Topic Based Segmentation of Classroom Videos

Tayfun Tuna, Mahima Joshi, Varun Varghese, Rucha Deshpande, Jaspal Subhlok, Rakesh Verma
Department of Computer Science
University of Houston

***Abstract*—Video of classroom lectures is a valuable and increasingly popular learning resource. A major weakness of the video format is the inability to quickly access the content of interest. The goal of this work is to automatically partition a lecture video into topical segments which are then presented to the user in a customized video player. The approach taken in this work is to identify topics based on text similarities across the video. The paper investigates the use of *screen text* extracted by Optical Character Recognition tools, as well as the *speech text* extracted by Automatic Speech Recognition tools. An automatic text-based segmentation algorithm is developed to identify topic changes and evaluated on a set of twenty-five lecture videos. The key conclusions are as follows. Screen text is a better guide to discovering topic changes than speech text, the effectiveness of speech text can be improved significantly with the correction of speech text, and combining screen text and accurate speech text can improve accuracy. Results are presented from surveys showing a high level of satisfaction among student users of automatically segmented videos. The paper also discusses the limits of automatic segmentation and the reasons why it is far from perfect.**

I. Introduction

Video is gaining popularity as a learning resource. Video recordings of classroom lectures are often made available as additional material for a conventional course, as the core of a distance/hybrid learning course, or posted publicly for community learning. Lecture videos are posted on a large scale on portals such as MIT OpenCourseware and Apple's iTunes University. In recent years MOOCs (Massive Open Online Courses) driven by video and other features have emerged as a potential disruptive technology for the delivery of education. There is a substantial body of research that has established that video is a versatile learning resource that is considered valuable by students and instructors [1], [4], [15], [17], [18]. The lecture videos that capture the overall classroom interaction provide an experience that mirrors the actual class to the students who are not able to attend. However, video is also commonly employed by students to access specific information, not just to replace missed lectures. In particular, review of the class content, e.g., for quizzes and exams, is an important use of video. Efficient retrieval of the appropriate information in a long lecture video is a major challenge with the video format. Therefore, dividing videos into topical segments is important for the advancement of video as a learning tool.

The research presented in this paper is in the context of the ICS (Indexed, Captioned, Searchable) Videos project at the University of Houston[18], [22]. The goal of the project is to ease navigation of lecture videos, making them a companion resource for learning, similar to a textbook. A video lecture is automatically partitioned into segments based on image and text analysis. We refer to this process as *indexing* or *segmentation*.[1] Video is searchable for keywords and concepts. Captions are developed for videos with speech recognition and crowdsourcing by students. All videos for an entire course (or department) are treated as a single "videobook" stream with global indexing and search capability. Several thousand students were surveyed and hundreds of students participated in focus groups during the project. Conclusions from this project relevant to the research presented in this paper are i) videos are a very valuable learning resource and ii) indexing enhances the value of videos significantly [3], [22].

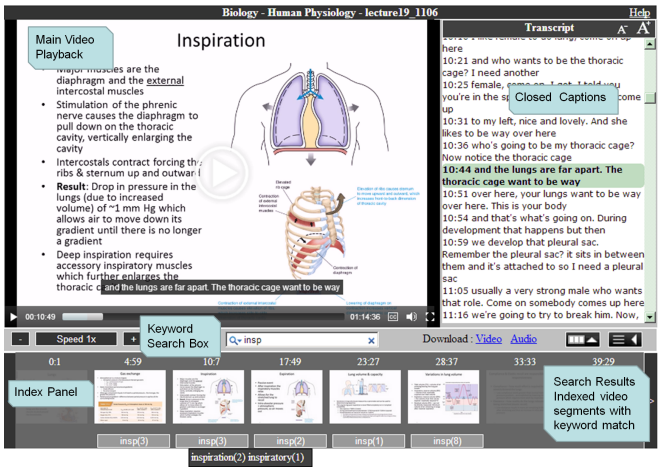


Fig. 1. ICS Video player with index points, search box, captions and transcript

The ICS Video player that encapsulates indexing, search, and captioning is illustrated in Figure 1. An index panel is situated on the bottom of the player; each index point represents a new topic in the form of a screen-shot of the video at that point of time. Users can navigate different topical segments of the video by clicking these index points. A search box is located below the main video playback for searching inside the video. Keyword search across videos is also supported. Captions are displayed as an overlay at the bottom of the main video playback window, in addition to a separate transcript window on the right side.

The goal of the research presented in this paper is to automatically partition a lecture video into topical segments, which can then be presented to the users as visual index points. However, automatically segmenting a video lecture by topic or

Contact email: jaspal@uh.edu

[1] In this work, the terms *indexing* and *segmentation* are used interchangeably.

subtopic is a very challenging problem as the precise meaning of a topic is subjective. The approach taken in this work to identify topics is based on text similarities across the video. A segmentation algorithm based on *cosine similarity*, a common metric to measure the similarity between two blocks of text, was employed. For topic-based segmentation, two approaches were investigated based on *screen text* or *speech text*. Screen text is the text that appears on the video frames, which typically corresponds to the viewgraphs used in teaching a class, but can also be from other sources such as web pages. Screen text is extracted with the help of optical character recognition (OCR) technology [20]. Speech text is the text corresponding to the audio in a lecture video, which includes everything spoken by the instructor as well as the interaction with students. Speech text is gathered by using an Automatic Speech Recognition (ASR) system. For selected videos, automatically generated speech text was corrected manually to remove errors in speech recognition.

Evaluation was done on a set of twenty-five lecture videos from courses in Computer Science and Biology and Biochemistry. The ground truth was established by asking the lecture instructor or another topic expert to manually identify topic transitions in the video. The segmentation obtained with screen text, speech text, corrected speech text, and combinations of these were evaluated for accuracy against the instructor generated ground truth. The inherent inaccuracy of human segmentation was also measured by asking multiple subject experts to segment the same videos by topic. The results show the accuracy of different approaches to segmentation as well as the limitations of the text based automatic segmentation process. The main reasons for errors in automatic segmentation are also presented based on a manual analysis of some of the videos employed for evaluation.

