Measures of User Performance in Video Retrieval Research

by

Meng Yang, Barbara M. Wildemuth, Gary Marchionini,
Todd Wilkens, Gary Geisler, Anthony Hughes, Rich Gruss, and Curtis Webster
Open Video Project, Interaction Design Lab,
School of Information and Library Science, University of North Carolina at Chapel Hill

SILS Technical Report TR-2003-02

June 2003

Table of Contents

Abstract

Browsing and searching for digital videos online is not as easy as it is with text documents. To address this problem, researchers have begun to create video surrogates to represent video objects. The purpose of this paper is to describe and provide preliminary data regarding six measures that can be used to evaluate the effectiveness of people's interactions with video surrogates. The six types of performance to be measured are:

- *Object recognition (with text stimuli);*
- *Object recognition (with graphical stimuli);*
- *Action recognition;*
- *Gist determination (free text);*
- *Gist determination (multiple choice); and*
- *Visual gist determination.*

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 with video surrogates. These measures will continue to be refined in studies conducted by the Open Video project; we also encourage others to employ them in video retrieval research.

1 Introduction

With the development of video compression and video retrieval technology, more and more digital videos can be found online, providing web users with more vivid and exciting information in addition to text materials. However, people also find that searching for and browsing digital videos may not be as easy as with text documents. To address this problem, video retrieval researchers have begun to create video surrogates that can represent, or stand in for, video objects. A number of such surrogates have been developed, and it is now necessary to evaluate their effectivness. The purpose of this paper is to describe and provide preliminary data regarding six measures that can be used to evaluate the effectiveness of people's interactions with video surrogates.

After providing some background information on video surrogates and the video retrieval research from which they were developed, six measures of user performance will be defined. These measures were developed as part of the Open Video project (www.open-video.org; Marchionini & Geisler, 2002), and have been used within the context of the project's research program. Data from Open Video studies will be used to explore the validity of the proposed measures.

2 Video Surrogates

Video retrieval researchers have begun to create representations of video objects intended to support users' search and browsing processes. These representations may be called video abstractions, video summaries, or video surrogates. The Informedia project generates three different kinds of *video abstractions*: poster frames, filmstrips and skims (Christel, Winkler & Taylor, 1997). The Digital Library Research Group at the University of Maryland has created two types of *video surrogates*: storyboards and slide shows to represent the whole video content (Tse, Marchionini, Ding, Slaughter, & Komlodi, 1998). *Video summaries* or *summarizations* are also used by other researchers (e.g., He, Sanocki, Gupta, & Grudin, 1999). *Video surrogate* is the term used in this paper, and is defined as a compact representation of the original video that shares major attributes with the object it represents.

Video surrogates can be classified into several groups, based on their medium: text, still image, moving image, audio, and multimodal–combinations of the other media (see Table 1). *Text surrogates* include all kinds of bibliographic information or metadata about the video, such as title, author, description and index terms. *Still image surrogates* represent the video content through extracted key frames. Key frames in videos are a natural analogue to keywords in text and can give viewers vivid and concrete information about the original video (O'Connor, 1985). Poster frame (Christel et al., 1997), storyboard, slide show (Tse et al., 1998) and video streams (Elliot, 1993) are all still image surrogates. A *moving image surrogate,* such as a fast forward, is more similar to the original video content since it contains action. A fast forward (Wildemuth, et al., 2002) mimics the fast-forward function of a VCR, playing the whole video content in a faster speed than normal without audio information. *Audio surrogates* use extracted audio information such as environmental sounds, music or people's dialogues to represent the video content. Multimodal surrogates which combine these textual, visual and audio information together prove to be more powerful than either text or visual surrogates (Ding, et al., 1999). An example is a video skim (Christel, Winkler, & Taylor, 1997), which summarizes the original video by concatenating significant subsets of video and audio data, similar to a movie trailer.

Table 1. Examples of video surrogates

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

As video retrieval researchers have put great efforts into generating better and better video surrogates and designing display structures, there is also a need to address the one fundamental issue of how to evaluate the effectiveness of various types of surrogates (Goodrum, 2001). In other words, what measures could (or should) be used to test how people perceive and understand video surrogates? This paper proposes six measures and tries to answer these questions.

