

# Effects of Audio and Visual Surrogates for Making Sense of Digital Video

**Yaxiao Song**

University of North Carolina at Chapel Hill  
100 Manning Hall  
yaxiaos@email.unc.edu

**Gary Marchionini**

University of North Carolina at Chapel Hill  
100 Manning Hall  
march@ils.unc.edu

## ABSTRACT

Video surrogates are meant to help people quickly make sense of the content of a video before downloading or seeking more detailed information. In this paper we present the results of a study comparing the effectiveness of three different surrogates for objects in digital video libraries. Thirty-six people participated in a within subjects user study in which they did five tasks for each of three surrogate alternatives: visual alone (a storyboard), audio alone (spoken description), and combined visual and audio (a storyboard augmented with spoken description). The results show that combined surrogates are more effective, strongly preferred, and do not penalize efficiency. The results also demonstrate that spoken descriptions alone lead to better understanding of the video segments than do visual storyboards alone, although people like to have visual surrogates and use them to confirm interpretations and add context. Participants were able to easily use the combined surrogates even though they were not synchronized, suggesting that synchronization of different media channels may not be necessary in surrogates as it is in full video. The results suggest that multimodal surrogates should be incorporated into video retrieval user interfaces and audio surrogates should be used in small display interfaces. The study also raises questions about the need to synchronize different information channels in multimedia surrogates.

## Author Keywords

Video surrogates, dual coding, multimedia.

## ACM Classification Keywords

H5.1. Information interfaces and presentation (e.g., HCI): Multimedia information systems.

## INTRODUCTION

Large collections of digital video are increasingly

accessible to people using devices ranging from desktop computers to PDAs and cell phones. Additionally, the range of video has expanded to diverse genres and lengths including feature-length films, news clips, instructional videos, and ephemeral ‘viral’ videos. This large volume and range of video demands good search tools that allow people to browse and query easily and to quickly make sense of the videos behind the result sets.

Popular search engines (e.g., MSN, Yahoo, and Google) provide textual ‘snippets’ in result sets to represent pages that may be relevant to a query. These ‘snippets’ (e.g., title, excerpt of text, URL) are *surrogates* for the webpage and facilitate rapid sense making and relevance judgments as searchers decide whether to examine the full webpage. Surrogates are essential components of good user interfaces for all search systems. Surrogates for video are even more crucial for video collections because video has more kinds of features that determine meaning (meaning is carried in multiple visual and multiple audio channels), it requires large storage or bandwidth (encouraging more judicious relevance judgments before downloading or streaming), and the human visual processing system makes rapid browsing a natural component of search (images and non-verbal sounds do not require decoding and thus support fast scanning of long lists of results). The bulk of surrogates in today’s video retrieval systems are text-based, although surrogates such as poster frames (single keyframe), storyboards, and fast forwards are increasingly available (e.g., Internet Archive, Open Video). Given the improvements in wireless broadband and the small displays of cell phones and PDAs, it is likely that audio surrogates will be added to video retrieval systems. A major goal of our ongoing research agenda is to create and evaluate effective video surrogates using different modalities and embed them in user interfaces for video digital libraries.

This paper reports results from a user study that compares the efficacy of three different video surrogates, one purely audio, one purely visual, and one combined audio and visual. We first review the role of surrogates in video retrieval research and development, then describe the user study design, present the results, and discuss the implications for theory and design.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

CHI 2007, April 28–May 3, 2007, San Jose, California, USA.  
Copyright 2007 ACM 978-1-59593-593-9/07/0004...\$5.00.

## VIDEO RETRIEVAL AND SURROGATES

In the information science literature, a surrogate is a condensed representation constructed to stand for a complete information object. The retrieval literature demonstrates how people make sense from the surrogates, which stand for the object so that the seeker can decide whether to retrieve it or not. During retrieval, surrogates enable decision-making by presenting search results in a uniform way, by supporting incidental learning and saving network capacity [5]. Regardless of the kinds of surrogates a system provides to users, surrogates can be applied in distinct and creative ways in practice. A good surrogate will have a primary use and many secondary uses. For example, surrogates that are primarily for finding may also be pathways to understanding for experienced users.

Although the literature on how people use surrogates and how indexers create them is quite rich for text resources, we are only beginning to understand what kinds of surrogates might be useful for videos. A number of researchers have investigated the kinds of surrogates people found useful for finding video and conducted empirical studies to determine their effectiveness [2, 3, 5, 8, 15, 19, 20, 22]. There are a variety of non-textual surrogates that may be used for video retrieval including the following:

- Poster frame: an image selected to represent the video, usually a single frame extracted from the video.
- Storyboard: a set of keyframes displayed in chronological order, usually in a tabular format.
- Slideshow: a series of keyframes presented for viewing one at a time for a few seconds each, i.e. as if viewing a slideshow.
- Collage: a dynamically-created, interactive image constructed of text and images from multiple videos, perhaps at different display sizes.
- Fast-forward: most simply created by selecting every Nth frame and displaying the frames at normal speed (30fps).
- Skim: a video clip 'abstract' created by compacting visual and audio information while preserving the original frame rate.
- Trailer: a pre-produced series of clips excerpted from a video.
- Spoken keywords or descriptions.

