Leveraging Video Interaction and Content to Improve Video Learning

Juho Kim
MIT CSAIL
Cambridge, MA 02139 USA
juhokim@mit.edu

Krzysztof Z. Gajos
Harvard SEAS
Cambridge, MA 02138 USA
kgajos@eecs.harvard.edu

Shang-Wen (Daniel) Li
MIT CSAIL
Cambridge, MA 02139 USA
swli@mit.edu

Robert C. Miller
MIT CSAIL
Cambridge, MA 02139 USA
rcm@mit.edu

Carrie J. Cai
MIT CSAIL
Cambridge, MA 02139 USA
cjcai@mit.edu

Abstract

Video has emerged as a dominant medium for online education, as witnessed by millions of students learning from educational videos on Massive Open Online Courses (MOOCs), Khan Academy, and YouTube. The large-scale data collected from students' interactions with video provide a unique opportunity to analyze and improve the video learning experience. We combine *click-level interaction data*, such as pausing, resuming, or navigating between points in the video, and *video content analysis*, such as visual, text, and speech, to analyze peaks in viewership and student activity. Such analysis can reveal points of interest or confusion in the video, and suggest production and editing improvements. Furthermore, we envision novel video formats and interfaces that automatically adapt to learners' collective watching behaviors.

Author Keywords

MOOCs; video learning; interaction peaks.

ACM Classification Keywords

H.5.1. Information interfaces and presentation (e.g., HCI): Multimedia Information Systems: Video.

Introduction

MOOCs often include hundreds of pre-recorded video clips. Research on MOOCs has shown that learners

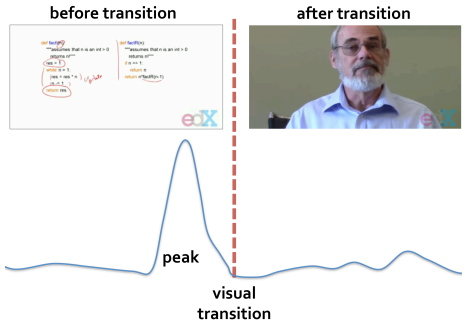

Figure 1. An example interaction peak. This peak represents students returning to see the code snippet slide that disappeared after transitioning into the talking head. An abrupt transition might not give students enough time to comprehend what's presented.

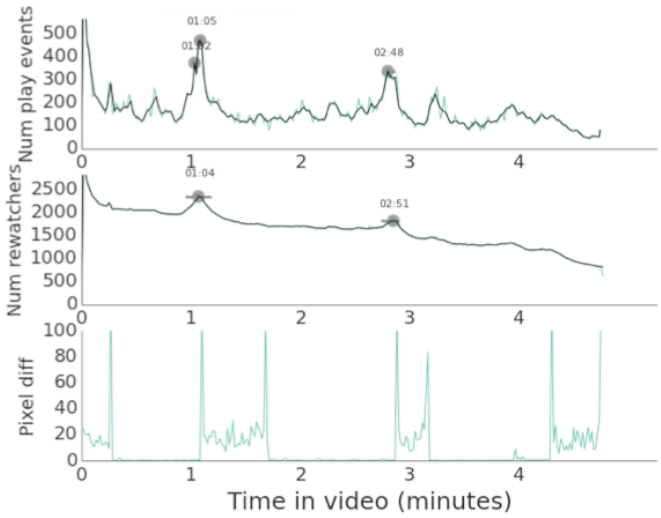

Figure 2. To analyze interaction peaks in a video, we visualize play events (top), re-watching sessions (middle), and pixel differences (bottom) over time. Detected peaks are marked with a gray point. In this example, the detected peaks coincide with a spike in pixel differences, which indicate a visual transition in the video.

spend a majority of their time watching videos [1] and their engagement with videos is affected by video length and production styles [2]. However, little research has focused on the click-level interactions *within* MOOC videos. With thousands of learners watching the same online lecture videos, video analytics can provide a unique opportunity in understanding how learners use video content and what affects their learning experience.

Our recent work [4] analyzed click-level interactions resulting from student activities within individual MOOC videos, namely playing, pausing, navigating to another point, replaying, and quitting. We analyzed video player interaction logs from four MOOCs offered on the edX platform (http://edx.org/) to identify temporal interaction patterns at the second-by-second level.

When a significant number of students interact with a common portion of a video, the resultant data can be binned to highlight peaks in the video timeline. **Peaks in viewership and student activity** can indicate points of interest for instructors and students. These spikes, hereinafter referred to as *interaction peaks*, can indicate student confusion, introduction of important concepts, engaging demonstrations, or video production glitches.

