Multimodal Surrogates for Video Browsing

Wei Ding
College of Library & Info. Services
University of Maryland
College Park, MD 20742
1-202-623-7038
weid@glue.umd.edu

Gary Marchionini
School of Info. & Library Science
University of North Carolina
Chapel Hill, NC 27599-3360
1-919-966-3611
march@ils.unc.edu

Dagobert Soergel
College of Library & Info. Services
University of Maryland
College Park, MD 20742
1-301-405-2037
ds52@umail.umd.edu

ABSTRACT

Three types of video surrogates visual (keyframes), verbal (keywords/phrases), and visual and verbal were designed and studied in a qualitative investigation of user cognitive processes. The results favor the combined surrogates in which verbal information and images reinforce each other, lead to better comprehension, and may actually require less processing time. The results also highlight image features users found most helpful. These findings will inform the interface design and video representation for video retrieval and browsing

Keywords

Multimedia information systems, abstracting methods, multimodal video surrogates, browsing, cognitive processes

1. INTRODUCTION

Use of digital videos on the Web is becoming more common for education and academic communication as well as for entertainment. As video collections grow, users must be provided with effective searching and browsing tools to facilitate quick and easy access. This requires appropriate surrogates to represent original video documents**.**

Surrogates (such as bibliographic citations, abstracts, and tables of contents) are used in most information retrieval systems. Browsing surrogates allows users to make quick decisions about whether to examine an information object in greater detail (result examination) and supports incidental learning (users can capture the most interesting information without reading/viewing the full document or video, information extraction). Surrogates, especially video surrogates, also save network capacity: In many cases the transfer of the full document or video can be avoided.

Creation and evaluation of surrogates are long-standing research issues in information science (Borko & Bernier, 1975). The nature of video data and users' approaches to video bring new challenging issues related to the design and display of video surrogates. Because pictures, motion, speech, and other audio all communicate important information in video, purely verbal video surrogates are insufficient; image-based surrogates are also required to express information that cannot be expressed through words. Several formats for visual video surrogates have been suggested, such as keyframes (e.g., O'Connor, 1985; Zhang, et al., 1995) and salient stills (e.g., Teodosio & Bender, 1993). Key frames, salient stills, and other types of surrogates can be extracted or produced automatically through visual or audio signal processing (e.g., color, luminosity, optical flow, texture) to detect features that might be used to cue human recognition and recall, and techniques and prototype systems have attracted much attention (for example, Christel, 1997, Ponceleon et al., 1998). However, we desperately need guidelines for determining the representation most appropriate in a given situation; such guidelines must be based on an understanding of the underlying cognitive processes that users bring to the representations.

Words and images serve different functions. For example, Cawkell (1995) predicts that pictures will not always outperform words because they do not always replace the descriptive power of words in expressing abstract concepts. Commercial still and moving image documents generally use textual descriptors to support retrieval (Enser, 1995). When one video clip is represented by multiple still images, processing of image sequences by users will involve additional issues: While the meaning of words as signs are generally agreed on

and made more specific by syntax, pictures are specific and made general by their context (Pryluck, 1976). Therefore, interpretation of visual surrogates that consist of a sequence of isolated images could be even more ambiguous than interpretation of verbal surrogates. O'Connor (1991) noted that still images represent only one small fragment of the time continuum represented by the moving image document. Some images together with some words may well be adequate to guide a user among many videos on similar topics. Turner (1994) posits that for access to video information, text and images are complementary and interdependent. A series of studies conducted by the authors and colleagues (summarized in Tse et al, 1998) with video surrogates composed only of keyframes show that keyframes support rapid identification of visual objects and adequate but not good comprehension of the video gist.

According to the research on information objects involving more than one modality (e.g., full motion videos), reinforcing rather than interfering effects can be expected if the information is well integrated. Redundancy theory suggests that the redundant information from different modalities provides cross-references to the target to be understood (Pryluck, 1976), overcome the limitations of single information modalities (Morgan & Welton, 1992), and increase possibilities for comprehension (Stone & Gluck, 1980). Based on brain lateralization and parallel processing theories, Kantowitz (1985) hypothesized that redundant information simultaneously perceived through two modalities (images and words) actually speeds up processing time. However, little empirical study has been done to test whether these findings on primary information objects are applicable to surrogates.

