Alternative Surrogates for Video Objects in a Digital Library: Users' Perspectives on Their Relative Usability

Barbara M. Wildemuth,[1] Gary Marchionini, Todd Wilkens, Meng Yang,
Gary Geisler, Beth Fowler, Anthony Hughes, and Xiangming Mu
Interaction Design Lab, School of Information and Library Science
University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-3360
[1]wildem@ils.unc.edu

Abstract. In a digital environment, it is feasible to integrate multimedia materials into a library collection with ease. However, it seems likely that non-textual surrogates for multimedia objects, e.g., videos, could effectively augment textual representations of those objects. In this study, five video surrogates were evaluated in relation to their usefulness and usability in accomplishing specific tasks. The surrogates (storyboards with text or audio keywords, slide shows with text or audio keywords, fast forward) were created for each of seven video segments. Ten participants, all of whom watch videos at least monthly and search for videos at least occasionally, viewed the surrogates for seven video segments and provided comments about the strengths and weaknesses of each. In addition, they performed a series of tasks (gist determination, object recognition, action recognition, and visual gist determination) with three surrogates selected from those available. No surrogate was universally judged "best," but the fast forward surrogate garnered the most support, particularly from experienced video users. The participants expressed their understanding of video gist as composed of three components: topicality, the story of the video, and the visual gist of the video. They identified several real-world tasks for which they regularly use video collections. The viewing compaction rates used in these surrogates supported adequate performance, but users expressed a desire for more control over surrogate speed and sequencing. Further development of these surrogates is warranted by these results, as well as the development of mechanisms for surrogate display.

1 Introduction

In a digital environment, it is feasible to integrate multimedia materials into a library collection with ease because they can be delivered as bit streams, just as textual materials can be delivered. This feasibility is demonstrated daily by the addition of collections of multimedia materials, such as video, to the World Wide Web and digital libraries. While delivery of digital video is possible over the Web, it is costly in download time. Thus, library users would benefit from the ability to assess the relevance of the available videos prior to downloading. While textual surrogates for the videos (such as the video title or a desription of its content) can assist in this process, it seems likely that non-textual surrogates for the available videos could

effectively augment textual surrogates. Unfortunately, because video is a relatively new medium, we have not yet developed surrogates analogous to the abstracts, reviews, tables of contents, prefaces, etc., that help people understand the gist of texts before reading the full objects. Although there is an enormous amount of technical research directed at creating such surrogates, there is little work on understanding how people can and will actually use such summaries. The emphasis to date has been on using the surrogates as metadata objects for the purposes of retrieval. We aim to understand broader issues such as how these surrogates are used to extract meaning and to make ongoing decisions that guide browsing within a library of digital videos. More specifically, in the current study we are concerned with understanding which surrogates are useful for which types of tasks and how people are able to incorporate these surrogates into their information processing strategies.

2 Background

Surrogates "stand for" objects. Abstracts, titles, and keywords are all familiar surrogates for complete documents or objects. Surrogates support both retrieval and gist extraction and have long been central to library and information science research (e.g.,[1,13]). In browsing, surrogates provide an important alternative to primary objects as they take far less time to examine and provide enough semantic cues to extract gist and allow users to assess the need for further examination of other surrogates and the primary object. In digital libraries and archives, surrogates are crucial for browsing large distributed collections that result from filtering programs or analytical queries of the data space.

Key frames are a natural analogue to keywords in text and much of the effort has focused on identifying and extracting key frames. O'Connor [20] suggested that key frames—still images representative of scenes extracted from the video itself—could be used to construct video abstracts or "contour maps." Rorvig [22] demonstrated the feasibility of creating visual surrogates based on extracted key frames and the creation of specific summaries has become an important aspect of the information retrieval research community's interest in multimedia (e.g., [17,25]).

Results of studies of people interacting with surrogates [8,11,29] have suggested that keywords and key frames each contribute unique elements to understanding and each reinforces the other. Text contributes iconographic, thematic information and images contribute preiconographic details and affective impact. These results parallel the Robertson et al. [21] work with thumbnail images for Web pages that showed that combining text and thumbnails led to fewer errors and failed trials than text or thumbnails alone. It seems logical that providing more and more varied information will lead to better results. However, an important question is how much the additional processing "costs" the browser and how this affects the overall browsing experience.

The temporal nature and multiple channels of video content exacerbate the need for search and browsing mechanisms that offer more representation facets than text. Some researchers have begun by developing structured interfaces for video surrogates (e.g., [7,10,32-34] and there has been excellent progress on the engineering of video surrogates and designing display structures (e.g., [9,18,27]), but there has been very

little work on studying how people interact with and use such surrogates. The Informedia Project is an important exception, having both constructed innovative interfaces for video retrieval and conducted studies of how people use alternative surrogate implementations [2,3,26]. In addition, Goodrum [12] has examined users' perspectives on the congruence between videos and several surrogates. Marchionini and his colleagues have conducted user studies to identify key parameters for video browsing [7,8,14,24,28] and have developed a set of methodologies that may be used in other video browsing efforts. However, these studies mainly focused on surrogates for specific video objects and item recall types of tasks. The current study focuses on the general tasks of reviewing and extracting salience (operationalized in gist determination, object recognition, action recognition, and visual gist performance tasks) from surrogates of the items retrieved from video collections.