To understand why these peaks occur, we manually inspected 80 videos from our set. One notable observation we made was that 61% of the peaks coincided with visual transitions in a video, such as switching from a slide to a classroom view, or from handwritten notes to a software screencast. Combining the interaction data with visual content analysis, we identified five student activity types that can lead to a peak: starting from the beginning of a new material, returning to missed content (Figure 1), following a tutorial step, replaying a brief segment, and repeating a non-visual explanation.

Data-Driven Video Analysis and Design

We believe analyzing interaction data has design implications for online video interfaces, which is the focus of our current work. This workshop paper builds on our recent work on interaction peaks [4] to propose two major research directions in data-driven video analysis and interaction design: 1) scalable and automatic methods to interpret interaction peaks by combining interaction data and content analysis, and 2) design of video interfaces that automatically adapt to collective learner interaction data. In the workshop, we hope to receive feedback on our ongoing work and share our vision in these directions.

1. Combining interaction data and content analysis.

Combining multiple data streams can lead to discovering meaningful video learning patterns otherwise not possible, as each stream brings in a

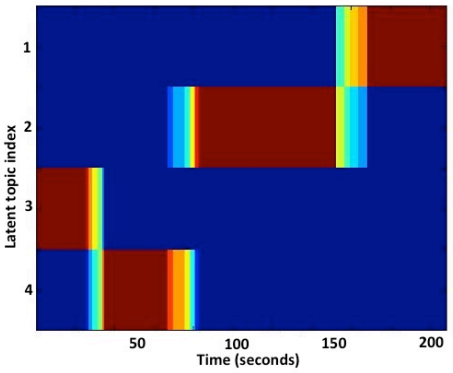

Figure 3. Latent topic distribution over time. The x-axis represents video time in seconds and the y-axis shows four latent topics. We use colors (red: 1, blue: 0) to indicate the probability of each topic every second.

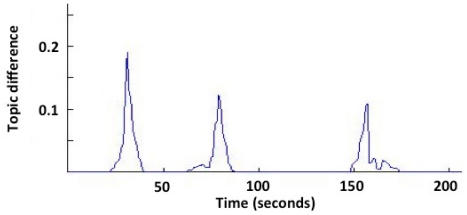

Figure 4. Topic transition likelihood over time. The x-axis represents video time in seconds and the y-axis shows distance between topics. Peaks represent topic transitions.

complementary perspective. We discuss techniques and challenges in working with various data streams.

1) Video interaction log

Video interaction logs typically include user name, time of access, video ID, event type (play, pause, etc.), and internal video time. Our data processing pipeline first reconstructs the watching history of each viewer and then aggregates the per-viewer history data to produce activity statistics for each second-long segment of the video. Specifically, the first step converts raw interaction log entries into watching segments. A watching segment keeps track of all continuous chunks of a clip watched by a user. It includes start and end time for every watched segment. The second step uses the segment information to create second-by-second counts of viewers, unique viewers, re-watching sessions, play events, and pause events. Re-watching sessions only consider a student watching a segment of a video twice or more. Play and pause events increment a bin count if the event is triggered within that bin. Finally, such information can be queried upon request for statistical analysis and further processing.

In the future, the pipeline can consider additional student interaction data. First, capturing interactions before or after watching a video might be useful. They will reveal students' learning paths beyond videos, providing additional context for understanding interaction peaks. Examples include students coming back to certain parts of a video while solving a problem, or reviewing videos while taking an exam. Also, analyzing video interactions other than play and pause, such as volume control, full screen, and video speed control, can help improve the quality of a video.

2) Visual content

To explore the connection between visual transitions and interaction peaks, we applied a visual analysis technique to complement the log analysis. We used an image similarity metric that computes pixel differences between two adjacent frames to quantify the amount of visual changes in the video. Our pipeline first samples a video frame every second, computes the image similarity using the standard technique, Manhattan distance, and finally stores the pixel distance value. Visualizing this data (Figure 2, bottom) shows that visual transitions often coincide with interaction peaks.

Visual content analysis can use more advanced scene detection algorithms to capture visual transitions more accurately. It can further apply Optical Character Recognition (OCR) to capture text inside a video frame. The captured text can be used to find important topics or summarize peaks. It can serve as another channel for the text analysis, which we describe next.