3 Background

3.1 Theoretical background

User performance is a broad construct and many types of performance can be defined. Constructs such as time to completion, performance accuracy, errors, and satisfaction are often addressed in usability studies (e.g., Nielsen, 1993). In information retrieval studies, recall and precision values are used to assess the final products of search. What is needed for video retrieval research is a new set of instantiations of these constructs that is appropriate to the tasks and the media. In order to define and operationalize performance measures appropriate to video retrieval research, the cognitive mechanisms by which viewers perceive images and motion pictures should be considered carefully.

Panofsky (1955) identified three levels of comprehension of visual images: pre-iconographical description, iconographical analysis, and iconographical interpretation or synthesis. Pre-iconographical description is possible if the viewer understands the factual subject matter shown in the image (e.g., a woman), which "can be identified… on the basis of our practical experience" (Panofsky, 1972, p.9). Iconographical analysis connects this factual understanding with the themes, concepts, or allegories related to the image (e.g., the image of a woman with a halo is understood to represent the Virgin Mary). Pre-iconographical understanding is necessary but not sufficient for this level of understanding; in addition, the viewer must have familiarity with the concepts or stories being represented. The third level, iconographical interpretation or synthesis, focuses on the symbolic meaning of the image, adding an emotional aspect to the viewer's understanding of the image. Cultural attitudes, as well as personal attitudes, contribut to this understanding. For example, a 15th century painting of the Virgin Mary kneeling in adoration before the Christ child reveals a new emotional attitude toward these two figures that became dominant during that period. Each of these levels of comprehension may also apply to video materials.

While Panofsky's definitions of these three levels of understanding have been criticized (Bann, 1996), roughly similar hierarchies are commonly found in studies of video and image retrieval (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).

Grodal (1997) takes a more cognitive approach, proposing a flow diagram with four main steps that describe viewers' processing of film. The first step consists of basic perception. The brain makes its first visual analysis of input such as textures, lines and figures. The second step consists of memory-matching. The brain searches its memory files for possible matches, aided by feelings of familiarity or unfamiliarity. Step three is the cognitive-emotional appraisal and motivation phase, which leads to step four, reactions at a high level of arousal, such as fear or happiness.

While Grodal's model does differ from the others in having four steps/levels, instead of three, it still resembles them in many ways. In all cases, the levels/steps can be classified into two general categories: low-level visual perception/identification and high-level cognitive understanding. Similar ideas can be found in theories explaining 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 performance measures will be defined to evaluate the effectiveness of video surrogates: *recognition tasks* and *inference tasks*.

3.2 Methodological background

Few studies have been done to learn how people interact with and use video surrogates. They include those conducted by the Informedia project at Carnegie Mellon University, by Goodrum (1997, 2001) and Rorvig, and by the Digital Library Research Group at the University of Maryland. They are briefly reviewed here, with special emphasis on the methods used to evaluate surrogate effectiveness.

3.2.1 The Informedia project, Carnegie Mellon University

The Informedia project conducted several studies of how people use alternative surrogate implementations (Christel, Smith, Taylor, & Winkler, 1998; Christel, Winkler, & Taylor, 1997; Smith & Kanade, 1998). In one user study, they compared three video surrogates: text lists, opening shot poster frames and query-based poster frames (Christel, Winkler, & Taylor, 1997). Each participant responded to twelve questions, four using each of the surrogates. Each question required the selection of a video from a set of 12 query results; a value (representing its relevance to the question) had been assigned to each selection by the researchers prior to the study. Three dependent variables were used to compare the three surrogates: participant scores on each question set, the time spent to answer each question set, and the subjective satisfaction as measured by several sections of the Questionnaire for User Interface Satisfaction (QUIS).

Two experiments were used to compare several video skims, differing in the methods used to select the "important" video and audio components to be included (Christel et al., 1998). The first experiment incorporated three dependent variables: a fact finding task, a gisting task, and subjective satisfaction (measured with the QUIS). In the fact finding task, subjects were given a question and asked to navigate to that region of a video that presented the answer. While seeking the answer region, they could toggle between the skim and the full video. In the gisting task, "subjects matched skims of a longer video with representative text phrases and single-frame images for that longer video" (p.172). In the second experiment, an image recognition task was used in place of the fact finding task. Subjects were presented with ten still images and asked whether they recognized each as being in a video they had just viewed.

