Video Digests: A Browsable, Skimmable Format for Informational Lecture Videos

Amy Pavel, Colorado Reed, Björn Hartmann, Maneesh Agrawala
University of California, Berkeley
{amypavel, cjrd, bjoern, maneesh}@cs.berkeley.edu

ABSTRACT

Increasingly, authors are publishing long informational talks, lectures, and distance-learning videos online. However, it is difficult to browse and skim the content of such videos using current timeline-based video players. Video digests are a new format for informational videos that afford browsing and skimming by segmenting videos into a chapter/section structure and providing short text summaries and thumbnails for each section. Viewers can navigate by reading the summaries and clicking on sections to access the corresponding point in the video. We present a set of tools to help authors create such digests using transcript-based interactions. With our tools, authors can manually create a video digest from scratch, or they can automatically generate a digest by applying a combination of algorithmic and crowdsourcing techniques and then manually refine it as needed. Feedback from first-time users suggests that our transcript-based authoring tools and automated techniques greatly facilitate video digest creation. In an evaluative crowdsourced study we find that given a short viewing time, video digests support browsing and skimming better than timeline-based or transcript-based video players.

Author Keywords

video digests; education; video presentation interfaces

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

INTRODUCTION

Informative videos such as classroom lectures, seminar talks, and distance-learning presentations are increasingly published online. For instance, websites such as edX [1], Khan Academy [2], and TED [3], offer thousands of informative video presentation on a wide variety of topics. Unlike live presentations, viewers can pause, replay, navigate, and alter playback speed to change the pace and structure of the video. The permanency of video also provides a referencable resource for later review.

Yet, one problem with video as a medium for informative presentations, is that it is difficult for viewers to browse and skim the underlying content. Viewers must scrub back-and-forth through a video to gain an overview of the presented topics or locate content of interest. Some platforms, such as edX and TED, let users search and navigate videos via a transcript; clicking a word in the transcript plays the video at that location. While transcripts do expose the content of the video, transcripts of informational videos often consist of long blocks of text and usually contain disfluencies and redundancies typical of speech. This makes the transcripts time-consuming to read and difficult to skim. Moreover, without a structured organization to the text, it can be difficult for viewers to browse the topics covered in the video or get a high-level overview of the content.

Recently, Bret Victor introduced a format for informational video presentations that is explicitly designed to helps viewer browse and skim a video presentation [41]. As shown in Figure 1, this format uses a textbook-inspired chapter/section structure to explicitly display the major themes in a presentation (the "chapters") as well as lower-level summaries of these themes (the "sections"). Specifically, each chapter corresponds to a topically-coherent segment of the video and consists of an embedded video player, a description (title) of the major theme in the segment, and a sequence of section elements. Each section element provides a short text summary and representative keyframe for a video segment within the chapter-level segment. We call this format a *video digest*.

The visual design of such video digests directly exposes the content of a video at a topical level: viewers can browse the chapter titles to obtain an understanding of the major themes in the presentation and skim the short summaries and keyframes to gain a finer-grained understanding of the presented content. This format encourages dividing informative presentations into short, topically-coherent video segments which, as indicated by prior work, aids knowledge transfer and decreases dropouts for educational videos [32, 20, 26]. However, creating a video digest is a time-consuming process: authors must segment the videos at multiple granularities (chapter/section), compose section summaries, select representative keyframes, and create the final output display.

In this paper, we present a set of tools to help authors create video digests by efficiently segmenting and summarizing the video through *transcript-based interactions.* The key insight of our approach is that much of the information in lecture videos is conveyed through speech. Therefore, our interface allows authors to navigate, segment and summarize the video using a time-aligned transcript of the speech. We also provide algorithmic tools for automatically segmenting the video

UIST 2014, October 5–8, 2014, Honolulu, HI, USA.
ACM 978-1-4503-3069-5/14/10.
http://dx.doi.org/10.1145/2642918.2647400

Figure 1. A video digest affords browsing and skimming through a textbook-inspired chapter/section organization of the video content. The chapters are topically coherent segments of the video that contain major themes in the presentation. Each chapter is further subdivided into a set of sections, that each provide a brief text summary of the corresponding video segment as well as a representative keyframe image. Clicking within a section plays the video starting at the beginning of corresponding video segment.

and a crowdsourcing pipeline for summarizing the resulting segments. Authors can further refine these auto-generated digests in the authoring interface if necessary.

We use our tools to generate video digests for several kinds of informational videos, ranging from a TED talk to an online MOOC lecture. We compare manually authored digests to auto-generated digests and find that they both distill the informational content to the most important topics. However, the auto-generated digests are more verbose than the manually authored digests. Feedback from first-time authors suggests that our transcript-based authoring tools combined with the automatic seeding greatly facilitate digest creation. We also conduct a crowdsourced study comparing video digests to timeline-based and transcript-based video player. We find that given 2 minutes to view lecture videos that are about 15 minutes long, viewers can recall up to twice as many of the key topics of the video using the digest format. At 8 minutes of viewing time, the differences between formats recedes. These results suggest that video digests support browsing and skimming of lecture videos better than the standard formats.

RELATED WORK

Our video digest format is inspired by Bret Victor's hand-crafted presentation design that he created for his talk *Media for Thinking the Unthinkable* [41]. Similarly, our work also draws from Jonathan Corum's slide-based presentation format that juxtaposes an image of each presentation slide with a transcript of the corresponding speech [15]. These examples were hand-crafted using a collection of off-the-shelf video editing, image retouching and HTML/Javascript coding tools. In contrast, our authoring tools combine techniques from prior work in video summarization, text summarization and media editing to facilitate the creation of video digests.