3 Research Questions

Our long-term goals are to understand surrogates within the context of digital library use. The current study focuses on particular characteristics of the surrogates and their effects on user preferences and performance. Specifically, this study addresses two research questions:

- What are the strengths and weaknesses of each surrogate, from the user's perspective?
- Are any of the surrogates better than the others in supporting user performance?

4 Methods

Five surrogates were created for each of seven video segments. The surrogates included two storyboards (one with text keywords and one with audio keywords), two slide shows (one with text keywords and one with audio keywords) and fast forward. Ten participants, all of whom watch videos at least monthly and search for videos at least occasionally, viewed the surrogates and provided comments about their strengths and weaknesses. In addition, they performed a series of tasks (gist determination, object recognition, action recognition, and visual gist determination) with three surrogates selected from those available. The methods used in the study are described in more detail below.

4.1 The Videos

The video segments were selected from the repository of the Open Video Project (http://www.open-video.org/), a shared digital video repository and test collection created at the Interaction Design Lab at the University of North Carolina at Chapel

Hill. The collection currently is more than a half terabyte and contains video and metadata for more than 1600 video segments.

For the current study, seven video segments were selected from the collection. They included both color and black & white videos and represented several genre (documentaries, educational, promotional). They were:

- Apollo, Segment 4006 (2:07),
- Chevrolet Promotional Videos: Master Hands, Segment 1 (4:54),
- Challenge at Glen Canyon, BOR03, Segment 2 (3:00),
- Educational Films: A Date with Your Family (9:59),
- Moon, Segment 2 (3:43),
- Hurricanes, Segment 1 (3:54), and
- New Indians, Segment 101 (2:11).

4.2 The Surrogates

With support from the Mpeg Encoded Retrieval and Indexing Toolkit (MERIT, http://documents.cfar.umd.edu/LAMP/Media/Projects/MERIT/) and customized Java and Perl programs, we manually created five surrogates to be evaluated in this study. Two of our surrogates were storyboards, each consisting of no more than 36 frames, laid out in a 6x6 grid. Users were allowed to view the storyboard for a limited amount of time, allowing 500 milliseconds per key frame. For example, if a storyboard included 20 key frames, the user was allowed to view it for 10 seconds. One of the storyboards was augmented with text keywords (consolidated from those independently assigned by two members of the research team), visible under the storyboard, at the bottom of the grid. The other was augmented with audio keywords (the same set); an audio recording of them was played during the viewing. The audio recording was generated by a speech synthesizer, so that standardization of pace and pronunciation could be ensured. Audio was repeated as necessary for the duration of the visual display. Two of the surrogates were slide shows incorporating the same set of key frames as were included in the storyboards. They were displayed at the rate of 250 milliseconds per frame. The entire set was displayed twice, with no pause between the two repetitions. The slide shows were augmented with textual and audio keywords, parallel to the augmentation of the storyboards. The fifth surrogate was a fast forward version of the video segment, mimicking the fast forward function of a VCR player. For this study, the target video segment was played at four times its original speed, so that it would "run" for about the same amount of time that the storyboard and slide show for the same video segment were displayed. Each participant viewed it once. No keywords augmented this surrogate.

4.3 Participants

Study participants were recruited through the distribution of flyers in several UNC-CH classes related to video production and the posting of the same flyers near the video collection in the UNC-CH libraries. Participants were included in the study only if they had some experience in searching video libraries or collections. The set of ten

participants included five men and five women ranging in age from their early 20's to early 40's. All of them used computers daily. The frequency with which they watch videos varied from daily to monthly and the frequency with which they search for videos varied from daily to occasionally.

4.4 Procedures

Each study session was conducted in three phases. First, the concept of a surrogate was explained to the participants as something that "might be used in place of viewing the whole video for certain purposes such as selecting a particular video for full viewing or sorting the videos into a certain order." During the first phase, the study participant worked with the Apollo and Master Hands segments (order of presentation was counter-balanced). Each participant first viewed the full segment, and then viewed three surrogates of that video. The surrogates were assigned to participants so that all participants could comment on all surrogates, with the medium of the keywords (text versus audio) counterbalanced to minimize order effects. While the participant worked with each surrogate, s/he was asked to identify its strengths and weaknesses, tasks for which it might be most appropriate, and its usefulness under time constraints.