Most of these surrogates are visual, although skims and trailers use both visual and audio modalities, and the spoken surrogates are strictly audio. This is mainly due to the fact that most of the video retrieval research and development efforts over the past decade focused on visual features (e.g., luminosity, color, texture, shapes) of video and the metadata and resulting surrogates were constructed to take advantage of these features. Five years of TREC Video results readily demonstrate the importance of linguistic data (in text format) for retrieval; 2005 was the first year that some groups showed better performance with visual

features than linguistic features [16]. However those results were probably due to very difficult linguistic conditions (multiple languages with automated translations). Nonetheless, several of the studies noted above have demonstrated that people like to have visual surrogates regardless of their performance effects. Given that visual and audio data as entry points for retrieval are increasingly practical with better broadband access, audio and video surrogates will surely become important components of video retrieval and sense-making.

Different surrogates have particular advantages and disadvantages and the unique advantages of different surrogates can be selectively applied in video retrieval systems. For visual surrogates these pros and cons are fairly well studied (e.g., [1, 3, 12]), but there is little work that examines audio surrogates even though they engage the user's natural ability to hear, require no training for sense-making, and may be alternatives to limited displays on small form factor devices. Furthermore, audio surrogates may be combined with visual surrogates to leverage multiple sensory channels.

Pavio's [17] dual coding theory has been used by instructional designers as a theoretical basis for multimedia materials that putatively accommodate multiple learning styles and improve retention through redundant coding of key concepts. For example, Mayer [13, 14] offers strong evidence for a 'modality effect'--accompanying narrated text with graphics rather than on-screen text. 'Audio accompaniment has also been found effective in virtual navigation. For example, Gunther et al. [9] note that "...auditory cues complement visual cues...by providing information redundancy."

Given the potential for visuals and spoken audio to aid in understanding primary information, it is reasonable to expect these effects to obtain in surrogates as well. The Informedia project created video 'skims' that were meant to be the most salient extracts from videos [2]. The skims were found to be effective but are difficult to create (several different feature sets were used to automatically extract the best short sequences from the video). The Open Video Project has demonstrated the usefulness of different visual surrogates (storyboards, fast forwards [20, 21]). Results from these two projects suggest that combining multiple surrogates of different modalities could be effective.

In this study, we examined the effectiveness for sense making of combining an audio narration of the video description together with a storyboard of keyframes. We were also curious about what the respective audio and visual elements add to sense making. To investigate these effects, we compared storyboards alone, audio narrations of descriptions (spoken descriptions), and the combination of storyboards and audio narrations of descriptions on an array of sense making tasks.

## METHOD

To investigate the effects of visual and audio surrogates for sense making, we conducted a within subjects experiment with 36 participants. Using a sample of comparable videos, three kinds of surrogates were created and participants completed five trials with each surrogate (one training and four test trials). The trials consisted of experiencing (viewing/listening) a surrogate, completing five kinds of sense making (gist) tasks, and providing self-report feedback during and after the trials. Repeated measures analyses of variance were conducted to evaluate the relationships among surrogate types and performance, confidence, time, and a suite of affective measures. Qualitative comments were also used to enrich the interpretation of results.

## Participants

Thirty-six participants were recruited by sending campus-wide mass mail invitation to participate. Participants were included in the study if they were native English speakers, used computers daily, watched videos at least on a monthly basis, and searched for videos at least occasionally. The recruited participants included 16 men and 20 women ranging in age from 19 to 47 and came from 24 different academic departments. Eleven of them were undergraduate students, 16 were graduate students, and 9 were university staff.

## Videos

The video segments used in this study were selected from the repository of the Open Video Project ([www.open-video.org](http://www.open-video.org)), which is a publicly accessible digital library. At the time of the study, the collection contained about 4000 video segments with textual metadata (e.g., titles, keywords, 1-2 sentence descriptions) and three different types of visual surrogates (storyboards, fast forwards, 7-sec excerpts) but no audio surrogates.

The fifteen video segments were selected from a set of NASA-produced videos in the digital library. The metadata, including the descriptions of those videos, were manually created by the Open Video Project staff. Three of the 15 segments were used for training purposes and were selected from the NASA Why Files series and the NASA Kids Science News series. To insure comparable videos as the basis for surrogation, the twelve video segments used for testing purposes were all selected from the NASA Connect series, which has a common structural format and conceptual level for different middle school science topics. The video segments used were (the ones with asterisks were training videos):

- NASA KSN – CDs (\*);
- NASA WF - Food Web (\*);
- NASA WF - Circuit Activity (\*);
- NASA Connect - Rocket Computer Simulation;
- NASA Connect - Rover and Experimental Robots;
- NASA Connect - Elliptical Orbit Activity;
- NASA Connect - Hurricanes and Computer Simulation;
- NASA Connect - Exercise and Nutrition;

- NASA Connect - First Flights;
- NASA Connect - Solar Flares;
- NASA Connect - Muscles;
- NASA Connect - Early Mariner Navigation;
- NASA Connect - Climate and Weather;
- NASA Connect - Science of Sound;
- NASA Connect - X-33 Scale Model Activity.

## Surrogates

We created three surrogates for each of the fifteen video segments: a storyboard, an audio (spoken) description, and a combination of these surrogates. The visual only surrogate was a storyboard (sometimes called a “filmstrip”), consisted of keyframes extracted from the video segment and displayed in chronological order in a tabular format. The storyboards we created in this study each consisted of 6 frames and were laid out in 1x6 grids.