3) Text from transcripts

Topic modeling: Recently we began investigating the relation between interaction peaks and the text content. For each video, we first computed the latent topic distribution over time, and discovered the point where salient topic changes occur. We used probabilistic latent semantic analysis (PLSA) [3] for topic modeling, which computes the latent topic distribution of each document by modeling the co-occurrence of word and document. We defined a document as each sentence along with its preceding

Example n-gram results

Examining sentences starting with the word "so," we observe that in many cases, the instructor is either initiating an explanation (e.g., "so let me spend a second on that," "so that means," "so why should we explore?"), inviting students to begin following along (e.g., "so let's start at the root", "so you see the apple"), or arriving at the take-home message of an explanation (e.g., "so in essence forward checking is just looking at the arcs…", "so we can get lots of information just from these five number summaries.") Similarly, the bigram "this is" frequently appears in contexts that suggest a visual explanation (e.g., "this is the double bonds here on this oxygen") or the naming of a particular concept (e.g., "and this is called a dislocation", "so this is sometimes called the first quartile").

and following five sentences. Then we computed the latent topic distribution for each sentence. The time codes in transcription are further used for mapping each sentence to the time axis. In the example shown in Figure 3, the instructor changed topics three times during this lecture (topic index 3->4->2->1).

We focus on times where a topic transition occurs. We treat the topic distribution in each second as a posterior vector, and compute the distance between two vectors from adjacent seconds to quantify the topic changes in the video transcription. We also visualize an example result in Figure 4, which shows three peaks for topic transitions. We analyzed the relation between user interaction and text channels based on this topic transition measure. The topic transitions explained 40% of the interaction peaks, where 25% of them (i.e., absolute 10%) are not explained by the visual channel.

N-gram analysis: Furthermore, we conducted an initial analysis of transcript content to explore potential linguistic patterns spoken by instructors during peaks in the clickstream. We define a sentence to be a *peak sentence* if any part of the sentence overlaps with the time range of an interaction peak. To perform analysis, we first divided 79 lecture transcripts into two corpora: 1118 peak sentences and 4787 non-peak sentences. We then computed counts of unigrams, bigrams, and trigrams for each corpus, where an n-gram refers to a contiguous sequence of n words in the corpus. Normalizing for the total number of n-grams in each corpus resulted in a percent score for each n-gram, indicating the relative contribution of each n-gram to the corpus. Finally, we examined n-grams with the largest differences in contribution between the two corpora.

The bigrams "<start-of-sentence> so" and "this is" were among bigrams with the largest differences in contribution, appearing more frequently in peak sentences than non-peak sentences. These preliminary findings are consistent with our understanding that interaction peaks often coincide with transitions or explanations. Though we have limited our analysis to 80 videos so far, language model computation on a larger set of data could potentially aid the automatic detection and categorization of peaks.

4) Speech and acoustic stream

Audio clues, such as lecturer's prosody, volume, pitch, and speaking rate, can also provide insight in explaining students' activity. We first examined the speaking rate and pitch. Several studies (e.g., [7]) point out the usefulness of these two factors in summarizing and understanding text and audio documents. However, as compared to the visual and text channel, changes in audio are much more subtle and noisy.

As an example, we displayed an instructor's pitch change from a lecture video in Figure 5. Using the audio track in the video, the pitch can be computed for every frame. While some trends can be observed, our current analysis has not found a meaningful connection between audio changes and interaction peaks. We plan to look for relationships between audio features and linguistic or visual features.

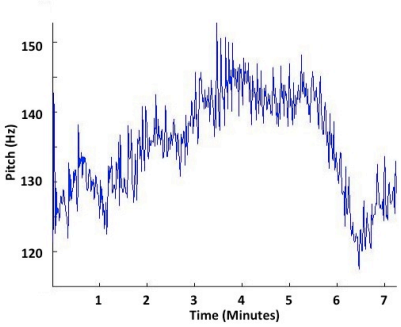

Figure 5. The instructor's pitch change over time within a video.

5) Automatic interpretation of peaks

Combining interaction data and content channels can enable the automatic detection and categorization of interaction peaks. In our earlier work, we algorithmically detected interaction peaks [6]. Then we manually categorized the detected peaks into five types, taking into account the existence and relative location of a visual transition. While the manual labeling provided insight and possible explanations for peaks, it is subjective and not scalable. Because our categorization relies only on the relative positioning between visual transitions and peaks, we believe it is feasible to build a classifier that automatically categorizes a peak. By aligning interaction peaks with visual and topic transitions, we observed how visual and text channels can complement each other, providing richer context and insight for interpreting interaction peaks. We are currently building technology for automatic, multi-channel, and scalable peak detection and categorization. It will enable us to design novel interaction techniques for educational videos that are driven by learner data.

A central assumption in analyzing video interaction data is that there will be a large number of students watching a video. But how many students do we need to start seeing a pattern in the learning behavior? Can we apply the same technique to on-campus classrooms where there are hundreds of students at maximum? Also, how do students with different learning intentions consume video? Are some learning patterns more salient in one group of students than others? Recent work on clustering learner goals [5] might help with a more in-depth analysis. These important questions can determine the scope of our approach and extend its applicability.

