ANALYSIS, USER INTERFACE, AND THEIR EVALUATION FOR STUDENT PRESENTATION VIDEOS

Alexander Haubold and John R. Kender

Department of Computer Science, Columbia University

ABSTRACT

In the domain of candidly-captured student presentation videos, we examine and evaluate approaches for multi-modal analysis and indexing of audio and video. We apply visual segmentation techniques on unedited video to determine likely changes of topics. Speaker segmentation methods are employed to determine individual student appearances, which are linked to extracted headshots to create a visual speaker index. Videos are augmented with time-aligned filtered keywords and phrases from highly inaccurate speech transcripts. An experimental user interface (UI) combines streaming videos, visual, and textual indices for browsing and searching. We evaluate the UI and methods in a large engineering design course. We report on observations and statistics collected over 4 semesters and 598 student participants. Results suggest that our video indexing and retrieval approach is effective, and that our continuous improvements are reflected in an increase in accuracy and completion rates of user study tasks.

1. INTRODUCTION

Video is a versatile medium, whose role in the classroom beyond lecture recordings has not yet been explored. They can be used effectively to record team interaction, student performance during presentations, or project work progress over the duration of a term. Video libraries of presentation videos over several semesters present a novel approach of archiving student work. One of the reasons for the reluctant acceptance of this medium for such classroom use is the time commitment required for production. In addition to equipment and camera operator expenses, the amount of time necessary for post-production and dissemination present a burden for instructors. Even when made available in its unedited format, means of finding information quickly do not exist without additional manual labor.

Instructional videos have received much attention, and were investigated in passive [1] and invasive environments [2]. Structuring and indexing of content is performed using visual cues [1,3] and textual cues within [4], and across [5]

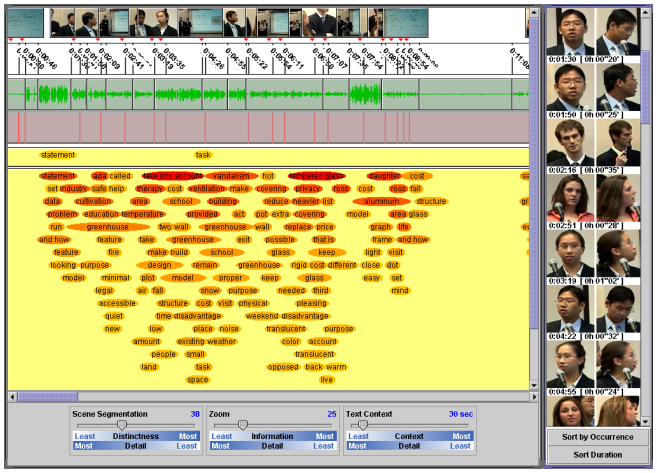

Figure 1: Video summary for presentation video. Indices include thumbnails, speaker and visual segmentation cues, keywords/phrases, and a visual speaker list.

lectures. Professional presentation videos by single speakers in controlled environments have also been considered [6].

In this paper, we focus on student presentation videos in a large university-level engineering course with more than 150 students per semester. Students work in teams of 4-8 on various projects, and hold two presentations summarizing their progress. This has yielded a massive database of 184 video sequences of approximately 162 hours over a period of 5 years. Characteristic of these videos is their low-quality production, which for their limited use does not require or merit elaborate setup or post-editing. Lighting and acoustics of the classroom remain unmodified, and exhibit large variations between recordings. A separate microphone is used to better capture presentations; however, variations in quality occur with varying speaking skills, volume, etc.

Indices and user interface for presentation videos must be sensitive to their intended use. Useful indices for student presentation videos include visual summaries of scenes, text indicative of content, and means of locating speakers (Fig. 1). Visual segmentation is performed by combining two methods that determine scene changes, one for abrupt and one for gradually changing content. Speaker segmentation determines individual student appearances, and is the foundation for extracting headshots for the visual speaker index. Highly inaccurate automatic transcripts are generated and filtered to produce relevant keywords and phrases [7].