3.2.2 Goodrum

Goodrum (1997, 2001), in her work supervised by Mark Rorvig at the University of North Texas and her later work at Drexel University, has examined users' perspectives on the congruence between video and several surrogates. The basic factor considered is the ability of a surrogate to enable users to make the same distinctions that they would make if they viewed the video. Four types of video surrogates (titles, keywords, salient still frames (i.e., poster frames), and multiple key frames) were compared under three task environments (no task, general task, specific task.). Volunteers were asked to render similarity judgments for all pairs of videos. Separate groups of volunteers were asked to render similarity judgments for all pairs of surrogates. Multidimensional scaling (MDS) was used to map the dimensional dispersions, describing the derived stimulus configuration for videos and surrogates.

3.2.3 The Digital Library Research Group, University of Maryland

The Digital Library Research Group conducted a series of studies of the effectiveness of video surrogates: slide shows at various rates of display (Ding, Marchionini, & Tse, 1997), simultaneous multiple slide shows (Slaughter, Shneiderman, & Marchionini, 1997), and dynamic versus static displays (Komlodi & Marchionini, 1998). Several measures were employed in these studies: gist determination, object recognition, action recognition, and user perception. Gist determination, object recognition and user perception tasks were used in all the three user studies to evaluate the video surrogates, and action recognition was only used in the study to compare dynamic and static displays.

Gist determination measures evaluate how well users can determine the overall meaning of a video from viewing only the video surrogate. Two gist determination measure were developed, based on two types of tasks: users writing a brief statement of the gist, or users selecting from a set of gist statements created by the researcher. The object recognition measure focuses on whether the user can remember whether particular objects appeared in a particular video surrogate. The task's stimulus objects may be represented linguistically (with object names) or graphically (with key frames). For the action recognition measure, viewers watched a short video clip and then wrote a sentence to summarize the gist. User perceptions were measured by questionnaires using a seven point Likert scale from slow (1) to fast (7), with a score of 4 indicating "neither". Subjects were asked about speed perception during both the object recognition and gist determination tasks.

4 Measures of User Performance

Our work attempts to improve on the measurement techniques developed by the University of Maryland team. Additionally, based on the cognitive theories about how people perceive movies, pictures and words, we propose two additional measures: action recognition and visual gist determination. After an introduction to the measures, putting them in context, each measure and its validity will be discussed in detail.

From a cognitive perspective, the six measures can be categorized as associated either with recognition or with inference. The three *recognition* measures include object recognition (textual) , object recognition (graphical) and action recognition (see Table 2). They correspond to the first two steps (visual analysis and memory matching) in Grodal's (1997) flow diagram. They test whether users remember seeing the stimulus words, frames or actions in the video surrogates they viewed. Performance on these measures will be dependent upon the users' pre-iconographical mental model of the objects in the video surrogates and, possibly, on their iconographical analysis of the objects seen (Panofsky, 1972).

Table 2. Measures of user performance while interacting with 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	

Inference imposes more cognitive load than recognition since it requires summarization, comparison and synthesis while recognition only requires perception and recall. Inference is also a higher level of video comprehension since it is based on how much thematic information users could obtain from browsing video surrogates and what "story" users construct based upon their video comprehension. The three *inference* measures include gist determination (free-text), gist determination (multiple-choice) and visual gist determination (see Table 2, above). They can be seen as the operationalization of the last two steps in Grodal's flow diagram (construction of narrative scene or universe and reaction). For the gist determination measures, viewers are asked to specify the gist of the video, either by writing a gist description or by selecting a gist description from a set of alternatives provided. The visual gist determination measure is intended to more explicitly take into account the visual and stylistic aspects of the videos represented by the surrogates, and asks viewers to infer the visual gist of the video and then to select key frames (not already viewed) that they believe belong to or come from the video. Performance on these measures will be dependent upon users' iconographical analysis and iconographical interpretation of the video surrogate (Panofsky, 1972).