We created an audio only surrogate--a spoken description--by simply recording the existing written descriptions taken from Open Video. Those descriptions were manually created by humans. Although automatic techniques to extract metadata and descriptions are proving to be effective (e.g., IBM MAGIC [11]) and will make implementing spoken surrogates more efficient and cost effective, we aimed to leverage the high-quality manually generated descriptions in this study. To ensure standardization of pace and pronunciation, a text-to-speech synthesizer<sup>1</sup> was used to generate the audio recording of the spoken descriptions. In the study, each audio recording was played at least once and participants could replay them.

In the combined surrogate, the storyboard is displayed and the spoken description is initiated upon display. Similar to the audio only surrogate, the audio portion of the combined surrogate could be replayed. In other studies that examined video surrogates (e.g., [20]), viewing time for the surrogates was limited, for example, allowing 500 milliseconds per key frame. It is possible that those users felt that the time they were allowed to spend on the surrogates was not long enough for them to make sense of the surrogates. In this study, we set no time limit for the participants to experience the surrogates.

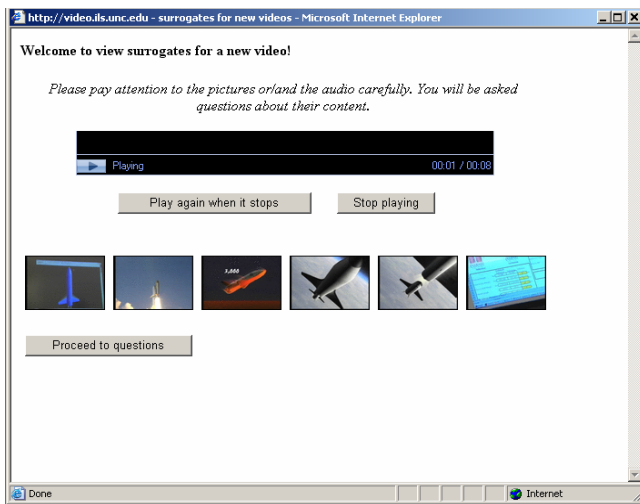
Figure 1 shows the online stimulus for the combined surrogate interface for one of the NASA Connect Videos included in this study. The visual surrogate stimulus showed exactly the same screen except the audio player and control buttons were not shown. Likewise, the audio surrogate condition showed only the audio player and not the storyboard.

## Tasks and Measures

Task development is crucial to user studies and there is little experience with sense-making tasks for video retrieval [6]. Sense making requires making inferences about

---

<sup>1</sup> AT&T Lab Text-to-Speech Online Demo with “Crystal” voice (<http://www.research.att.com/~ttsweb/tts/demo.php>)



**Figure 1. Screen shot for Combined Surrogate**

meaning based on evidence at hand<sup>2</sup>. We adopted a set of tasks related to gist that have been used in several studies emanating from the Open Video group [20, 21, 22], and added a “title selection task” for this study. These tasks require people to recall or recognize linguistic gist and determine visual gist. The five types of tasks used in the study were:

*Written Gist Determination (open-ended question):* Participants write a summary of the video from the surrogate. The summaries were scored by two researchers independently on a three point scale (0 is wrong, 1 is partially correct, and 2 is correct). The correlation between the respective scores was 0.76, so the two sets of scores were averaged for each trial to yield final scores in the 0-2 range.

*Keyword Recognition (multiple choices):* Participants select keywords that are appropriate for the video from a set of 8 or 9 words. Some of the keywords came from the keyword field for the video used by the Open Video Project (i.e. were correct), and others were selected from keywords for other videos in the Open Video repository (i.e., were wrong). Some of the words were more concrete (e.g., “aircraft” for the First Flight video) and some were more abstract (e.g., “visualization” for the Hurricanes and Computer Simulation video). To avoid setting a pattern of right/wrong (e.g., half of each on every trial), we slightly varied the number of keywords (8-9) and the number of correct keywords (3-5), however the total number of keywords participants saw using each surrogate condition was the same. The number of keywords correctly identified as correct or wrong across each set of four trials was normalized to the 0-1 range for comparison across trials.

<sup>2</sup> Russell et al. [18] provide a model for sense making as extracting information from primary documents retrieved; we focus here on sense making based on the surrogates.

*Title Selection (single choice from four alternatives):* Participants select the most appropriate title for the video segment that the surrogate represents. The correct title was the title of the video segment used in the Open Video repository, and the wrong ones were selected from titles for other videos in the same video collection. This task was scored as correct or incorrect, thus the score was either 0 or 1 for each trial and the sum taken across the four trials and then divided by four to yield a normalized score between 0 and 1.

*Keyframe Recognition (multiple choices):* Participants select from a set of 6-12 keyframes those that are appropriate for (or they think that come from) the video. Some of the key frames were selected from the storyboard of the video segment in the Open Video repository (i.e., were correct), and other were selected from storyboards of other videos in the same collection (i.e., were wrong). As with the keyword recognition task, the number of keyframes and correct keyframes were varied slightly across trials and the total number of keyframes they saw in each surrogate condition remained the same. The number of key frames correctly identified was normalized to the 0-1 range for comparison across surrogate conditions.

*Gist Recognition (single choice from four alternatives):* Participants select from a set of four descriptions the most appropriate one for the video segment that the surrogate represents. This was scored as correct or incorrect with the correct response the 1-2 sentence description from the video in the Open Video repository, and the distractor responses taken from descriptions for other videos in that series. The task was scored as correct or incorrect, thus the score is either 0 or 1 for each trial and the sum taken across the four trials and then divided by four to yield a normalized score between 0 and 1.

