THE EFFECT OF TEXT IN STORYBOARDS FOR VIDEO NAVIGATION

Michael G. Christel and Adrienne S. Warmack
Carnegie Mellon University
Pittsburgh, PA 15213 USA
{christel, asw}@cs.cmu.edu

ABSTRACT

A storyboard is a presentation scheme for abstracting information in a digital video clip based on imagery. This paper describes a series of storyboard interfaces with added transcript text features. These interfaces are used in a controlled experiment focusing on the utility of transcript text in storyboards for news video navigation. We wished to explore whether such text resulted in improvements in video navigation, and, if so, whether the amount of text and its synchronization with video imagery affected the navigation task. The text-augmented storyboards performed significantly better than storyboards with no text. Full transcript text produced benefits when presented as a block, whereas reduced contextual text descriptions produced benefits when aligned with storyboard image rows.

1. INTRODUCTION

Digital data is proliferating exponentially, as witnessed by the accumulating amount of content on the World Wide Web. In recent years, this data has included video, but video is a difficult media type to access efficiently, requiring significant amounts of viewing time to play through linearly and tremendous patience to download when network congestion or legacy hardware limits one's download bandwidth. In light of this situation, many multimedia interface researchers have focused on developing alternate representations for video, i.e., *surrogates*, enabling users to quickly assess whether a video clip is worthy of further inspection and providing quick navigation within the clip itself.

A surrogate ideally preserves and communicates the essential content of a source video or audio in a compact representation. Examples include brief titles and individual "thumbnail" images. Another common approach presents an ordered set of representative thumbnail images simultaneously on a computer screen [1, 7, 13, 14], referred to here as *storyboards* and shown in Figures 1 and 2 augmented with additional transcript text. The storyboard acts as a pictorial overview by representing each shot of a video clip with a thumbnail image, where a shot is a single video sequence collected from a camera. It acts as a navigation aid by processing mouse clicks within its borders as seek commands on the video. For example, clicking on the thumbnail showing the Seattle Space Needle causes the video clip to open (if it is not open already), and then begin playing at the shot where the Space Needle image occurs. By examining the utility of various surrogates for news video, we provide empirical evidence for future design improvements of digital video sites, focusing on open questions concerning text within storyboard surrogates.

2. STORYBOARD MULTIMEDIA SURROGATE

Storyboards are common to most digital video libraries, including CAETI [13], Pictorial Transcripts [3], and the Baltimore Learning Community [4] perhaps in part due to the amount of attention the image processing community has given to the automatic breakdown of video into component shots [3, 7, 9, 13, 14]. The storyboards we use are created by first identifying shots based on color histogram changes and black frame detection, and then representing each shot with an image from the video based on camera motion, and the detection of overlaid text or faces. Shot detection accuracy has been measured at greater than 90% [10].

For the Informedia CNN library, over 1 million shots are identified with an average length of 3.38 seconds. The clips, i.e., single news stories, average 110 seconds in length, resulting in an average image count for clip storyboards of 32.6. An area of active multimedia processing research attempts to reduce the number of shots represented in a storyboard to decrease screen space requirements [1, 7, 13]. In our research here, we did not include any shot reduction strategies, choosing to maintain our focus on the contribution of text to an unabridged image storyboard where all shots are represented.

Fairly accurate time-aligned transcripts exist for the test video collection via capturing their closed-captioning, determining when each word was spoken through an automatic alignment process using the Sphinx-III speech recognizer [12], and filtering the text into a mixed upper and lower case presentation. This timed text enables the creation of the text-augmented treatments of Figure 1. Within this figure, shot processing may indicate that the second storyboard row spans from 0:42 to 1:35 of the clip, and time alignment provides us with the means of getting the transcript text associated with the video clip's contents in the range 0:42 to 1:35.