Two user studies were conducted using these measures. The first study (Wildemuth, et. al., 2002) was conducted in fall 2001 and examined the effectiveness of five different kinds of video surrogates: storyboard with textual keywords, storyboard with audio keywords, slide show with textual keywords, slide show with audio keywords, and fast forward. Ten participants first interacted with all the surrogates. They then selected one surrogate to view for each of three video segments (for a total of 30 viewings of the surrogates). The fast forward surrogate was selected 14 times, the slide show with audio keywords 6 times, the storyboard with text keywords 6 times, and the storyboard with audio keywords 4 times. All six performance measures were completed by each participant after viewing each surrogate. The second study (Wildemuth, et.al., 2003) conducted in spring 2002 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. Forty-five subjects participated in this study and each interacted with four video surrogates. In total, this study included 180 observations for each measure. Each measure, along with relevant data from these two studies, is described below.

4.1 Object recognition measures (textual, graphical)

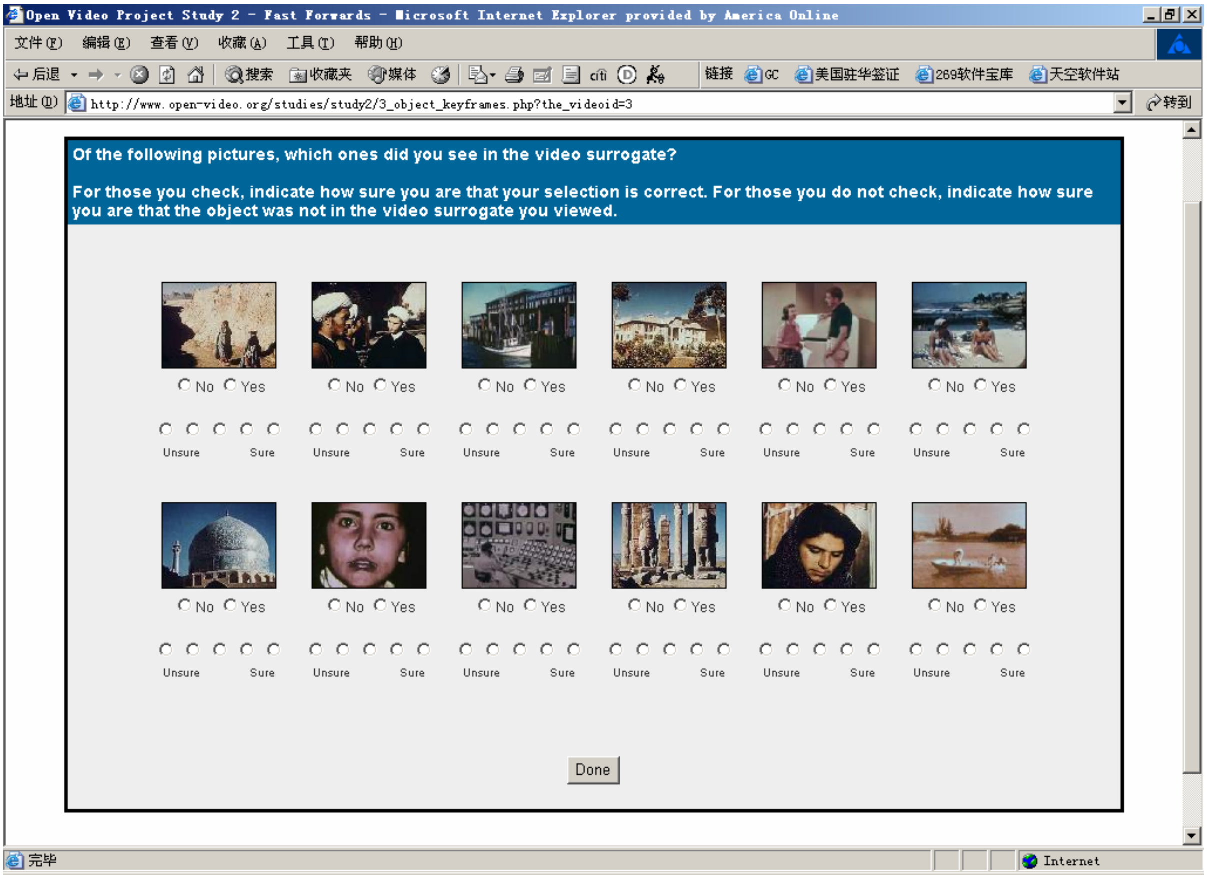
4.1.1 Rationale

Object recognition tasks, in which the participant is provided with a set of stimuli and asked which of them were seen in the surrogate, are intended to test the user's ability to recall which objects were seen in a video surrogate recently viewed and which objects were not in the video surrogate. The rationale for including this task in an evaluation of video surrogates is that it is closely related to the users' real-world task of selecting particular frames for later re-use—a task described by a sample of users participating in Open Video Project studies. If a person performs well in the object recognition task, it can be argued that the video surrogate supports the task of frame selection well.

As people view film/video, they conduct visual analysis and then compare what they see with what is available in their memories (Grodal, 1997). It can be expected that particular objects within the video will be identified and remembered. Potter and her colleagues (Potter & Levy, 1969; Potter & Kroll, 1987; comments by Clark, 1987) posit a conceptual model integrating people's understanding of pictures and words representing the same objects. This model suggests that viewing an object's name activates the viewer's pre-existing conceptual representation of it, while viewing an "imaginal" representation of an object may activate the viewer's verbal representation of it (which, in turn, activates the conceptual representation) or it may directly activate the viewer's conceptual representation of the object. Based on this model, one would expect that object recognition tasks could be based on either object names or images of the object (i.e., key frames) as stimuli.

4.1.2 Stimuli

