

Detecting Significant Events in Lecture Video using Supervised Machine Learning

Christopher Brooks¹, Kristofor Amundson, Jim Greer

Laboratory for Advanced Research in Intelligent Educational Systems (ARIES)
Department of Computer Science, University of Saskatchewan, Canada
cab938@mail.usask.ca, kta719@mail.usask.ca, greer@cs.usask.ca

Abstract. This paper describes work we are doing to identify significant events in video captures of academic lectures. Unlike other approaches which tend to define per-image comparison threshold values based on intuition or empirically derived results, we use supervised machine learning techniques to automatically determine appropriate image characteristics based on end-users understanding of what constitutes an important event. This makes our approach more adaptable to different kinds of content, and still provides a substantial level of agreement with human experts.

Keywords. Course casting, video capture, supervised machine learning, significant events, chaptering, indexing, multimedia

Introduction

The recording and broadcasting of lectures in higher education has seen rapid interest and growth as high speed networks become ubiquitous and portable media players become more widely available. Computer-based capture and transcoding systems (also called *vodcast* or *course casting* systems), such as echo360², Replay³, and Recollect⁴ provide the ability to capture video of interactions happening in class (e.g. video of the instructor or audience) as well as the slides or desktop that is being shared. They typically use dedicated hardware frame grabbers that sit between the presenter desktop and a digital projector and, as such, can capture both static content (e.g. PowerPoint) as well as dynamic content (e.g. YouTube video clips, web browsing, etc.)⁵. The audio and video from the capture sources is then mixed together and encoded for a particular output platform, which is most often a proprietary web-based video player.

The creation of video content for online learning holds benefits for both educational institutions and learners alike. Educational institutions that have not yet begun to supporting distance learning can now leverage the expertise they have in creating rich interactive classroom experiences, and can do so for a low cost and with

¹ Corresponding Author.

² See <http://www.echo306.com>

³ See <http://www.replay.ethz.ch/>

⁴ See <http://www.recollect.ca>

⁵ Other course casting solutions (e.g. Panopto, available at <http://www.panopto.com>) exist which do not use hardware capture devices for output signal, however, these systems are few in number and generally less full featured than the ones listed.

little change to their current teaching practices. In turn, students at a distance receive some of the benefits of the natural consequential communication that happens in the course which may increase motivation as well as retention and completion rates, a historically important problem in distance education [12].

Regardless of these and other benefits, rich media learning content faces a number of unique challenges. In this paper we are principally interested in exploring one of these challenges, namely, how can we automatically partition the lecture video so that end-user browsing of content is simplified? Overviews of a whole video are difficult to provide because of the high bandwidth, continuous, and temporal ordering of image data. For instance, generating thumbnails for chaptering of a video is often done by giving the video to a subject expert who then identifies chapter points as they watch it. This kind of post-processing increases the cost of course casting production, so most automated software implementations in the market today create thumbnails either at fixed intervals or with simple heuristics. The result of this is a potentially long list of thumbnails that allow for coarse grained navigation throughout a lecture but require some guess work when a learner is trying to find the beginning of a relevant section or topic.

We argue that the automation of high quality chaptering is possible by applying supervised machine learning techniques to video content using still-frames from the video as attributable elements. We begin this paper by visiting the issue of significance in educational video, where we focus specifically on what significance means and how end-users perceive chaptering. Section 2 presents a novel approach to the identification of significant events using supervised learning approaches, and compares this to the heuristic method described by others in [3]. We conclude the work in section 3 with the identification of techniques that we see as promising continuations of our research.

1. Understanding Significant Events

1.1. What is Significance?

Earlier work in this area has dealt primarily with finding “significant” points or events in a video. The videos we are studying are primarily screen captures of lecture presentations, similar to the data used in [3, 5], but detecting significant events in video has traditionally focused on photorealistic domains such as LifeLogs (e.g. [1, 2]). Regardless, our approach has similar end goals to these and other related works [1, 2, 3, 5] - to index video by some notion of significance to enable efficient referencing and browsing of video by end-users.