Pilot testing with university staff and students gave overwhelming support for displaying interleaved text below

Figure 1. Storyboard image with interleaved transcript undergoing no reduction

the image row, in agreement with most pictorial figure text captions appearing beneath the imagery. These pilot tests resulted in the choice of MS Sans Serif 8 point font, where a 1024 x 768 resolution 19" monitor was used. We investigated interleaving on a per-shot (per-image) basis, but pilot users expressed frustration at the unevenness of text distribution across shots (some shots may have a lot of dialogue and others none at all), and the fragmentation of the transcript into tens of small text pieces. This is in agreement with earlier work on video skims that found piecing together small bits of dialogue was not as effective as piecing together longer, phrase-based pieces [2]. Interleaving is hence done on a storyboard row basis rather than an image-by-image basis.

We followed the same contextual strategy used in prior video skim work to reduce the text display requirements down to one line per image row, as shown in Figure 2. When the storyboard is for a clip returned from a query, use the query terms to pinpoint words of interest in the transcript. Include those matching words as part of the image row text. If space allows, expand each word to its enclosing phrase and put the phrase into the image row text, in agreement with a past video skim result that increased grain size (favoring phrases over words) produced more effective surrogates [2].

If space still exists per row, pick the most important remaining word from the row's full transcript text, expand it to its enclosing phrase if possible, and add to the row text.

Figure 2. Storyboard with interleaved text, compressed to fit as a single line caption for each image row

Importance is judged via a term frequency-inverse document frequency measure (TF-IDF). The TF-IDF of a word is its frequency in a clip divided by its frequency in a corpus (~1000 hours of CNN broadcasts). A high TF-IDF indicates a word that marks a clip well by appearing often within it but rarely in the rest of the corpus.

The space-filling algorithm for collecting a single line of text per row requires an automated way to determine phrase boundaries. We break phrase boundaries based on punctuation found in the closed-captioned source, prepositions, and conjunctions. We make use of CMU's Link Grammar Parser to find prepositional phrases [11].

On each line, the text is ordered based on its video alignment, in agreement with image ordering that is also based on the video time. Each text line is a sequence of time-ordered phrases. Phrases that naturally follow one another in the full transcript are kept as is; otherwise, a semi-colon and space are added as punctuation between the phrases to indicate that some dialogue text was dropped. The process is automatic, and so some errors, including those introduced by closed-captioning and upper/lower case conversion, remain in the storyboard.

The research reported here directly addresses the role of text synchronization and reduction within storyboards, specifically:

1. Does the presence of transcript text displayed with the storyboard affect navigation performance?
2. Does the alignment of transcript text with storyboard image rows matter?
3. Does the reduction of transcript text based on query context matter?

3. EXPERIMENT

Twenty-five participants (14 male, 11 female) were recruited for this study from the Pittsburgh community via electronic bulletin boards and paper flyers on university bulletin boards. Participants made use of "CNN World View" and "CNN World Today" news video broadcast

from October through December 1999. This video was automatically processed by Informedia technology to break down the broadcasts into clips and to create storyboards for every shot for every clip [12]. 4246 clips were identified, with an average of 30 shots, and hence 30 thumbnails in the storyboard, per clip.

We measured the navigation task by presenting the participants with a graphic menu into the top 12 clips returned by the Informedia search engine for a particular question. The participants' goal was to play the section of the clip (from the set of 12) that answers the question as quickly as possible. No other cues were presented in the menu, i.e., no titles other than the generic numeric identifier, to better assess the storyboard's utility. Participants did not have to issue queries nor examine result sets of more or less than 12 results, because we wanted to assess the value of different storyboards, not the participants' skill in authoring queries.

To avoid biasing the experiment to favor imagery over text or vice versa through our choice of questions, we made use of an outside information source representative of news query systems: the Learning Resources site [5]. This site provided questions tied to particular CNN clips, and we picked the first 6 clips that matched our data (Oct.-Dec. 1999), and the first question for each clip (from the 5 or so on the web site) which produced at least 12 potential matches (to create the set of 12), but could only be answered by a single clip. Thus, the task under consideration was known item fact retrieval.

A within subjects multiple Latin Squares design was used [8], with 5 treatments:

- NoText: thumbnail images only, with no text
- AllByRow: shown in Figure 1
- All: same text as shown in Figure 1, but presented in a single text block beneath the storyboard imagery
- BriefByRow: shown in Figure 2
- Brief: same text as in Figure 2 but presented in a single text block of 4 lines beneath the imagery

The questions were always presented in the same order. Each participant experienced all 5 treatments, with order counterbalanced by the Latin Square design [8]. Participants began with a set of on-line instructions, and then received untimed practice with the interface. For the subsequent 5 questions, their performance was timed, with the running timer visible and a cash incentive provided for quick yet accurate retrieval.