This paper is organized as follows. Section II discusses prior work related to segmentation of lecture videos. Section III discusses the extraction of screen text and speech text. Section IV presents the text based automatic indexing algorithm. Section V discusses the evaluation methodology and presents the results of indexing algorithm employing screen text and speech text. Section VI presents the results of evaluation of indexing by student users based on survey results. Section VII discusses the reasons for indexing errors with slide text and speech text. Section VIII contains conclusions.

II. Related Work

In general, video segmentation or indexing requires the detection of key frames or labels that indicate a change of content in a video [6], [8], [11], [14]. A multitude of methods have been developed that use low-level image properties, such as color and texture, to group contiguous video frames and provide reasonable automation while lacking the ability to provide topical segmentation [2], [7], [12], [16]. The work presented in this paper focuses on classroom lecture videos or screencasts. We employ similar techniques as a preprocessing step for detecting the slides in lecture videos.

Topic based segmentation of lecture videos requires processing the screen text extracted by OCR, and/or speech text extracted by ASR. Various methods have been developed that use both OCR and ASR data for content-based video retrieval, semantic multimedia retrieval, and meta-data generation [10], [13], [25]. Extraction of segments and keywords from both OCR and ASR methods and ranking the keywords is discussed in [25]. Comparing the speech text segments for similarity to determine the topic boundaries is studied in [10] employing a dictionary-based approach that compares selected features among segments. However, human supervision is required for customizing the dictionary for a particular subject area. The indexing in our work is different as the video indexing method is unsupervised and fully automated.

In summary, the main directions of related research are indexing of movie videos, segmentation based on visual properties, and extraction and analysis of OCR and ASR keywords. These are complementary to the work presented in this paper. The main subject of this paper is how speech and text compare as the input for segmentation of videos, if they can be used together, and the reasons why these approaches often fail.

III. Extracting Text From Videos

The main research objective of this paper is segmentation of video lectures based on textual content of the lectures. Here we discuss how different types of text are extracted from a video.

A. Screen Text

Screen text is obtained by applying Optical Character Recognition (OCR) tools to video frames. Typically this text corresponds to viewgraphs employed during the lecture but can also include other content such as web sites or files displayed during a lecture. One of the premises of this research is that an analysis of screen text can provide guidance on topic transitions in a video lecture.

After a comprehensive analysis of available OCR tools, we opted to use the MODI (Microsoft Office Document Imaging) tool set. We found that OCR tools generally have limited effectiveness at recognizing text in the presence of 1) certain text and background color and shade combinations, 2) text mingled with colorful shapes, and 3) small and exotic fonts. To increase the detection efficiency of text on video frames, we used simple image processing techniques for image enhancement (IE) prior to the application of OCR tools. IE operations employed include segmentation of text, enlargement with interpolation, and color inversion. The process of obtaining screen text from videos employed in this research is detailed in [23]. Typically an accuracy of well over 90% is obtained with this enhanced OCR extraction framework. Hence this work does not consider manual correction of OCR errors.

B. Speech Text

Spoken text is simply the text corresponding to the audio in a recorded lecture. It primarily consists of the lecture from the instructor but may also include student interaction. Speech text can provide important information that determines topic changes in a video.

Various ASR(Automatic Speech Recognition) tools are commercially available and we experimented with *Dragon Naturally Speaking*, *Windows Speech Recognition*, and *YouTube*. In the end, YouTube was employed based on an

analysis discussed in [5]. The accuracy of speech recognition varies widely based on the instructor and lecture content. The average accuracy in our experiments was only around 68%.

C. Hybrid Text

Hybrid text is simply the union of screen text and speech text. In order to utilize the strengths and topic-related keywords from both speech and screen text, we employed a hybrid text type for video indexing purposes. It should be noticed that the volume of speech text typically far exceeds the volume of screen text.

D. Corrected Speech Text

All ASR tools generate significant errors when employed on the speech component of classroom videos. There are various reasons for errors, such as a heavy accent, technical vocabulary, poor recording, and the colloquial nature of a classroom lecture. The speech text was corrected manually using a crowdsourced caption editor [5] in order to evaluate the impact of ASR errors on topic based video segmentation. The average speech text accuracy on selected videos was improved from 68% to 99% with this correction process.[2]

IV. Text-based Indexing

Indexing is the task of dividing a lecture video into segments that contain different topics. A video is composed of a sequence of thousands of images (or frames). In order to process video data efficiently, a video segmentation technique should detect scene changes and find the unique images. Therefore, video segmentation task involves two steps as depicted in Figure 2. First step is preprocessing to identify all transition points, i.e., places where the image on the video changes significantly. Subsequently, a subset of these transition points are selected as index points representing topic change based on text analysis. The assumption is that topic transitions happen at transition points which typically represent slide changes in a lecture.

A. Preprocessing: Identifying Transition Points

Identification of transition points is based on a comparison of successive frames in the video. Frames are commonly recorded in 24-bit RGB representation; color value for each pixel is encoded in 24 bits where three 8-bit unsigned integers (0 through 255) represent the intensities of red, green, and blue. Corresponding pixels in successive frames are considered different if they differ by a minimum RGB threshold when the RGB values of the pixels are compared. The threshold value is chosen empirically after evaluation of a large number of diverse lectures. Details of the process of identification of transition points is discussed in [23].

[2]About 1% of the words were not correctly identified by students making the corrections manually. The ground truth is the instructor's version of the transcript.