A large obstacle in this area of research is the noise that occurs in the frames reduced from the videos. This noise comes from a variety of sources, and the most common method of dealing with it is to set thresholds, such as the ones described in [3] that are used to preprocess “...transitional capture noise, analog variation noise, shift and fuzziness noise, and signal noise”. Setting thresholds needs to be done on a per-installation basis, as many sources of this noise are different based on environmental aspects (e.g. interference). Further, as more multimedia and interactive technologies move into the classroom, traditional “slide shows” move toward more active document formats such as simulations, web browsers, and videos. Algorithms that don't use human expertise are generally at a disadvantage when trying to provide indexing for these kinds of presentations.

Instead of providing a functional definition of significance as has been done by others, we elicited a demonstrative definition based on the end goals of systems users may be using. We asked six subjects who were unfamiliar with video capture systems to go through four lectures of different topics and identify where significant events were. For the purposes of this study, we described a significant event as "...the position in which a sub-chapter marker would be useful for students like [themselves] in quickly navigating through the lecture[s] for review or study". We also provided a screenshot (Figure 1) of the user interface of the Recollect system to demonstrate the intended usage of output for a player that supported sub-chaptering.

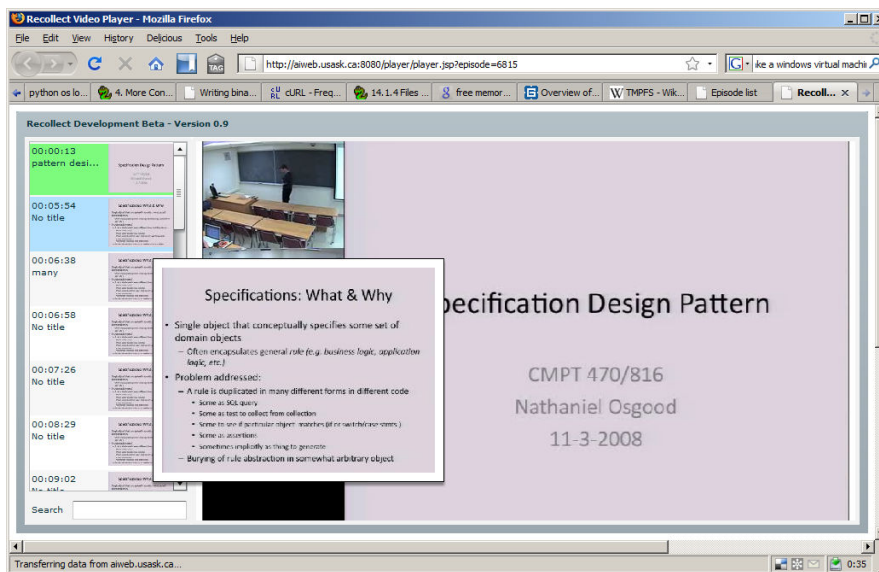


Figure 1. Screen shot of the Recollect system interface provided to study subjects demonstrating the usage of sub-chaptering in lecture video systems. Sub-chapters are shown as thumbnails on the left indexed by time, and provide a larger view when the mouse pauses over thumbnail

The tool provided to study subjects allowed for navigating through a video linearly both forwards and backwards in 1, 5, 10, and 30 frame increments (each frame was equivalent to one second of video). Subjects were also able watch the video at various speeds, though none used this option. No audio was provided as it was not being considered part of the candidate feature space. Subjects were both observed and given post-study questionnaires to elicit feedback.

The lecture videos chosen for this study included two senior level undergraduate courses which used Microsoft PowerPoint as the main delivery mechanism, and included either mostly plain black on white lettering (*Video A*) or more graphical slides with significant moments of Web-browsing during the presentation (*Video B*). One graduate course was also included (*Video C*), as well as an introductory undergraduate course which used handwritten lecture notes obtained by capturing instructor

interaction with the SMART Symposium system⁶ (*Video D*). All lectures were in the domains of Computer Science or Mathematics.

