Measuring User Performance During Interactions with Digital Video Collections

Meng Yang, Barbara M. Wildemuth, Gary Marchionini, Todd Wilkens, Gary Geisler, Anthony Hughes, Richard Gruss, and Curtis Webster

Open Video Project, Interaction Design Lab, School of Information and Library Science, CB #3360, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-3360

With more and more digital videos found online, video retrieval researchers have begun to create various representations or surrogates for digital videos, such as poster frames, storyboards, video skims and fast forwards. How to evaluate the effectiveness of these video surrogates has become an issue for researchers. This paper proposes two general classes of user tasks—recognition tasks and tasks requiring inference—for which performance measures were developed. The measures include graphical object recognition, textual object recognition, action recognition, free-text gist determination, multiple-choice gist determination and visual gist determination. The preliminary results from two user studies applying these six measures are also discussed in this paper.

INTRODUCTION

With the development of video compression and video retrieval technology, more and more digital videos can be found online. Video retrieval researchers have, therefore, begun to create various representations, or surrogates, of video objects, intended to support users' searching and browsing. We define a video surrogate as a compact representation of the original video that shares major attributes with the object it represents. Within the context of digital video collections, varied types of video surrogates have been created and their use studied. They include poster frames, filmstrips and skims (Christel, Winkler & Taylor, 1997); storyboards and slide shows (e.g., Tse et al., 1998; Wildemuth et al., 2002); fast forwards (Wildemuth et al., 2003); and automatically-generated summaries (He et al., 1999).

Video surrogates can be classified based on the medium in which they are represented: text surrogates, still image surrogates, moving image surrogates, audio surrogates, and the combination of these different types – multimodal surrogates (see Table 1, next page). *Text surrogates* include all kinds of bibliographic information or metadata about the video. *Still image surrogates* represent the video content by extracting key frames, a natural analogue to keywords in text. Poster frames (Christel et al., 1997), storyboards, slide shows (Tse et al., 1998) and video streams (Elliot, 1993) are all still image surrogates. A *moving image surrogate* is more similar to the original video content since it contains action. Video skims (Christel et al., 1997) and fast-forward surrogates (Drucker et al., 2002; Wildemuth et al., 2003) are moving image surrogates. *Audio surrogates* use extracted audio information such as environmental sounds, music or people's dialogues to represent the video content. Multimodal surrogates combine textual, visual and audio information (Ding, Soergel, & Marchionini, 1999).

As video retrieval researchers have developed these surrogates, there is also a need to evaluate their effectiveness (Goodrum, 2001). However, the methods used to evaluate surrogates in textual databases are inappropriate, since the measures are also text-based and so are limited in their ability to tap the multimedia characteristics of video surrogates. The purpose of this paper is to describe and discuss the strengths and weaknesses of six measures developed in evaluating the effectiveness of video surrogates in terms of user performance.

Table 1. Examples of video surrogates

Type of surrogate	Examples
Text surrogate	Title, keyword, description
Still image surrogate	Poster frame, storyboard/filmstrip, slide show, video stream, key-frame-based table of contents
Moving image surrogate	Skim, fast forward
Audio surrogate	Spoken keywords, environmental sounds, music
Multimodal surrogate	Text surrogate + still image surrogate, still image surrogate + audio surrogate

THEORETICAL BACKGROUND

In order to define and operationalize performance measures appropriate to video retrieval research, we must take into account the purposes for which people might

interact with video surrogates. These purposes might include selecting a video for viewing or showing to someone else, or selecting particular clips or frames from the video for inclusion in another video they are producing (Wildemuth et al., 2002). To accomplish these purposes, they need to be able to perform specific tasks during their interactions with the video surrogates. For example, it can be argued that being able to determine the gist of a video is useful in selecting a video for viewing, while the ability to see and remember particular objects in a surrogate is useful in selecting frames for re-use. Because tasks such as gist determination and object recognition can be more concretely defined and operationalized than the user's general purposes, measuring people's performance of these tasks can be used to evaluate the relative quality of different video surrogates. In order to select tasks appropriate for further development as the basis for performance measurement, the cognitive mechanisms by which viewers perceive images and motion pictures should be considered carefully.

