit algorithm that distributes words in dense regions of audio activity. Regions with relatively small amplitude spikes are assigned a low weight, while those with many spikes are assigned a high weight. The linear transcript text is then fitted to these weights. We found this method to be a good approximation for words-to-time mapping, with an error range of no more than 60 seconds with respect to ground truth. We are aware of more accurate means of mapping, but await the results of further studies to determine the criticality of word placement precision in browsing and retrieval. One alternative to the discussed automatic processing and alignment of text and index generation is through costly and unscalable, yet accurate manual services.

## 6. INTERFACE

The interactive user interface is modeled as a linear time line (see Figure 6), horizontally spanning the screen (see Figure 1). This provides an overview of the presentation video’s structure and contents.

Audio and video segmentations are included in the user interface as visually segmented activity graphs. The activity curve for audio represents audio amplitude, and for video the amount of change between two adjacent frames or clips. Activity on the video track is particularly interesting, as it provides clues about the amount of action at any given point. A more or less steady horizontal line indicates a video segment with conversational qualities, while a prolonged fluctuating line points out intense motion, e.g. in a film screening or an interactive demonstration.

A timeline combines separate audio and video segmentations. Thumbnails for sufficiently long video segments are placed above the timeline, and theme and topic phrases, if available, are placed below the audio segmentation graph.

For further exploration of the video, a zoom feature has been implemented that can be used to stretch the graph from 100 video frames/pixel to 1 frame/pixel (see Figure 8). Stretching the video has several benefits. Overall, it enables more precise temporal browsing by viewing a shorter period of the video summary. Areas of densely overlapping thumbnails are expanded, allowing the viewer to distinguish better between presentation units. In addition, the large collection of phrases in the yellow rows distributes horizontally over time, thus decreasing the depth of phrase nesting. As a result, the text becomes more manageable to read, as outlined in Figure 8. Clicking on thumbnails reveals their original size. We plan on extending the interface to allow audio/video playback from any point in the graph.

The interface has been modeled on informal observations of instructors and students accessing videos with more standard tools. The subjects in our classroom tend to have some familiarity with video editing, leading to the design of a row-media layout. We intend this view to be especially helpful for the viewing of video clips, while the text augmentation rows serve as search indices.

## 7. USER STUDY

### 7.1 Experiment

We have evaluated our methods of visual summarization and indexing in the preliminary interface (see Figure 1) with 176 students of varying knowledge about the contents of the videos (see Table 2). We have set up a recorded experiment in which students were asked to use the interface to answer between 5 and 6 specific video-related questions (see Table 3). The study was conducted with no more than 7 students at a time. Before the experiment, participants were given an introduction to the tool for 10-15 minutes by a proctor. During that time, they actively experimented with all of the features while verbal explanations were given by the proctor. After the experiment, participants had the option of playing with the interface at their leisure. In a follow-up survey we asked the students about their experience and how they liked working with the individual parts of the interface.

The goal of the user study was to determine what features of the browsing tool were most important to the students, how quickly they could answer specific search-related questions using these features, and the amount of benefit these indexing features and video summaries enabled.

To get a better sense of the added benefit of video summaries, we have integrated into the interface a media player, but only enabled it for half the participants: 84 of the students had only the summary, while 92 of them were able to use the summary with the video. Students who were unable to use the video were thus forced to use the summary only. This served as a clearer baseline for measuring the effectiveness of the summaries.

**Table 2. User study participants in 4 groups (group results are in Figure 9)**

Group	Description	#	Age Group
A	Students in class	79	17-19
B	Students in class	73	17-19
C	External students familiar with teams and projects	17	19-21
D	External students unfamiliar with teams and projects	7	30-50

**Table 3. Questions for recorded user study. Some types of questions were asked more than once per user study.**

Type of question	# Questions
Find your appearance during the presentation	152
Find beginning of your team's presentation	169
Find your team's discussion on topic X	169
Find presentation X (X = one they have no knowledge of temporally or contextually)	193
Summarize presentation between time A-B using only text augmentation	183

## 7.2 Results

We evaluated the results separately for video and non-video and present the comparisons here.