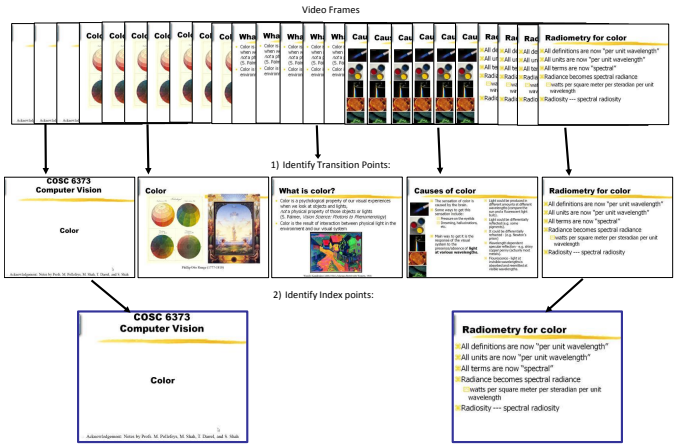


Fig. 2. Indexing framework steps: 1) Transition points (unique video frames) detected by RGB Color difference. 2) Index points representing different topics are selected among the transition points.

B. Text Similarity Metric: Cosine Similarity

The core idea of text based segmentation is that different topics are represented by different groups of words. Comparing the frequencies of different words in blocks of text establishes how similar they are in content and topic. Intuitively, a video splits into different topical segments at the point where the mix of words being used in video frames changes significantly. And this change can be detected by a comparison of the similarity of two text blocks. While many different text similarity metrics have been discussed in literature, we used *cosine similarity*, a well known and proven metric in information retrieval and text mining [9], [19]. It is a measure of similarity between two vectors, calculated by the dot product of the vectors divided by the product of their norms as shown by the formula below. The vectors A and B correspond to the frequency of words in the context of text based segmentation.

$$cos(\theta) = \frac{A.B}{\|A\|.\|B\|} = \frac{\sum_{i=1}^{n} A_i.B_i}{\sqrt[2]{\sum_{i=1}^{n} A_i^2}.\sqrt[2]{\sum_{i=1}^{n} B_i^2}}$$

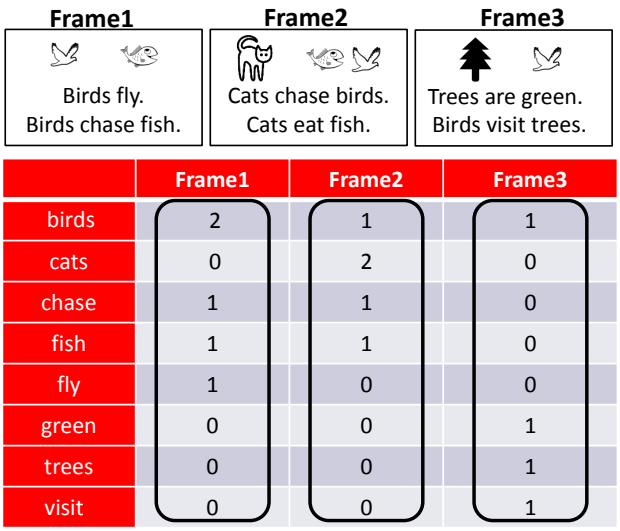
An example of text similarity calculation is depicted in Figure 3. Three frames and their word frequency vectors are listed. The cosine similarity between the vectors representing adjacent frames is computed as follows.

$$\text{cosine\_similarity}(Frame_1, Frame_2) = 0.57$$
$$\text{cosine\_similarity}(Frame_2, Frame_3) = 0.19$$

This matches the intuitive judgment that $Frame_1$ and $Frame_2$ are more similar to each other than $Frame_2$ and $Frame_3$. The implication is that any topic change inside this sequence should start with $Frame_3$. Cosine similarity measure is normalized with respect to document length as it compares the relative frequency of common words.

C. Text-based Indexing Algorithm

The main purpose of the indexing algorithms is to partition a lecture video so that each segment represents a topic. Before the indexing phase, the lecture video is divided into transition segments [21], [24]. The segmentation algorithm repeatedly



	Frame1	Frame2	Frame3
birds	2	1	1
cats	0	2	0
chase	1	1	0
fish	1	1	0
fly	1	0	0
green	0	0	1
trees	0	0	1
visit	0	0	1

Fig. 3. A sequence of frames and their frequency of word vectors is computed to determine similarity text similarity

merges the smallest segment in the video to the segment on the right or left, based on cosine similarity with a group of segments on the left, and a group of segments on the right, respectively, as illustrated in Figure 4. An empirically selected value of *Grouping Duration (480 seconds)* determines the number segments on the left and right that are included for text comparison. The algorithm is explained as follows.

Data: A list of transition points ;
Required number of index points (N);
Grouping duration in seconds;
Result: N index points that are a subset of given transition points;
repeat
  Select transition segment with smallest duration;
  if *the similarity is more towards right group* **then**
    merge right;
  else
    merge left
  end
until *Number of transition points == Required number of index points*;

Algorithm 1: Text-based indexing algorithm

A pictorial example of the algorithm is provided in Figure 4. In this example, the similarity of the smallest segment K is compared with the left as well as the right group and merged with the most suitable neighbor depending on the similarity value.

We have employed a simple indexing algorithm that assumes a fixed number of index points. A detailed comparison of different algorithms is included in [21]. However, the goal of this paper is to compare speech text and screen text as the input for indexing, and we believe this algorithm is adequate for this purpose.

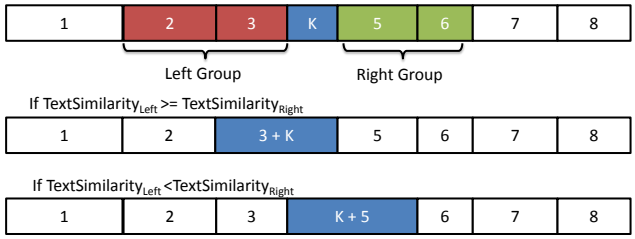


Fig. 4. Text based indexing algorithm: Shortest segment compared to left and right group of neighbor segments and merged based on similarity

V. Evaluation

The objective of evaluation is to measure the accuracy of segmentation based on screen text and speech text.