In the second phase, the participants interacted with the Glen Canyon and Date with Your Family segments. The procedures were the same as the first phase, except that the full video segment was ***not*** viewed. Thus, participant comments and understanding were based solely on the surrogates.

In the third phase, each participant was asked to complete several assigned tasks while interacting with each of three surrogates. Each participant selected the surrogate with which to perform the tasks, for each of three video segments: Moon, Hurricanes, and New Indians. The participant was free to use the same surrogate each time, or to select a different surrogate for each segment. In the 30 trials (three for each of the 10 participants), the storyboard with audio keywords was used four times, the storyboard with text keywords and the slide show with audio keywords were each used six times, and the fast forward surrogate was used 14 times.[1] Participants were asked to think aloud while completing the assigned tasks.

Some of the assigned tasks have been used in previous studies, while others were created for this study. All of the tasks are closely related to reviewing and extracting salience from the surrogates—a task that is typical in digital library settings. First, the participant completed two gist determination tasks. In the first, the participant was asked to write a brief summary (a few sentences) of the video content; in the second, the participant was asked to select the summary that best represented the video content from five brief text summaries presented. The gist statements generated in the first task were scored using the method reported in Tse et al. [28]. The statements were independently scored by two members of the research team and any differences were resolved by a third member of the team. Next, the participant completed two object

[1] The participants' selection of surrogates for use in phase 3 should not be interpreted as an indicator of their preferences for those surrogates. Several of the participants explicitly stated that they selected a surrogate to use in phase 3 because they did *not* like it and wanted to confirm their negative judgment of it.

recognition tasks. In the first, the participant was presented with a list of 12 object names and asked to mark those objects seen in the surrogate. Six of the 12 had been seen (i.e., were correct). Three of the correct and three of the incorrect were concrete objects, e.g., "astronaut" for the Moon video, while the remaining objects were more abstract, e.g., "space program" for the Moon video. In the second object recognition task, 12 key frames were presented to the study participant. Six of them were randomly selected from the key frames used in the creation of the storyboard and slide show surrogates (i.e., were correct), three were selected from other segments of the same video, and three were selected from other videos in the Open Video repository. The participant was asked to mark those frames seen in the surrogate. An action recognition task was newly developed for this study. Six mini-segments (each 2-3 seconds long) were displayed to the study participant. Two of the mini-segments were from the video segment represented in the video surrogate (i.e., were correct), two were from another segment of the same video, and two were from a different video. In response to each mini-segment, the participant was asked whether s/he believed it to be from the same video segment as represented in the surrogate. The final task was developed for this study, and is intended to assess the participant's ability to determine visual gist: a combination of topic, story line, and style. Twelve frames were displayed to each study participant, ***none*** of them from the surrogate. Three were other frames from the video segment (considered correct); three were selected from other segments of the same video (also considered correct); three were selected from other videos of a similar style (e.g., black and white versus color), and three were selected from videos of other styles. The study participant was asked to select those frames that s/he believed "belong" in the video segment for which the surrogate was seen. For the object recognition, action recognition and visual gist determination tasks, the total score was the sum of the correct items marked and the incorrect items not marked.

All phases of the study were videotaped, using the usability workstation in the UNC-CH Interaction Design Lab. The videotape included a face shot, a keyboard/mouse shot, and screen capture. Participant comments from the videotapes were transcribed and analysed by inductively identifying themes in the comments and categorizing the comments by these themes. The phase 3 data were analysed quantitatively; descriptive statistics will be reported here. In addition, the effects of surrogate and video segment on participant performance were investigated with analysis of variance.

5 Results and Discussion

After the participants' preferences for particular surrogates are summarized, this section will focus on participants' understanding of what gist might encompass, and their ability to determine the gist of the video from viewing the surrogate. This will be followed by a discussion of the tasks for which surrogates might be used and the relationship between those tasks and the phase 3 performance tasks, the efficiency gained through use of the surrogates, and the effects of differences in the video segments on participant performance.

5.1 User Preferences

At the end of each phase of data collection, each of the participants was asked which surrogate s/he preferred. While not all participants expressed a preference, most did. Participant preferences tended to change over time, and some preferred particular surrogates for particular purposes. Some participants (David, Ling)[2] preferred the fast forward surrogate, and several additional participants (Matt, Ryan, Cheryl, Pam) suggested that they would prefer a fast forward surrogate if audio keywords were added. The others who stated a preference were divided among storyboard with audio keywords (Jan) and the slide show with audio keywords (Bettie, Steven). These results indicate that the fast forward surrogate should be further developed with the addition of audio keywords.