The goals of this research were to better understand the roles of visual and verbal information in representing multimedia documents, to explore the cognitive processes involved in video surrogate examination, and to identify various decision patterns and impact factors behind the user behavior. It investigated the information representation power of different modalities in video data based on user performance (accuracy and response time) and user preference

The research was conducted in two parts: A quantitative, experimental study (for detailed results see Ding, 1999) and a qualitative study that probed more deeply into the results, to be reported here..

Both studies are built upon the Baltimore Learning Community project (BLC) (Marchionini et al, 1997), which allows public school teachers access a web-based multimedia database and link relevant instructional resources to their lesson plans and class presentations. This digital library includes texts, websites, still images, and segmented educational documentary videos from the Discovery Channel, Maryland Public Television, and the National Archives. The interface provides both visual (keyframes) and verbal (bibliographic information and abstracts) surrogates as video previews.

A brief description of the quantitative studies provides the background for the report on the qualitative study that is the focus of this paper. Twenty-three graduate students were recruited from the master's and doctoral programs at the College of Library and Information Services of the University of Maryland. Among the participants, there were 5 males and 18 females. The average age was 34.7, ranging from 21 to 53 years old. Participants were randomly assigned to each task. 11 participated in the comprehension task, the other 12 did the visual gisting task. Incomplete data from one participant for the comprehension task and two for the visual gisting task were dropped for data analyses. The surrogates tested were twelve keyframes in a storyboard (KF), six keywords (KW), and a combination (KF+KW).

It was hypothesized that the combined surrogates (KF+KW) would generate significantly better accuracy performance than either of the single-modal surrogates. It was expected that processing the combined surrogates might take longer, but in terms of the adjusted performance (the ratio of accuracy performance to processing time), the combined surrogates would still be significantly better than the other two surrogates. The results showed that although the KF+KW group consistently did better than the other surrogate groups, there was no significant difference between surrogate types in the accuracy performance based on the true/false judgment task that presented to participants six phrases/sentences (some true, some false) describing the video for which they just examined the surrogate and asked them to indicate true or false. In terms of the adjusted performance (ratio of accuracy to processing time), there was no significant difference found between the combined surrogates (KF+KW) and the keyframe (KF) surrogates, nor did it take significantly longer to process the combined surrogates than the keyframe surrogates. KW did significantly better in adjusted performance, a result of six keywords taking significantly less time to process.

When the accuracy was measured by the number of iconographical concepts identified in three free-form sentences participants wrote about each video, participants with the combined surrogates did significantly better than participants with the keyframe surrogates. The results show that participants using the combined surrogates, the keyword surrogates, or the full motion videos tended to use more iconographical concepts than preiconographical concepts, while the participants with the keyframe surrogates used more preiconographical concepts than iconographical concepts.

Consistently, user preference order was combined surrogates, keyframe surrogates, and keyword surrogates.

The results of the quantitative study were consistent with some expectations but not with others. A deeper understand-

ing was sought through a qualitative study designed to address the research questions detailed in the next section.

2. RESEARCH QUESTIONS

Based on the results of the quantitative comparison of surrogates, this study focused on people's cognitive processing while browsing video surrogates. It addressed a series of research questions organized around three constructs: the usefulness of the multimodal (combined) surrogates, the decision-making processes in surrogate examination, and the cognitive mechanisms in video comprehension.

Usefulness of the combined surrogates

What do images and words each contribute to the comprehension of the video surrogate? What time-accuracy tradeoffs were involved? Why did participants prefer combined surrogates with keyframes and keywords most of all and visual surrogates over verbal surrogates?

Cognitive processes in video surrogate examination

How does the cognitive process evolve during the interaction between users and video surrogates? What cognitive processes were involved in visual gisting and verbal comprehension tasks? How do users process the combined surrogate? How does verbal and visual information reinforce each other? How might understanding these processes inform the design of video representations?