Video Summarization

Automatically summarizing a video to include only the most salient content is a long-standing research problem. Truong and Venkatesh's [39] survey of work on this problem divides video summarization methods into two main approaches; *keyframe methods* identify a sequence of static frames that together represent the salient video content [40, 8, 4, 22], while *video skim methods* shorten the input video by removing non-essential content [36, 23, 38, 14]. Keyframe methods primarily focus on conveying the visual content of a video in static form and are not designed to expose any of the information content contained in the vocal audio track. Video skims concatenate important segments of a video into a shorter video, but still rely on opaque, timeline-based navigation of the video's content. In contrast, our work focuses on presenting the informational content of a video in a hierarchically organized chapter/section structure that supports browsing and skimming.

An alternative to algorithmic video summarization is to crowdsource the summarization task. Adrenaline [5] uses a

crowdsourcing pipeline to extract representative keyframes from video segments. EpicPlay [37] uses viewers' interactions on social media to identify important moments in sports video. Lasecki et al. [29, 30] use non-expert crowds to transcribe videos and to describe activities in videos. Most related to our technique is the work by Kim et al. [26] which annotates steps in how-to videos. They use a Find-Verify-Expand technique to label steps in a how-to video and associates before and after images of each actionable step. Our work similarly relies on crowdsourced judgments to extract information from a lecture video.

Text Summarization

Text summarization can be divided into two alternative approaches [33]: *extractive* summaries concatenate existing text fragments from the document, while *abstractive* summaries generate new language to convey the main topics of the text. As lecture transcripts often contain disfluencies and redundancies typical of speech, directly concatenating portions of the transcript into an extractive summary often generates incoherent results. Therefore we focus on generating abstractive summaries for video digests, but because current algorithmic techniques cannot produce human quality abstractive summaries [21], we use crowdsourcing to generate them.

Our work builds on several previous crowdsourcing techniques for abstractive summarization of text. Soylent [6] employs human language understanding to generate abstractive summaries to shorten text: different sets of crowd workers identify lengthy sentences, rewrite to shorten these sentences, and vote on the best edits. However, edits are limited to single sentences and the resulting summaries are not always coherent across different edits. Burrows et al. [9] develop a crowdsourcing pipeline for summarizing text documents that includes an automatic classifier designed to filter out poor summaries. Researchers have also presented techniques for using crowdsourced summaries of text documents to improve machine translation pipelines [10, 16]. We are inspired by these prior techniques and introduce a new crowdsourcing method to generate high-quality abstractive summaries based on text and video information.

Media Editing Tools

Our tools for generating video digests also build upon prior work on transcript-based video and audio editors. Several systems have used time-aligned transcripts to support audio/video editing through text manipulation operations [43, 11, 7, 35]. Such tools provide, for example, clip segmentation through editing markers in the transcript. Other metadata, such as annotations of actions and steps, can facilitate automatic shortening of how-to videos [12]. Drawing from this body of work, we have focused the video digest editing interface on transcript-based interactions.

CREATING A VIDEO DIGEST

A video digest uses a textbook-inspired chapter/section structure to make the video easier to browse and skim (Figure 1). The chapter elements correspond to topically-coherent segments of the video that present a major theme, and the section elements segment the chapter into lower-level topic shifts.

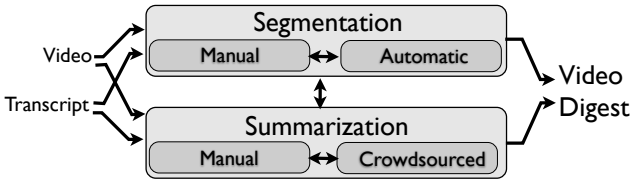

Figure 2. Given an input video and transcript, our authoring interface provides users with the ability to manually segment/summarize the content, automatically segment the content and crowdsource the summaries, or apply any combination of these two approaches.

Each section provides a brief summary and representative keyframe for its corresponding video segment (Figure 1).

To produce such a digest, an author must complete the following tasks: (1) segment the video into chapters, (2) title the chapters, (3) segment the video into short sections, (4) compose a text summary for each section and (5) select a keyframe for each section. Although most of these tasks can be interleaved, in practice we have found that video digest authors sometimes work bottom-up and create sections first (tasks 3-5), before grouping them into chapters (tasks 1-2), while at other times they work top-down creating chapter segments first (tasks 1-2), and then breaking them further into sections (tasks 3-5). Cycling between these two strategies is common [42].

Regardless of the strategy, each of the five tasks is time-consuming with current tools. Segmenting a video by topic often involves watching the video several times and scrubbing back-and-forth to find topic boundaries. Composing section summaries also typically requires re-watching a segment multiple times to make sure the main points are fully captured in the summary.

We have developed a set of tools to facilitate video digest creation (Figure 2). Our tools take a video and a corresponding transcript as input and lets users segment and summarize the video using a combination of manual, automatic, and crowdsourcing techniques. Users can interleave segmentation and summarization steps in any order. We first describe our video digest authoring interface and then present the algorithmic methods underlying this interface.

VIDEO DIGEST AUTHORING INTERFACE

As shown in Figure 3, our video-digest authoring interface consists of two main panes: an *Aligned Transcript pane* (right) lets authors read the speech content, click on a word to navigate to the corresponding point in the video player, and mark chapter/section start points, while a *WYSIWYG Editor pane* (left) lets authors specify chapter titles, sections summaries and keyframes.