A. Evaluation Framework

A suite of 25 video lectures listed in Table I was selected for evaluation. The subject areas were Computer Science and Biology and Biochemistry. The sources of the video were lectures recorded at the University of Houston and the Coursera website. The textual content of the videos was obtained by using OCR methods and YouTube as discussed in Section III. For a subset of the videos, the text obtained from YouTube was manually corrected for ASR errors for evaluation.

TABLE I. List of source of courses used for evaluation

Source	Major	Course Name	Lecture Count
UH	Computer Science	Introduction to Computing	4
UH	Computer Science	Computer Organization and Programming	5
UH	Computer Science	Digital Image Processing	2
UH	Computer Science	Computer Architecture	2
UH	Biology	Human Physiology	3
Coursera	Computer Science	Compilers	3
Coursera	Computer Science	Cryptography	2
Coursera	Computer Science	Machine Learning	2
Coursera	Computer Science	Probabilistic Graphical Models	2
Total			25

A major difficulty in evaluating an automatic segmentation algorithm is that the ground truth, i.e., the optimal set of index points, is often not obvious even to the instructor of a course. It is very challenging to decide if a transition point is the start of a subtopic or not. The creator of each lecture video (normally the instructor teaching the course) was asked to rate every transition point on its appropriateness to be an index point based on the extent to which it represented a change in the topic. The following scale was used for ranking:

- Definitely Index Point (+2)
- Probably Index Point (+1)
- Probably Not Index Point (-1)
- Definitely Not Index Point (-2)

However, the output of the segmentation algorithms is binary, i.e., each transition point is determined to be an index point (1) or not an index point (-1). The quality of the set of index points identified by an automatic indexing algorithm is determined as follows. Suppose the ground truth for a transition point is "Definitely Index Point". Then if the algorithm correctly identifies it as an index point, +2 is scored, while if it is incorrectly identified as not an index point, then -2 is scored. Now suppose the ground truth for a transition point is "Probably Index Point". Then if the algorithm correctly identifies it as an index point, +1 is scored, while if it is incorrectly identified as not an index point, then -1 is scored. Similarly, +2 or -2 is scored for segments rated as "Definitely Not Index Point" and +1 or -1 for segments rated as "Probably Not Index Point". The scoring mechanism is illustrated in Figure 5. The sum of all individual scores is added to determine the raw indexing score for a video that we label as the Video Indexing Score (VIS).

Suppose the video lecture contains n transition points. Each transition points will have a ground truth score and an algorithm score. If G_i and A_i are the ground truth score and the algorithm score, respectively, of transition point i then the overall Video Indexing Score is represented as:

$$\text{VIS} = \sum_{i=1}^{n} (G_i * A_i)$$

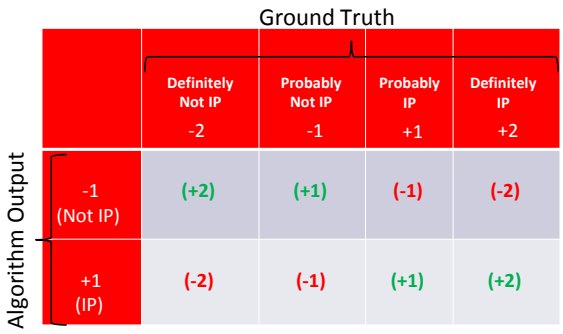
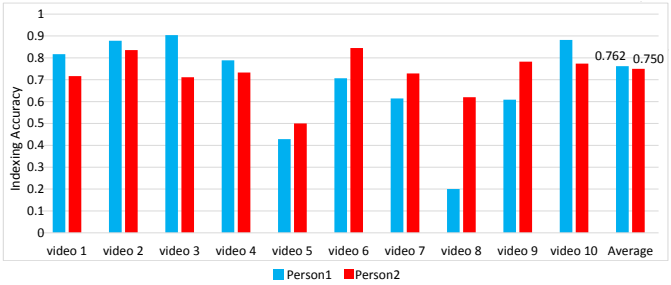


Fig. 5. Video indexing scoring for different ground truth values and algorithm results

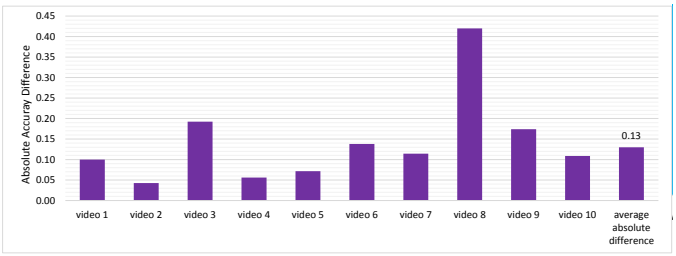
Finally the accuracy score of an algorithm for a video is computed as a percentage of the theoretical maximum VIS score for the video corresponding to theoretically optimal indexing. It should be noted that this metric is designed for comparing algorithms but not necessarily an indicator of absolute accuracy; the accuracy score drops in a non-linear fashion with errors in indexing, and can theoretically be negative.

B. Human Accuracy

The ground truth employed for evaluating indexing algorithms is the information on index points provided by the instructor. However, it is important to note that experts familiar with the subject matter are likely to come up with different ground truths. In order to validate this, an experiment was conducted where two subject matter experts were asked to index a set of videos and the results were compared against the ground truth provided by the instructor. The results are tabulated in Figure 6. Figure 6 (a) shows that the two experts have different accuracy on different videos, although their average accuracies are very close; 0.750 vs 0.762. Figure 6 (b) shows that the the difference in accuracy between the two experts varies between 4% (video 2) and 42% (video 8) with an absolute average difference of 13%. The implication is that further enhancements could improve the performance of video indexing algorithms, but it may be impossible to achieve perfect accuracy because of the uncertain nature of the ground truth. In the results of this paper, we also plot *human relative indexing accuracy* which is the accuracy achieved by an algorithm as compared to the average accuracy of our human experts.