Overall, we find strong evidence that our tool and methods of summarization lead to more efficient searching, while at the same time retaining the accuracy of results of traditional linear search methods. We have compiled statistics on the usage and accuracy of the 6 questions from Table 3, and present the results in Figure 9 a-f. It is clear from these graphs that the percentage of correct answers stays roughly the same irrespective of having video available or not. Accuracy here is defined as answering a question within a range of acceptable values, e.g. to "find" a particular portion of the video, its start time must be within 1 minute of the ground truth.

While accuracy is at a similar level, the time required to answer a question was reduced by 20% on average when participants were *unable* to use the video. The main difference between the two groups lies in the usage of video versus keyframes, summarized in Figure 9. We have observed that when video is available, students tend to spend more time watching video than necessary to complete a task of searching. Since keyframes are by themselves uninteresting to watch over a prolonged period of time, they tend to be used more efficiently. Figure 9 points out that in almost all cases, an increased usage of keyframes leads to a faster response rate. Moreover, video playback takes significant time for each question.

While response time is relatively faster without the usage of video, the absolute reduction in time by using any version of the tool over linear search tools (i.e. reverse and fast forward on VCR or in Media Player) will be much greater, although we have not investigated this comparison.

## 7.3 Additional Observations

We note several side effects which may have influenced the user study and introduced some statistical bias. The library of presentation videos spanned 5 sections of a course in Engineering Design. Depending on the size of each section, presentations were either recorded on one DV (digital video) tape, or split between two. We may possibly expect the time to search over two separate videos to be somewhat higher for longer sections; however, these flaws were equally distributed among the two groups of experiments.

Participation in the 30-minute user study was mandatory for all participants, and students received course credit for completion. Actual participation is 98.3%, given that 3 students inadvertently did not start the time-logged experiment.

## 7.4 Recommendations

Presentation videos follow an inherent structure, which can be identified and used in visualization. In our corpus of material, a given video contains between 1 and 5 presentations, each of which follows the structure of: Setup, Introduction, Main body, Question and Answers, and Exit. The video browsing tool would benefit from such additional higher-level segmentations to enhance searching and browsing. Our experiments clearly show the dramatic difference in finding a video that is relatively close to the beginning of a video (Figure 9a) compared to one that is in the middle of a video (Figure 9b): accuracy drops to below 50%, and time required for the search doubles. Results are even more dramatic when video, rather than thumbnails are being used. Identification and visualization of higher-level structures may improve the searching and browsing experience for the user.

Our survey reveals that students would be in favor of seeing more still pictures, while also suggesting a better organization than what the present tool offers. Because of the outlined efficiency boost of using summary keyframes, a keyframe player (see previous work on instructional video [8]) would serve as an alternative to a streaming video player, but with additional time savings and fast browsing.