Most studies of video/image retrieval suggest that people interact with images/videos at three levels (Eakins & Graham, 1999; Greisdorf & O'Connor, 2002). At the most basic level, primitive features of the image (e.g., color, shape) are perceived. At a second level, logical features (e.g., people, things, places, actions) are perceived. At this level, people draw on their existing knowledge to identify the objects perceived. The third level requires inductive interpretation of the image/video, with inferences being made about its abstract attributes, including emotional cues and atmosphere (Greisdorf & O'Connor, 2002). This three-level hierarchy is remarkably similar to Panofsky's earlier (1955, 1972) description of three levels of comprehension for visual images: pre-iconographical, iconographical, and iconological.

A similar perspective is represented by Grodal's (1997) proposal that a viewer's processing of film consists of four main steps. The first step is basic perception, i.e., the initial visual analysis of textures, lines and figures. The second step is memory matching, aided by the viewer's familiarity or unfamiliarity with the content of the film. Step three is the cognitive-emotional appraisal and motivation phase, and leads to step four, reactions at a high level of arousal, such as fear or happiness.

From these earlier theories of people's interactions with images/video, we can conclude that there are two general categories of perception: low-level visual perception/identification and high-level cognitive/affective understanding. Similar theories have been used to explain people's reading process: "Simply stated, reading involves an array of lower-level rapid, automatic identification skills and an array of higher-level comprehension/interpretation skills" (Grabe, 1991, p. 383). Therefore, two general classes of tasks can be defined to evaluate the effectiveness of video surrogates: *recognition tasks* and *tasks requiring inference*.

PRIOR EMPIRICAL WORK

Some prior work has been done to learn how people interact with and use video surrogates, primarily by the Informedia project at Carnegie Mellon University, by Goodrum and Rorvig (Goodrum, 1997, 2001) at the University of North Texas, and by the Digital Library Research Group at the University of Maryland. The methods they used to evaluate surrogate effectiveness are reviewed here.

The Informedia project (Christel et al., 1997) compared three video surrogates (a text list, opening shot poster frames and query-based poster frames). To evaluate user performance, they used three variables: scores on a question set, the time spent to answer the question set, and subjective satisfaction as measured with the Questionnaire for User Interface Satisfaction (QUIS; Chin et al., 1988). In a comparison of several video skims (Christel et al., 1998), they also utilized a fact finding measure and a gist determination measure. In the fact finding measure, subjects were given a question and asked to navigate to that region of the video presenting the answer. In the gist determination measure, subjects chose from text-phrase and thumbnail-image menus those items that best represented the material covered by the skim.

Goodrum and Rorvig's (1997, 2001) work evaluated the ability of a surrogate to enable users to make the same distinctions that they would make if they viewed the full video. Four types of video surrogates (titles, keywords, salient still frames (i.e., poster frames), and multiple key frames) were compared under three conditions representing three levels of search specificity. Subjects were asked to render similarity judgments for all pairs of videos. Multidimensional scaling (MDS) was used to map the dimensional dispersions, and the maps stimulated by each surrogate could be compared with those stimulated by the full video.

Studies conducted by researchers at the University of Maryland (Ding, Marchionini, & Tse, 1997; Komlodi & Marchionini, 1998; Slaughter, Shneiderman, & Marchionini, 1997) compared a number of still image surrogates and variations on their display. These studies incorporated measures of gist determination, action identification, object recognition, and user perceptions of slide show speed. Gist determination was defined as users' ability to determine the overall meaning of a video from viewing only the video surrogate. Gist determination was measured in two ways: (1) users wrote a gist description themselves (which was also analyzed for a measure of action identification) and (2) users selected a gist description from a set of statements created by the researchers. Object recognition was defined as the users' ability to remember whether particular objects appeared in the video surrogate. Two methods were used to measure object recognition: one in which the stimulus objects were represented linguistically (with object

names), and the other in which the stimuli were represented graphically (with key frames). User perceptions of slide show speed were measured by questionnaires using a seven point Likert scale from too slow (1) to too fast (7).