5.2 Gist from a User's Perspective

Determining the gist of the video was the most important function of the surrogates, from the users' perspective. Three different understandings of gist were present in their discussions of using the surrogates to determine gist. The first was the view that the surrogate could help the user to understand what the video was about, i.e., the topic of the video. This understanding is parallel to the concept of topical relevance [5,23]. Participants referred to this concept by describing the surrogate as providing an "overview," describing the "content" of the video, or "what it's about." This function of the surrogates was the one for which keywords played the most prominent role. In the words of Jan, as she was using the fast forward surrogate (with no keywords): "If you don't have the [key]words, you'll think it's about something else, not as it was supposed to be." The participants found the keywords that were proper nouns to be particular useful, as Cheryl commented: "The common noun keywords are debatable, but the proper noun keywords would be really good to have."

Secondly, the participants found the surrogates most useful when they told the "story" of the video or had a narrative structure. As Cheryl commented: "It's in chronological order, so that makes sense and you can kind of follow the story line through." Ryan criticized one of the storyboard surrogates for lacking this information: "I didn't get much of a sense of the flow of the story." This desire for a narrative structure is most likely associated with the temporal nature of video. In addition, the users' interactions with the surrogates are consistent with van Dijk and Kintsch's [30,31] model of discourse comprehension. In it, they postulate that readers use macrostrategies to form an initial hypothesis about the gist of a text based on initial cues from the text, and then interpret additional cues from the text in light of their initial hypothesis concerning its gist. This cognitive process was most apparent while participants viewed the video segment, A Date with Your Family, produced in 1950. The keywords included the term "date," yet the video depicted a family meal and was oriented toward etiquette. A number of the participants were disturbed by the conflict between the expected scenario of a date (i.e., their initial hypothesis of the video's gist) and the family scene they were viewing. As expressed by Pam, "When the voice said

[2] Names used in the paper are pseudonyms assigned to protect the participants' confidentiality.

'date,' there was a picture of 'dad,' kind of crazy to me. It wasn't an image of 'date' to me." A more positive example supporting van Dijk and Kintsch's theory was Ryan's viewing of the Apollo segment: "I got, pretty early on, that it was training for the mission to the moon. Once I had that, I could figure out what each of the pieces were in the training."

The third understanding of gist presented by the study participants is what we are calling visual gist. Based on the participant comments, it is a combination of topicality, narrative structure, and visual style. While this concept needs additional clarification, participant comments clearly indicated that they formed a more holistic view of gist, beyond topic and narrative. Most of the positive comments related to visual gist were associated with the fast forward surrogate, such as Matt's comments: "The motion really added a lot… I have a stronger sense of what the movie's like… It definitely gave it a whole different feel… It gave me more of a sense of what to expect from watching [the entire video]."

During the third phase of the study, participants performed two gist determination tasks: one in which they generated a statement of the gist of the video and one in which they selected a gist statement from five presented to them. Scores on the participant-generated statements could range from 0 to 3. The participants' mean score was 1.68 (s.d. = 0.75); their actual scores covered the entire range possible. There were no differences between surrogates in this score ($F=0.39$ with 3, 26 df,[3] $p=0.7643$). Further comparisons with analysis of variance indicated that there were no differences by the basic form of the surrogate (storyboard v. slide show v. fast forward; $F=0.58$ with 2, 27 df, $p=0.5653$) or by the medium with which the keywords were presented ($F=0.35$ with 2, 27 df, $p=0.7050$). On the second gist determination task, participants were correct on 80% of their statement selections. Again there were no differences by surrogate ($p=0.85$, Fisher's exact test), by the basic form of the surrogate ($p=0.83$, Fisher's exact test), or by keyword medium ($p=0.72$, Fisher's exact test).

5.3 Tasks for Which Video Surrogates Might Be Used

One of the weaknesses in our basic knowledge of people interacting with digital video libraries is in our understanding of the tasks for which video collections might be used. As the study participants used the various surrogates, they were asked about possible uses for which the surrogate might be useful. The results of these interviews revealed a number of tasks associated with video use.

The task originally envisioned in the research design (as expressed in the definition of surrogates provided to the participants) was the classic information retrieval task of selecting videos from the collection. Jan described the possibilities of looking for particular content or particular visual techniques. When working with the fast forward surrogate, Cheryl said, "You get a pretty good sense of whether or not you wanted to actually see the whole thing and take the time to pursue it further." Matt made a more fine-grained distinction, noting that he was more likely to use the

---

[3] For this and other analyses of the effects of the surrogates, there are only 3 degrees of freedom associated with the model, since participants used only four of the available surrogates during the third phase of the study.

surrogates for school projects and reference tasks than for entertainment-oriented searching. Bettie commented that surrogates would be useful for determining whether a particular video would be appropriate for children.

In addition to selection decisions related to the entire video, a number of participants pointed out the possibility of using the surrogates for selecting particular frames or clips for later use. Participants noted differences between the surrogates for these types of tasks. For example, Ryan contrasted the utility of the storyboard with text for identifying a particular image, but would prefer the slideshow if he was looking for particular visual elements. David, a very experienced video user, preferred the fast forward surrogate for "picking out highlights", but most of the participants believed that the storyboards would be most useful for identifying particular frames or sections of the video. Participants also pointed out that "sometimes you need to compare images" (David), for which the storyboard would work most effectively. David also noted that, "ultimately they're going to have to organize what they're viewing," and that users will need ways to manipulate portions of a segment.