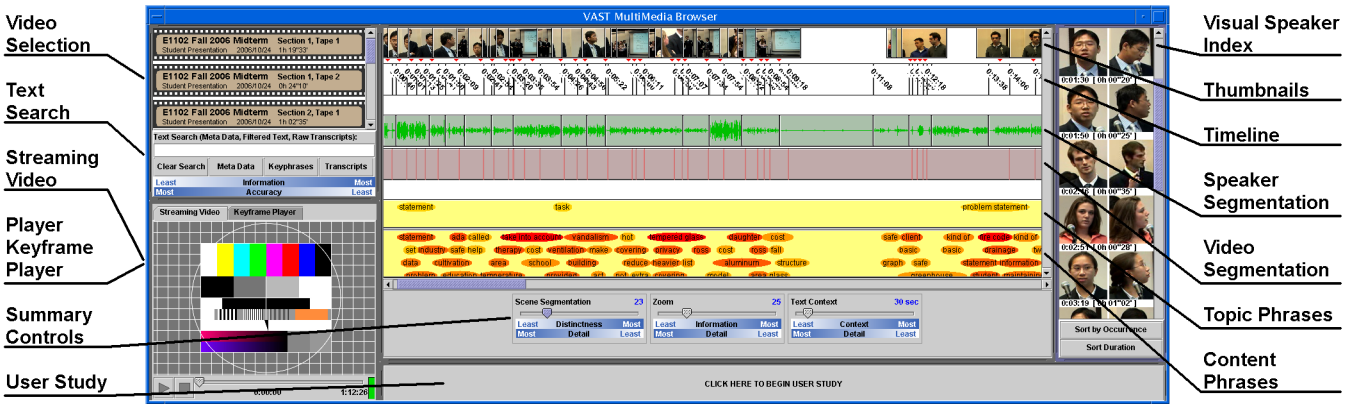

Figure 2: User Interface featuring video summaries in the multi-modal domain.

2. VISUAL SEGMENTATION

Unedited presentation videos do not feature scene cut production cues. Also, the recording environment does not clearly separate between stage and audience. We apply two methods that generate visual segmentation to account for the noisy video data. We then allow the user to select the granularity of visual change in the user interface.

When presentations make use of electronic slides and the camera captures their content, slide changes are used as a cue to indicate an interesting visual change. Abrupt visual activity of this kind is similar to scene cuts in edited video. We use a windowed approach that detects significant visual change V in the immediate neighborhood of a point P in the video. We divide a video frame into a 10x10 grid and compute consecutive frame differences using pixel intensity change between sub-regions of two frames. Intensity change is defined by a threshold value of 30 (out of 255) beyond which the difference contributes a unit amount. Only up to 10% of absolute visual change in a sub-region contributes to the global aggregate measure of difference between two frames. We impose this sub-region measure to attenuate the absolute contribution of visual activity in small regions. Using this approach, the global intensity change is clipped at 10% of the maximum amount of change that can occur between two frames, e.g. a complete change of color. Significant visual change is detected between two video frames (point P) when their value deviates significantly from the visual change prior to or after P. We define:

V_{left}: Visual activity in the set $\{V(P_{left\text{-}2sec}), \ldots, V(P_{left})\}$

V_{right}: Visual activity in the set $\{V(P_{right}), \ldots, V(P_{right+2sec})\}$

Significant visual change is detected when:

$$\max(V_{left}) < V(P) - \mu(V_{left}) \wedge \max(V_{right}) < V(P) - \mu(V_{right})$$

Our second method of detecting interesting visual events is based on gradually changing content in the video, such as camera pans, zooms, entering and/or leaving of a person with respect to the camera view. We employ a windowed approach to comparing histogram differences between more distant frames, in our case 4 seconds apart. We have determined experimentally that a color histogram with 2 bits per color resulting in 64 bins is sufficient in capturing visually significant events. Due to the pair-wise comparison of distant frames, a change in visual content is determined by a significant drop in the similarity measured between the left and right windows. Such a point characterizes the start of calm visual activity after a sequence of motion.

In a final step, visual activity from both methods is combined. Resulting events are identified by their intensity, which are used as tunable parameters in the user interface.

3. VISUAL SPEAKER INDEX

Next to presentation content, students and their performance are the main focus in student presentation videos. The duration of an individual student's appearance is relatively short and ranges from 5 seconds to 5 minutes. A typical 80-minute presentation video contains about 60 student appearances, a large number of which are repetitions. A common task for users of video summaries is to locate the appearance of a particular individual.

An effective index into these videos must accommodate the intended use of searching for presenters. In the absence of manually annotated names, a visual index of speaker faces is an alternative. We rely on automatic speaker segmentation [7] to create a list of individual speakers. We then manually extract regions from video key frames, which best portrait the individual presenters. Contrast normalization adjusts the highly varying lighting conditions in the recorded video.

We have designed four slightly different versions of a visual speaker index (Fig. 3). In order of search and retrieval performance, they are: (1) large (75x75 pixel) head/shoulder and profile shots, (2) small (50x50 pixel) headshot, (3) large head/shoulder shot, and (4) small head/shoulder and profile shots. As part of our user study, we have measured the effect of each configuration.