Observations and survey results from the subject participants identified that they used two distinct mechanisms for identifying significant events; visual structure (e.g. slide advancement in PowerPoint, or extending the page in the Symposium) and semantics (topics being taught in the slides). Out of the six participants, five used primarily visual structure to identify events, while the sixth used the semantics of the lecture material. All participants indicated frustration at being unable to find good event times for the handwritten lecture (*Video D*) in part because the video was under constant change and because the ideal thumbnail for a sub-chapter is not available until the end of the content it describes (e.g. a thumbnail from time t best describes the content starting at time $t-\delta$ and ending at t). Table 1 gives the Fleiss' kappa [8] results.

Table 1. Inter-rater reliability measures for significant events in four lecture videos. Kappa is calculated with $n=6$ for the first column, $n=5$ for the second column after adjusting for the one outlier (semantic annotator), and $n=6$ when including Dickson's algorithm with the outlier adjusted values.

	κ	κ'	$\kappa' p_{\text{pos}}$	$\kappa' p_{\text{neg}}$	κ' with Dickson [3]
Video A	0.65	0.86	0.87	0.99	0.55
Video B	0.49	0.65	0.68	0.99	0.42
Video C	0.51	0.65	0.66	0.99	0.44
Video D	0.14	0.18	0.21	0.99	0.11

The effect of including the semantic rating used by the sixth annotator is clear throughout all videos. Using the categories provided by Landis and Koch [9], we see that for one of the videos (*Video A*) agreement increases to almost perfect ($\kappa=0.86$) when the semantic annotations are removed, and increases from moderate to substantial agreement for two of the other videos (*Video B* and *Video C*).

In addition to determining the quality of significance between raters, we included Dickson's algorithm [3] as another reference⁷. This algorithm is a multi-pass image processing function that examines both pixel and block characteristics of video to determine stable events. It is particularly relevant to this problem as it was designed for educational lecture recordings, and uses similar frame grabbing hardware as our test set. The last column of Table 1 provides inter-rater reliability values describing the level of agreement between the result of Dickson's algorithm and the outlier adjusted set of human raters. We note that including Dickson's algorithm reduces agreement by a significant amount, as agreement between Dickson and any other rater is quite poor.

We have included positive and negative agreements calculated using [15] on the advice of Feinstein and Cicchetti [13, 14] who argue that binary ratings can provide for misleading low kappa values. As $\kappa' p_{\text{pos}}$ values (e.g. the amount of agreement that a slide is a sub-chapter marker) are nearly equivalent to κ' , this supports the notion that κ' is a reasonable statistic for inter-rater reliability.

⁶ See <http://smarttech.com>

⁷ The original algorithm runs the second pass multiple times while our simplification runs this pass only once. Details about the bounding area for spatial difference calculations were not available in time for submission, but analysis of the results suggest these would have minimal effect given our input data.

1.2. Summary of Significance

The study presented here has illuminated a number of interesting artifacts about significant events in sub-chaptering of video. First, there are two ways that humans use to chapter lecture video; visual structure and content semantics. Of these, there is high agreement in the former, especially if the video being presented is prototypical PowerPoint slides. This suggests that supervised machine learning may be an appropriate technique in that it initially leverages human expertise and correlates this with attributes of the content being examined to create rules for automatic mining. In addition, humans are resilient to motion video being captured (e.g. web browsing) and maintain reasonably levels of high agreement ($\kappa=0.65$).