Sense making and video comprehension

How did participants describe/make sense of videos? Did they treat all the available information equally or not? What kind of information (e.g., text in images, captions) captured more attention and is more important in video sense making? How did participants take advantages of those kinds of information?

3. RELATED WORK

Compared to the effort involved in the content-based video indexing and retrieval technology, there is a dearth of evaluation studies with users' participation. How to leverage the available video visualization techniques and accommodate user needs into effective interfaces for video browsing and access continues to be an important issue.

Despite the strong research interest in the information and the cognitive processes users use in making relevance judgments (e.g., Wang & Soergel, 1998), studies particularly focusing on video documents are rather sparse. Goodrum (1997) compared four different video surrogates (key frames, salient still, title and key words) for 12 10-second video clips (without sound) from the CNN environmental unit. Participants were to judge the similarity between different videos based on examining on type of surrogate at a time, and also to compare the utility of videos in answering an informational question. She found that with visual surrogates participants had a higher agreement in their similarity judgments than with text-based (title and key words) surrogates. For utility judgments with respect to specific queries, text-based surrogates resulted in higher agreement among participants than visual surrogates. The hypothesis that visual surrogates have advantages over text-based ones for generic queries was not supported. Christel et al. (1997) compared different video result presentations (poster frame menu vs. text title menu), and found that poster frames, when chosen based on the query, lead to significantly faster location of the relevant video (fact-finding) by a user over the presentation of only a text title menu.

An earlier study investigated the pattern of information uses for video selection in manual video libraries (Cohen, 1987). Those results highlight the importance of subject (topicality) information for video selection.

4. METHODOLOGY

This study took an exploratory approach as a follow-up to the quantitative study that showed that users preferred combined surrogates, although performance did not differ across image-only, text-only, and combined surrogates (Ding, in progress). Data were collected through observations and user's thinking aloud while performing recognition and comprehension tasks.

4.1 Materials

Fourteen 2-3 minute video clips were selected from a collection of 24 one-hour Discovery documentaries in the BLC database. For each video clip, three types of video surrogates were created.

Visual surrogates From the keyframes automatically extracted by a scene-change-based segmentation program (MERIT, Kobla et al., 1997), the twelve best keyframes were selected and organized as a story board.

Verbal surrogates Six keywords (words or phrases)were manually picked from the audio channel, or assigned based on the overall meaning of the video.

Combined surrogates Six keywords were listed at the top of the storyboard of 12 keyframes.

4.2 Participant tasks

We developed two user tasks verbal comprehension and visual gisting, which are needed in authentic information seeking activities for quick video browsing and result examination in a real video database. These tasks and techniques to measure them have been used in several previous studies and refined here (summarized in Tse et al, 1998).

4.2.1 Verbal Comprehension

Verbal comprehension is the extent to which the user can get the main idea of the video clip from the surrogates. Better comprehension would allow more accurate relevance judgments and video selections. There were two sub-tasks: free-form writing, and true/false judgments on phrases/sentences. For the writing task, participants were asked to write 2-3 sentences summarizing what the video clip was about. For the true/false judgment task, there were 6 summary statements /phrases for each video clip, some correct and some not (distractors), in random proportion and sequence. The statements were shown on the screen one at a time, with participants being asked to make judgments by clicking on the corresponding button. The testing statements were based on the transcripts, answers from the pilot participants, and the abstracts available in the Baltimore Learning Community (BLC) digital library.

4.2.2 Visual Gisting

The purpose of the visual gisting task was to test the surrogate's adequacy in representing visual information. It investigated to what extent users could perform "visual closure" by watching surrogates: A surrogate carries essential information of the video, and leaves blanks and uncertainties as well. When viewing the surrogate, users need to fill in the blanks and imagine what this video would look like. Participants were shown 10 test images, some from the video and some not (distractors) in random proportion and sequence, one at a time, and asked whether the image belongs to the video. The test images did include images from the surrogate. Distractors were selected from other videos by the researcher and another BLC staff member and approved by a panel. This task allows users to demonstrate comprehension through images rather than through linguistic devices and, for the keyword-only surrogate, assesses their ability to make inferences across media where the images are the target rather than verbalized meanings.