To support transcript-based navigation, segmentation and summarization, our interface relies on a time-aligned text transcript of the input video. When possible we obtain transcripts from the video source. For example, edX and TED provide transcripts of their talks online. Otherwise we use the crowdsourcing transcription service `rev.com` which accepts an audio file as input and returns a verbatim transcript for $1.25 per minute. We then time-align the transcript to the

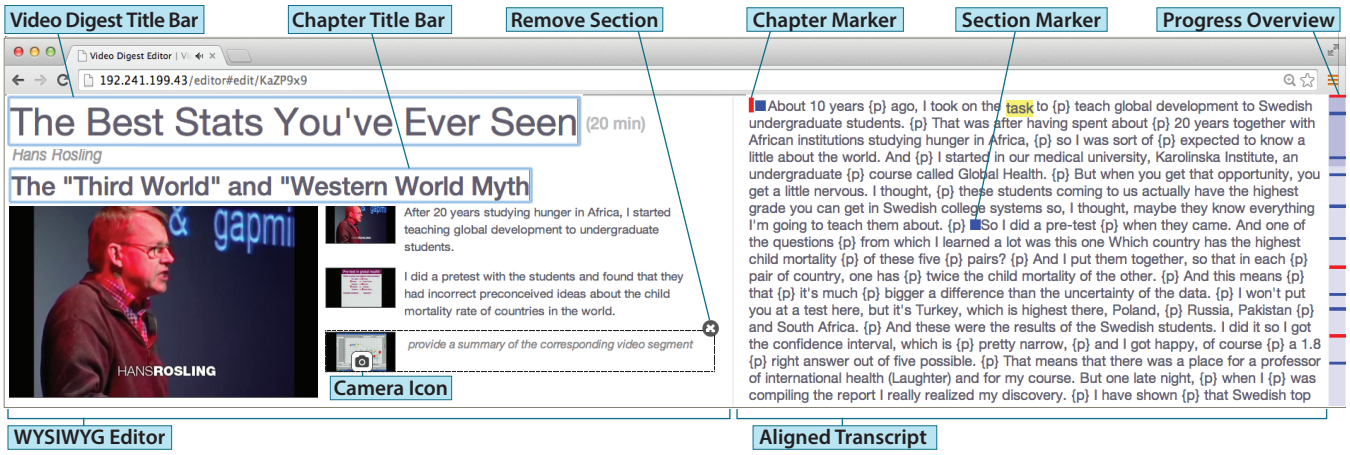

Figure 3. Our interface facilitates creating and editing video digests. The interface consists of two main panes: (1) An Aligned Transcript pane for navigating, segmenting and summarizing the talk and (2) a WYSIWYG Editor pane for adding chapter titles, summaries and keyframes for each section. Additionally, a Progress Overview scrollbar allows authors to view their segmentation progress and return to areas for refinement.

audio track of the video using the phoneme estimation and mapping technique of Rubin et al. [35, 44].

Chapter and Section Segmentation

To segment the video into topically-coherent chapters and sections, authors place chapter markers (red) and section markers (blue) in the Aligned Transcript pane using mouse clicks with modifier keys ('ctrl+alt' for chapter, 'ctrl' for section). Chapter markers denote the start of a major theme in the presentation while section markers denote less-significant topic changes within each chapter. Dragging these markers to different locations in the transcript changes the starting point of the chapter/section. In addition, clicking a section marker with the chapter modifier key creates a new chapter at that location, and the clicked section as well as all remaining sections in the original chapter are automatically moved into the new chapter. This operation splits the original chapter into two chapters at the clicked section marker. Conversely, clicking a chapter marker with the section modifier key removes the original chapter and appends all of its sections to the preceding chapter. This operation merges the clicked chapter with the preceding chapter. The Progress-Overview scrollbar on the right side of the Aligned Transcript pane represents the entire length of the video as a vertical bar. It shows the locations of all chapter/section markers and allows the author to quickly assess areas that need segmentation or refinement.

When the author places a new chapter marker in the transcript, our interface generates a new chapter in the WYSIWYG Editor with its video element cued to the location of the chapter marker. Similarly, when the author places a new section marker, our interface generates a new section keyframe and summary box in the WYSIWYG Editor pane at the appropriate location. By default, it fills the keyframe with the first frame of the corresponding video segment and places the cursor in an empty adjoining summary box. The WYSIWYG Editor automatically updates when the author drags a section/chapter start point to a different location or splits/merges chapters by clicking on chapter or section markers with the opposite modifier keys. Authors can delete sections by clicking the "Remove Section" button that appears when hovering over or modifying a section. Removing all sections from a chapter deletes the chapter.

In addition to these manual segmentation operations, the author can select a portion of the transcript and then invoke our automatic segmentation algorithm (see Algorithmic Methods Section) on the selected text, using a right-click menu.

Section Summaries and Keyframes

Authors compose section summaries directly in the summary boxes of the WYSIWYG Editor. Clicking on a summary box scrolls the transcript view to the corresponding text segment for quick reference. The author can replace the default keyframe by navigating the chapter's video player to the desired location and clicking a camera icon beneath the desired keyframe. Alternatively, an author can right-click on the summary box to invoke our crowdsourcing summarization pipeline (see Algorithmic Methods Section). The pipeline returns a crowd generated section summary and keyframe which the author can then refine if necessary. Finally, the author can set the video digest title and chapter-level titles by directly editing the text in the respective title bars.

ALGORITHMIC METHODS

Our system provides automated techniques for segmenting a video into topically-coherent units and obtaining summaries of such segments via a crowdsourcing pipeline.

Algorithmic Segmentation

Automatic text segmentation is a well-studied problem in natural language processing [24, 13, 17]. Eisenstein et al.'s [18] Bayesian topic segmentation (BSeg) algorithm is one of the leading techniques for segmenting speech-based text. BSeg is designed to group sequences of lexically cohesive text fragments into a segment – the text fragments can be any user-defined sequence of words in the text such as phrases, sentences or paragraphs.