4. RESULTS AND DISCUSSION

As expected, most participants found the video where the answer to the given question was revealed: 115 of 125 questions were answered correctly. We observed a statistically significant difference between the task completion times for the storyboard treatments, $F(4, 76) = 2.78$, $p < .05$. Consideration of Figure 3 indicates that the fastest navigation occurred with BriefByRow, and slowest with NoText. The standard error of a treatment mean here

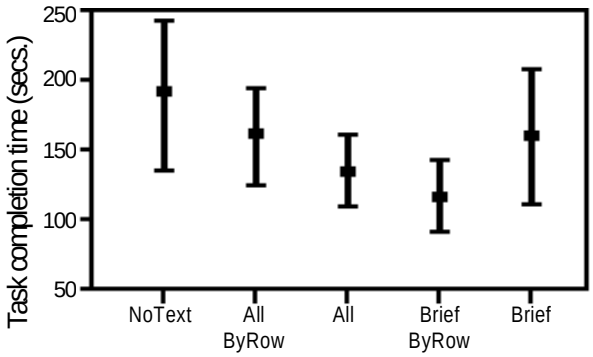

Figure 3. Task times with 95% confidence intervals

is 16.98 seconds. Based on obtaining a significant difference, we investigated the treatments further, using orthogonal contrasts between the treatments. One such contrast is that NoText has no text but the other treatments all do. Other contrasts include brief vs. all text, interleaved or not, and storyboards that follow traditional print media heuristics (short text captions beneath figures; paragraphs of text kept together as a block) versus those that do not.

This deeper analysis revealed that storyboards with text produced significantly better task times than storyboards without text, $F(1, 76) = 6.42$, $p < .05$. In addition, storyboards that follow traditional print media heuristics (BriefByRow, All) produced significantly better task times compared to storyboards that have more novel presentations (AllByRow, Brief), $F(1, 76) = 4.00$, $p < .05$. No significant difference was found between storyboards presenting text in a block vs. interleaving it within the imagery. Likewise, no significant difference was found between storyboards that collapsed the text to a line per row vs. those that kept all of it.

Table 1. Mean ranking for treatments (1 = favorite, 5 = least favorite), by participants at end of experiment

NoText	AllByRow	All	BriefBy Row	Brief
4.60	1.12	3.04	2.52	3.72

There was overwhelming agreement between participants in ranking the treatments, e.g., 24 of 25 rated AllByRow best, and 21 of 25 rated NoText worst. Using the chi-square goodness of fit test, the hypothesis that treatments would all end up with average ranks of 3, i.e., no user preference, is strongly rejected, $X^2 = 57.04$, $p < .00001$. Table 1 shows average ranking for treatments.

Storyboard surrogates clearly improved with the addition of text, in agreement with Ding's conclusions on the benefits of surrogates having both imagery and text [4]. Participants preferred the storyboards with text and achieved faster task times with them, and the text modality was rated very highly in terms of its importance in the storyboard. These conclusions are interesting in that the text was derived automatically from phrase partitioning, captioning, and capitalization processes that are imperfect,

and hence the quality of the output text was lower than the proofread text used in Ding's research. The experiment here shows that automatically produced storyboard text, even with its imperfections, results in improved video surrogates.

This study extends Ding's work by examining the questions of text layouts and lengths in storyboards. Participants favored interleaved presentation: the interleaved treatments were voted the top two storyboard schemes. If our analysis considered only the subjective ratings, then we would conclude participants want all the text interleaved (as in Figure 1) and therefore should be given it. However, the whole purpose of the surrogate is to enable more efficient activity on the video clip being represented, and the task portion of the experiment showed that navigation efficiency is best served with reduced interleaved text. Despite users' preference to have all the text available, they were able to accomplish the information finding task in less time with the interleaved reduced text format.