4.3 Sampling of Participants

This experiment took a "purposeful sampling" approach (Patton, 1990, p. 169). Participants were selected deliberately in order to provide important information that can not be gotten as well from other choices (Maxwell, 1996). Twelve volunteers were recruited: 4 schoolteachers, 4 graduate students in education (teachers to be), and 4 other graduate students with different majors (computer science, law, audiology, and biology).

4.4 Data Collection Methods

Think aloud (Ericsson and Simon, 1993) was adopted as the major method for data collection and supplemented by observation and post hoc interviews. The different methods supplemented each other and overcome some of the respective deficiencies in single methodological approaches. For example, through observation and participant's think aloud, the researcher was able to compare what was said and what was done by the participant so that the subjectivity of verbalization could be minimized, information that was implicit or not recordable could be captured. Through the post hoc interview, misunderstandings or confusion can be clarified or dismissed. All the performance data were tracked by computer, and all sessions for thinking aloud and interviews were audio-tape recorded.

4.5 Experimental procedure

The study was conducted at the University of Maryland from July to October, 1998. Participants came to the researcher's office and used the same computer (PC, Pentium 200, 17-inch monitor with resolution of 1024x768). Participants were randomly assigned to either the visual gisting task or the verbal comprehension task. After a practice session, the participant was exposed to each surrogate treatment (keywords, keyframes, and the combination) in a random sequence, each with two sets of video surrogates. For each set of surrogates, the participant talked aloud about what s/he saw in the surrogate, what the video was about, and then proceeded to the user task. S/he would click on a yes/no button to indicate whether the test picture belonged to the video (visual gisting) or the verbal statement correctly reflected the meaning of the video, while talking aloud the decision steps. Participants were instructed to speak out "everything that runs across your mind" .

Transcribed audio tapes and field notes were jointly coded and analyzed using NUD.IST (a program for analysis of non-structured data).

5. RESULTS AND DISCUSSION

5.1 Usefulness of Combined Surrogates

All participants except one found the combined surrogates more useful than the single-modal surrogates, and most of them said "It depends" when asked about their preference between visual-only and verbal-only surrogates. Images and words each provided unique information and served different purposes that might not be provided otherwise; images and words reinforced each other, facilitated information integration and video sense making. In addition, some users are more visually oriented while others are more verbally oriented; combined surrogates can meet the needs of both. Key findings with brief discussions and examples follow.

> Words tell the "aboutness" or the meaning of the image sequence (keyframe storyboard) and supplement the images.

Although participants could get some basic idea of what the video is about from the keyframes, the images were often so broad as to fit into many different topics., while keywords

often explicitly demonstrated the subject matter itself. This theme was mentioned by several participants and is best illustrated by one subject who noted: "Without the words I would have just seen a bunch of birds. Keywords tell a little bit about the story." Further, when there was a conflict between the images and the keywords, participants tended to depend more on the words.

Uncertainty about visual details may hinder users from interpreting the video. Participants tended to ignore information that was fuzzy, uncertain or confusing to them, and drew conclusions based only on information they were sure about. The video "Dry Season Animal Survival Strategy" describes how birds and monkeys share food during the dry season and their roles in the food chain in the rainforest. One participant, based on the keyframe surrogate, did not recognize the animal as a monkey, nor realize that this animal was a pollinator for the flowering trees. So he ignored the part of the monkey in the video, and thus missed the theme about "food chain" and "share." His conclusion was drawn based only on the activities of birds. In contrast, participants with the combined surrogate captured more themes of the video, and figured out more details from the keyframes. A more accurate and sound inference was made when the participant saw the keywords "food chain", "sharing", "flowering trees", and "monkey".

It is likely that the verbal information led to the construction of a story scheme which participants used to make sense of the video. The visual surrogates provided specific details for the story scheme. Also, the verbal information appeared to serve as a label for images or a container of the visual contents, which shaped the conceptual theme or outline, the main idea of the video.