The strength of BSeg is its ability to incorporate a variety of features such as cue phrases (e.g. "in conclusion", "therefore", "so", etc.) that might signal topic transitions. In pilot experiments[1] we found BSeg to outperform several other modern text segmentation algorithms [24, 31]. As a result, we use BSeg to automatically obtain both section- and chapter-level segments in our system.

BSeg produces a *linear* segmentation of the input text rather than the *hierarchical* chapter/section segmentation needed by our system. To obtain a hierarchical segmentation, we apply BSeg twice. First, we use the sentences from the original transcript as the input text fragments and BSeg returns a sequential grouping of these sentences into section-level elements. Next, we apply BSeg again, but we treat the output section-level elements from the first application as the input fragments to the second BSeg application. If summaries of the sections are available, we instead use these summaries as the input fragments to the second BSeg application. In either case this second application of BSeg groups the section-level segments into topically-coherent units that form the chapters of our digest.

Crowdsourced Section Summaries

We have developed a crowdsourcing pipeline for obtaining section summaries. In our pipeline, one set of crowdworkers compose a summary and select a representative keyframe for each of the input sections. A second set of crowdworkers rank the summaries and keyframes based on quality. We then return the top-ranked summary for each section to the authoring interface.

In order to create browsable and skimmable video digests, each summary must concisely summarize the main topic of the corresponding section and use the same grammatical person and tense as the surrounding sections. We emphasize these goals in a set of guidelines we provide to the crowdworkers. We tell them that the summary should: (1) convey the main point(s) of the section, (2) omit non-essential details, (3) use the same tense (e.g. past, future) and grammatical person (e.g. first person, third person) as the section transcript, and (4) use concise wording, free of grammatical errors. In early experiments with the task, we found that workers often generated overly-detailed summaries. Based on these experiments, we added the guideline that (5) the summary should be less than three sentences in length.

For each summary task, we provide workers with the video and the aligned transcript cued to the section start point (Figure 4). We give workers access to the complete video and transcript so that they can build additional context when needed (e.g. to resolve ambiguous pronoun references in the section). We ask at least three crowdworkers to provide summaries and keyframes for each section. To ensure high-quality summaries, we then pipe these summaries into a ranking stage, where a different set of crowdworkers rank the quality of the summary-keyframe pair from the set of such pairs for each section. We ask these crowdworkers to base

[1] See supplementary material for details on these experiments and our technique for setting BSeg's parameters.

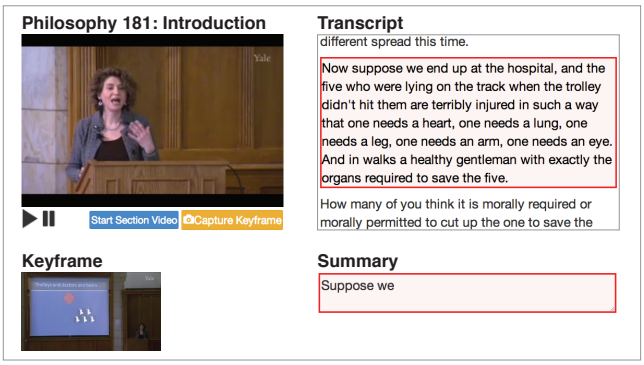

Figure 4. Crowdworker task for summarizing a section of the video. Workers can navigate the video using the timeline-based player (left) or the aligned transcript (right). The video is initially cued to the beginning of the section that must be summarized and the corresponding portion of the transcript is highlighted in red. Workers can select a keyframes by clicking on the *capture keyframe* button area, and they can write a text summary in the *summary* textbox.

their rankings on the summary writing guidelines and to provide a short justification for their top-ranked selection. To reduce the cognitive load required to understand each section, we ask each crowdworker to summarize or ranks three consecutive sections.

We pay crowdworkers $0.60 to write three section summaries and select the corresponding keyframes. To incentivize high-quality work we offer $0.10 bonuses to the worker who writes the top-ranked summary for each section. Similarly we pay crowdworkers $0.60 to rank the summary-keyframe pairs for three sections.

RESULTS

Figure 5 shows manual and auto-generated video digests produced using our tools. The complete interactive results can be viewed at `http://vis.berkeley.edu/papers/videodigests`. These results were generated using the following set of input videos (Table 1):

- *Philosophy 181: Introduction* by Tamar Gendler [19]: a 13.3 minute recording of an in-class philosophy lecture at Yale University (Figure 5A,B)
- *The Power of Prototyping* by Scott Klemmer [28]: a 13.8 minute introductory lecture from Coursera that uses a slide-based presentation (Figure 5C,D)
- *US History Overview: Jamestown to the Civil War* by Salman Khan [25]: a 18.5 minute overview of U.S. history from Khan Academy that uses a pen-based screencast presentation (Figure 5E,F)
- *The Best Statistics You've Ever Seen* by Hans Rosling [34]: a 19.9 minute high-production-quality, stage-based presentation (Figure 5G,H)

One of the paper authors created the manual video digests, and we created the auto-generated digests by combining our automatic segmentation algorithm on the entire transcript with our crowdsourced summarization pipeline. Table 1 shows the total time and cost required to create the manual