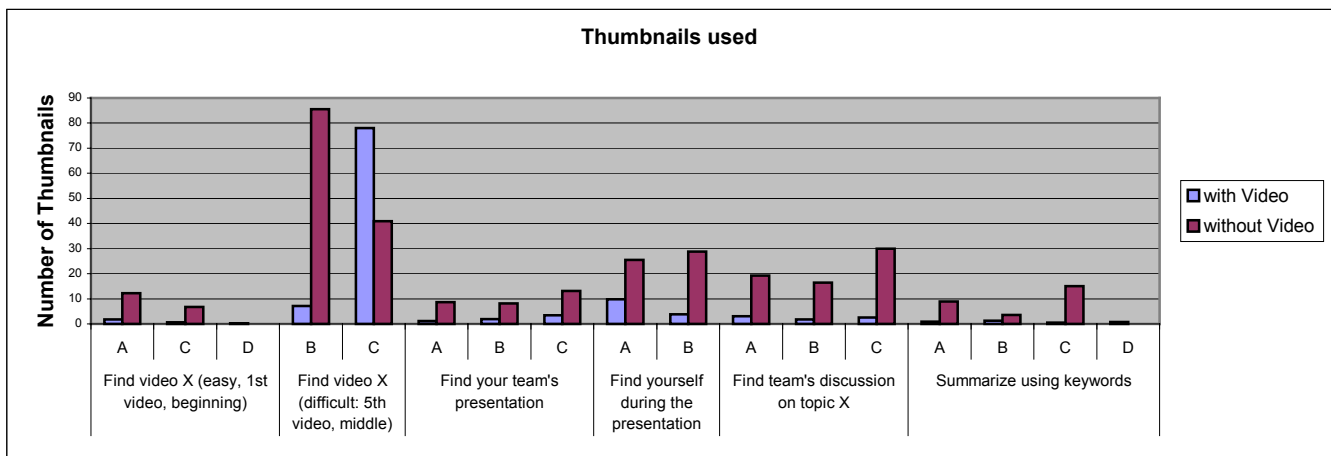
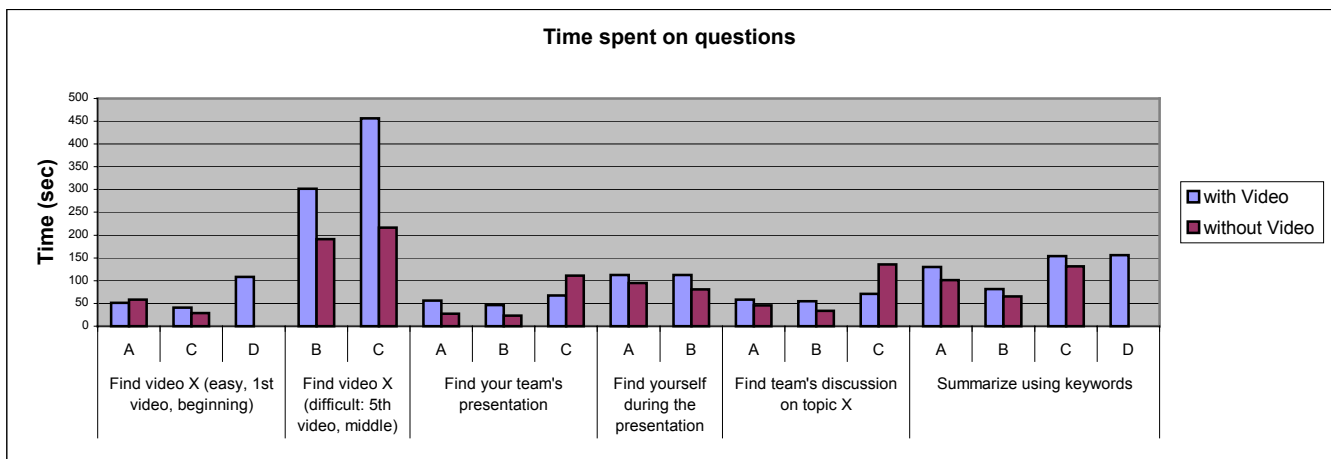
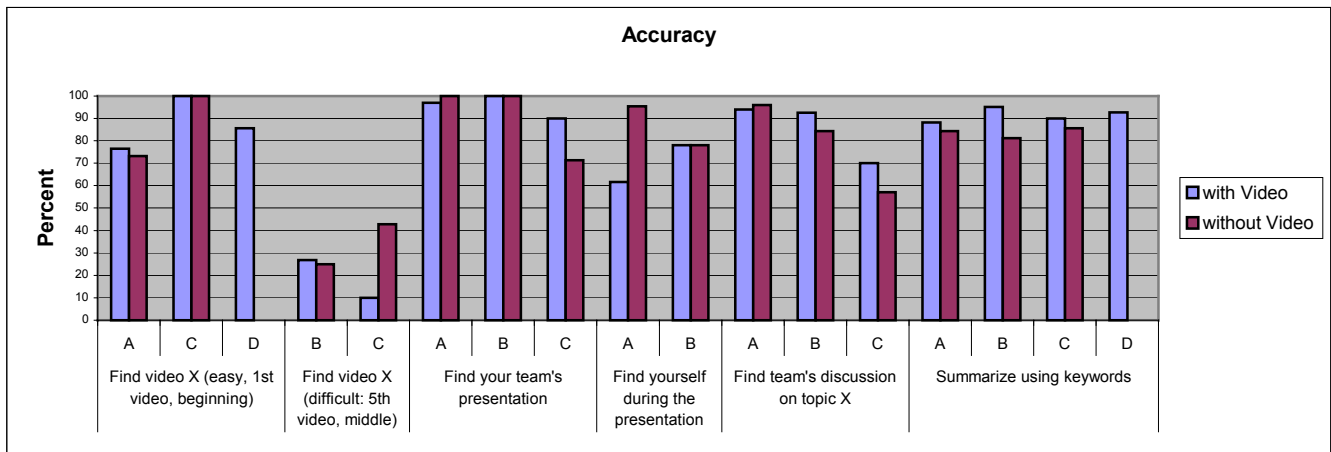
Text graphs were used extensively by the participants; however, they commonly cited problems of having to deal with too many phrases. This problem of organization, while addressed with the zoom feature, requires additional attention. Possibilities for remedies include previous work (see [9]).