The graphical object recognition task used a set of 12 images (key frames extracted from a video). 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 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 some of its stylistic characteristics, 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 color, all 12 key frames serving as stimuli were color. Figure 1 shows one example of this task. In this example, the first, second, and fourth images in the first row and the first, fourth, and fifth images in the second row were from the video surrogate being evaluated ("Iran: Between Two Worlds"), and so were correct if selected. The third image in the first row and the second and sixth images in the second row were distractors (i.e., incorrect items) selected as "similar" in style, while the last two images in the first row and the third image in the second row were selected as distractors that were "different" in style.


Open Video Project Study 2 - Fast Forwards - Microsoft Internet Explorer provided by America Online
文件(F) 编辑(E) 查看(V) 收藏(A) 工具(T) 帮助(H)
后退 搜索 收藏夹 媒体
链接 GC 美国驻华签证 269软件宝库 天空软件站
地址(D) http://www.open-video.org/studies/study2/3_object_keyframes.php?the_videoid=3
转到
○ No ○ Yes
○ ○ ○ ○ ○
Unsure Sure
Done
完毕
Internet

Of the following pictures, which ones did you see in the video surrogate?

For those you check, indicate how sure you are that your selection is correct. For those you do not check, indicate how sure you are that the object was not in the video surrogate you viewed.

Figure 1. Stimuli for graphical object recognition measure

In the <u>textual</u> object recognition task, a set of 12 object names served as the stimuli. Of the 12 object names, six were selected from frames seen in the video surrogates being evaluated and six were names of objects not seen in the video surrogate (i.e., they were distractors). Within each set of six, three were concrete objects and three were abstract objects. For example, the list of stimuli for the video, "Iran: Between Two Worlds," included the following:

Table 3. Stimuli for textual object recognition measure

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

4.1.3 Performance data

Performance scores on the object recognition measures can range from 0 to 12. A score of 12 would be achieved if the six frames or object names selected from the target video were selected and the six frames or object names from other sources were not selected (i.e., for each of the 12 stimulus objects, the user can receive a score of 1, correct, or 0, incorrect).

The 10 participants in the first study scored an average of 8.97 (from the 12 possible) on the textual object recognition task, with a standard deviation of 1.61. They also scored an average of 8.97 (s.d.=1.81) on the graphical object recognition task. In the second study, the average score of the 45 participants was 8.6 (s.d. = 1.35) in the textual object recognition task and 9.7 (s.d. = 1.65) in the graphical object recognition task. The scores on these two measures were not affected by the differences in fast forward surrogate speed—the surrogate attribute being evaluated in the second study.

4.1.4 Validity of the object recognition measures

Across both studies, the object recognition scores were relatively high, indicating that this task was not particularly difficult for the study participants. The results concur with Shepard's (1967) finding that people had remarkably accurate recognition memory for pictures (96.7%).