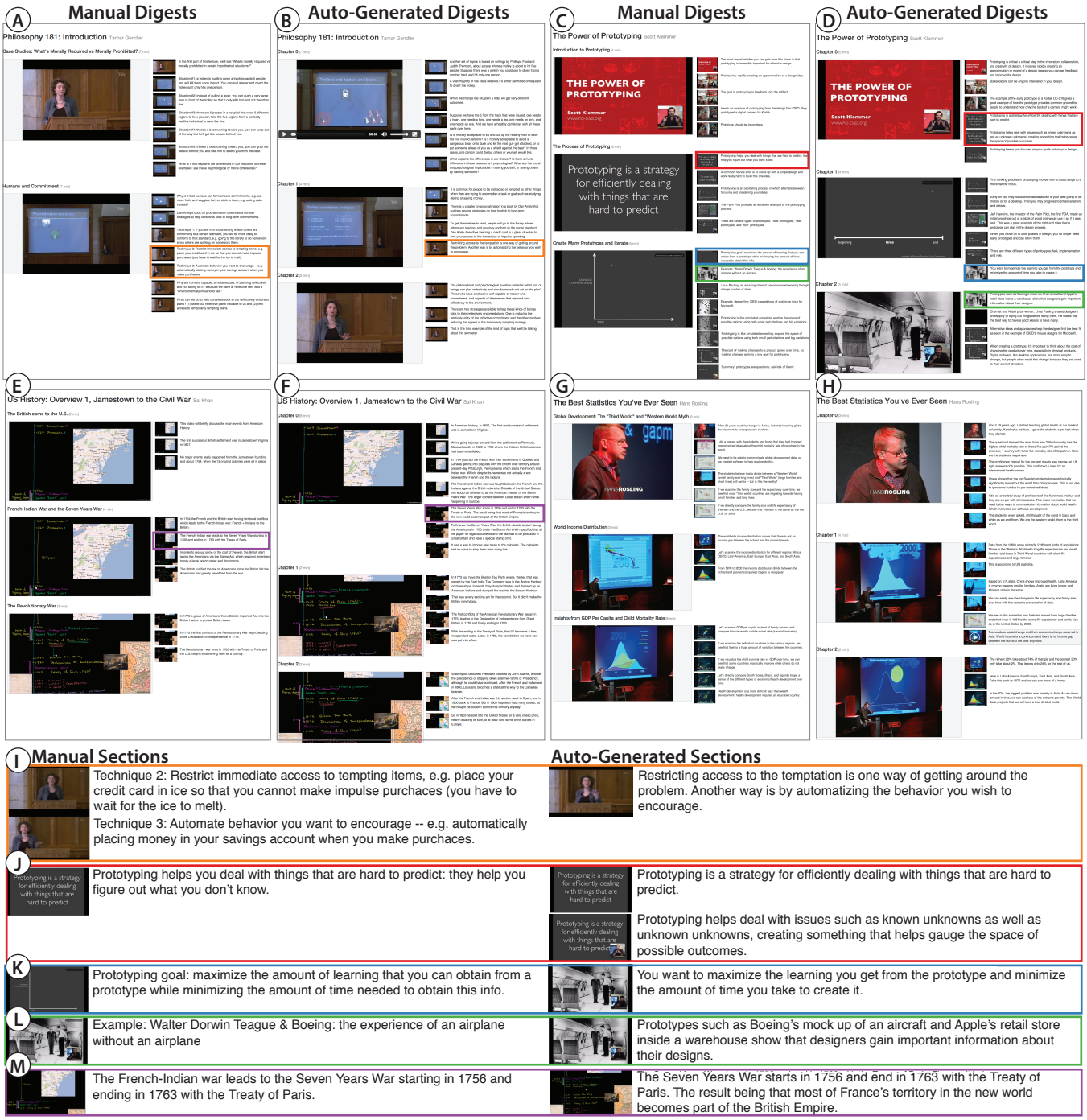

Figure 5. Manual and auto-generated video digests for four lecture videos Gendler (A,B), Klemmer (C,D), Khan (E,F) and Rosling (G,H). Differences between the manual and auto-generated results are highlighted below (I-M). Example (I) and (J) show differences in segmentation where two sections in the manual digest are combined into one section in the auto-generated digest and vice-versa. Example (K) shows how sections summaries can be very similar between the manual and auto-generated digests. However, examples (L) and (M) show that the manual digests often include more succinct summaries than the corredponding auto-generated digests.

(M) and auto-generated (A) digests respectively. We instrumented our authoring tools to record the time spent performing each subtask involved in digest creation. In the manual case, we spent 48% of the total time reviewing the transcript and lecture video, 42% of the time composing section summaries, 6% of the time writing chapter titles, and the remaining 4% of time performing all other operations including placing segments. Overall, the time to create the digests manually was about 3-4 times the length of the input lecture. In the auto-generated case we recruited three unique crowdworkers to generate summaries and keyframes for each section and three more to rank each summary-keyframe pair. The total crowdsourcing cost was between about $0.50 and $1.00 per minute of the input lecture.

	Gendler		Klemmer		Khan		Rosling	
	M	A	M	A	M	A	M	A
Time to create	40m	–	34m	–	41m	–	62m	–
Crowdwork cost	–	$19	–	$23	–	$35	–	$54
Num. of chapters	2	3	3	4	6	6	4	8
Num. of sections	20	26	14	13	17	16	20	40
Compression ratio	4.8	3.7	10.0	6.3	7.7	4.6	7.1	3.3

Table 1. Creating manual (M) and an auto-generated (A) video digests for four lecture videos required either authoring time (M) or payments to crowdworkers (A). We report the number of chapters, sections and compression ratios for each of the resulting digests.

Table 1 also shows that the manual and auto-generated digests contain similar chapter and section counts for all of the lectures except Rosling. The Rosling lecture contains many short anecdotes that each use different vocabulary. In the auto-generated digest, BSeg segments the lecture based on the frequent vocabulary changes and produces twice as many chapters and sections as in the manual digest. Unlike BSeg, the manual digest author grouped together the anecdotes into higher-level concepts. Despite the differences in chapter/section counts, Figures 5G and 5H show that the two Rosling digests cover the same topics, but at different granularities. Other differences in the section-level segementations are shown in Figures 5I and 5J.