Surprisingly, the answers on the utility of interleaving text and reducing text for storyboards must be qualified. Interleaving text with the imagery is not universally better, as AllByRow had relatively poor task times. Reducing text is not universally better, as Brief had relatively poor task times and poor subjective ranking. Based on their majors and job title, the participants for this experiment have vast experience with printed textual materials like books and have likely learned how to efficiently skim through such text. That experience helps users when skimming the block text of the All treatment, but fails them when the text already has been reduced as in the Brief treatment. Similarly, the image rows may add interference to the text block skimming in the AllByRow treatment, slowing down task completion time.

Rather than just conclude that text in a block is best as it leverages from our text skimming abilities, the experiment did reveal that automatic collapsing of text can be just as efficient for navigation as long as interleaving is performed. If the reduced text is presented as a block, then users may expect to skim that block of text like typical, readable text with lots of redundancy, and become frustrated when that is not the case. Hence, the Brief treatment received the second worst subjective ranking. With interleaving, though, a space-efficient storyboard is produced, acceptable to users and with great utility as a surrogate. This result is highly relevant to the designs for portable information appliances where display space is a limited resource. The BriefByRow treatment produced the best task times, better than the All and AllByRow treatments, with less required display space.

If interleaving is done in conjunction with text reduction, to better preserve and represent the time association between lines of text, imagery and their affiliated video sequence, then a storyboard with great utility for information assessment and navigation can be constructed. Such combined transcript sequencing and compression could become part of the feature set offered by next generation digital video players. These conclusions are based on surrogates for a fact-finding task against a particular genre of video: news. Video preview features that work well for one genre may not be suitable for a different type of video. A recent study found that shot images were used most frequently and rated most useful for news, travel, and sports, with lowest ratings and least use for classroom lecture and conference presentations [6]; it did not address transcript text. Future work includes examining the utility of storyboards with text for open-ended browsing tasks, and with other genres where the visual content may not be as rich and varied.

5. ACKNOWLEDGMENTS

This material is based on work supported by the National Science Foundation under Cooperative Agreement No. IRI 9817496. Special thanks go to Melissa Keaton and Bryan Maher for their efforts in experiment preparations.

6. REFERENCES

1. J. Boreczky et al. "An Interactive Comic Book Presentation for Exploring Video," *Proc. CHI '00*, pp. 185-192, 2000.
2. M. Christel et al. "Evolving Video Skims into Useful Multimedia Abstractions," *Proc. CHI '98*, pp. 171-178, 1998.
3. R. Cox et al. "Applications of Multimedia Processing to Communications," *Proc. of IEEE*, pp. 754-824, 1998.
4. W. Ding et al. "Multimodal Surrogates for Video Browsing," *Proc. ACM Dig. Lib.*, pp. 85-93, 1999.
5. Learning Resources, http://www.literacynet.org/cnnsf.
6. F. Li et al. "Browsing Digital Video," *Proc. CHI '00*, pp. 169-176, 2000.
7. R. Lienhart et al. "Video Abstracting," *Comm. ACM, 40*, 12, pp. 54-62, 1997.
8. R. Petersen. *Design and Analysis of Experiments.* New York: Marcel Dekker, 1985.
9. D. Ponceleon et al. "Key to Effective Video Retrieval: Effective Cataloging and Browsing," *Proc. ACM Multimedia Conf.*, pp. 99-107, 1998
10. M. Smith, and T. Kanade, "Video Skimming and Characterization through the Combination of Image and Language Understanding," *Carnegie Mellon University CS Tech. Report CMU-CS-95-186R*, May 27, 1996.
11. D. Temperley, D. Sleator, and J. Lafferty, Carnegie Mellon University Link Grammar Parser 3.0 (August 1998), http://bobo.link.cs.cmu.edu/index.html/.
12. H. Wactlar et al. "Lessons Learned from the Creation and Deployment of a Terabyte Digital Video Library," *IEEE Computer, 32*, 2, pp. 66-73, 1999.
13. B.-L. Yeo and M.M. Yeung, "Retrieving and Visualizing Video," *Comm. ACM, 40*, pp. 43-52, 1997.
14. H.J. Zhang et al. "Automatic Parsing and Indexing of News Video," *Multimedia Sys., 2*, pp. 256-266, 1995.