For each of the tasks, the accuracy of each trial was recorded as participants worked and the mean score for the four trials was used as the unit of analysis for making comparisons. Additionally, the time those participants spent experiencing the surrogates and the time they used to complete each task were also recorded. We were curious about whether confidence in answers to the tasks were related to the surrogates and thus required a confidence rating for each task trial. Figures 2 and 3 show the online stimuli for the keyframe recognition and gist recognition tasks respectively for the same NASA Connect video mentioned previously.

In addition to performance and confidence measures, we asked participants to rate their affective responses to each surrogate. After completing all the trials in a surrogate condition, participants completed a questionnaire that included thirteen 5-point Likert scaled statements related to usefulness (e.g., Using this system helps me better estimate the gist of videos) and usability (e.g., I found this system easy to use). This instrument was adapted from the scales in Davis [4]. These responses were combined into a usefulness

score and a usability score for each participant- surrogate combination.



Figure 2. Task 4: Keyframe Recognition

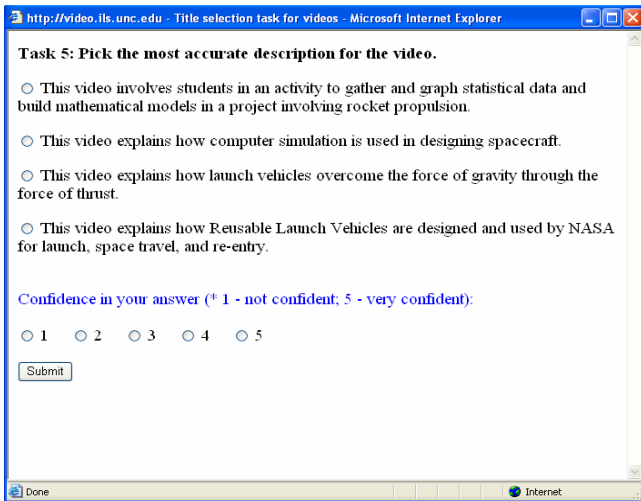


Figure 3. Task 5: Video Gist Recognition

Two 7-point semantic differential scales adapted from Ghani et al. [7] with five word pairs that focused on engagement and enjoyment (e.g. Using the video surrogates is not interesting / interesting; How you felt using the video surrogates: attention was not focused / attention was focused) were also completed and responses combined into engagement and enjoyment scores for analysis. Upon completion of all three surrogate conditions, participants completed a final questionnaire that asked them to compare the three different surrogates using open-ended response forms.

### Procedure

The 36 participants were scheduled over a two week period and were run in groups of 4-10 depending on participant availability. Participants were seated at alternating workstations in a laboratory with 30 identical Dell

workstations with 17" flat-panel displays so that they could not see other screens. Headphones were also provided. Each test session ran for up to 1.5 hours. An online protocol system was developed to administer the study and collect data. This system also managed the counterbalancing scheme across the surrogate condition orderings.

Upon arrival, participants read and signed the consent form and then were briefed as a group on the general purpose of the study by the investigator(s) and given a short introduction to the three surrogate conditions and the types of tasks they would perform in the study. We made it clear to the participants that we were not trying to evaluate their abilities in completing the tasks, but wanted to learn how different types of video surrogates help people understand the content of videos. We noted that it was not a race to complete the tasks quickly and encouraged participants to work at their own pace. Once these preliminaries were completed, participants were instructed to wear headphones for the remainder of the experimental session and all subsequent activity was individualized and controlled by the online protocol system. Participants first filled out a pre-session demographic questionnaire about themselves and their experiences with computers and video usage. The system then initiated the first counterbalanced surrogate condition assigned to each participant.

For each of the three conditions, participants experienced a surrogate and performed five tasks without going back to the surrogate. They completed five trials for the surrogate, with the first trial serving as a practice trial and not included in the data analysis. For each task, the time to view the surrogate and the time to complete the task were recorded by the online protocol system as were task responses and confidence ratings. For all trials, the same task order was used: written gist determination, keyword recognition, title selection, keyframe recognition, and gist recognition. The order of the tasks was important. For example, the open-ended gist writing task was completed first so that participants did not gain extra information from the other tasks (especially the gist recognition task that included actual descriptions). For similar reasons, the "back" button in the browser was disabled. Once they chose to go to the next page, they were no longer able to get back to the previous page. It is important to note that actual video segments were not played at any time during the study so that participants had to make sense of the videos merely by studying the surrogates. In the surrogates with audio (i.e., audio surrogates and combined surrogates), participants had the option to stop the audio or play it again, and the numbers of times the participants replayed and stopped the audio descriptions were recorded as well.

When participants had finished all five trials for whichever surrogate condition they were assigned first, they completed the usefulness, usability, engagement, and enjoyment questionnaires. They then were presented with a new set of five trials using the second surrogate type and repeated all

these same steps. They then repeated the process for the third type of surrogate.

After the participants had worked with all three surrogate conditions, they filled out a final questionnaire that focused on comparing their experiences with the three surrogate conditions. Upon completing the final questionnaire, participants were thanked and given \$20 for their participation.

### Research Questions

Although dual coding theory explains why people perform better on understanding multimedia content that carefully coordinates the different information channels, it is not obvious that this obtains for surrogates that are highly abbreviated and may not be temporally coordinated. Thus, our main research hypothesis for this study was that the combined surrogate would outperform the individual surrogates on all measures and not take significantly more time to use overall. The measures of interest were performance on the tasks, time to experience the surrogate and time to complete the tasks, confidence in task accuracy, ratings of usability, and ratings of engagement and enjoyment.