Finally we note that the manual digests usually contain less summary text overall than the auto-generated digests. Table 1 reports the compression ratio – the number of words in the original transcript divided by the number of words in the digest – for the lectures. Although the manual digests achieve higher compression ratios than the auto-generated digests, both condense the information compared to the transcript. The examples in Figures 5L and 5M suggest that crowdworkers tend to put more context into their section summaries which makes them longer than manually authored summaries. Because crowdworkers only summarize a small part of the lecture and cannot see the surrounding summaries, they may be compensating by repeating contextual information. In contrast, the author of a manual digest has access to all of the summaries and can eliminate such redundancies. Nevertheless, as demonstrated in Figure 5, the crowdsourced summaries do capture the the main concepts of each section similar to the manual summaries.

INFORMAL USER FEEDBACK

To gauge the utility of our video digest creation tools, we conducted an informal evaluation with two users (U1 and U2). We asked them to manually create a video digest and also to refine an auto-generated digest using our authoring tools. We examined the time they spent editing the digests and the number/type of edits they made. We also conducted a post-evaluation interview to gather qualitative feedback.

In the evaluation, U1 manually created a digest for Scott Klemmer's *Power of Prototyping* lecture and refined an auto-generated digest for Tamar Gendler's *Philosophy 181* lecture. U2 completed the opposite tasks for these two lectures. Both of these input videos are approximately the same length. Before starting the tasks, we presented each user with an example of a video digest, explained the chapter/section structure, and demonstrated the authoring interface. For the refinement task, we instructed the users to refine the digest so that it matched the quality of their manually created digest.

	User 1		User 2	
	M	E	M	E
Lecture	Klemmer	Gendler	Gendler	Klemmer
Time to create	54m	37m	30m	20m
Num. of sections	19	13	14	16
Num. of sections edited	–	6	–	4
Summary-edit keystrokes	3812	1355	1011	317

Table 2. In an informal evaluation two users created a manual video digest (M) and edited an auto-generated digest (E). We report the number of sections created for both (M) and (E), but only include the number of sections each user edited in the latter case.

Table 2 shows that both users spent less time refining the auto-generated digest than creating the manual digest and edited fewer than half of the auto-generated section summaries. Both users performed fewer keystrokes when editing the auto-generated summaries, than when creating the summaries manually from scratch. When editing the auto-generated digests, modifying the sections was the dominant form of interaction. U1 focused on improving the flow between section summaries, while U2 mainly adjusted section boundaries. They rarely modified default keyframes or adjusted chapter boundaries when they were refining the auto-generated result. Although both users did edit the auto-generated summary text, they also provided positive feedback on the auto-generated summaries: U1 stated that the auto-generated summaries were "summarized in a way that I wouldn't think of myself, and I think what they did was correct and great." U2 noted that the auto-generated summaries were "on-target." U1 noted encountering a single incorrect summary, while U2 found one unnecessary segment boundary.

STUDY: DO DIGESTS SUPPORT BROWSING/SKIMMING?

We performed a comparative study to test the hypothesis that video digests afford browsing and skimming better than alternative formats. In our experiment we asked crowdworkers to watch one of four lectures (Gendler, Klemmer, Khan, Rosling) using one of the following formats:

- ***Manual:*** a manually created video digest using our tools.
- ***Auto:*** an auto-generated video digest.
- ***Video:*** a timeline-based video player.
- ***Script:*** a transcript-based video player.

We gave the crowdworkers a fixed length of viewing time (either 2, 5 or 8 minutes) and asked them to "quickly browse and skim" the content of the lecture. We then hid the lecture and asked them to provide "an approximately 5 sentence summary" of the main points covered in the lecture. We purposely did not give the crowdworkers enough time to watch the entire lecture so that they had to browse and skim its content to write a complete summary.

We asked 4 unique crowdworkers to summarize each combination of independent variables (lecture, format, viewing time) yielding 192 total summaries (4 lectures $\times$ 4 formats $\times$ 3 viewing time $\times$ 4 crowdworkers). We paid each crowdworkers \$0.90 for the summary plus a \$1.00 bonus if the

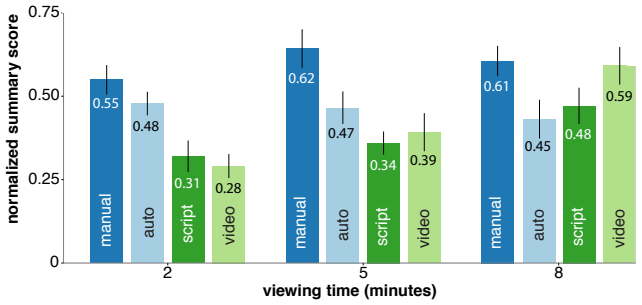

Figure 6. Crowdworkers viewed a video lecture in one of four formats (manual, auto, script, video) for 2, 5 or 8 minutes and then wrote a summary of the presentation. We scored these summaries using a gold standard topic list. Each bar shows the mean and standard error of the scores for each condition.

worker obtained the highest summary score for the (lecture, format, viewing time) condition.

To evaluate the five sentence crowdworker summaries the first two authors of this paper worked together to manually build a *gold standard* list of key topics discussed in each of the four lectures. We then used these lists to score the number of topics covered in each crowdworker summary. Finally, we normalized the scores based on the total number of gold standard topics for each lecture.