In the graphical object recognition measure, the correct items were selected from the video surrogate, and the incorrect items were selected from videos that were similar or different in style. Those items drawn from videos of a different style were easier than those drawn from similar videos or the target video (chi square (2df) = 76.7179, p<0.0001). Study participants (study 2) were able to perform most accurately on items selected from dissimilar videos, getting 92% of these items correct (i.e., *not* selecting them). On the items from the target video and similar videos, performance was essentially the same: 77% correct on the target video items (i.e., the correct items) and 76% correct on the (incorrect) items selected from similar videos. This measure might be more valid if the items from dissimilar videos were not included as distractors. Future studies should investigate this possibility for making this measure even more discriminating than it was in our first two studies.

In the textual object recognition measure, we included both concrete and abstract/conceptual terms among the correct and incorrect options. Study participants (study 2) performed better on the abstract/conceptual items (chi square (1df) = 161.0969, p<0.0001), getting 83% of them correct, versus 61% of the concrete items correct. It seems likely that participants are "guessing" correctly on some of the abstract items, based on inferences they are making about the video's content. Since this measure is intended to measure only object recognition, with little or no reliance on inference, it may be more valid to include only concrete items among the stimuli in future studies.

The relationship between these two measures can be investigated with correlation analysis: if they are measuring the same underlying construct, object recognition, then they will be highly correlated. In the first study, performance on the two measures was moderately correlated (r=0.34, p=0.0633). Results from the second study, however, do not indicate any relationship (r=-0.09, p=0.2512). Thus, it is not clear that the two measures are equivalent. Additional studies using these measures should investigate the possibility of making the two sets of stimuli more equivalent, e.g., by naming the objects shown in the graphical stimuli for use as textual stimuli.

4.2 Recognition of actions

4.2.1 Rationale

An action recognition task was newly developed for use in the Open Video Project studies. Rather than being asked to recognize whether a particular object appeared in the video surrogate, the study participant is asked to recognize whether a particular short action sequence appeared in the video surrogate.[1] Like the object recognition task, the rationale for evaluating action recognition is grounded in users' reports of their need to select particular video clips for various types of re-use.

4.2.2 Stimuli

The action recognition task, as implemented for our studies, uses mini-segments, each 2-3 seconds long, as stimuli. Six mini-segments were displayed to the study participant. Of these six, two were selected from the video segment represented in the video surrogate (i.e., were correct), two were from another segment/video of a style similar to that of the target video, and two were from another segment/video of a style that was different from that of the target video. In response to each mini-segment, the participant would be asked whether s/he believed it to be from the same video segment as represented in the surrogate.

4.2.3 Performance data

The scoring for this task was comparable to that used with the object recognition task. A point was scored each time the participant selected one of the two clips selected from the target video and each time the participant rejected one of the four clips selected from a different video/segment.

The 10 study participants from the first study scored an average of 4.6 (out of a possible 6 points), with a standard deviation of 1.0. The results of that study indicated that the fast forward surrogate was more effective in supporting this task than the other surrogates (mean=5.1 for the fast forward versus means ranging from 4.0 to 4.3 for the other three surrogates; F=3.36 with 3df, p=0.0340). Based on this data, this measure was not related to any of the other performance measures. In the second study, the average score of the 45 participants was 4.5 (s.d. = 0.93). Action recognition performance was affected by the surrogate speed (F=3.62 with 3, 176 df, p=0.0112); performance on the slowest speed was better than on the two highest speeds.

[1] It should be noted that the user will not have actually seen any of the action clips used as stimuli in this task, since no surrogates include clips from the original videos. All the surrogates are created by sampling key frames from the videos; thus, the clip's frames that appear between key frames will not have been seen.

4.2.4 Validity of the action recognition measure

As noted earlier, the video clips used as items for this measure were drawn from the target video, similar videos, and dissimilar videos. As with the measure of graphical object recognition, performance on the items from dissimilar videos was significantly higher than on the correct items or the incorrect items drawn from videos similar to the target video (chi square (2df) = 31.0795, $p<0.0001$). If the distractor was from a dissimilar video, 85% of the responses were correct (i.e., the clip was *not* selected as being from the target video). If the distractor was from a similar video, 71% of the responses were correct, and if the item was drawn from the target video, 73% of the responses were correct. Future studies should investigate whether drawing all the distractors from similar videos might make this measure even more valid and discriminating.