Table 2. Metrics to evaluate the effectiveness of video surrogates

	Text	Still image	Action
Recognition	Object recognition (text)	Object recognition (graphical)	Action recognition
Inference	Gist determination (free text) Gist determination (multiple-choice)	Visual gist determination	

The measures originally developed at Maryland form the basis for several of the measures described and discussed here. These measures have been developed for user studies as part of the Open Video Project (see Marchionini & Geisler, 2002, for an overview of the project). This project (http://www.open-video.org/) aims to develop and test surrogates to inform the design of user interfaces for digital video environments and to understand the nature of how people make sense of video content. Results from two new studies using these measures are presented, along with descriptions of modifications made to the measures. Also, the initial development of a measure of visual gist determination is described, as well as results from its use.

MEASURES OF USER PERFORMANCE WITH VIDEO SURROGATES

As noted above, our performance measures are related to two general categories of tasks: recognition and inference (see Table 2). The *recognition* tasks include object recognition (textual or graphical) and action recognition, and correspond to the first two steps (visual analysis and memory matching) in Grodal's (1997) flow diagram. These tasks are defined as users' ability to remember seeing particular objects or actions in the video surrogates, and our measures of task performance use, respectively, textual, still image, and moving image stimuli. The *inference* tasks include gist determination (free-text or multiple-choice) and visual gist determination (incorporating stylistic as well as topical considerations), and correspond to the last two steps in Grodal's (1997) flow diagram (construction of narrative scene or universe, and reaction). To measure performance on these tasks, study participants are asked to determine the gist or visual gist of the video, based only on viewing the surrogate. Stimuli for these gist determination measures include both text and still images.

Two user studies were conducted based on these measures. The first study (Wildemuth et. al., 2002) examined the effectiveness of five different kinds of video surrogates. Ten participants each interacted with one surrogate (selected by the participant) for each of three video segments. The surrogates included in the evaluation were: storyboard with text keywords (6 observations), storyboard with audio keywords (4 observations), slide show with audio keywords (6 observations), and fast forward (14 observations).[1] In total, this initial trial data included 30 observations for each of the performance measures. The second study (Wildemuth et al., 2003) tested different speeds of the fast forward surrogate. Four fast forward speeds (1:32, 1:64, 1:128 and 1:256) were examined for four video clips. Each of the 45 participants in this study interacted with four video surrogates. In total, this study included 180 observations for each measure. The remainder of this section describes the six measures and discusses results based on their use in these two studies.

Object recognition (graphical and textual)

Object recognition is defined as the user's ability to recall which objects were seen in a video surrogate recently viewed. The rationale for including this task in an evaluation of video surrogates is that it is closely related to the users' real-world purpose of selecting particular frames or segments for later re-use—a purpose described as important by participants in the first phase of Wildemuth et al. (2002). If a person performs well on measures of object recognition, it can be argued that the surrogate adequately supports frame selection.

Two parallel versions of the object recognition measure were used: one using textual stimuli and the other using graphical stimuli. In each version, a set of stimuli was presented to the study participant, who was asked to mark whether each object had been seen in the surrogate or not.

In the textual object recognition measure, the stimuli were 12 object names. Of these, six were selected from frames seen in the video surrogate being evaluated; six (distractors) were names of objects not in the video surrogate. Within each set of six, three were concrete objects and three were abstract objects. For example, Table 3 (next page) shows the stimuli used for a video segment of the documentary video "Iran: Between Two Worlds".

The graphical version of this measure used a set of 12 key frames as stimuli. Of these, six were selected from the video surrogate being evaluated; three were selected from a different video that was similar in style (though not necessarily content) to the target video; and three

[1] The fifth surrogate, a slide show with text keywords, was not selected by any participant for further use during this second phase of the study.

were selected from a different video that was different in style from the target video. Though the last set of distractors was selected from a video that was different in style, the color status (color versus black and white) of the frames serving as stimuli was held constant, e.g., if the target video was in black and white, all 12 key frames serving as stimuli were black and white. Figure 1 illustrates the stimuli used for a documentary video titled "Iran: Between Two Worlds".