One participant emphasized the importance of words in linking keyframes, and suggested that there should be two kinds of words some at a higher level to connect the images together, and others to match the images individually:

> No, because I think that does not really help in a sense that if you already have an image, of course it is going to extract some word by itself per se, you don't need to tell me what the image is about. That could help, but they are not as important as the words that connect the images together. Where would you place the phrase "ancient wonders of the world"? You can't place it anywhere, or you can probably place it anywhere for that matter. But these words give you additional information There are two types of words, one is the words that deal with the subject matter. They need to be there. And then words supplementing the images, I mean the words corresponding to the images respectively. That type of keywords is somewhat different from the keywords themselves.

This quote further differentiates the role that words and images are expected to assume in the surrogate. Images are visually specific, and often verbally ambiguous. Keywords should impart the overall meaning of the video, specify, disambiguate, or supplement the information the images communicate. Whenever it is necessary, individual images should have captions that supply specific information not obvious form the image alone

Keyframes add unique details to aid understanding

Although participants frequently mentioned the importance of verbal information in making sense of the videos, often the usefulness of the verbal information was built upon the key frames. The keywords facilitated the understanding of the keyframes and the keyframes helped add detail and substance to the words. Participants commented that the uniqueness of images, such as concreteness, vividness, impressiveness, and realism, cannot be easily or effectively conveyed by words. Images illustrated abstract concepts by providing "what it looks like". They gave detailed information including settings, emotion and background as well as the main focus. Images were more interesting and fun to watch, and invited exploration and association. They could also be repurposed for multiple uses and interpretations. A teacher participant who strongly preferred keyframe surrogates to the others mentioned:

> Sometimes when you have the images, you don't have to focus on what the actual topic is. You can put it in another setting or create another setting with those images yourself, so that would give you a sense. Oh, I could use them this way, and I would have never thought about using it this way if I had heard the whole clip or the sound.

Images and words reinforce each other.

The semantics of verbal information and the uniqueness of images reinforce each other. Participants explained why the combined surrogate was superior to the single modal surrogates from different perspectives: In terms of information sufficiency, text-only and images-only both constrained full understanding of the video content. Putting them together could optimize the structure of the representation in that words and images each provided different kinds of information and supported each other. From a user's perspective, the combined surrogate could better accommodate different user needs or different types of users. From a cost-effectiveness perspective, well-integrated surrogates allow effective examination per time investment.

Text in images drew special attention.

Text in images (e.g., images with captions or graphics) was not studied as a specific surrogate type in this research. However, the key frames that contained such text clearly drew participants' attention, presumably because they conveyed

more information than regular keyframes. Two kinds of text appeared in the keyframes in this study: subtitles and graphics. In these cases, subtitles were captions in a different language from what was spoken in the video. In describing these keyframes, participants used phrases like conveying "a lot of information", "indicative", "crucial" and "attention capturing". Text in images may be especially useful because it is well integrated and matched with the corresponding images even beyond the general role of keywords in the combined surrogates. Also, participants did not have to match the text and images themselves, which sometimes may cause ambiguity or confusion otherwise.

"Information sufficiency" depends on the user's background.

Several participants observed that a user with sufficient background knowledge could interpret images correctly without words (for example recognizing Big Ben and knowing it is London or a specific animal and knowing it lives in the rain forest) or, conversely, picture in his or her mind the images that are implied by words. This reinforces the idea that clearly the success of video surrogates depends on the proper match between the user and the surrogate. This is an area that needs more research attention.

Participants preferred having video surrogates

For the purpose of screening and selection, participants liked the idea of quickly browsing video surrogates to get the gist before spending time and effort watching the full motion video. All participants except one claimed that they preferred the combined surrogates, citing information abundance and accommodation of different user needs and user characteristics. This further confirmed the results from the quantitative study. Furthermore, participants agreed that processing the combined surrogates did not take any longer than either of the separate surrogates. It appears that the time to make sense of the surrogate is made up of perceptual processing and cognitive inferencing components; while presenting an integrated surrogate may increase perceptual processing time, it seems to decrease the time needed for cognitive inferencing.