4.3 Gist determination: inferring meaning from the video surrogate

4.3.1 Rationale

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. If the surrogate supports this task well, the user is able to make accurate relevance judgments or selection decisions about videos, thus having a successful and efficient browsing experience.

Participants in the fall 2001 study provided many comments concerning their ability to determine the gist of the video based on viewing the surrogate, and considered this the most important function of the surrogates. 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 them understand what the video was about, i.e., the topic of the video. Secondly, the participants found the surrogates most useful when they told the "story" of the video or had a narrative structure. 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 (1978, 1983) 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. The third understanding of gist presented by the study participants is what we are calling visual gist, and will be discussed in the next section.

4.3.2 Stimuli

Two measures of gist determination were developed for our current studies: free-text writing and multiple-choice selection. The free-text writing measure is described first, along with the methods for scoring the gist descriptions generated. Then, the multiple-choice measure is described.

The *free-text gist determination measure* asks the user to generate a gist description after viewing the video surrogate. Specifically, the instructions were, "Please write a brief summary of the video." Once the gist descriptions were generated by the study participants, they had to be scored. The first study employed a scoring procedure developed by Tse, Marchionini, Ding, Slaughter, and Komlodi (1998). It was based on each statement's "depth of comprehension and degree of involvement of external knowledge." The possible scores were:

0 – not correct
1 – correct literal objects or events
2 – correct general thematic information or ‘common sense’ judgments
3 – accurate thematic information.

For the first study, two members of the research team independently scored each of the gist descriptions. On two of the three videos, there was 70% agreement between the scorers; on the third, there was only 30% agreement. Clearly, the scoring procedures needed further clarification in order to be applied consistently. For the first study, a third member of the research team resolved the differences in the assigned scores.

The following examples illustrate the gist descriptions generated by the study participants related to the video “Moon,” and how the scoring procedures were applied.

> “Astronauts training for mission to moon. Several tests, including underwater and parachuting excercises.” (Score: 3)
>
> “I think this surrogate is focusing on man's first experience going to space. It focuses on how he was trained and presents his accomplishments of going to the moon. (Neil Armstrong)” (Score: 2)
>
> “The meaning of this video is interviewing the astronaut, Neil Armstrong.” (Score: 1)

Due to the low agreement between the two scorers in the first study, a new scoring system was developed in the second study. It included two scores (correctness/accuracy and level of detail) on each of two dimensions (objects/events and higher-level perspective). A brief description of this scoring procedures is included in Box 1.

OBJECTS/EVENTS

- Correctness:
  0 (No correct objects/events are listed) to
  2 (Of objects/events listed, >60% are correct)
- Detail:
  0 (0 or 1 correct objects/events are listed) to
  2 (5 or more correct objects/events are listed)

HIGHER-LEVEL PERSPECTIVE (E.G., THEME OR PLOT)

- Correctness/accuracy:
  0 (No correct higher-level perspective is included) to
  2 (Higher-level perspective is mostly/completely accurate)
- Detail:
  0 (No perspective, or a single phrase is provided) to
  2 (Very detailed/complete perspective is provided)

Box 1. Scoring of free-text gist determination measure, Study 2

Two members of the research team were trained on the scoring procedures and then independently scored each of the gist descriptions. The correlation of the scores between these two members was 0.76 and it was concluded that the interrater reliability was satisfactory.

The following examples illustrate the gist descriptions generated by the study participants related to the video "Iran," and how the scoring procedures were applied.

> "This seems to be a documentary or tourism video about a middle eastern country, maybe Iran. It first shows the prehistory of the place, with a lot of archaeological sites. Then it shows some of the crafts, like carpet making, and then some people living in traditional ways. Then it shows some of the modern cities and industry." (Score: 8)
>
> "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." (Score: 5)
>
> "This video seems to be a documentory about some country in far East or North Africa." (Score: 1)

The *multiple-choice gist determination measure* is based on gist descriptions generated by members of the research team. The study participants were asked to select which best described the video for which they had seen the surrogate. For example, the candidate gist descriptions for "Moon" were:

> It describes the background and astronaut training for the Apollo mission. (correct)
> It shows us some pilot training in the air force.
> It describe that some people are visiting the Kennedy Space Center.
> It shows us a science fiction movie of the 1960's.
> It describes the design of experimental aircraft.