Table 3. Examples of textual object recognition stimuli

Objects seen in video surrogate		Objects not seen in video surrogate	
Concrete	**Abstract**	**Concrete**	**Abstract**
Power plant	Archaeology	Jeep	Warfare
Wall carving	Craftsmanship	Pottery	Storm
Fountain	Middle east	Pyramid	Computer technology

Table 4. Results on object recognition measures

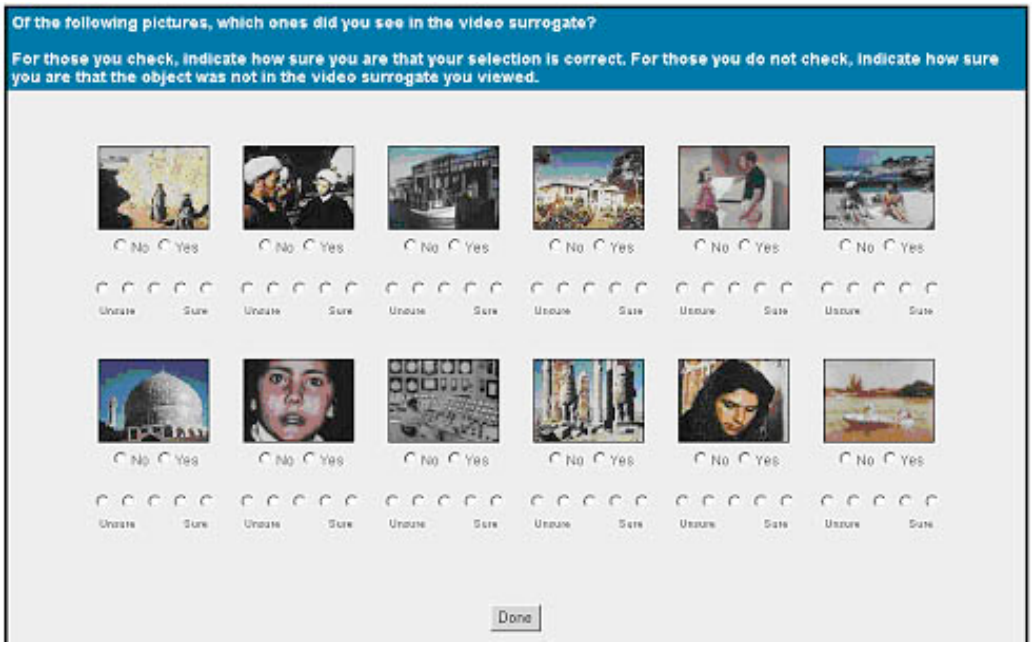
	Study 1: 10 participants, 30 observations		Study 2: 45 participants, 180 observations	
	Mean	**s.d.**	**Mean**	**s.d.**
Object recognition (textual)	9.0	1.61	8.6	1.35
Object recognition (graphical)	9.0	1.81	9.7	1.65

Performance scores on the object recognition measure can range from 0 to 12, since each yes/no response could be right or wrong. Performance scores on the two object recognition measures in the two studies are reported in Table 4. The relatively high scores indicate that this measure was not particularly difficult for the study participants. The results corroborate Shepard's (1967) finding that people have very accurate recognition memory for pictures (96.7%).

Preliminary evidence that these two measures are tapping the same underlying construct (i.e., object recognition) is mixed. In the first study, the two measures were correlated (r=0.34, p=0.0633), but in the second study they were not (r=-0.09, p=0.2512). Thus, until these two measures are better understood, they cannot be considered redundant.

Action recognition

Action recognition is defined as the study participants' ability to recognize whether a particular short action sequence appeared in the video surrogate. Like the object recognition task, the rationale for evaluating action recognition is grounded in users' reports of their goal of selecting particular video clips (Wildemuth et al., 2002). Therefore, a measure of action recognition was newly developed for use in the Open Video Project studies.



The stimuli for the action recognition measure were six brief clips from the original video, each 2-3 seconds long. Of the six, two were selected from the video represented in the video surrogate (i.e., were correct), two were from another video of a style similar to that of the target video and two were from a another video of a style that was different from that of the target video. In response to each clip, the participant was asked whether s/he believed it to be from the video segment represented by the surrogate.

The scoring for this measure was comparable to that used with the object recognition measure; each yes/no response to a clip was scored as right or wrong. The maximum possible score was 6. The 10 study participants from the first study scored an average of 4.6 (s.d.= 1.0). In the second study, the average score of the 45 participants was 4.5 (s.d. = 0.93). Like the object recognition measures, it can be concluded that achieving success on this measure is not particularly difficult. These scores were not correlated with either of the object recognition measures, suggesting that action recognition can be distinguished from object recognition.