(a) Indexing accuracy of human experts



(b) Absolute difference in accuracy of human experts

Fig. 6. Evaluation of indexing by human experts

C. Results

The text based indexing algorithm was employed to segment a suite of twenty five videos with screen text, speech text, and hybrid text; the latter simply being the union of screen text and speech text. The accuracy was measured in relation to the ground truth. Additionally, the relative accuracy as compared to human indexing was also computed based on the discussion earlier in this section, and is represented as *Human Relative Indexing Accuracy*. The premise is that an algorithm can at best achieve human accuracy. The results are presented in Figure 7.

We observe that the accuracy of segmentation with screen text is somewhat higher than that with speech text, while the accuracy of segmentation with hybrid text is in between the two. The accuracy varies in the range between 82.8% and 86.3% as compared to human accuracy. However, screen text is not the best choice for every video; 19 videos showed better segmentation with screen text while 6 videos showed better segmentation with speech text. We speculate that the reason for overall higher accuracy of screen text is that speech text has errors and screen text is sparse but is still likely to contain the keywords that define topic transition. The hybrid approach did not improve over the screen text, possibly because it is

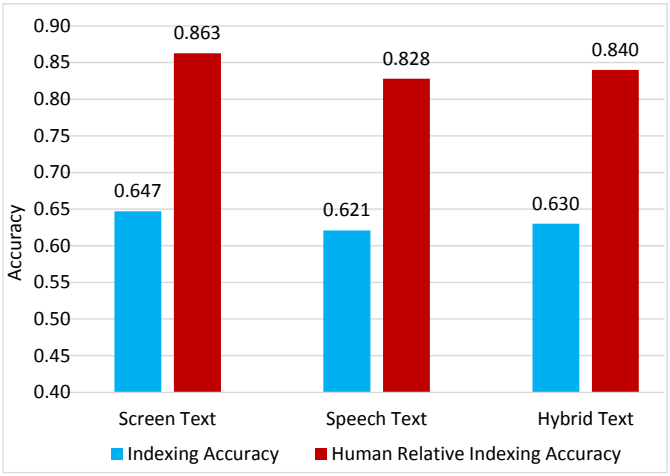


Fig. 7. Automatic indexing accuracy for screen text, speech text and hybrid text

dominated by speech text as the sheer volume of speech text far exceeds screen text. Perhaps better ways of combining speech text and screen text can lead to results superior than what can be achieved individually.

Speech text is automatically generated from lecture audio by YouTube. It typically had many errors because of the weakness of automatic speech recognition. Further, the quality of speech text varied significantly among videos. Figure 8 plots the accuracy of automatic indexing for different ASR error rates. It is clear that the accuracy of speech recognition is an important factor in automatic indexing.

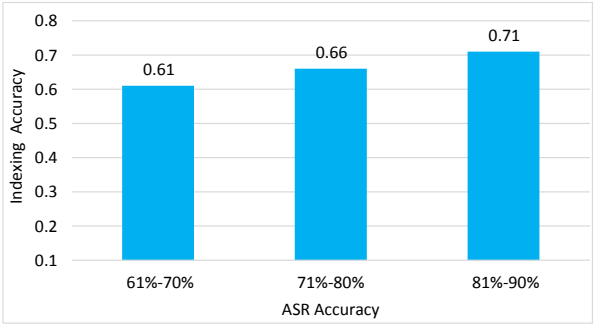


Fig. 8. Average indexing accuracy in relation to accuracy of automatic speech recognition

Additional experiments were conducted to determine any relationship between the human judged quality of speech text and the corresponding accuracy of automatic indexing. A scale from zero to five was developed to rate the quality of speech text:

- 5- Excellent
- 4- Very Good
- 3- Good
- 2- Average
- 1- Poor
- 0- No Text

Each video lecture was heard for 10-15 minutes by one of the authors in order to assign a quality rating to the speech text. No videos were rated 0 or 5 in this process. Subsequently the segmentation accuracy was measured for each group separately for analysis. The results are presented in Figure 9. The figure again shows a positive correlation between the quality of speech text and the quality of indexing.

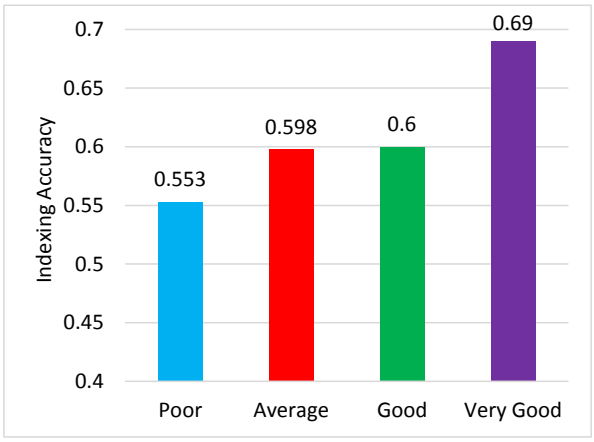


Fig. 9. Average indexing accuracy in relation to human judged quality of automatic speech recognition

To further explore the relationship between the speech recognition quality and indexing effectiveness, we performed an evaluation using manually corrected speech text. Speech text from 11 of the videos was manually corrected with the help of the ICS captioning tool. The accuracy of segmentation with speech text, corrected speech text, and screen text for these 11 videos is displayed individually in Figure 10 and summarized in Figure 11. Corrected speech text leads to significantly better segmentation accuracy as compared to (uncorrected) speech text and performs better than screen text.