It was surprising to us to find that Dickson's algorithm had minimal agreement with human raters. Having used Dickson's in deployment for roughly a year, it has provided a seemingly reasonable set of indices for online lectures being deployed through Recollect. There are a number of explanations for this; sometimes Dickson's gets the index screenshot correct, but has it time shifted by a small amount (since there tends to be more artifacts and noise occurring when near a transition). Further, it wasn't designed to deal with significant changes on the screen over time (e.g. scrolling), it will instead treat each of these scroll events as negligible, but over time the change will be drastic. Human raters were able to record the number of these events and then decide when the sum of the tiny scroll events became a major event. Nonetheless, this provides a good example as to the challenges being faced in developing automated lecture indexing systems.

2. Supervised Machine Learning

The strength of supervised machine learning is that it leverages similarities in data sets in domains that are difficult to empirically evaluate. The output of a supervised machine learning process can be a set of rules that can be applied to attribute data for relatively fast classification. In our domain the artifacts of interest (images of a presentation) can change depending on instructor pedagogical practice, however, the understanding of what significance is is strong in lectures we have tested that used primarily PowerPoint style slides (e.g. *Video A*, *Video B*, and *Video C*).

To test the hypothesis that supervised machine learning can produce useful indexing, we had one human rater annotate a term worth of video data for *Video A* as *Course A* and *Video B* as *Course B* using the same tool as the subjects in our previously study. He was familiar with the project, and instructed to use visual structure to do his annotations, and he annotated roughly 59 hours of presentation. The frames from these videos (210,637 in total) were encoded as high quality JPEG images, and we collected fifteen image statistics on each frame (described in Table 2). While there are many kinds of attributes that could be included, we chose these as previous experience and work in other domains [1] suggested they had a high probability of being appropriate.

Table 3 summarizes the results of the data mining process using the J48 decision tree classifier provided with the Weka toolkit. While a number of classifiers were tried, the J48 classifier provided good results and provided them in a form that makes them easy to codify into our existing video application framework. We compared the classifiers performance to that of each video individually, as well as the union set of both videos.

Table 2. Image attribute definitions used to compare frames from Videos A and B. Each attribute compares two images, the candidate for the significant event and the previous image.

Attribute	Description
Pixel Contrast	The difference of contrast between pixels in each image as per Rec. 601. [10]
Pixel Histogram Intersection	The minimum values of the histograms of two images summed across all bands of the image.
Block Contrast (16 ²)	Similar to Pixel Contrast but the image is first partitioned in 16x16 pixel sized blocks.
Block Histogram Intersection (16 ²)	Similar to Pixel Histogram Intersection but the image is first partitioned into 16x16 pixel sized blocks.
Block Contrast (32 ²)	Similar to Block Contrast (16 ²) but with 32x32 pixel sized blocks.
Block Histogram Intersection (32 ²)	Similar to Block Histogram Intersection (16 ²) but with 32x32 pixel sized blocks.
Block Contrast (64 ²)	Similar to Block Contrast (16 ²) but with 64x64 pixel sized blocks.
Block Histogram Intersection (64 ²)	Similar to Block Histogram Intersection (16 ²) but with 64x64 pixel sized blocks.
Difference	The mean of the difference of pixel values in each band of two images.
Difference ²	Similar to the above but it is the difference between the squares of the differences for each image band.
Mean	The average pixel value across all bands of one image compared to its neighbor.
Root Mean Squared	The difference between the root of the square of the means of each image across each band.
OCR Set Difference	Each image was run through an Optical Character Recognition (OCR) engine ⁸ . Previous experience suggested that non-English words and words less than three characters long are often the result of miss-recognitions by the OCR software. This attribute is the number of words included in the OCR output for this set that are longer than three characters, in the English dictionary, and not in an adjacent images word set. ⁹
OCR Set Difference 2	Similar to the above, but uses the previous images' word set.
OCR Intersection	The number of English words greater than three characters that exist in the intersection of two image word sets.

The results indicated a substantial agreement ($\kappa > 0.60$) for all tests, indicating that the results of the classification were a set of rules that produced output that correlated well with the human subject's values. Including p_{pos} values suggests that these kappa values are appropriate reflections of actual confidence.