The motion in a video and the video's style were also attributes about which the participants wanted information in the surrogates. They differentiated between "long shots" and "zooms," and described a desire to search for particular "film techniques" and "camera angles." In all these cases, seeing the movement in a video was critical. As David pointed out, "With motion pictures, you want to see how it moves at some point in the process."

In summary, participants expected the surrogates to provide the capability to select videos from the collection, to select and organize particular frames or clips, and to evaluate the style or identify particular stylistic techniques. For selecting videos, different surrogates were judged to have particular advantages or disadvantages. There was general agreement that storyboards were most useful for selecting and organizing particular frames or clips and that the fast forward surrogate was most useful when focusing on the video's style.

5.4 Performance Tasks in Relation to User-Defined Tasks

During the third phase, in addition to gist determination, the participants performed object recognition, action recognition, and visual gist tasks. Object recognition, in which the participant is provided with a set of stimuli and asked which were seen in the surrogate, is related to the user-defined task of selection of particular frames. If a person performs well in the object recognition task, it can be argued that the surrogate supports the task of frame selection well. There is a parallel relationship between action recognition and the selection of particular clips. The visual gist performance task, which asks users to predict whether a particular frame is from the same video segment as shown in the surrogate, incorporates both content and stylistic considerations. This performance task is most closely related to users' desire to evaluate the movement or style in a video.

Participants performed two object recognition tasks. For the first, the 12 stimuli were names of objects that may have been represented in the surrogate viewed. The mean score was 9.0 (s.d. = 1.6), and scores ranged from 6 to 12. Analysis of variance indicated that the effect of the surrogate approached significance ($F=2.70$ with 3, 26 df,

p=0.0664). A post hoc Duncan's multiple range test indicated that the effect was associated with the difference between the storyboard with text keywords (mean = 10.2) and the storyboard with audio keywords (mean = 7.5). For the second object recognition task, the stimuli were 12 video frames. The mean score was 9.0 (s.d. = 1.8) and scores ranged from 3 to 12. The effect of the surrogate was not significant (F=0.36 with 3, 26 df, p=0.7823). The mean score on the action recognition task was 4.6 (s.d. = 1.0), and scores ranged from 2 to 6. The effect of the surrogate was statistically significant for this task (F=3.36 with 3, 26 df, p=0.0340). A post hoc Duncan's multiple range test indicated that the fast forward surrogate outperformed the rest of the surrogates. The means, by surrogate, are shown in Table 1. For the visual gist task, the participants were presented with 12 frames and asked to select the frames that "belonged" in the target video segment. The mean score was 9.7 (s.d. = 1.4), and scores ranged from 7 to 12. The effect of the surrogate was not significant (F=0.08 with 3, 26 df, p=0.9709). For two of the performance tasks, a particular surrogate seemed to provide better support than others: the fast forward surrogates in support of the action recognition task and the storyboard with text keywords in support of the object recognition task (with textual stimuli). In terms of user performance, it is reasonable to view at least these two surrogates as promising for future investigation.

Table 1. Action recognition performance, by surrogate*

Surrogate	n	Mean score
Fast forward	14	1.6
Storyboard with audio keywords	4	1.0
Storyboard with text keywords	6	0.8
Slide show with audio keywords	6	0.8

* No participant selected the slide show with text keywords for use during phase 3 of the study.

Of more import for future research is the validation of the performance measures used in this study. Of the two gist determination tasks presented earlier, each has its strengths and weaknesses. The scoring method used for the user-generated gist statements still needs refinement. While there was a significant amount of interrater agreement on the scoring (Cohen's kappa = 0.354 [4]), it was far from ideal. We expect that the problems experienced in scoring may be minimized if a clearer distinction is made between scores of 2 and 3 and if scorers are clearly instructed to base their judgments on their viewing of the entire video segment. The multiple-choice gist determination task can be objectively scored, but care must be taken in developing distractor statements that are at an appropriate level of difficulty. A similar challenge exists for the remaining performance tasks–selecting distractors that are appropriate. In this study, specific criteria were established for selecting a set of distractors that were of varying levels of difficulty. As more is known about people's interactions with video material, it is likely that these criteria can be further refined. In spite of these challenges, it is fair to conclude that the measures developed and used in this study provide a basic set of valid approaches to the measurement of performance in interacting with video materials. Their psychometric properties seem reasonable (means just above the midpoint of the possible range of scores; appropriate variability in the scores). In addition, they are clearly related to the real-world tasks that users expect to perform with video collections.