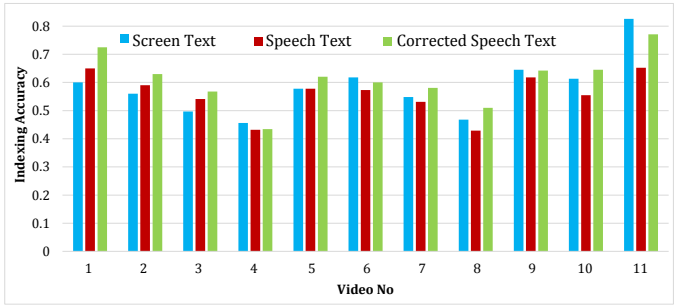


Fig. 10. Indexing accuracy with corrected speech text for selected lecture videos

In summary, the screen text received from OCR tools was better for segmenting lecture videos than speech text generated by ASR tools. However, the quality of speech text is important for accuracy and corrected (and hence virtually error free) speech text is better than screen text for segmentation. Simple hybrid text obtained by combining speech text and screen text did not perform any better than screen text alone. Experiments were not performed for corrected screen text because the automatically derived screen text was fairly accurate; usually over 90%. However, this will be a subject for future work.

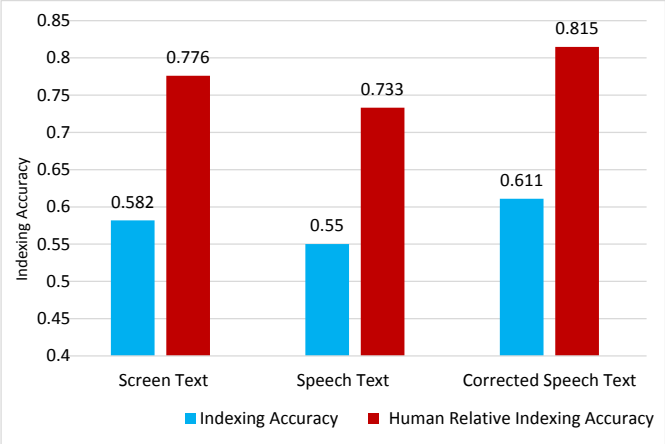


Fig. 11. Average indexing accuracy with screen text, speech text and corrected speech text for selected videos

VI. Survey Results

Indexed Captioned Searchable (ICS) Video usage is assessed to develop an understanding of the overall perceived value of the video lectures as well as the value of video indexing. Surveys were administered over 5 years in more than 10 semesters [3]. Figures 12 and 13 show the response of approximately 120 students from Spring 2013 and Fall 2013 semester to a forced-answer question about the usefulness and value of the indexing. Figure 12 shows that well over 90% of respondents agreed, that the video indexing was helpful, that the placement of index points in the video timeline was appropriate for the lectures, that the layouts of the index images made the index feature easy to use, and that the index points separated a lecture into logical segments. In this figure "Disagree strongly", "Disagree" and "Disagree slightly" is merged to "Disagree***" due to the low number of responses.

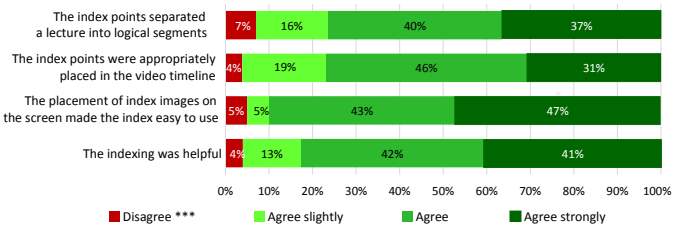


Fig. 12. Quality of video indexing

Responses to additional questions on the value of indexing are presented in Figure 13. Students are strongly supportive of the statements that the index feature functioned well, that the index points provided enough information to identify video segments of interest, and that the index made it easy to navigate the video. The statement that index points represented the start of a new subtopic had somewhat weaker support than the other assertions. It is important to note that even imperfect indexing is perceived as very valuable by the students.

In open-ended comments, students reported several benefits from using the index including (a) saving time, for example one student wrote, "I did not have to wade through the rest of the lecture just to answer one question"; (b) skipping through material the student was familiar with to get to the

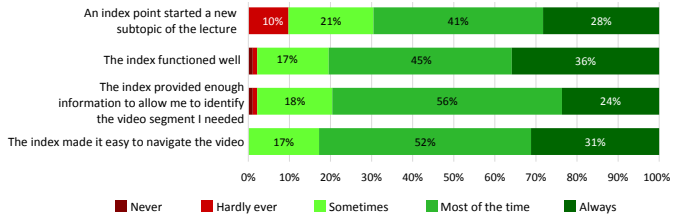


Fig. 13. Value of video indexing

challenging sections; and (c) returning to a section of the lecture if an interruption occurred. For example, one student wrote, "Sometimes I would have to pause the lecture to take care of other responsibilities that I had to attend to, and when I was ready to come back to the lecture I'd pick up exactly where I was at. It was great!". Another student said, "The indexing feature, in my opinion, is one of the best parts regarding this video player. It separated the lecture into reasonably sized sections and made it easy to know where to pick a lecture back up if I had to stop watching for a while."

VII. Discussion

Several videos were manually analyzed to understand why the screen text and speech text based algorithms sometimes provided incorrect index points in lecture videos. We illustrate some of the reasons with examples.

A. Speech Text Limitations

Figure 14 summarizes the reasons for the errors in segmentation with speech text.

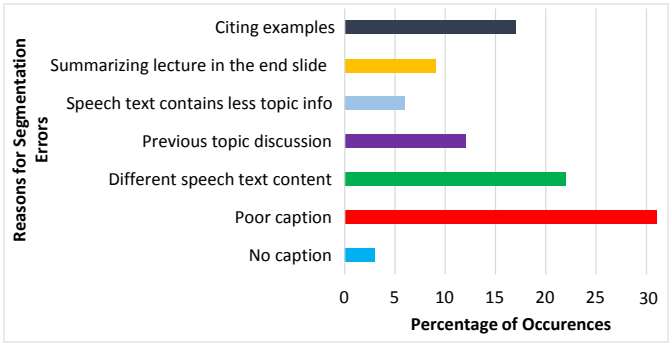


Fig. 14. Causes of segmentation errors with speech text

The most common reason for erroneous results in most of the lecture videos for speech text segmentation is the *poor quality of caption text* that leads to unrecognized text, incomplete sentences, incorrect technical words representing topic information, etc. Reduced audio quality predictably degrades the caption quality and segmentation accuracy as well. One possible solution is to manually correct the speech text but the process is labor intensive. *Different speech text content* such as instructor talking about weather, exams, or assignments, that are irrelevant to the topic flow is another leading cause of poor segmentation. This can potentially be minimized in the future by only using a glossary of subject terms for indexing. Other causes include switching to a topic away from the main flow of the lecture, such as citing examples and review of a topic discussed earlier.