Figure 6 shows the normalized scores for each format and viewing time aggregated across the four lectures. Using a Kruskal-Wallis test, we found significant difference in these scores when compared across viewing times (2 minutes: $\mu = 0.408$; 5 minutes: $\mu = 0.457$; 8 minutes: $\mu = 0.530$. $\chi^2(2) = 9.92$, $p = 0.007$). Further analyzing each viewing time, we found a significant difference in scores when comparing across the formats for 2 minutes ($\chi^2(3) = 18.31$, $p < 0.001$) and 5 minutes ($\chi^2(3) = 16.23$, $p = 0.001$) , but not for 8 minutes ($\chi^2(3) = 6.77$, $p = 0.080$). Pairwise Mann-Whitney tests with Holm-Bonferroni correction indicated significant differences in the following format pairs: At 2 minutes of viewing time, manual-video ($U(15) = 32$, $p = 0.002$), manual-script ($U(15) = 44.5$, $p = 0.009$), and auto-video ($U(15) = 58.5$, $p = 0.037$) are significantly different. At 5 minutes of viewing time, the manual-video ($U(15) = 56.5$, $p = 0.03$) and manual-script ($U(15) = 33.5$, $p = 0.002$) were significantly different.

Further analyzing the formats, we found a significant increase in scores when comparing across the viewing times for only the video format ($\chi^2(2) = 13.26$, $p = 0.001$). Pairwise Mann-Whitney tests with Holm-Bonferroni correction find significant differences in the following viewing time pairs: 2 minute-8 minute ($U(15) = 34.5$, $p = 0.001$) and 5 minute-8 minute ($U(15) = 67.5$, $p = 0.047$).

In short, at viewing times of 2 and 5 minutes, the video digest formats (manual and auto) allowed viewers to recall up to twice as many of the key topics than the video player formats (script and video). However, this effect diminished for the longest viewing time of 8 minutes. For the roughly 15 minute long lecture videos we tests, viewers could successfully recall many key topics after only 2 minutes of viewing time using the video digest; giving extra viewing time yielded little improvement that was not statistically significant. In contrast, with the standard timeline-based video player, summarization performance was low when viewers were given only 2 minutes, and gradually improved with additional time. Together these results suggest that both of the video digest formats – manually authored and auto-generated – facilitate browsing and skimming of informational lecture videos.

CONCLUSION AND FUTURE WORK

We have presented a set of tools for creating video digests; a new format for informational talks that exposes the structure of the content via section-level summaries and chapter-based grouping. We provide a transcript-based authoring interface and explore techniques for automatically segmenting and summarizing an input video. An informal evaluation suggests that our tools make it much easier for authors to create video digests, and a crowdsourced experiment indicates that the video digest format affords browsing and skimming better than alternative video presentation interfaces. As more and more informational videos are published online, we believe these tools will make it easier for people to browse and skim the underlying content and identify topics of interest. We see several promising directions for future work.

Ensuring consistency across crowdsourced summaries. Our current crowdsourcing pipeline does not ensure consistency between section summaries produced by different workers. Separate workers may include redundant information or use pronouns that are ambiguous given previous summaries. One solution might be to include an additional stage in the crowdsourcing pipeline where new crowdworkers check multiple consecutive summaries for overall consistency. These crowdworkers could also provide titles for the video digest chapters as our current system does not produce such titles automatically.

Support for highly-technical content. We tested our automatic pipeline on four lectures that are accessible to a broad, well-educated audience. However, we have not tested our tools with lectures that require specialized knowledge, or highly-technical content where crowdworkers may not have the necessary background to write summaries. One fruitful direction for MOOC style lectures may be to ask students who choose to watch a lecture to write the summaries, e.g. students in a graduate-level quantum mechanics course. Creating summaries may help the students learn the material and also generate high-quality summaries for technical material.

Use video data in algorithmic segmentation. Our segmentation algorithm only uses the transcript to segment the video. Future work could incorporate visual and audio information to improve video segmentation. It may also be possible to use viewer interaction data to automatically infer segmentation points in the video in the manner of Kim et. al [27].

ACKNOWLEDGMENTS

We thank Bret Victor for showing us how he manually created his video digest and helping us think through the interface design for creating video digests. This work was partially

supported by a SanDisk fellowship, an NDSEG fellowship and NSF grants IIS-1210836 and IIS-1252819.