A secondary research question related to the relative effectiveness of audio and visual surrogates for different types of tasks. The video retrieval community has repeatedly found that the main semantic gist is carried in the words associated with a video, however, people like to have visual surrogates and claim that they offer added value [16, 22]. Thus, we hypothesized that the audio surrogate would outperform the visual surrogate on linguistic gist tasks, but that the visual surrogate would be at least equally preferred on the affective measures.

### RESULTS

To investigate the primary and secondary research questions, we compared the three surrogate types on measures of performance and confidence across the five types of tasks, measures of time to study the surrogate and time to complete the tasks, and measures of usefulness, usability, engagement, and enjoyment.

Table 1 presents the means and standard deviations for the five task measures for each of the three surrogates, the overall main effect *F* values and probabilities of statistical reliability. The four trial results for each task-surrogate pair were averaged and these mean scores were used as the unit of analysis for each participant. Thus, the means in the cells represent the means of these mean scores for all 36 participants. Repeated measure ANOVA (SPSS Release 13) was used to test the main effects across the three different surrogates. When the main effects were found to be statistically reliable at less than the 0.05 level, pair wise contrasts based on estimated marginal means were conducted and are reported in the text discussion for each task.

Table 1 shows the almost uniformly strong effects of using the combined surrogates on the different tasks. In all cases, except keyframe selection, the combined surrogates yielded the highest mean values and lowest variability. Audio surrogates alone were generally more helpful than the visual surrogates alone except in the case of visual gist keyframe selection, which was expected, and the case of keyword selection (recognition), which was surprising. Everyone did well on the title selection task and thus there was little variance to detect.

Strong effects of different surrogates were shown for the write gist task and the gist selection task (both at  $p < 0.001$  level). For the gist writing task, pairwise contrasts showed statistically reliable differences in the visual and combined (mean difference = .865,  $p < .001$ ) and in the visual and audio (mean difference = -.822,  $p < .001$ ), but no differences between the audio and combined (mean difference = -.043,  $p = .182$ ). For the gist selection task, pairwise contrasts likewise showed statistically reliable differences in the visual and combined (mean difference = -.118,  $p < .001$ ) and in the visual and audio (mean difference = -.095,  $p = .002$ ), but no differences between audio and combined (mean difference = -.023,  $p = .169$ ).

The keyword recognition result was surprising. Pairwise differences were found between the audio and combined surrogates (mean difference = -.039,  $p < .001$ ) but not between other pairings. People in this study were generally quite good at making inferences about the gist of the videos given rather impoverished surrogates of any kind. The number of choices and fine granularity of keywords made this task a measure of specific token recognition rather than a measure of gist. Additionally, the fact that the distractors were selected from keywords for related videos in this series made them all plausible. Alternatively, the synthetic spoken audio may have made it difficult to understand specific words but do well understanding the overall gist in the description because they could use surrounding words to establish context. Comments about audio quality made by participants in the open-ended questions lend some credibility to the latter explanation.

We expected that people would do better on the keyframe selection with the visual surrogate than the audio surrogate, but there was no difference. Although people did reasonably well on the keyframe selection tasks, they had the lowest overall means among all the tasks. There may have been a novelty effect here as people are not used to this type of task. Table 2 summarizes the results for participant reports of self confidence for each task-surrogate combination. After completing each task, participants checked a 5-point confidence scale and as with the other data, these values were averaged for the four trials for each task-surrogate combination. Overall, people were statistically reliably more confident about their sense

Surrogate	N	Write Gist		Keyword Recog.		Title Select		Keyframe Select		Gist Select	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Visual	36	1.06	0.37	0.89	0.06	0.95	0.10	0.83	0.06	0.88	0.16
Audio	36	1.88	0.23	0.86	0.07	0.98	0.08	0.82	0.05	0.97	0.09
Both	36	1.92	0.13	0.9	0.06	0.99	0.06	0.83	0.08	0.99	0.04
		F=171.8, p<.001		F=6.1, p=.006		F=1.80, p=.174		F=0.208, p=.813		F=12.90, p<.001	

**Table 1. Performance on Five Tasks by Surrogate**

Surrogate	N	Write Gist		Keyword Recog.		Title Select		Keyframe Select		Gist Select	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Visual	36	3.54	0.91	3.97	0.86	4.01	0.87	3.73	0.91	3.97	0.9
Audio	36	4.20	0.81	4.26	0.80	4.61	0.69	3.84	0.81	4.75	0.71
Both	36	4.21	0.82	4.32	0.83	4.60	0.80	3.89	0.85	4.71	0.76
		F=20.3, p<.001		F=6.35, p=.003		F=20.47, p<.001		F=1.15, p=.319		F=40.05, p<.001	

**Table 2. Confidence on Five Tasks by Surrogate**

making with the combined surrogates than with the visual surrogates. In the main, participants were generally confident about their performance, with the keyframe selection task yielding the lowest confidence levels overall, which reinforces the novelty interpretation for the accuracy results.