B. Screen Text Limitations

Figure 15 summarizes the reasons for the errors in segmentation with screen text.

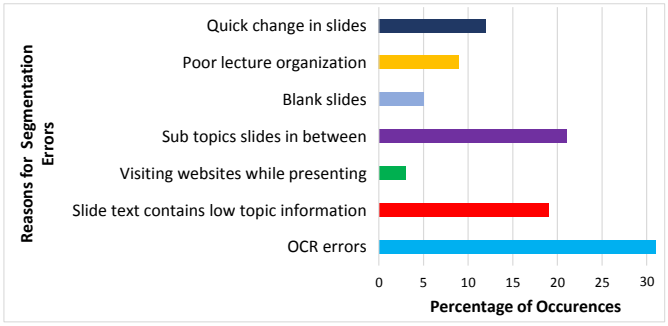


Fig. 15. Causes of segmentation errors with screen text

The largest source of errors in segmentation is *OCR errors* that lead to incorrect text data as a result of failure to recognize the text characters accurately, even though the OCR based text retrieval is overall fairly accurate. There could be various reasons for this, such as the size of the characters, presence of mathematical formula, or handwritten texts in a slide; an example of which is shown in Figure 16. Accuracy with manual correction of OCR errors is worth investigating but not a practical solution.

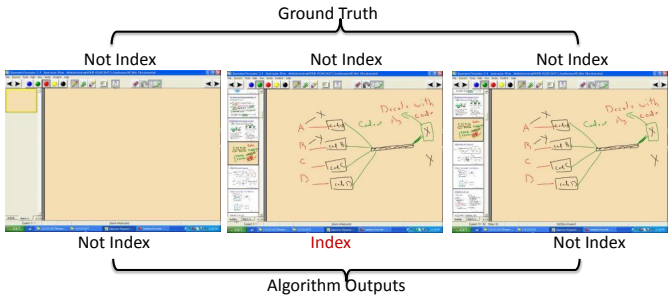


Fig. 16. Hand writing leads to OCR detection errors

Another problem with screen text is the scenario where the *screen text contains low topic information*. An important underlying reason is visual content with little textual information, such as the example shown in Figure 17. A hybrid approach of combining the text, image, and audio data could be a possible solution to solve this problem.

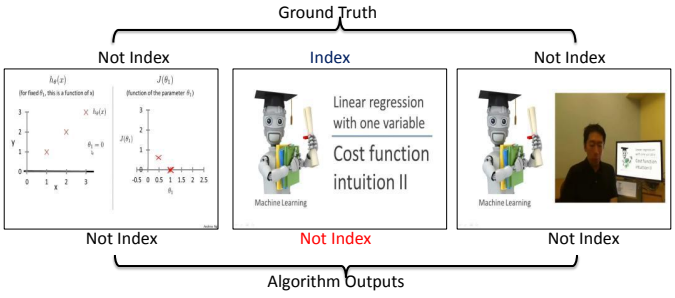


Fig. 17. Video frames with low volume of text lead to inaccurate indexing

Other reasons for errors that were discovered include visiting websites that leads to irrelevant text recognized by OCR, outline/subtopic slides in lectures, poor lecture organization such as browsing in a word file or switching windows, which again lead to irrelevant text recognition by OCR for the segmentation algorithm.

VIII. Concluding Remarks

The ability to automatically segment videos based on topics can significantly enhance the value of classroom lecture video as a learning resource. This paper investigates the use of screen text obtained with the help of Optical Character Recognition (OCR) tools and speech text obtained with Automatic Speech Recognition (ASR) tools to drive a text based segmentation process. The results show that screen text led to more accurate segmentation of videos in comparison with speech text from ASR tools, in large part because the errors in speech recognition far exceeded the errors in text recognition. Manually corrected speech text provided better data for indexing than screen text. Manual correction of screen text is not analyzed in this work. However, it should be noted that manual correction of screen or speech text are not practical options.

Screen text is typically based on instructor's viewgraphs and hence is well prepared and focused. Speech text, on the other hand, is improvised and not as focused, but the amount of text is plentiful. The conclusion is that screen text and speech text both contain useful information for lecture indexing. We believe that it should be possible to jointly use speech text and screen text for improved segmentation, but that is a subject for future research; our simple experiments did not show benefits of using them together over the better individual method.

More research is needed to achieve consistently good topic based segmentation. In this paper we have used a simple text based algorithm for video segmentation. Other algorithms, particularly those based on machine learning, hold significant promise towards achieving accurate topic based segmentation, perhaps close to what can be achieved by humans. Future improvements in OCR and ASR will have a great beneficial impact on the accuracy of topic based segmentation. Even for the topics addressed in this work, a set of twenty-five video lectures is not enough to derive firm conclusions and can only be considered preliminary work.

Finally student surveys in this project have shown clearly that classroom videos are an important learning resource and that segmentation by topics is very valuable. We hope future research will address the challenges involved, and lecture videos will be made widely available to students.

Acknowledgments

We would like to acknowledge the contributions of several members of the ICS Videos group that developed the infrastructure for the project. We thank Olin Johnson, Nouhad Rizk, Shishir Shah, and Chad Wayne; the instructors who participated in the evaluation, along with the students in their classes. We thank Lecia Barker, who developed the evaluation mechanisms employed in this work. We would also like to thank the anonymous reviewers who helped improve the paper significantly.