Gist determination: inferring meaning from the video surrogate

One of the goals of a video surrogate is to support the viewer's ability to infer the gist of the full video from viewing only the surrogate. This is particularly important for searching and browsing in general and is especially crucial for video repositories, where the video files are very large and so require a long time to download. If the surrogate supports this task well, the user is able to make accurate relevance judgments or selection decisions about videos, thus saving time for the user. Study participants (Wildemuth et al., 2002) highlighted gist determination as the most important function of video surrogates. However, there was not consensus on the meaning of "gist", with three different understandings presented in their comments: gist as aboutness, i.e., the topic of the video; gist as the "story" of the video or its narrative structure; and a third understanding that we are calling visual gist, to be discussed in the next section. Two measures of gist determination were developed for our current studies: free-text and multiple-choice.

The free-text version of the measure asks the user to "write a brief summary of the video". Generally, the descriptions provided by the first study participants were quite short (the longest was 55 words; the shortest was 3 words). In the second study, they tended to be somewhat longer, but still varied in length (from 3 to 176 words). An example of the gist statements generated for "Iran: Between Two Worlds" was: "A possible documentary about somewhere in the Middle East, maybe Egypt based on some of the artifacts that were shown at the beginning. Then it moved on to talk more about the daily life of the people living in that place". Note that, in spite of only viewing a few key frames or a moving image surrogate, people are able to fill in considerable details about a video using their personal knowledge and inferential abilities.

Once these gist descriptions are generated by the study participants, they must be scored. The first study employed a simple 3-point scoring procedure developed by Tse et al. (1998), but it was found to be unreliable (70% agreement between two independent raters on two of the videos, but only 30% agreement on the third video). A new scoring scheme was developed for our second study; it included two scores (correctness/accuracy and level of detail) on each of two dimensions (objects/events and higher-level perspective). Two members of the research team independently scored the 180 gist statements from the second study, and their scores were strongly correlated ($r=0.76$, $p<0.0001$); this level of reliability was considered satisfactory.

The second gist determination measure gives the user five candidate gist descriptions written by members of the research team and asks him/her to select the one that best describes the video represented by the surrogate. Note that considerable discussion led to these statements. They were drafted by individuals and discussed by the group before adoption. For example, the candidate gist descriptions for "Iran: Between Two Worlds" were:

- It shows a documentary style look at a Middle Eastern country. Past history of art and culture is examined. Modern day commercial and social activities are presented. (correct)
- It shows an overview of Middle Eastern royalty. Royal customs and architecture are explained and documented.
- It tells us how Middle Eastern people live in poverty. It describes their hardships in trying to cope with modern society.
- It documents a Middle Eastern family's life. Daily depictions of their usual activities are shown. It is revealed how desperately they need humanitarian aid.
- It tells us about the modern commercial life of a Middle Eastern country. Business and industrial processes are examined. Their impact on the environment is noted.

On the free-text gist determination measure, results are reported only for the second study, since scores from the first study are considered unreliable. Participants averaged 2.9 (out of a possible 8 points; s.d.=1.72). On the multiple choice gist determination measure, participants got 80% correct in the first study and 46% right in the second study. Based on these data, the multiple choice measure was easier, as would be expected. There is some indication that performance on these two measures is related (based on data from the second study). Those who selected the correct statement on the multiple choice measure also scored higher on the free text measure (mean = 3.1) than those who were incorrect on the multiple choice measure (mean = 2.7; $t=1.79$ with 178 df, $p=0.0759$). Although these data are only

suggestive, this relationship does merit further investigation.

Visual gist determination

Visual gist is currently defined as the viewer's overall understanding of the video, including both its content and its cinematic style. Based on the comments of the first study participants, it is a combination of topicality, structure/form, and visual style. These elements combine to provide a visceral 'feel' for the video. While this concept needs additional clarification, participant comments clearly indicated that they formed a more holistic view of gist, beyond topicality. Having this more complete understanding of a video will support the user in making accurate selection decisions after viewing only the surrogate. It will also help us to create better surrogates to support user goals such as finding segments for difficult to index constructs such as "entertaining" or other affective characteristics.