While audio and video segmentations were used implicitly in the video summaries, their respective graphs did not receive much attention during the user study. Because they were only intended for visualization purposes at this point, participants pointed out that an interactive implementation of these graphs would benefit the browsing tool. Potential features include filtering of video information, and playback of speakers' audio samples.

Finally, students would like to see this tool be used for browsing lecture videos, citing the benefit of having a full visual record and being able to revisit material at their own pace.

## 8. SUMMARY AND FUTURE WORK

We have presented methods of segmentation, text augmentation, and visualization of presentation videos. Our approach of combined analysis and visualization of audio and video shows that the two media are neither inclusive nor exclusive, but complementary.



**Figure 9 a-f:** Statistics collected from 176 user study participants in 4 groups (A, B, C, D), summarizing the overall improvements of the tool. We compare students having used the tool with and without access to the video stream. While accuracy of answers has stayed roughly the same (a), the speed at which questions were answered improved in almost all cases when using only the video summaries (b). Thumbnails extracted during video segmentation were used as a substitute when the video stream was not available (c), which accelerated the search queries.

We enhanced the segmentation by using index phrase filtering to provide further cues for browsing and searching of presentation video content.

We have evaluated our methods and the experimental tool, and find good evidence to support its benefits for browsing and searching. In the immediate future, we plan on integrating styles of indexing and visualization from previous work to fit the genre of presentation video. In [8], we have introduced an abstract topological view for lecture videos, which students found very useful. In [9], we have used imperfect text transcripts to build an interface that visually grouped lecture videos from an entire semester (15-40 videos). We will be investigating these methods of visual abstraction and video library generation for presentation videos, to address the need for showing video structure and providing a video library browser. We hope to introduce our new browsing tool to the classroom in a more permanent setting and to evaluate the effects of having access to such a tool for the improvement of student presentation skills.

We have begun to investigate gesture classification, and speaker clustering and labeling to extract additional structure from presentation videos. From the opposite point of view, manual annotation of the presentations by instructors with grades, comments, etc., have been cited as desirable additions, whose usefulness will need to be investigated.

Finally, a library search beyond merely displaying a list of videos requires closer attention. With the addition of hundreds of presentations a year, scalability of the video archive becomes an interesting topic of research.

## 9. ACKNOWLEDGMENTS

We would like to thank the Dean's office of the School of Engineering and Applied Science at Columbia University for the support of this research and for the opportunity of conducting the extensive user study among the student body.