2. Designing video interfaces that adapt to collective learner behaviors

What does our interaction data analysis mean for video interfaces? We believe our analysis has implications for the design of improved video learning experience. First of all, there are lessons for video recording and editing. For example, a peak where students return to content shortly after a cut indicates that the cut might be abrupt. Video editors can use such information to make data-driven decisions about where to add scene cuts. Instructors can address student confusion inferred from the peak analysis during virtual office hours or in discussion forums.

However, we are even more excited about developing new video interfaces that better respond to collective learning behaviors on-the-fly. We envision a world in which the experience of interacting with a learning video keeps improving automatically as more and more learners interact with it. A common interaction pattern we observed in our data is non-sequential and selective watching. Re-watching students tend to actively seek their points of interest, in contrast to first-time watchers who watch more linearly. This suggests a need for improved tools to help learners navigate through the video more effectively. Capturing and displaying a representative frame for each interaction peak can visually summarize a video (Figure 6). The highlights are dynamically updated as more learner behaviors are collected. To further support learning, content analysis techniques can add more interactivity to videos. For example, text processing can label a peak with relevant keywords and index them for

enhanced search and browsing. Finally, our work thus far has opened a gateway for exploring novel interaction techniques to enable a much more dynamic video-learning experience. For instance, video segments near interaction peaks could be automatically slowed down to accommodate longer comprehension time, paused to present in-video quizzes as a checkpoint, or temporarily duplicated to keep the most recent content (a slide or a note) visible even when the main video stream transitions to the talking head of the instructor.

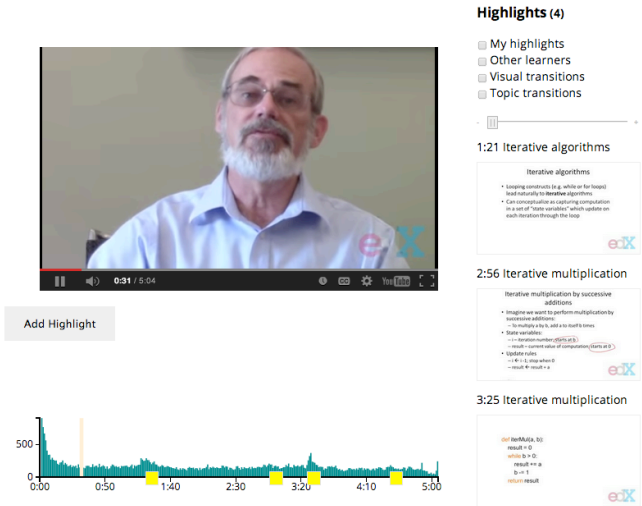

Figure 6. A video interface prototype with highlights (right) that are dynamically created from interaction peaks. The timeline (bottom left) shows play events over time, and highlights are marked in yellow.

Conclusion

Large-scale data generated from students' interaction with video enables a deeper understanding of how students use video in their learning trajectory. We believe there are many unanswered questions around how to interpret and understand the video interaction data. Such analysis can help improve existing videos by revealing points of interest and confusion. Furthermore, HCI researchers and platform designers can take a data-driven approach to designing the next-generation video content and interfaces. We hope to seek feedback on our approach, and invite more researchers to share our vision.

References (suggested ones with links)

[1] Breslow, L., Pritchard, D.E., DeBoer, J., Stump, G.S., Ho, A.D., and Seaton, D.T. Studying learning in the worldwide classroom: Research into edX's first MOOC. *RPA 8* (2013), 13-25. http://www.rpajournal.com/studying-learning-in-the-worldwide-classroom-research-into-edxs-first-mooc/

[2] Guo, P.G., Kim, J., Rubin, R. How Video Production Affects Student Engagement: An Empirical Study of MOOC Videos. In *Learning at Scale 2014, to appear*. http://juhokim.com/files/LAS2014-Engagement.pdf

[3] Hofmann, T. Probabilistic Latent Semantic Indexing. In *SIGIR'99*, ACM(1999), 50-57.

[4] Kim, J., Guo, P.G., Seaton, D.T., Mitros, P., Gajos, K.Z., Miller, R.C. Understanding In-Video Dropouts and Interaction Peaks in Online Lecture Videos. In *Learning at Scale 2014, to appear*. http://juhokim.com/files/LAS2014-Peaks.pdf

[5] Kizilcec, R.F., Piech, C., and Schneider, E. Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. In *LAK'13*, ACM (2013), 170–179.

[6] Marcus, A., Bernstein, M.S., Badar, O., Karger, D.R., Madden, S., and Miller, R.C. Twitinfo: aggregating and visualizing microblogs for event exploration. In *CHI'11*, ACM (2011), 227–236.

[7] Xie, S., Hakkani-Tur, D., Favre, B., and Liu, Y. Integrating prosodic features in extractive meeting summarization. In *ASRU*, 2009.