Clearly, operationalizing a construct that is in such an early stage of being defined is a challenge. In our studies, we provided participants with a set of 12 stimuli (i.e., key frames), none of which actually appeared in the surrogate viewed by the study participants. Participants then received the following instruction: "Of the following frames, check the ones you think belong in this video." The interviewer also read this instruction to each participant, to ensure that the participant distinguished this task from the earlier graphical object recognition task.

Six of the key frames were selected from the target video (but had not been seen in the surrogate). Of the remaining six key frames, three were selected from a different video of a similar style and three were selected from a different video of a different style. The scoring method was comparable to that used for the recognition measures.

The mean score in the first study was 9.7 (of a possible 12; s.d.=1.36), and in the second it was 8.4 (s.d.=1.41). Thus, as in the other gist determination measures, subjects were mainly successful in identifying key frames that 'belonged' with the viewed surrogate. As might be expected, there is some evidence that this score may be related to performance on the free-text gist determination measure. Scores on the two measures (study 2) were positively correlated (r=0.31), though this correlation did not reach statistical significance (p=0.0977). No statistically significant relationship with the multiple-choice gist determination measure was found in either study. Visual gist determination was related to object recognition (textual) in the second study (r=0.20, p=0.0064).

By looking at the relationships between the visual gist determination measure and the other measures, we find mixed evidence concerning the value of this measure of performance. This evidence can be interpreted using Grodal's (1977) level of perception as a guide. The visual gist scores were generally not related to the action or object recognition scores (with the exception of one relationship in the second study). This lack of relationship may be attributable to the fact that the recognition tasks require the study participants only to recall what they have seen, while the visual gist task requires them to make inferences about what they have seen. The possibility of a relationship between the visual gist task and the free-text gist determination task seems more likely. While statistical significance was not achieved in these studies, data from study 2 suggests that further refinement of the scoring procedures on the gist determination measure may reveal that visual gist determination and the more traditional gist determination constructs are related, both requiring the study participants to make inferences. As these measures will be used in future studies, their psychometric properties will continue to be investigated and the measures themselves refined based on these findings.

DISCUSSION

Additional considerations in using the proposed measures

As these measures were implemented in the studies described here, a variety of issues arose. While they were resolved for the purposes of these studies, they may be resolved differently within the context of different studies. The issues, discussed below, include the order in which the measures should be administered, the effects of user confidence on performance, and the effects of video characteristics on performance.

Order of administration. For the first study, the two gist determination measures were administered first (free-text before multiple-choice), because they were assumed to be the most important in relation to user interactions with video. Next, the three recognition measures were presented, and finally the visual gist determination measure. Some problems were found with this ordering. In particular, the multiple-choice gist determination measure, by providing alternative hypotheses about the gist of the video, could affect participants' understandings of the video, thus affecting their performance on the later measures. This problem can be interpreted in light of Van Dijk and Kinsch's model of discourse comprehension (1978, 1993), in which 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. Although this model studies how people interact with text information, the same strategy also likely applies to video viewers. Therefore, in the second study, the multiple-choice gist determination measure was moved to the very end, after subjects finished all the other measures. Although this seemed like a better ordering in the second study, the issue of learning as users complete multiple measures is important to keep in mind when designing studies and interpreting results. Because there are so many other primary variables to try and control (e.g., video content,

surrogate type, user characteristics), making the type of measure an independent variable seems an extreme expense to incur even with counterbalancing techniques.

Effects of differences in confidence levels. During the first study, a number of participants commented on their confidence in completing the performance measures, particularly the object recognition measures. Specifically, sometimes they felt quite sure that they saw the frame or object when they watched the surrogate and sometimes they could not easily decide whether they had seen it or not. The second study incorporated a direct measurement of each subject's confidence on each selection, using a five-point rating scale with each yes/no selection in each measure. The confidence data from the spring 2002 study will be analyzed to determine whether confidence level is associated with the correctness of the selection (or non-selection), the act of selecting (versus non-selecting), characteristics of the surrogate (i.e., the speed of the fast forward being evaluated), or characteristics of the video (i.e., whether the video had a narrative or a categorical structure/form).