## 10. REFERENCES

- [1] Abowd, G.D., Atkeson, C.G., Feinstein, A., Hmel, C., Kooper, R., Long, S., Sawhney, N., Tani, M. Teaching and Learning as Multimedia Authoring: The Classroom 2000 Project. In *Proceedings of the ACM Multimedia conference (MM '96)* (Boston, MA, November 18 – 22, 1996). ACM Press, New York, NY, 1996, 187-198.
- [2] Amir, A., Berg, M., Chang, S.-F., Hsu, W., Iyengar, G., Lin, C.-Y., Naphade, M., Natsev, A., Neti, C., Nock, H., Smith, J.R., Tseng, B., Wu, Y., Zhang, D. IBM Research TRECVID-2003 Video Retrieval System. In *Proceedings of the TREC Video Workshop (TRECVID '03)* (Gaithersburg, MD, November 17 – 12, 2003)
- [3] Amir, A., Srinivasan, S., Ponceleon, D., Petkovic, D. CueVideo: Automated video/audio indexing and browsing. In *Proceedings of the ACM International Conference on Research and Development in Information Retrieval (SIGIR '99)* (Berkeley, CA, August 15 – 19, 1999). ACM Press, New York, NY, 1999, 326.
- [4] Blei, D.M., Moreno, P.J. Topic Segmentation with an Aspect Hidden Markov Model. In *Proceedings of the ACM International Conference on Research and Development in Information Retrieval (SIGIR '01)* (New Orleans, LA, September 9 – 13, 2001). ACM Press, New York, NY, 2001, 343-348.
- [5] Chaisorn, L., Chua, T.S., Lee, C.H. The segmentation of News Video into Story Units. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME 2002)* Lausanne, Switzerland, August 26 – 29, 2002). IEEE Computer Society Press, 2002, Volume 1, 73-76.
- [6] Chen, S.S., Gopalakrishnan, P.S. Speaker, environment and channel detection and clustering via the Bayesian Information Criterion. In *Proceedings of the 1998 DARPA Broadcast News Transcription and Understanding Workshop* (Landsdowne, VA, February 28 – March 3, 1998). 127-132.
- [7] Gauvain, J.L., Lamel, L., Adda, G. Audio Partitioning and Transcription for Broadcast Data Indexation. *Multimedia Tools and Applications, Vol. 14, Issue 2* (June 2001), 187-200.
- [8] Haubold, A., Kender, J.R. Analysis and Interface for Instructional Video. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME '03)* (Baltimore, MD, July 6 – 9, 2003). IEEE Computer Society, 2003, Volume 2, 705-708.
- [9] Haubold, A., Kender, J.R. Analysis and Visualization of Index Words from Audio Transcripts of Instructional Videos. In *Proceedings of the IEEE International Workshop on Multimedia Content-based Analysis and Retrieval (MCBAR '04)* Miami, FL, December 15, 2004). IEEE Computer Society Press, 2004, 570-573.
- [10] He, L., Sanocki, E., Gupta, A., Grudin, J. Auto-Summarization of Audio-Video Presentations. In *Proceedings of the ACM Multimedia conference (MM '99)* (Orlando, FL, October 30 – November 5, 1999). ACM Press, New York, NY, 1999, 489-498.
- [11] Mukhopadhyay, S., Smith, B. Passive Capture and Structuring of Lectures. In *Proceedings of the ACM Multimedia conference (MM '99)* (Orlando, FL, October 30 – November 5, 1999). ACM Press, New York, NY, 1999, 477-487.
- [12] Ponceleon, D., Srinivasan, S. Automatic discovery of salient segments in imperfect speech transcripts. In *Proceedings of International Conference on Information Knowledge Management (CIKM '01)* (Atlanta, GA, November 5 – 10, 2001). ACM Press, New York, NY, 490-497.
- [13] Smith, M.A., Kanade, T. Video Skimming and Characterization through the Combination of Image and Language Understanding. In *Proceedings of the International Conference on Computer Vision (ICCV '98)* (Bombay, India, January 4 – 7, 1998). Narosa Publishing House, 1998, 61-70.
- [14] Wilson, A., Bobick, A. Hidden Markov models for modeling and recognizing gesture under variation. *Hidden Markov models: Applications in Computer Vision*. World Scientific Publishing Co., 2001, 123-160.