5.5 Efficiency as a Function of Viewing Compaction

One of the goals of creating surrogates is to allow the user to review retrieved items and make decisions about their relevance more quickly than if they were required to review the complete item [15]. For video materials, this concept can be instantiated in the human-centric idea of viewing compaction, i.e., the ratio of time to view the full video segment to the time to view the surrogate. In the current study, the viewing compaction rate was approximately 15:1 (ranging from 8:1 for the fast forward surrogates for Moon and New Indians to 29:1 for the storyboards and slide shows of A Date with Your Family). These compaction rates were expected to be acceptable for users to be able to determine the gist of the video segment, based on Ding et al.'s [8] work with slide shows.

From the phase 3 data, we can conclude that this viewing time was reasonable. Mean scores on all performance measures were above the midpoint of the range of possible scores. However, participant comments often related to a desire for spending more time with the surrogate. Ryan was one of the people who commented on the brevity of his view of the storyboard: "For the amount of time, it was a lot of pictures – I couldn't take it all in." Almost all of the participants felt the slide show was too fast, e.g., Kevin's comment: "It still goes too fast, I think. It's too much information in too little time." Opinions of the speed of the fast forward were more mixed, and tended to be related to the participant's level of prior experience in using video. David, an experienced video user, spoke positively about the speed: "Another strength is that it's fast – you can see what it is and move on." Ling, a less experienced video user, believed that selection decisions would suffer because of the speed of the fast forward surrogate: "It's so fast, I think a lot of people [might be] interested in this [video], but it's shown so fast, you cannot [be] sure." The discrepancy between performance and satisfaction has often been reported in usability studies [19], and this study is no exception. Future studies should investigate surrogate use in a more naturalistic setting, with adequate user control over the speed of each surrogate and the number of times each surrogate can be viewed and where other contextual cues are present such as titles or an articulated query.

5.6 Effects of Video Content and Style

For this study, video segments were purposively selected from the Open Video repository to represent a variety of genre and styles. Based on participant comments and some of the performance data from phase 3, both user perceptions and performance can be affected by characteristics of the video segment itself. Of the three videos used in phase 3, New Indians was least well represented by its surrogates. There was a significant difference in the mean gist determination scores ($F=7.07$ with 2, 27 df, $p=0.0034$) and in the visual gist scores ($F=8.01$ with 2, 27 df, $p=0.0019$). In both cases, New Indians had lower performance than the other two videos (see Table 2).

Table 2. Differences in phase 3 performance, by video

Video	Mean gist determination score	Mean visual gist score
Moon	2.2	10.1
Hurricanes	1.7	10.5
New Indians	1.1	8.6

Two themes came out in participants' comments about video characteristics and their effects on use of the surrogates. The first theme was concerned with the variability in the key frames derived from the video segment. If the video was relatively homogeneous visually, none of the surrogates were very effective. For example, David, contrasted Master Hands, "It was easier to deal with [Master Hands] because there was more variety in the images, I think," with Apollo, "The interesting thing about this video is that the footage for this video is all very similar in value, so those frames are kind of hard to distinguish." It is likely that the homogeneity of content in the key frames from New Indians accounted for at least some of the difference in performance during phase 3. The second theme expanded on the first, as a stronger preference for keywords when the visual portions of the surrogate were not as useful. The low performance of the New Indians surrogates in phase 3 may also be attributable to the quality of the keywords associated with those surrogates. Matt pinpoints the importance of keywords for some videos in his comments about Apollo and Glen Canyon: "You're pretty much looking at so much of the exact same thing – I mean, it's not the exact same thing but really similar things – water going down or a plane flying around – that it would really help to have some additional idea of what's happening… Like more than just a visual stimulus – like audio [keywords] on top." It may be possible to take these differences in video style into account as surrogates are created, e.g., through some frame similarity adjustment analogous to IDF.

6 Conclusion

While the current study is an early and exploratory effort to understand how people interact with video collections and surrogates of video objects, its findings will be useful in shaping further research. Though no surrogate triumphed as the "best," the fast forward surrogate garnered substantial support from the study participants, particularly from experienced users of videos and video collections. The participants expressed their understanding of video gist as composed of the content or topic of the video, the story or narrative structure of the video, and the visual gist of the video (a combination of topicality, story, and visual style). The participants were successful in using the surrogates to determine gist, recognize objects and actions they had seen in the surrogates, and identify frames that "belonged" in a particular video (i.e., determine visual gist). Participants were able to identify several tasks for which surrogates would be useful, such as selecting videos from a collection, selecting and organizing frames and clips from a particular video, and identifying particular visual techniques used in a video. The compaction rates used in the surrogates allowed users to efficiently interact with them, but users expressed a preference for slowing them down or controlling their

viewing in other ways. Participants also noted some differences in which types of surrogates might be most useful for particular types of videos.