Effects of video characteristics. In the first study, participants commented on the videos' structure/form (e.g., narrative vs. categorical structure/form)[2] and style (e.g., color vs. black and white) and their effects on perceptions of video content. Subjects preferred color to black and white segments, and they said they could determine the topic of a film organized by categories much easier than a narrative film. In addition to these video attributes, other aspects of the video's style (e.g., its pace or the style of the audio track) may affect a person's ability to interact effectively with a surrogate of the video. To begin to investigate some of these effects, the second study controlled for video structure/form and color status through the selection of the target videos: two color and two black and white; two narrative and two with categorical structure. Each subject completed all six measures for each of these four videos. The video's structure/form affected scores on gist determination (multiple choice); people found it easier to determine the gist of videos with categorical structure/form. The video's color status affected performance on gist determination (multiple choice) (performance was higher on color videos), object recognition (graphical) (performance was higher for black and white videos), and action recognition (performance was higher for color videos). Addition studies using additional videos are needed to more fully understand how the video's characteristics affect performance as people interact with the video surrogates.

[2] Categorical films use subjects or categories as a basis for their syntactic organization, typically basing each segment of the film on one category or subcategory. Narrative films use cause-effect, time, and space as a basis for syntactic organization (Lindley and Nack, 2000; Bordwell and Thompson, 1997).

Limitations of the measures

While we believe these measures provide a strong starting point for achieving the goal of measuring user performance when interacting with video surrogates, they do have limitations. For example, they are performance measures, and do not measure people's preferences in relation to the characteristics of alternative video surrogates. Future studies will be augmented with the addition of satisfaction measures such as the Questionnaire on User Interaction Satisfaction (QUIS). Secondly, there are some interactions between these measures whenever they are used in combination (i.e., participants learn from the stimuli used in each measure, as well as learning from the surrogate being evaluated). While changing the order in which the measures are administered helps to alleviate some of these problems, it may be necessary to select just a few of the measures for any particular study. This approach will focus attention on fewer measures for each study, but will minimize contamination from using multiple measures. Most importantly, it is important to consider that the tasks represented in these measures are limited by their isolated and laboratory application. The tasks were isolated from a larger search or browsing episode in that surrogates were presented without the surrounding context of a query and set of results. This was quite purposeful in these studies, as we aimed to isolate the effects of surrogate quality from the effects of other aspects of the search context. However, laboratory study results can be applied to naturalistic settings only with caution. Our next studies will continue to be conducted in the usability lab but will be embedded within the context of complete search episodes and will provide some opportunities for users to define their own search objectives.

CONCLUSION

This paper describes and presents initial data from the use of six measures that evaluate people's performance when interacting with alternative video surrogates. They correspond to two general cognitive processes: *recognition* (textual object recognition, graphical object recognition, and action recognition) and *inference* (free-text gist determination, multiple-choice gist determination, and visual gist determination). This categorization of these measures is consistent with the cognitive processes through which viewers perceive and understand images and videos (Eakins & Graham, 1999; Greisdorf & O'Connor, 2002; Grodal, 1997; Panofsky, 1955). The three recognition measures require that the user recognize objects or actions that appeared in the video surrogates viewed, and are associated with the users' needs to select video frames or clips for re-use. The gist determination measures require that the user infer an understanding of the full video from only the surrogate, and are associated with the users' needs to select videos that are relevant for particular purposes. While some additional development of the measures is needed, their initial field testing indicates that they are

practical and can differentiate multiple levels of performance. These measures will continue to be refined as they are used in studies conducted by the Open Video project. We also encourage other researchers to employ them in video retrieval research.

ACKNOWLEDGMENTS

This paper is based upon work supported by the National Science Foundation under Grant No. IIS 0099638.

REFERENCES

Bordwell, D., & Thompson, K. (1997). *Film Art: An Introduction.* 5th ed. New York: McGraw-Hill.

Chin, J.P., Diehl, V.A., & Norman, K.L. (1988). Development of an instrument measuring user satisfaction of the human-computer interface. *CHI '88 Conference Proc. (Washington, DC, May 15-19, 1988)*, 213-218.

Christel, M., Smith, M., Taylor, C. R. & Winkler, D. (1998). Evolving video skims into useful multimedia abstractions. *Proc. of CHI '98(Los Angeles, April 18-23, 1998),* 171-178.

Christel, M., Winkler, D. & Taylor, C. R. (1997). Improving access to a digital video library. Paper presented at INTERACT97, the 6th IFIP Conference on Human-Computer Interaction (Sydney, Australia, July 14-18, 1997).