One of the goals of creating surrogates for videos is to allow users to review the surrogates and make sense of the video or make decisions about their relevance quickly [10]. The ratio of time to view the full video to the time to view a surrogate is termed the ‘compaction rate’ [2, 20]. High compaction rates are generally desirable as long as people are able to make sense of the full object. Static surrogates such as storyboards give people control over how long they experience the surrogate and thus compaction rates are under user control. Dynamic surrogates minimize user control by offering replay options at best. For example, Informedia skims offer compaction rates of 10:1, Open Video fast forwards offer compaction rates of 64:1, and Internet Archive slide shows offer compaction rate of 30:1 and all these allow people to replay the surrogate. Wildemuth et al. [21] suggested that although people were able to determine the gist of the video segment at high compaction rates (e.g. 500 milliseconds per key frame for storyboards or 64:1 for fast forwards), they often had a desire for spending more time with the surrogate. More generally, usability guidelines suggest that people want to have control over their user interfaces. For spoken audio surrogates it is possible to use compressed speech to speed up the surrogate a bit, however, the surrogates we created did not use compressed speech and the compaction rate depends on the actual time to read the descriptions. The NASA video segments used in this study were in the 1.5-5 minute range (average run length was 4 minutes 8 seconds) and the spoken descriptions we created averaged 10.7 seconds (range 4-16 sec.), thus yielding an idealized

average compaction rate of about 23:1, although as we discuss below, actual rates were lower due to replays.

Table 3 shows the mean time the participants spent experiencing each surrogate and the mean time per trial that they spent completing the five tasks using each surrogate. The means in each cell thus represent the mean time per trial for each of the conditions. People spent statistically reliably less time experiencing visual surrogates (mean = 19.47 sec) than audio surrogates (mean = 27.22 sec) and combined surrogates (mean = 28.81 sec). Thus, people spent almost 50% more time to ‘consume’ the audio or combined surrogates. They spent less time on audio surrogates than on the combined surrogates, but the differences were not statistically significant. A considerable amount of this difference is explained by the number of times participants replayed the surrogates in the audio and combined conditions. Twenty-three of 36 participants replayed the audio surrogates at least once, with one person replaying 16 times in the four trials. Twenty-five of the 36 replayed the combined surrogate at least once with a different participant replaying 16 times in the four trials. The mean numbers of replays for the audio and combined conditions were 3.64 and 3.28, respectively, which a paired sample t-test showed were statistically reliably different. This effectively reduces the compaction time for these surrogates by a factor of almost four. It was also interesting that once people started to replay, they allowed the audio to complete; even though they could stop play (stopping play only occurred 15 times in the more than 500 audio plays). Thus, when people have control over surrogates they may take much more time to use them and the audio alone condition motivated more of this behavior.

Although the text-to-speech synthesizer we used to create the audio is one of the best ones that are freely available online, many participants complained that the spoken

Surrogate	N	Mean Time (SD) to Experience Surrogate (sec)	Mean Time (SD) to Complete Task (sec)
Visual	36	19.47 (8.55)	90.75 (37.72)
Audio	36	27.22 (15.28)	87.93 (27.39)
Both	36	28.81 (15.68)	88.09 (31.94)
		F=7.27, p=.001	F=0.113, p=.894

**Table 3. Time to Experience Surrogate and to Complete Task, by Surrogate**

descriptions sounded too mechanical, and they had to play the audios again to recognize some of the words. For example, for one audio description, several people misheard “How ear works” as “How air works”. As the quality of text-to-speech synthesizers gets better and better, we expect the differences between idealized compaction rates and actual rates may move closer together.

These very small time differences between audio only and combined conditions (less than half a second on average) suggest that people are able to quickly integrate unsynchronized audio and visual information. Apparently, people are able to manually integrate these distinct surrogates like they do with coordinated channels of primary information

Time to complete tasks was not dependent on surrogate conditions. Participants completed the tasks most quickly using audio surrogates (mean = 87.93), were a little bit slower using combined surrogate (mean = 88.09), and were slowest using visual surrogates (mean = 90.75), but these differences were not statistically reliable. Further analysis of time differences across the five tasks showed that the times to complete the title and gist selection tasks were statistically reliably different with pairwise contrasts showing that the participants in the visual conditions took longer to complete these tasks than when they used the other two surrogates. Together with the better overall performance that the combined surrogates yield, these results demonstrate that time is not a penalty for the added value that having both kinds of surrogate offers.

Data from the questionnaires strongly reinforce the efficacy of the combined surrogates and the general effectiveness of the audio surrogates. Table 4 summarizes the results for participant reports of subjective measures for each surrogate. After completing tasks for each surrogate, participants rated their experience with the surrogate. Usefulness and usability were rated on a 5-point Likert scale, while engagement and enjoyment were rated on 7-point scales. As with the other data, these values for each were averaged for all 35 participants who provided usable responses.

Combined surrogates consistently lead to the highest positive levels on all four of the affective measures, while visual surrogates consistently yielded the lowest levels.

Pairwise contrasts were statistically reliably different for visual and combined (mean differences and p values were -1.16 and  $p < .001$ , -.65 and  $p < .001$ , -.64 and  $p = .012$ , -1.11 and  $p < .001$  respectively for usefulness, usability, engagement, and enjoyment). Pairwise contrasts favored combined over audio at statistically reliable levels for usefulness (mean difference -.37,  $p = .05$ ), and enjoyment (mean difference -.81,  $p < .001$ ). Pairwise contrasts favored audio over visual at statistically reliable levels for usefulness (mean difference -.80,  $p = .003$ ) usability (mean difference -.58,  $p = .002$ ) and engagement (mean difference -.64,  $p = .012$ ). Clearly, the combined surrogate offers much better affective advantage. The expectation that the visual surrogates would be highly engaging was not born out, perhaps because the audio surrogates were also engaging and somewhat novel, and/or the storyboards are less engaging than other kinds of visual surrogates such as fast forwards or skims.