5.2 Cognitive Process in Video Surrogate Examination

An understanding of the information processing mechanisms involved in combined surrogate examination may explain, from a different point of view, why participants preferred the combined surrogates. It would also shed light on techniques for presenting video surrogates in real video browsing settings.

Participants differed in the strategies they used for making sense of a video based on the combined surrogates. Some participants claimed that they simply followed the sequence in which the information was organized in the surrogate. Some first looked at the image sequence as a whole briefly, and then went to the words, and finally went back to examine the images individually and carefully. Others first read the words, and then went to the images. Several factors influenced the choice of a surrogate processing strategy.

First, a processing strategy was related to a participant's preference of modality. For example, Participant 7 contended that once the keywords were carefully and properly chosen, images may not be necessary as part of the video surrogate. Accordingly, he always went to the words first. Participant 11, on the other hand, highly preferred images to words. She said that her eyes were first drawn to the images when she processed the first combined surrogate, and she thought that was automatic.

Second, the processing strategy was also dependent on the viewer's familiarity with the information in the surrogate. When the topic was not familiar, participants tended to first resort to the verbal information. Although Participant 11 processing the first combined surrogate looked at the images first, she could not get much information from there and had to switch to the keywords. So, in the second trial she went to the keywords first. But then she commented that for the second surrogate (of a video on the Cherokees), the words did not help much. Participant 2 also ended up going to the words first in processing the second combined surrogate.

Third, another factor that could direct the processing is the visual appeal/attractiveness of images. It is possible that attractiveness resulted not from familiarity, but novelty (e.g., special attention paid to unusual scenes), or emotional response.

Thus, the surrogate information processing strategy depends on factors such as the user's modality orientation, the user's knowledge about the surrogate content, and the visual appeal in the surrogate. These factors were intertwined from situation to situation, and caused participants to take different steps to reach their goal contextually.

Regardless of whether they started with words or images, participants adopted two main generic processing strategies: sequential vs. selective. With the first strategy, participants basically followed the sequence as the information was presented. The second strategy was more dynamic and proactive. Participants first built up a quick scheme of a story, and then tested whether the scheme still held when further details or specificity were added. In this regard, it seemed that words were better suited to scheme building, and images were better for confirmation and illustration.

Participants addressed different uses of the sequence of the images. When there was no additional information available except for the images, they had to make full use of the sequence. When there was verbal information in addition to the images, they tried to absorb the visual information based on the verbal cues.

There seemed to be a lot of visual/verbal integration by participants. Time was not recorded in this experiment, but in the first experiment we found that there was no significant difference in processing time between the keyframe surrogate and the combined surrogate. It again suggests that the information integration is faster and easier even though the perceptual part of the processing may in fact take a bit longer (but not a perceptively recognizable amount longer to the participant).

Participants liked the combined surrogates because they could cross-examine the surrogates and build up and further refine a scheme to make sense of the video. Going to the images or the words first was mainly dependent on personal orientation to the modality. Most participants seemed to first quickly scan the words or image sequence as a whole, and then switched to the other modality. No matter whether the participant processed the information more sequentially or selectively, they tended to rely more on the words than the images to set up the baseline of the story in the video. The intertwining of the multiple impact factors could cause the participant to adapt strategies and tactics to each specific situation.

5.3 Sense Making and Video Comprehension

This section answers two main questions: what information most captured participants' attention? What information in the surrogates was most used to comprehend the video?

5.3.1 Attention-getting Information

Participants did not attend to all the stimuli equally, instead, something usually first captured their attention, especially when keyframes were available in the surrogate. With the verbal surrogate, participants in the verbal comprehension task tended to make up a sentence by including all the keywords provided or by paraphrasing them. With surrogates that included visual elements, participants often paid more attention to keyframes with one of the following features:

Text in pictures Captions or graphics gave "voice" to images;

Interaction information Action scenes attracted more interest than static scenes;

Symbols Icons and stereotypes attract attention and cued visual gisting;

Novelty Unusual scenes gained attention and might lead to inferences;

Emotion Scenes that evoked strong emotional response attracted attention.