Ding, W., Marchionini, G. & Tse, T. (1997). Previewing video data: browsing key frames at high rates using a video slide show interface. *Proc. of the International Symposium on Research, Development and Practice in Digital Libraries (Tsukuba, Japan),* 151-158.

Ding, W., Soergel, D., & Marchionini, G. (1999). Performance of visual, verbal, and combined video surrogates. *Proc. of the 62nd ASIS Annual Meeting (Washington, DC, October 31-November 4, 1999)*, 651-664.

Drucker, S., Glatzer, A. De Mar, S. & Wong, C. (2002). SmartSkip: Consumer level browsing and skipping of digital video content. *Proc. of CHI '02 (Minneapolis, MN April 20-25, 2002),* 219-226.

Eakins, J. P., & Graham, M. E. (1999, Jan.) Content-based image retrieval: a report to the JISC Technology Applications Programme. Institute for Image Data Research, University of Northumbria at Newcastle. http://www.unn.ac.uk/iidr/report.html. Last visited May 14, 2003.

Elliot, E. (1993). Watch, grab, arrange, see: thinking with motion images via streams and collages. MSVS thesis document. Cambridge, MA: MIT Media Lab.

Goodrum, A. (1997). Evaluation of text-based and image-based representations for moving image documents. Unpublished doctoral dissertation, University of North Texas.

Goodrum, A.A. (2001). Multidimensional scaling of video surrogates. *Journal of the American Society for Information Science, 52*(2), 174-182.

Grabe, W. (1991). Current developments in second language reading research. *TESOL Quarterly*, *25*(3), 375-406.

Greisdorf, H., & O'Connor, B. (2002). Modelling what users see when they look at images: a cognitive viewpoint. *Journal of Documentation, 58*(1), 6-29.

Grodal, T. (1997). *Moving Pictures --- A New Theory of Film Genres, Feelings, and Cognition*. Oxford: Clarendon Press, 59-61.

He, L., Sanocki, E., Gupta, A., & Grudin, J. (1999). Auto-summarization of audio-video presentations. *Proc. of the 7th ACM International Conference on Multimedia (Part 1),* 489-498.

Komlodi, A. & Marchionini, G. (1998). Key frame preview techniques for video browsing. *Proc. of the ACM Digital Libraries Conference '98 (Pittsburgh, PA, June 24-26, 1998)*, 118-125.

Lindley, C. A., & Nack, F. (2000). Hybrid narrative and categorical strategies for interactive and dynamic video presentation generation. *The New Review of Hypermedia and Multimedia, 6.* London: Taylor Graham.

Marchionini, G. & Geisler, G. (2002). The Open Video Digital Library. *dLib Magazine, 8*(12). http://www.dlib.org/dlib/december02/marchionini/12marchionini.html.

Panofsky, E. (1955). *Meaning in the Visual Arts: Meanings in and on Art History.* Doubleday.

Panofsky, E. (1972). *Studies in Iconology: Humanistic Themes in the Art of the Renaissance.* New York: Harper & Row.

Shepard, R. N. (1967). Recognition memory for words, sentences, and pictures. *Journal of Verbal Learning and Verbal Behavior*, *6*, 156-163.

Slaughter, L., Shneiderman, B., & Marchionini, G. (1997). Comprehension and object recognition capabilities for presentations of simultaneous video key frame surrogates. *Research and Advanced Technology for Digital Libraries: Proc. of the First European Conference (EDSL '97, Pisa, Italy),* 41-54.

Tse, T., Marchionini, G., Ding, W., Slaughter, L., & Komlodi, A. (1998). Dynamic key frame presentation techniques for augmenting video browsing. *Proc. of AVI'98: Advanced Visual Interfaces (L'Aquila, Italy, May 25-27, 1998),* 185-194.

Van Dijk, T. A., & Kintsch, W. (1983). *Strategies of Discourse Comprehension*. New York: Academic Press.

Van Dijk, T. A., & Kintsch, W. (1978). Toward a model of text comprehension and production. *Psychological Review, 85*(5): 363-394.

Wildemuth, B. M., 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. Presented at the European Conference on Digital Libraries (ECDL), September 2002.

Wildemuth, B. M., 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., Joint Conference on Digital Libraries* (in press).