## DISCUSSION

In spite of a few anomalies on the performance measures, the quantitative data clearly demonstrate that combined visual-audio surrogates are effective, are strongly preferred, and do not penalize efficiency. These results were strongly reinforced by the open-ended comments of participants and suggest that people are able to integrate two distinct sets of surrogates that use different sensory channels but are not temporally coordinated. When asked which of the three surrogates they liked most, 31 of the 36 participants selected the combined storyboard, with 3 and 2 selecting the audio and visual respectively. One participant noted: “Together you have two things to compare, so you can confirm your thoughts on the topic of the video. With each one separately, it is possible to figure it out, but it is much harder.” Another said: “With the two together, the surrogate is more efficient, and understanding the surrogates becomes simpler than when they are apart.”

The results also reinforce the power of words in carrying semantic information in video. There are many caveats of course: these videos are educational documentaries with high semantic information content; the tasks were mainly concerned with linguistic gist; and people are quite familiar with summarizing media of all types with words. Nonetheless, there were many comments about the effectiveness of spoken audio. Several pointed out the directness and conciseness of the spoken description, for example, one participant noted, “The audio takes the guess work out of the content of the video,” and another said, “It gives you a concrete outline of what is going on in the video. It focuses your attention so that your mind won’t wander.” Several participants made comments about the continuity of the audio surrogate. One noted: “Even though a picture is worth a thousand words, a few selected pictures cannot explain the deeper meaning of the subject--audio connected the dots,” and another said “[audio] adds a description so the user doesn’t have to put together a puzzle from the pictures.”

Surrogate	N	Usefulness		Usability		Engagement		Enjoyment	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
Visual	35	2.79	1.09	3.43	0.82	4.16*	1.29	3.61*	1.52
Audio	35	3.59	1.13	4.00	0.80	4.56	1.24	3.90	1.47
Both	35	3.96	0.95	4.08	0.68	4.80	1.19	4.71	1.31
		F=20.91, p<.001		F=16.38, p<.001		F=4.02, p=.022		F=13.11, p<.001	

**Table 4. Subjective Measures, by Surrogate**

These results do not imply that visual surrogates do not add value to sense making. The positive aspects of the visual surrogates mentioned by participants were varied. Several said that the visual surrogates added context to help in sense making. One said “They anchor the content. It was almost as if the brain needs something to attach to while an outside stimulus such as audio content is being fed to it.” Other participants discussed the feel of the video and one also pointed out that the visual elements will help the viewer understand appropriate audiences: “I think the visual surrogates aided in my understanding the video a lot better than the audio. The visual showed segment clips from the actual video so you could get a better feel of the depth of the content of the video as well as what kind of audiences.” Another participant noted the motivational aspects of the visual surrogates: “The storyboard provides the 'catch' that might draw a person to learn what the video is about. The storyboard was fun and engaging. It didn't necessarily make it easier for me to understand the video, but it made me more curious to find out what the video was really about.” Another participant noted that images would aid memory: “Images that show striking depictions of a topic tend to linger in my mind a bit longer, helping me be more focused.” Thus, there are several advantages that visual surrogates bring to sense making, some of which augment semantic gist inferences and some add new kinds of visual gist information that aids in sense making.

Participants were asked about what they disliked most about any of the surrogates. About one third of the participants made some kind of comment about the audio quality (hard to understand, too mechanical, too fast or slow) and about one-third made some kind of comment about the visual quality (keyframes too small (86 by 59 pixels), not meaningful).

This study, as expected shows that using multiple channels generally leads to better performance (more information is better); however, it also raises the question of channel synchronicity. The literature shows superior effects when the channels are coordinated/synchronized and it is assumed that this will also hold for surrogates. The combined surrogates in this study were clearly not synchronized and were effective. Creating synchronized surrogates (e.g., excerpts) is costly and difficult and an important question raised by this study is whether the costs of synchronization are worth it for surrogates. In

fact, it may be more beneficial to sample from different channels and as long as it is clear to users that the channels are not coordinated, there may be more information content. One participant out of the 36 noted it was difficult to integrate the two kinds of information, saying “Not seeing the video play like usual makes the pictures distracting as you try to listen to the audio because you are trying to make sense of both seemingly unrelated things.” The remaining participants did not mention difficulty recognizing that the channels were not coordinated and some comments spoke to the ease of integration. One said: “They are infinitely more useful together. The audio tends to be a very broad description and having a storyboard accompany it gives a better sense of scale and detail of the contents of the video.” Another noted: “They work together and complement each other. The audio tells you straight up what you need to know but the visual with it allows you to process the information more and remember it much more easily.” The question of channel synchronization in surrogates bears future investigation.

These results have implications for the design of user interfaces for digital video retrieval systems. First, they strongly support using combined surrogates rather than single-channel surrogates. Although the compaction rates drop a little with the addition of audio, the semantic value seems well worth the effort. Using multimodal surrogates is especially recommended for full screen displays.

Second, the results support adding spoken descriptions as audio surrogates to video retrieval systems, alone if not in combination with visual surrogates. This seems especially appropriate for small display devices where screen real estate is minimal. Audio has great potential for mobile devices especially given limited screen estates, storage and bandwidth, and as earphones and earpieces (wireless or wired) allow private listening and control.