Table 3. Comparison over one term of instruction from two classes, *Course A* and *Course B*, as well as the union of these datasets. Each video was encoded by a single rater, and mined using the J48 decision tree provided with Weka using ten-fold testing to minimize chances of Type III errors.

	K_{J48}	$K_{J48} P_{pos}$	$K_{J48} P_{neg}$	$K_{Diskson}$	$K_{Diskson} P_{pos}$	$K_{Diskson} P_{neg}$
Course A	0.72	0.75	0.99	0.36	0.01	0.99
Course B	0.66	0.63	0.99	-0.28	0.01	0.99
Course A \cup Course B	0.67	0.66	0.99	0.37	0.01	0.99

⁸ We used the GOCR framework available at <http://jocr.sourceforge.net>.

⁹ All of our lectures were presented in English, though they may have contained non-English acronyms, words, or related formula that may have been discarded by our text cleaning process. We used the 5 desk dictionary wordlist available at <http://wordlist.sourceforge.net>.

While it is unfair to compare this agreement directly to the κ' values from Section 2 because of the difference in data set sizes (κ' examined only a single lecture), it is interesting to note that values fall roughly in the same range of significance outlined by Landis and Koch [9].

In addition to comparing the supervised machine learning results to a rater, we calculated the inter-rater reliability values of Dickson's algorithm using Fliess' Kappa metric. Similar to previous results, we saw poor agreement between this algorithm and our human rater. When slides included minimal interactivity (e.g. *Course A*) confidence was much higher, but still not significant. That p_{neg} is so high may be why Dickson's algorithm has seemed reasonable to most users - it doesn't find "correct" indices, but is able to discount most of the "incorrect" frames appropriately.

3. Conclusions

An important factor of our work is that instead of making a model of what constitutes a significant event through top down processes, we are building this model bottom-up based on individual users expertise. The definition of significance is thus described by what human raters provide as examples of this definition, and not through a semantic of our own. Instead of ontological reasoning, we rely on data mining and prototypical reasoning to identify what frames in a video are of significance for sub-chaptering.

There are a number of natural directions this work should go in to increase the effectiveness of our approach. Principle among these is that more image characteristics should be generated for comparisons. There is a large body of work that has had success when considering sharpness, noise, and entropy in digital video. While we considered contrast and histograms extremely important because of the content of our video (e.g. written slides), these other measures may be equally important. We also believe that audio segmentation (e.g. gap detection) may provide interesting attributes, and are working on methods to incorporate this in future experiments.

In addition, to make results applicable to our end application we must be able to generate image characteristics quickly. The two lectures presented here took roughly 12 hours to generate using 20 modern desktops running in parallel. This is at least two orders of magnitude slower than what would be appropriate for a fully deployed system. By reducing the number of attributes to just those that were most effective and performance tuning attribute gathering implementations we anticipate this problem is still tractable, however it is likely that real-world deployments will see tradeoffs between accuracy and speed.

Whether we should consider end-user usage data in our algorithms for evaluating the effectiveness of sub-chaptering is an interesting question in itself, and one we are most interested in pursuing. If students viewing lectures navigate to a time offset in the lecture using our index and then navigate away quickly this may indicate that the index isn't useful. Similarly, if they navigate to a certain location and continue to watch the video this may indicate that the index is appropriate. In addition, we are motivated by other work [11] that suggests that end-users are willing to help provide metadata for video lectures as they are watching them. Providing methods of indexing within the end-user interface may be an effective method of obtain course-specific training data, which may increase the reliability of indices.

The weakest aspect of the study that we have presented here is that we only compare our outcomes against one rater. While section 2 attempts to mitigate this by providing evidence that raters are in high agreement, a more ideal situation would be to train and test the supervised machine learning algorithm against multiple raters. The amount of time it takes an individual to identify events in a whole term of data has made this cost prohibitive, and ten-fold testing on smaller sets of data is likely to generate poor results. We believe that collecting end-user data as described previously may help in increasing the validity of our results, as wider-scale testing can take place.