Partial support for this work was provided by the National Science Foundation's Division of Undergraduate Education under Course, Curriculum, and Laboratory Improvement (CCLI) program with Award No. DUE-0817558. NSF support was also provided under grants DGE-1241772 and CNS-1319212. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] G. D. Abowd, "Classroom 2000: An experiment with the instrumentation of a living educational environment," *IBM Systems Journal*, vol. 38, pp. 508–530, 2000.

[2] F. Arman, A. Hsu, and M.-Y. Chiu, "Image processing on compressed data for large video databases," in *Proceedings of the first ACM international conference on Multimedia*. New York, NY, USA: ACM Press, 1993, pp. 267–272.

[3] L. Barker, C. L. Hovey, J. Subhlok, and T. Tuna, "Student perceptions of indexed, searchable videos of faculty lectures," in *Proceedings of the 2014 IEEE Frontiers in Education Conference (FIE)*, Madrid, Spain, Oct 2014.

[4] S. Chandra, "Lecture video capture for the masses," *SIGCSE Bull.*, vol. 39, no. 3, pp. 276–280, Jun. 2007.

[5] R. Deshpande, T. Tuna, J. Subhlok, and L. Barker, "A crowdsourcing caption editor for educational videos," in *Proceedings of the 44th Annual Frontiers in Education Conference*, Madrid, Spain, 2014.

[6] H. Haberdar and S. Shah, "Change detection in dynamic scenes using local adaptive transform," in *British Machine Vision Conference 2013*. BMVA Press, 2013, pp. 6–1.

[7] H. Haberdar and S. Shah, "Video synchronization as one-class learning," in *Proceedings of the 27th Conference on Image and Vision Computing New Zealand*. ACM, 2012, pp. 469–474.

[8] A. Hampapur, R. Jain, and T. E. Weymouth, "Production model based digital video segmentation," *Multimedia Tools Appl.*, vol. 1, no. 1, pp. 9–46, 1995.

[9] A. Huang, "Similarity measures for text document clustering," in *New Zealand Computer Science Research Student Conference*, J. Holland, A. Nicholas, and D. Brignoli, Eds., Apr. 2008, pp. 49–56.

[10] M. Lin, J. Nunamaker, J.F., M. Chau, and H. Chen, "Segmentation of lecture videos based on text: a method combining multiple linguistic features," in *System Sciences, 2004. Proceedings of the 37th Annual Hawaii International Conference on*, Jan 2004, pp. 9 pp.–.

[11] D. Ma, B. Xie, and G. Agam, "A machine learning based lecture video segmentation and indexing algorithm," in *Proc. SPIE*, vol. 9021, 2013, pp. 90 210V–90 210V–8.

[12] A. Nagasaka and Y. Tanaka, "Automatic video indexing and full-video search for object appearances," in *Proceedings of the IFIP TC2/WG 2.6 Second Working Conference on Visual Database Systems II*, Amsterdam, The Netherlands, 1992, pp. 113–127.

[13] J. Nandzik, B. Litz, N. Flores-Herr, A. Lhden, I. Konya, D. Baum, A. Bergholz, D. Schnfu, C. Fey, and J. Osterhoff, "CONTENTUS technologies for next generation multimedia libraries," *Multimedia Tools and Applications*, vol. 63, no. 2, pp. 287–329, 2013.

[14] C.-W. Ngo, F. Wang, and T.-C. Pong, "Structuring lecture videos for distance learning applications," in *in Proceedings Fifth International Symposium on Multimedia Software Engineering*, 2003, pp. 215–222.

[15] H. Odhabi and L. Nicks-McCaleb, "Video recording lectures: Student and professor perspectives," *British Journal of Educational Technology*, vol. 42, no. 2, pp. 327–336, 2011.

[16] K. Otsuji and Y. Tonomura, "Projection detecting filter for video cut detection," in *MULTIMEDIA '93: Proceedings of the first ACM international conference on Multimedia*, New York, NY, USA, 1993, pp. 251–257.

[17] S. K. A. Soong, L. K. Chan, C. Cheers, and C. Hu, "Impact of video recorded lectures among students," *The 23rd Annual Conference of the Australasian Society for Computers in Learning in Tertiary Education*, pp. 789–793, 2006.

[18] J. Subhlok, O. Johnson, V. Subramaniam, R. Vilalta, and C. Yun, "Tablet pc video based hybrid coursework in computer science: Report from a pilot project," in *Proceedings of the 38th SIGCSE Technical Symposium on Computer Science Education*. Covington, Kentucky, USA: ACM, 2007, pp. 74–78.

[19] S. Tata and J. M. Patel, "Estimating the selectivity of TF-IDF based cosine similarity predicates," *SIGMOD Rec.*, vol. 36, no. 2, pp. 7–12, Jun. 2007.

[20] T. Tuna, "Search in classroom videos with optical character recognition for virtual learning," Master's thesis, University of Houston, 2010.

[21] T. Tuna, "Automated lecture video indexing with text analysis and machine learning," Ph.D. dissertation, University of Houston, 2015.

[22] T. Tuna, J. Subhlok, L. Barker, V. Varghese, O. Johnson, and S. Shah, "Development and evaluation of indexed captioned searchable videos for stem coursework." in *Proceedings of the 43rd SIGCSE Technical Symposium on Computer Science Education*. ACM, 2012, pp. 129–134.

[23] T. Tuna, J. Subhlok, and S. Shah, "Indexing and keyword search to ease navigation in lecture videos," in *Applied Imagery Pattern Recognition*, 2011, pp. 1–8.

[24] V. Varghese, "Development and evaluation of text-based indexing for lecture videos," Master's thesis, University of Houston, 2014.

[25] H. Yang, F. Grnewald, M. Bauer, and C. Meinel, "Lecture video browsing using multimodal information resources," in *Advances in Web-Based Learning ICWL 2013*, J.-F. Wang and R. Lau, Eds. Springer Berlin Heidelberg, 2013, vol. 8167, pp. 204–213.