Third, designers must take care when creating the audio and visual surrogates to make them usable. Visual surrogates must be easily viewable without consuming the entire screen. Perhaps adjustable image size will help as long as the controls require low effort. Likewise, audio must be clearly articulated and people should be able to replay it easily. It may also be useful to give people control over the type of voice and the speed of play, although those parameters were not examined here.

## CONCLUSIONS

As people gain more experience with digital video, they will come to expect surrogates that are not only text-based. This study demonstrates the efficacy of spoken descriptions and especially the combination of those descriptions with visual surrogates. We will continue to investigate audio surrogate design parameters, including ways to combine spoken metadata with other kinds of visual surrogates such as fast forwards.

## ACKNOWLEDGMENTS

We thank the study volunteers for their participation and the anonymous reviewers for good suggestions. This work was partially supported by a grant from the NSF (IIS 0455970) and an IBM Faculty Research Award.

## REFERENCES

1. Christel, M et al. Collages as Dynamic Summaries for News Video. *Proc. Multimedia '02*, ACM (2002), 561-69.
2. Christel, M. Smith, C.R. Taylor, and D. Winkler, Evolving Video Skims into Useful Multimedia Abstractions. In *Proc. CHI '98*, ACM (1998), 171-178.
3. Christel, M. and Warmack, A. The Effect of Text in Storyboards for Video Navigation. In *Proc. IEEE Int'l Conf. Acoustics, Speech and Signal Processing* (2001), 1409-1412.
4. Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3), 319-340.
5. Ding W., Marchionini G., & Soergel, D. Multimodal Surrogates for Video Browsing. In *Proc. the Fourth ACM Conference on Digital Libraries* (1999), 85-93.
6. Fluhr, C., Moellic, P., & Hede, P. (in press). Usage-oriented multimedia information retrieval technological evaluation. *Multimedia Information Retrieval Workshop*. ACM Press (2006).
7. Ghani, J. A., Supnick, R., & Rooney, P. (1991). The experience of flow in computer-mediated and in face-to-face groups. *Proceedings of the Twelfth International Conference on Information Systems (December 16-18, 1991, New York)*, 229-237.
8. Goodrum, A. (1997). Evaluation of Text-Based and Image-Based Representations for Moving Image Documents. Unpublished doctoral dissertation, University of North Texas.
9. Gunther, R., Kazman, R, and MaccGregor, C. (2004) Using 3D sound as a navigational aid in virtual environments. *Behaviour and Information Technology*. 23(6), 435-446.
10. Li, F., Gupta, A., Sanocki, E., He, L., Rui, Y (2000). Browsing Digital Video. *Proc. CHI 2000*, ACM Press (2000), 169-176.
11. Li, Y., Dorai, C. and Farrell, R. Creating MAGIC: System for generating learning object metadata for instructional content. In *Prof. ACM Multimedia* (2005), 367-370.
12. Marchionini, G., Wildemuth, B., & Geisler, G. (2006). The Open Video Digital Library: A Mobius strip of theory and practice. *Journal of the American Society for Information Science and Technology*. 57(2), 1629-43.
13. Mayer, R. (2003). Elements of a science of E-learning. *Journal of Educational Computing Research*, 29(3), 297-313.
14. Mayer, R., & Moreno, R. (1998). A split-attention effect in multimedia learning: Evidence for dual processing systems in working memory. *Journal of Educational Psychology*, 90(2), 312-320.
15. O'Conner, B. (1985). Access to moving image documents: Background concepts and proposals for surrogates for film and video works. *Journal of Documentation*, 41(4), 209-220.
16. Over, P., Kraaij, W., & Smeaton, A. (2005). TRECVID 2005: An introduction. *Proc. TRECVID 2005* (Gaithersburg, MD), 1-14. [http://www.cdvp.dcu.ie/Papers/TRECVID2005\\_Overview.pdf](http://www.cdvp.dcu.ie/Papers/TRECVID2005_Overview.pdf).
17. Pavio, A. (1986). *Mental representations: A dual coding approach*. Oxford: Oxford U. Press.
18. Russell, D., Stefik, M., Pirolli, P., & Card. S. (1993). The cost structure of sensemaking. *Proc. the SIGCHI Conference on Human Factors in Computing Systems*. ACM Press, 269-276.
19. Turner, J. (1994). Determining the subject content of still and moving image documents for storage and retrieval: An experimental investigation. Unpublished doctoral dissertation, University of Toronto.
20. Wildemuth, B., Marchionini, G., Wilkens, T., Yang, M., Geisler, G., Fowler, B., Hughes, A., & Mu, X (2002). Alternative Surrogates for Video Objects in a Digital Library: Users Perspectives on Their Relative Usability. In *Proc. the 6th European Conference on Digital Libraries* (2002), 493-507.
21. Wildemuth, B., Marchionini, G., Yang, M., Geisler, G., Wilkens, T., Hughes, A., & Gruss, R. (2003). How fast is too fast? Evaluating fast forward surrogates for digital video. *Proc. JCDL* (2003), 221-230.
22. Yang, M. & Marchionini, G. (2005). Deciphering visual gist and its implications for video retrieval and interface design. *Conference on Human Factors in Computing Systems: CHI '05 extended abstracts on Human factors in computing systems*. NY: ACM Press, 1877-1880.