People People and, to a lesser degree, animals captured more attention than inanimate objects.

5.3.2 Sense-making Strategies

Participants tended to use the most attention-capturing cues to build up the theme of the video, and then used the other information to reinforce, confirm, or adjust the story. From the way participants described/imagined what the video was (should be) about, it can be seen that comprehension was centered on people and their activities. Also they tried to describe the video with as many specific terms as possible they tried to be specific with the name, location, and events. This was consistent with results from the first experiment that showed that iconographical concepts were more frequently used than pre-iconographical concepts in summarizing the full motion video.

Participants showed a strong people-orientation (consistent with Valva's findings, cited in Massey and Bender, 1996). They tended to make up a story of a video by putting particular person(s) seen in the video at the core, even though in many cases that might not actually be accurate. Maybe it was easier to come up with a story involving people. For example, one video actually shows how the tribe lives in harmony with nature and how they keep animals as pets, but participants paid more attention to the people. "So my story about this video is how the boy spent his leisure time with animals in the village." In the video, "Early Trains and Railroads", there was only one picture of a man demonstrating how the first telegram worked. The story line interpreted by one participant was "I guess that guy is a railroad timekeeper so that they could use the telegram to see whether or not the train is on time and to ensure there won't be accidents ".

Participants would target other contents if people were not available. "We seem to have lots of birds and tropical habitats. It doesn't seem to be about people." "This video describes the forbidden city showing details of the exterior and interior. Other than that, there are no people here anywhere. Very deserted."

Thus, viewer knowledge about the domain represented in the video and the presence of people in the images are powerful orientations for sense making. Specificity-orientation requires that the viewer have sufficient prior knowledge. Providing proper verbal information (e.g., keywords) might compensate for the lack of knowledge. People-orientation seems to be a plausible strategy to make sense of the video, especially when available information is limited, as in processing video surrogates. Future keyframe extraction techniques may need to take these factors into consideration.

6. IMPLICATIONS FOR DESIGN

The results presented here and more detailed results found in Ding (1999) suggest some guidelines for the design of video surrogates, including the extraction of keyframes.

Include both images and text in video surrogates (multimodal surrogates).

Make sure that images and text are coordinated and complement and reinforce each other.

Include two levels of text:

Thematic: Text describing the video as a whole its theme.

Specific: Text keyed to individual images (captions).

Provide thematic and specific iconographic information (who, what, where, when, how, why, to what purpose/ with what effect) through text.

When selecting keyframes, pay particular attention to the following

Frames with people, particularly people interacting with each other, with animals, or with other elements of the environment.

Frames including text, icons, or other symbols (see, for example, Lienhart 1996).

Frames with vivid colors.

Frames with unusual/novel scenes or objects

Frames evoking emotion.

7. CONCLUSION

Our study shows that users strongly prefer video surrogates that combine verbal information (text in a generic sense) and images. Each modality makes a unique contribution to the comprehension of a video, and in combination they reinforce each other. Verbal information helps users get the overall meaning of the video and specify or clarify the thematic information described in the visual surrogates, such as who, where, when and how. Put differently, verbal information conveys the iconographical meaning, and supports users' understanding of the meaning of the contents of images. On the other hand, visual information is concrete, vivid, detailed, and more real; it is more apt to convey affect, emotion, and excitement and to draw attention. Often verbal information helps the user to extract more meaning from images. The combined surrogates integrated verbal and visual information and facilitate information processing so that it actually may take less time to process the larger amount of information present in combined surrogates as compared to purely visual or purely verbal surrogates. Further studies could investigate whether short abstracts instead of keywords and/or the audio presentation of the verbal part would bring still greater benefits.

While image-based surrogates, particularly those that can be prepared automatically, have received much attention, the message to designers is clear: Provide video surrogates that integrate images and text. Text may be descriptive of the video as a whole or label a specific image; both kinds of text are helpful. In selecting images for surrogates, such as in keyframe extraction, consider images that depict interaction, especially of people, images that evoke emotions, images that contain text or symbols, and images that are novel and attractive these kinds of images seem to help people most in making sense of a video.