REFERENCES

1. edX. `http://www.edx.org`.
2. Khan Academy. `http://khanacademy.org`.
3. TED. `http://www.ted.com/`.
4. Barnes, C., Goldman, D. B., Shechtman, E., and Finkelstein, A. Video tapestries with continuous temporal zoom. *ACM Trans. Graph. 29*, 4 (July 2010), 89:1–89:9.
5. Bernstein, M. S., Brandt, J., Miller, R. C., and Karger, D. R. Crowds in two seconds: Enabling realtime crowd-powered interfaces. In *UIST*, ACM (2011), 33–42.
6. Bernstein, M. S., Little, G., Miller, R. C., Hartmann, B., Ackerman, M. S., Karger, D. R., Crowell, D., and Panovich, K. Soylent: a word processor with a crowd inside. In *Proc. of the 23nd annual*, ACM (2010), 313–322.
7. Berthouzoz, F., Li, W., and Agrawala, M. Tools for placing cuts and transitions in interview video. *ACM Trans.Graph. 31*, 4 (2012), 67.
8. Boreczky, J., Girgensohn, A., Golovchinsky, G., and Uchihashi, S. An interactive comic book presentation for exploring video. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '00, ACM (New York, NY, USA, 2000), 185–192.
9. Burrows, S., Potthast, M., and Stein, B. Paraphrase acquisition via crowdsourcing and machine learning. *ACM Transactions on Intelligent Systems and Technology (TIST) 4*, 3 (2013), 43.
10. Buzek, O., Resnik, P., and Bederson, B. B. Error driven paraphrase annotation using mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, Association for Computational Linguistics (2010), 217–221.
11. Casares, J., Long, A. C., Myers, B. A., Bhatnagar, R., Stevens, S. M., Dabbish, L., Yocum, D., and Corbett, A. Simplifying video editing using metadata. In *Proceedings of the 4th conference on Designing interactive systems: processes, practices, methods, and techniques*, ACM (2002), 157–166.
12. Chi, P.-Y., Liu, J., Linder, J., Dontcheva, M., Li, W., and Hartmann, B. Democut: generating concise instructional videos for physical demonstrations. In *UIST*, ACM (2013), 141–150.
13. Choi, F. Y. Advances in domain independent linear text segmentation. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, Association for Computational Linguistics (2000), 26–33.
14. Christel, M. G., Smith, M. A., Taylor, C. R., and Winkler, D. B. Evolving video skims into useful multimedia abstractions. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM Press/Addison-Wesley Publishing Co. (1998), 171–178.
15. Corum, J. Storytelling with Data. `http://style.org/tapestry/`, February 2014.
16. Denkowski, M., Al-Haj, H., and Lavie, A. Turker-assisted paraphrasing for english-arabic machine translation. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, Association for Computational Linguistics (2010), 66–70.
17. Du, L., Buntine, W., and Johnson, M. Topic segmentation with a structured topic model. In *Proceedings of NAACL-HLT* (2013), 190–200.
18. Eisenstein, J., and Barzilay, R. Bayesian unsupervised topic segmentation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics (2008), 334–343.
19. Gendler, T. Philosophy 181: Introduction. `http://oyc.yale.edu/philosophy/phil-181/lecture-1`, Spring 2011.
20. Guo, P. J., Kim, J., and Rubin, R. How video production affects student engagement: An empirical study of mooc videos. In *Proceedings of the first ACM Learning@ scale conference*, ACM (2014), 41–50.
21. Gupta, V., and Lehal, G. S. A survey of text summarization extractive techniques. *Journal of Emerging Technologies in Web Intelligence 2*, 3 (2010), 258–268.
22. Haubold, A., and Kender, J. R. Augmented segmentation and visualization for presentation videos. In *Proceedings of the 13th annual ACM international conference on Multimedia*, ACM (2005), 51–60.
23. He, L., Sanocki, E., Gupta, A., and Grudin, J. Auto-summarization of audio-video presentations. In *Proceedings of the seventh ACM international conference on Multimedia (Part 1)*, ACM (1999), 489–498.
24. Hearst, M. A. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational linguistics 23*, 1 (1997), 33–64.
25. Khan, S. Us history overview: Jamestown to the civil war. `https://www.khanacademy.org/humanities/history/history-survey/us-history/v/us-history-overview-1--jamestown-to-the-civil-war`, April 2011.
26. Kim, J., Nguyen, P., Weir, S., Guo, P. J., Miller, R. C., and Gajos, K. Z. Crowdsourcing step-by-step information extraction to enhance existing how-to videos. In *Proceedings of the 2014 ACM annual conference on Human factors in computing systems*, ACM (2014).

27. Kim, J., Shang-Wen, L. D., Cai, C. J., Gajos, K. Z., and Miller, R. C. Leveraging video interaction data and content analysis to improve video learning. In *CHI'14 Extended Abstracts on Human Factors in Computing Systems*, ACM (2014).
28. Klemmer, S. The power of prototyping. **https://class.coursera.org/hci/lecture**, 2012.
29. Lasecki, W., Miller, C., Sadilek, A., Abumoussa, A., Borrello, D., Kushalnagar, R., and Bigham, J. Real-time captioning by groups of non-experts. In *UIST*, ACM (2012), 23–34.
30. Lasecki, W. S., Song, Y. C., Kautz, H., and Bigham, J. P. Real-time crowd labeling for deployable activity recognition. In *Proceedings of the 2013 conference on Computer supported cooperative work*, ACM (2013), 1203–1212.
31. Malioutov, I., and Barzilay, R. Minimum cut model for spoken lecture segmentation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, Association for Computational Linguistics (2006), 25–32.
32. Mayer, R. E., and Moreno, R. Nine ways to reduce cognitive load in multimedia learning. *Educational psychologist 38*, 1 (2003), 43–52.
33. Nenkova, A., Maskey, S., and Liu, Y. Automatic summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts of ACL 2011*, Association for Computational Linguistics (2011), 3.
34. Rosling, H. The best statistics you've ever seen. **http://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen**, February 2006.
35. Rubin, S., Berthouzoz, F., Mysore, G. J., Li, W., and Agrawala, M. Content based tools for editing audio stories. In *UIST*, ACM Press (2013), 113–122.
36. Smith, M. A., and Kanade, T. Video skimming and characterization through the combination of image and language understanding. In *Content-Based Access of Image and Video Database, 1998. Proceedings., 1998 IEEE International Workshop on*, IEEE (1998), 61–70.
37. Tang, A., and Boring, S. # epicplay: crowd-sourcing sports video highlights. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM (2012), 1569–1572.
38. Taskiran, C. M., Pizlo, Z., Amir, A., Ponceleon, D., and Delp, E. J. Automated video program summarization using speech transcripts. *Multimedia, IEEE Transactions on 8*, 4 (2006), 775–791.
39. Truong, B. T., and Venkatesh, S. Video abstraction: A systematic review and classification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP) 3*, 1 (2007), 3.
40. Uchihashi, S., Foote, J., Girgensohn, A., and Boreczky, J. Video manga: generating semantically meaningful video summaries. In *Proceedings of the seventh ACM international conference on Multimedia (Part 1)*, ACM (1999), 383–392.
41. Victor, B. Media for thinking the unthinkable. **http://worrydream.com/MediaForThinkingTheUnthinkable**, April 2013.
42. Victor, B. Personal communication, December 2013.
43. Whittaker, S., and Amento, B. Semantic speech editing. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM (2004), 527–534.
44. Yuan, J., and Liberman, M. Speaker identification on the scotus corpus. *Journal of the Acoustical Society of America 123*, 5 (2008), 3878.