4. USER INTERFACE

Our user interface includes various visual summarization tools, means for searching the video library, and a streaming

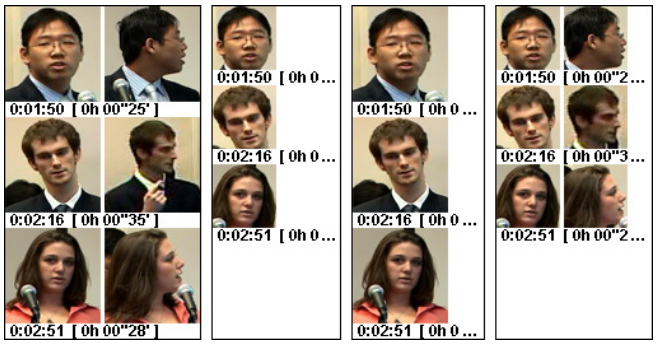

Figure 3: Four versions of speaker indices in order of search and retrieval performance. Left to right: (1) Large head/shoulder shot and profile shot, (2) Small head shot, (3) Large head/should shot, (4) Small headshot and profile shot. Performance is measured by duration and completion rate for a user study search task.

video player to play back full contents (Fig. 2). Video summaries include thumbnails, timeline, speaker and visual segmentations, and keyword and key phrase graphs. Mouse movement over thumbnails enlarges them to full-size images. Keywords and phrases are represented by blips, whose background color and vertical position denotes their importance. Words and phrases with stronger meaning are colored red and are located closer to top.

Three controls are included for the user to customize the summaries. "Scene Segmentation" varies the granularity of visual segmentation, which also drives the number of displayed summary thumbnails. A lower value decreases distinctiveness of scenes, but increases detail. "Zoom" allows the user to vary the duration of the video summary visible per unit screen space. A lower value decreases the duration of video summary and amount of information visible on the screen, but increases the level of detail. Higher values are preferred for obtaining a superficial overview of video content, while a lower value is more useful for exploring sections of the video, for example short presentations. "Text Context" creates a temporal context between phrases by combining repeated usage of phrases. Increasing text context groups similar words temporally, expanding their blips horizontally to mark the duration over which the represented text is used in the video.

We include an experimental visual speaker index in lieu of cumbersome annotation of student names. While headshots are presently extracted manually from the underlying speaker segmentation, existing research shows that automatic extraction is feasible [8,9].

5. USER STUDY

We periodically administer user studies in our large engineering design course to evaluate the usefulness of our tools. Task-based experiments measure the effects of various UI components, focusing on duration, completion, and accuracy of student responses. We have collected data from 598 participants over a 4-semesters period.

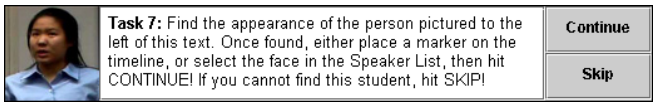

Figure 4: User study task for search and retrieval of a presenter. Given the visual cue of the person, students must find the appearance in a set of videos.

	Large Head / Shoulder & Profile	Small Head	Large Head	Small Head / Should & Profile
Completion	97%	97%	91%	91%
Duration (sec)	86.72	126.12	137.12	155.94
Participants	31	30	35	33

Table 1: User study results for visual speaker index in order of decreasing performance.

Adjustments and improvements to the video browser are made after each term, taking into account results from our user studies and suggestions from surveys.

We have designed a set of 5-7 tasks related to search and summarization of video content, which students must complete. These tasks are comparable to typical queries performed on a set of videos containing many presentations:

1. Find your own appearance in the video.
2. Locate the portion of the video in which your team discusses topic XYZ.
3. Find the beginning of your team's presentation.
4. Find the presentation on subject XYZ (titled ABC).
5. Using the available keywords for the presentation located between TIME1 and TIME2, summarize the project's goals as best as possible.

We define several criteria for evaluation. Completion rate denotes the number of tasks completed properly. A lower value indicates that tasks were skipped often, whether due to frustration of not finding the answer, or advertently/ inadvertently skipping tasks. Accuracy measures the temporal distance between a user's selection and the correct answer for tasks related to searching. Finally, we measure the temporal duration of a task.

We have observed very positive developments with continuous improvements to our video browser. Overall, task completion rates have improved from 82% to 92% over 4 semesters. For the most characteristic search task of locating an unfamiliar presentation in a set of several videos, the completion rate has improved from 58% to 73%. Accuracy, too, has increased overall, but we note an interesting "trust" effect. In the absence of a text search feature, which is particularly useful when locating unfamiliar material, students apply more care in locating the correct response in the entire set of videos, which requires more time but increases accuracy. Surprisingly, if a text search engine is used, accuracy drops significantly from 4 to 229 seconds, while completion increases from 52% for 73%. Analysis shows that the average is due to a number of

outliers with high off-target answers, while the remaining 80% of students still mark the correct answer within an error margin of 4 seconds. We believe that the high off-target responses are due to students trusting the search results, which correctly narrow the search domain to one video, but do not identify the approximate location of the results. In lieu of an exact response, the next best answer is to select a random location in the video. In the next iteration of the tool, we are including a feature by which search terms are highlighted in their matching locations in the video.