Determining appropriate sub-chapter points for video of a lecture is a difficult problem. The effects of software being used and the activities being done (e.g. using a Symposium system, or web browsing) have a significant effect both on human detection of events as well as algorithmic detection of events. Nonetheless, we have shown that videos that have a minimal amount of such activity can be semi-automatically annotated with indices using a J48 decision tree. We validated this approach with ten-fold testing showing that for these kinds of lectures, substantial agreement with mined rules can be achieved. This is important, as techniques currently in use provide comparatively poor results (e.g. Dickson's, or fixed time intervals).

References

- [1] A.R. Doherty, D. Byrne, A.F. Smeaton, G.J. Jones and M. Hughes. Investigating KeyFrame Selection Methods in the Novel Domain of Passively Captured Visual Lifelogs. (2008) In CIVR 2008: ACM International Conference on Image and Video Retrieval, Niagara Falls, Canada, 7-9 July 2008.
- [2] A.R. Doherty and A.F. Smeaton. Automatically Segmenting LifeLog Data Into Events.(2008) In WIAMIS 2008: 9th International Workshop on Image Analysis for Multimedia Interactive Services, Klagenfurt, Austria, 7-9 May 2008.
- [3] P. Dickson, W.R. Adrion and A. Hanson. (2006) Automatic Capture of Significant Points in a Computer Based Presentation. In: Proceedings of the Eighth IEEE International Symposium on Multimedia. San Diego, CA: IEEE Computer Society, 2006:921-926.
- [4] P. Ziewer. Navigational Indices and Content Interlinkage On The Fly. (2006) In: Proceedings of the Eighth International Symposium on Multimedia, 2006: 915-920
- [5] T. Syeda-Mahmood and S. Srinivasan. Detecting topical events in digital video. (2000) In ACM Conference on Multimedia, 2000.
- [6] I. H. Witten and W.Frank (2005) Data Mining: Practical machine learning tools and techniques, 2nd Edition, Morgan Kaufmann, San Francisco, 2005.
- [7] W. Cheng, Y. Chuang, B. Chen, J. Wu, S. Fang, Y. Lin, C. Hsieh, C. Pan, W. Chu, and M. Tien. Semantic-Event Based Analysis and Segmentation of Wedding Ceremony Videos. In proceedings of the 9th ACM SIGMM International Workshop on Multimedia Information Retrieval, September 28–29, 2007, Augsburg, Bavaria, Germany
- [8] J. L. Fleiss. (1971) Measuring nominal scale agreement among many raters. Psychological Bulletin, Vol. 76, No. 5 pp. 378–382
- [9] J. R. Landis, and G. G. Koch. (1977) The measurement of observer agreement for categorical data, in Biometrics. Vol. 33, pp. 159–174
- [10] International Telecommunications Union, ITU-R BT.601
- [11] C. Munteanu, Y. Zhang, R. Baecker, G. Penn. (2006) Wiki-like Editing of Imperfect Computer-Generated Webcast Transcripts.CSCW06, November 4-8, Banff Canada.
- [12] A. Parker. (1999) A Study of Variables that Predict Dropout from Distance Education. International Journal of Educational Technology, v1 n2 p1-10 Dec 1999
- [13] A. R. Feinstein and D. V. Cicchetti. (1990) High Agreement But Low Kappa: I. The Problems of Two Paradoxes. Journal of Clinical Epidemiology. Vol 43., No6., pp. 543-549.
- [14] D. V. Cicchetti and A. R. Feinstein. (1990) High Agreement But Low Kappa: II. Resolving the Paradoxes. Journal of Clinical Epidemiology. Vol 43., No6., pp. 551-558.
- [15] G. P. Samsa. (1996). Sampling Distributions of p_{pos} and p_{neg} . Journal of Clinical Epidemiology. Vol. 49, No. 8, pp. 917-919.