8. Acknowledgments

This research is supported by the US Department of Education Technology Challenge Grant (#R303A50051) to the Baltimore Learning Community Project. We thank Discovery Communications for the use of the educational videos. Thanks are also due to the participants in this study.

9. REFERENCES

[1] Borko, H. & Bernier, C. L. (1975). *Abstracting Concepts and Methods*. Academic Press.

[2] Cawkell, A. E. (1995). *A Guide to Image Processing and Picture Management*. Gower.

[3] Christel, M.; Winkler, D.; Taylor, C. R. (1997). Improving access to a digital video library. *In Proceedings of the 6th IFIP Conference on Human-Computer Interaction*. (Sydney, Australia)

[4] Cohen, A. A. (1987). Decision making in VCR rental libraries: information use and behavior patterns. *American Behavioral Scientist*. 30(5), 495-508.

[5] Ding, W. (1999, in progress) *Cogintive processing of multimodal surrogates for video browsing*. Doctoral dissertation. University of Maryland.

[6] Enser, P. G. (1995). *Pictorial information retrieval. Journal of Documentation.* 51(2), 126-170.

[7] Ericsson, K. A.; Simon, H. A. (1993). *Protocol analysis: verbal reports as data. Cambridge*, MA: The MIT Press..

[8] Lienhart, R.(1996). Automatic text recognition for video indexing. In *Proceedings of the ACM Multimedia* '96. 11-20.

[9] Marchionini, G.; Nolet, V.; Williams, H.; Ding, W.; Beale, J.; Rose, A.; Gordon, A.; Enomoto, E.; Harbinson, L.(1997). Content + Connectivity =>Community: Digital Resources for a Learning Community. *Proceedings of the 2nd ACM International Conference on Digital Libraries*. 212-220.

[10] Massey, M.; Bender, W. (1996). Salient stills: process and practice. *IBM Systems Journal*, 35, 3-4, 557-573.

[11] Maxwell, J. A. (1996). *Qualitative research design: an interactive approach*. Sage Publications.

[12] Morgan, J.; Welton, P. (1992). *See what I mean an introduction to visual communication*. 2nd ed. Edward Arnold. 150p.

[13] O'Connor, 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.

[14] O'Connor, B.C. (1991). Selecting key frames of moving image documents: A digital environment for analysis and navigation. *Microcomputers for Information Management*, 8(2), 119-133.

[15] Patton, M. Q. (1990). *Qualitative evaluation and research methods* (2nd ed.). Newbury Park, CA: Sage.

[16] Ponceleon, D.; Srinivasan, S.; Amir, A.; Petkovic, D.; Kiklic, D. (1998). Key to effective video retrieval: effective cataloging and browsing. In *Proceedings of ACM Multimedia '98*. 99-107.

[17] Pryluck, C. (1976). *Sources of Meaning in Motion Pictures and Television*. Arno Press. 241p.

[18] Teodosio, L.; Bender, W. (1993). Salient stills from video. *Proceedings of ACM Multimedia* '93. 39-46.

[19] Tse, T.; Marchionini, G.; Ding, W.; Slaughter, L.; Komlodi, A. (1998) Dynamic Key Frame Presentation Techniques for Augmenting Video Browsing. In *Proceedings of the Advanced Visual Interfaces '98 Conference*.

[20] Turner, J. (1994). *Determining the subject content of still and moving documents for storage and retrieval: an experimental investigation*. Unpublished Ph.D. Dissertation. University of Toronto.

[21] Wang, P.; Soergel, D. (1998). A cognitive model for document use during a research project. Study I: document selection. *Journal of the American Society for Information Science*, 49(2), 115-133.

[22] Zhang, H. J.; Low, C. Y.; & Smoliar, S. W. (1995). Video parsing, retrieval and browsing. In M. T. Maybury (Ed.) *Intelligent Multimedia Information Retrieval*. 137-158. MIT Press.