In our latest user study, we have evaluated the visual speaker index for the most effective configuration. Students were presented with the task of locating an unfamiliar face in a set of videos (Fig. 4). Our 129 participants were randomly assigned one of the four speaker indices (Fig. 3). Results from this task are presented in Table 1. The highest performance is exhibited by the speaker index with the most visual detail, namely large head/shoulder and profile shots. The success of this configuration cannot, however, be correlated to the use of a head/shoulder shot for the user study task, because the second highest performance for search and retrieval is evidenced by the small headshot.

Analysis of time required to fulfill tasks shows an overall decrease compared to earlier versions of the interface. Average time for a search or summarization task is 100 seconds. Search tasks for unfamiliar content far outweigh all other tasks with an average of 5 minutes. However, we should note that the correct response for a question of this type is found in a window of 5-10 seconds from video footage with duration 23,380 seconds (6.5 hours). For summarization tasks we observed only a nominal increase in duration, which is due to the shift from multiple choice to entry responses in our latest iteration of the user study. For these comparative results over 4 semesters, we have only considered user studies administered in comparable settings.

For three consecutive semesters (469 participants) we have measured the effect of making available the actual video as part of the video summaries. Half of the participants had access to the video, while the other half was unable to view video or listen to the audio track. We have found that for tasks of search and summarization, availability of video was counterproductive. With similar accuracy and completion of tasks, students with access to video required 50% more time to complete their assigned tasks. We believe that with access to video, students get "stuck" watching extraneous material without making effective use of summarization tools.

In our latest iteration we have evaluated the effect of environment among other factors. Half the class completed the user study in lab under supervision, while the other half completed it as homework. An on-line tutorial replaced in-class instructions for home users. Statistics of usage vary greatly between the two groups, while completion rate was marginally better for in-class students (by 4%). On average, home users took one third as much time to become acquainted with the video browser (344 vs. 947 sec), and required twice as much time to complete tasks (196 vs. 100 sec). They used significantly more streaming video (24 vs. 11 sec per task) and less time adjusting user controls (11 vs. 20 sec per task). In part this discrepancy can be explained by the difference of introduction to the tool.

In general, we can conclude that our methods of video analysis and our tools for searching and visualization are effective for information retrieval in video libraries. User studies have helped identify shortcomings and strong points, and addressing them in subsequent improved versions resulted in improvements in search and retrieval. Our selection and enhancement of visual, speech, text cues and their UI components indicate that our automatic analysis of video is effective for increasing accuracy and completion, and decreasing duration of search and summarization tasks.

6. CONCLUSION

We have presented approaches for segmentation and indexing of presentation video. Methods of analysis and design of user interfaces were successfully evaluated in user studies with 598 students over four semesters. Performance of tasks related to video search and retrieval has increased with the introduction of improved methods. In future work, we will be further analyzing visual speaker indices and their automatic creation.

7. REFERENCES

[1] S. Mukhopadhyay, B. Smith, "Passive capture and structuring of lectures," *MM '99*, ACM, Orlando, FL, pp. 477-487, 1999.

[2] G.D. Abowd, C.G. Atkeson, A. Feinstein, C. Hmelo, R. Kooper, S. Long, N. Sawhney, M. Tani. "Teaching and learning as multimedia authoring: The classroom 2000 project," *MM '00*, ACM, Boston, MA, pp. 187-198, 1996.

[3] A. Haubold, J.R. Kender, "Analysis and Interface for Instructional Video," *ICME '03*, IEEE Press, Baltimore, MD, pp. 705-708, 2003.

[4] M. Lin, J.F. Nunamaker, M. Chau, H. Chen, "Segmentation of lecture videos based on text: a method combining multiple linguistic features," *HICSS '04*, Big Isl., HI, pp. 3-11, 2004.

[5] Haubold, A., Kender, J.R, "Analysis and Visualization of Index Words from Audio Transcripts of Instructional Videos," *MCBAR '04*, IEEE, Miami, FL, pp. 570-573, 2004.

[6] L. He, E. Sanocki, A. Gupta, J. Grudin, "Auto-summarization of audio-video presentations," *MM '99*, ACM, Orlando, FL, pp. 489-498, 1999.

[7] A. Haubold, J.R. Kender, "Augmented segmentation and visualization for presentation videos," *MM '05*, ACM, Singapore, pp. 51-60, 2005.

[8] H.L. Wang, S.F. Chang, "A Highly Efficient System for Automatic Face Region Detection in MPEG Video", *IEEE Transactions on Circuits System for Video Technology, Vol. 7, No. 4*, pp. 13, 1997.

[9] R. Cutler, L. Davis, "Look who's talking: speaker detection using video and audio correlation," *ICME '00*, IEEE, New York, NY, pp. 1589-1592, 2000.