From these findings, several conclusions can be drawn. First, all the surrogates tested in this study are candidates for further development. The weakest was the slide show with text keywords (not preferred by anyone, rarely spoken of positively, not selected by anyone for use in phase 3), so its development is of lowest priority. Of highest priority is the further development of a fast forward surrogate (or surrogates) with audio keywords. In addition, research is needed to determine which information compaction rates result in the best viewing compaction rates for users (taking into account the tradeoff between viewing time and level of understanding).

The role of keywords is another area warranting further research. The participants in this study used the keywords for several purposes: to understand the content of the video (as expected), as advance organizers for viewing the visual portion of the surrogate, and as a source of ideas for terms to use in future searches. These last two uses were most clearly stated by Bettie, "It was telling me 'Neil Armstrong' and 'astronauts,' pointing me to what to look for, what to grasp" and by Kevin, "It tells you some of the topics, then you could go look those up online or in an encyclopedia or something." These surrogates are a hybrid of verbal and non-verbal information and the value of each type of information in representing video materials is worthy of additional study.

Once a suite of useful surrogates has been developed, the next step is to develop mechanisms with which users can control the display of the surrogates. First, users would like to have control over the display of each surrogate, e.g., the starting, stopping, and speed of the fast forward and slide show surrogates and the display time for the storyboard. As David said, "It all comes down.. to flexibility and control. You need to do different things at different moments." In addition, users would like to be able to move from surrogate to surrogate. They viewed different surrogates as being more or less useful for different types of tasks, and so would like to move from one to another easily. While some participants had a particular sequence in mind (e.g., David expressed the desire to search on text indexing, then move to storyboard, then move to fast forward), others expected the sequence of surrogate use to vary from situation to situation. It is in response for this need for flexibility that we are pursuing the development of the AgileViews user interface framework [16]. This framework defines several different views of a collection (including overviews, previews, reviews, peripheral views, and shared views), as well as control mechanisms that facilitate low-effort actions and strategies for coordinating the views. This and similar work, when implemented, will provide digital video library users with the tools they need for effectively retrieving, reviewing and extracting salience from video materials.

Acknowledgements

The research team would like to thank David Doermann, University of Maryland, for his assistance with the use of the Merit software and Curtis Webster for the additional support he has provided for this project. This research was supported by grant NSF IIS-0099638 from the National Science Foundation.

References

1. Borko, H., Bernier, C.: Abstracting Concepts and Methods. Academic Press, New York (1975).
2. Christel, M., Smith, M., Taylor, C. R., Winkler, D.: Evolving Video Skims into Useful Multimedia Abstractions. In: Proceedings of CHI '98: Human Factors in Computing Systems (Los Angeles, April 1998). ACM Press, New York (1998), 171-178.
3. Christel, M., Winkler, D., Taylor, C. R.: Improving Access to a Digital Video Library. Paper presented at the Human-Computer Interaction: INTERACT97, the 6th IFIP Conference on Human-Computer Interaction, Sydney, Australia, July 14-18, 1997.
4. Cohen, J.: A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement **20** (1960), 37-46.
5. Cooper, W. S.: A Definition of Relevance for Information Retrieval. Information Storage & Retrieval **7** (1971), 19-37.
6. Ding, W.: Designing Multimodal Surrogates for Video Browsing and Retrieval. Unpublished doctoral dissertation, University of Maryland (1999).
7. Ding, W., Marchionini, G., Soergel, D.: Multimodal Surrogates for Video Browsing. In: Proceedings of Digital Libraries '99: the Fourth Annual ACM Conference on Digital Libraries (Berkeley, CA, August 1999). ACM Press, New York (1999), 85-93.
8. Ding, W., Marchionini, G., Tse, T.: Previewing Video Data: Browsing Key Frames at High Rates Using a Video Slide Show Interface. In: Proceedings of the International Symposium on Research, Development and Practice in Digital Libraries (Tsukuba, Japan, 1997), 151-158.
9. Elliot, E.: Watch, Grab, Arrange, See: Thinking with Motion Images via Streams and Collages. MSVS thesis document., MIT Media Lab, Cambridge MA (1993).
10. England, P., Allen, R. B., Sullivan, M., Bianchi, M., Heybey, A., Dailianas, A.: Ibrowse: The Bellcore Video Library Toolkit. In: Proceedings of the SPIE Photonics West '96: Electronic Imaging Science and Technology '96: Storage and Retrieval for Image and Video Database IV (San Jose CA, January 1996).
11. Goodrum, A.: Evaluation of Text-Based and Image-Based Representations for Moving Image Documents. Unpublished doctoral dissertation, University of North Texas (1997).
12. Goodrum, A. A.: Multidimensional Scaling of Video Surrogates. Journal of the American Society for Information Science **52** (2001), 174-182.
13. Heilprin, L.: Paramorphism versus Homomorphism in Information Science. In: Heilprin, L. (ed.): Toward Foundations of Information Science. Knowledge Industry Pub., White Plains NY (1985), 115-136.
14. Komlodi, A., Marchionini, G.: Key Frame Preview Techniques for Video Browsing. In: Proceedings of the ACM Digital Libraries Conference '98 (Pittsburgh, PA, June 24-26, 1998). ACM Press, New York (1998).
15. Li, F., Gupta, A., Sanocki, E., He, L., Rui, Y.: Browsing Digital Video. In: CHI 2000 Conference Proceedings: Human Factors in Computing Systems (The Hague, Netherlands, April 3-6, 2000). ACM Press, New York (2000), 169-176.
16. Marchionini, G., Geisler, G., Brunk, B.: AgileViews: A Human-Centered Framework for Interfaces to Information Spaces. In: ASIS 2000: Proceedings of the 63rd ASIS Annual Meeting, Volume 37 (Chicago, November 12-16, 2000). Information Today, Medford, NJ (2000), 271-280.
17. Maybury, M.: Intelligent Multimedia Information Retrieval. MIT Press, Cambridge MA (1997).
18. Mills, M., Cohen, J., Wong, Y.: A Magnifier Tool for Video Data. In: Proceedings of CHI '92: Human Factors in Computing Systems (Monterey, CA, May 3-7, 1992). ACM Press, New York (1992), 93-98.

19. Nielsen, J., Levy, J.: Measuring Usability: Preference vs. Performance. Communications of the ACM **37** (April 1994), 66-75.
20. O'Connor, B.: Access to Moving Image Documents: Background Concepts and Proposals for Surrogates for Film and Video Works. Journal of Documentation **41** (1985), 209-220.
21. Robertson, G., Czerwinski, M., Larson, K., Robbins, D., Thiel, D., van Dantzich, M.: Data Mountain: Using Spatial Memory for Document Management. In: Proceedings of the 11th Annual ACM Symposium on User Interface Software and Technology (San Francisco, November 1998), 153-162.
22. Rorvig, M. E.: A Method for Automatically Abstracting Visual Documents. Journal of the American Society for Information Science **44** (1993), 40-56.
23. Saracevic, T.: The Concept of "Relevance" in Information Science: A Historical Review. In: Saracevic, T. (ed.): Introduction to Information Science. Bowker, New York (1970), 111-151.
24. Slaughter, L., Shneiderman, B., Marchionini, G.: Comprehension and Object Recognition Capabilities for Presentations of Simultaneous Video Key Frame Surrogates. In: Research and Advanced Technology for Digital Libraries: Proceedings of the First European Conference (EDSL '97, Pisa, Italy, 1997), 41-54.
25. Smeaton, A.: Content-based Access to Digital Video: The Físchlár System and the TREC Video Track. Paper presented at MMCBIR 2001 - Multimedia Content-based Indexing and Retrieval (INRIA, Rocquencourt, France, September 2001). http://www.cdvp.dcu.ie/Papers/MMCBIR2001.pdf. Last accessed January 26, 2002.
26. Smith, M., Kanade, T.: Video Skimming and Characterization through the Combination of Image and Language Understanding. In: Proceedings of the 1998 IEEE Workshop on Content-based Access of Image and Video Databases (Bombay, India, January 1998). IEEE, Los Alamitos CA (1998), 61-70.
27. Tonomura, Y., Akutsu, A., Otsuji, K., Sadakata, T.: VideoMAP and VideoSpaceIcon: Tools for Anatomizing Video Content. In: Proceedings of INTERCHI '93: Human Factors in Computing Systems (Amsterdam, April 1993), 131-136.
28. Tse, T., Marchionini, G., Ding, W., Slaughter, L., Komlodi, A.: Dynamic Key Frame Presentation Techniques for Augmenting Video Browsing. In: Proceedings of AVI '98: Advanced Visual Interfaces (L'Aquila, Italy, May 1998), 185-194.
29. Turner, J.: Determining the Subject Content of Still and Moving Image Documents for Storage and Retrieval: An Experimental Investigation. Unpublished doctoral dissertation, University of Toronto (1984).
30. van Dijk, T. A., Kintsch, W.: Strategies of Discourse Comprehension. Academic Press, New York (1983).
31. van Dijk, T. A., Kintsch, W.: Toward a Model of Text Comprehension and Production. Psychological Review **85** (1978), 363-394.
32. Wactlar, H., Christel, M., Gong, Y., Hauptmann, A.: Lessons Learned from Building a Terabyte Digital Video Library. Computer **32** (2, 1999), 66-73.
33. Yeo, B.-L., Yeung, M.: Retrieving and Visualizing Video. Communication of the ACM **40** (December 1997), 43-52.
34. Zhang, H. J., Low, C. Y., Smoliar, S. W.: Video Parsing and Browsing Using Compressed Data. Multimedia Tools and Applications **1** (1995), 89-111.