The multiple-choice measure resulted in a binary score; the user's response was either correct or it was not.

4.3.3 Performance data

Generally, the free-text gist descriptions provided by the study participants were quite short. In the first study, the shortest was three words; in the second, it was one word. The longest description in the first study was 55 words; 176 words in the second study.

On the free-text gist determination measure, the ten participants in the first study averaged 1.68 (out of a possible 3 points), with a standard deviation of 0.75. On the multiple choice gist determination measure, they got 80% correct. Based on these data, the multiple choice form of the task was easier, as would be expected. It would be expected that scores on these two measures would be related to each other; however, in the first study, there was only weak evidence of a relationship. Those who were correct in their selection of a multiple choice gist description averaged 1.79 on the free-text gist determination task, while those who were incorrect on the multiple-choice measure averaged only 1.25 on the free-text measure. These differences were not statistically significant (t=1.62 with 28df, p=0.1165). This lack of a statistically-significant relationship may be due to the small sample size or to the lack of reliability in the scoring method. It is also possible that some people are good at recognizing the correct gist statement, but not at expressing the video's gist on their own (or vice versa).

As improvements in the scoring procedures are made, this relationship should be examined again.

In the second study, the 45 participants averaged 2.9 (of a possible 8 points), with a standard deviation of 1.72. On the multiple-choice gist determination measure, they got 46% correct. As in the first study, it appears that the multiple-choice measure was easier. The relationship between the two gist determination measures was again investigated. Those who were correct on the multiple-choice measure scored, on average, 3.1 on the free-text measure; those who were incorrect on the multiple-choice measure scored, on average, 2.7 on the free-text measure. This difference was only marginally significant (t=1.79 with 178 df, p=0.0759). Since the scoring procedures were improved and supported reliable scoring and the sample size was also quite large, it must be concluded that the relationship between these two measure is not strong.

4.3.4 Validity of the gist determination measures

Some of the recognition measures, already discussed, were related to the gist determination measures. Action recognition performance was related to both the free-text and the multiple-choice gist determination measures. It had a weak but statistically-significant correlation with the free-text measure (r=0.17, p=0.0227). Those who responded correctly to the multiple-choice gist determination measure scored, on average, 4.7 on the action recognition measure; those those responded incorrectly on the multiple-choice gist determination measure score, on average, 4.4 on action recognition. This difference, though small, was statisically significant (t=2.49 with 178 df, p=0.0136). One possible explanation for these weak but statistically-significant relationships is that people's ability to remember small pieces of the action in a video is weakly related to their understanding of the entire video.

The other statistically-significant relationship was between the graphical object recognition measure and the full-text gist determination measure. These two scores were moderately correlated (r=0.38, p<0.0001). The most likely explanation for this relationship is analogous to the proposed explanation for the relationship between action recognition and gist determination: that a person's ability to remember graphical images from a video surrogate is related to their general understanding of the entire video.

4.4 Visual gist determination

4.4.1 Rationale

Visual gist is 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, 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: "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]." Having this more complete understanding of a video will support the user in making accurate selection decisions when choosing which videos from a collection may be useful for particular purposes.

4.4.2 Stimuli

Clearly, operationalizing a construct that is in such a preliminary stage of being defined is a challenge. In our first two studies, we provided participants with a set of stimuli (i.e., key frames) and gave them 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, and ensured that the participant distinguished this task from the earlier graphical object recognition task.

The stimuli for this task consisted of 12 key frames, *none* of which actually appeared in the surrogate viewed by the study participants. 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 graphical object recognition task. The item was counted as correct if (1) it was selected and from the set of six frames from the target video or (2) it was not selected and was not from the target video.

4.4.3 Performance data

The 10 participants in the first study averaged 9.7 (of a possible 12), with a standard deviation of 1.36. The 45 participants in the second study averaged 8.4, with a standard deviation of 1.41. It can be concluded that this measure was not difficult for the study participants; all of them got at least 5 of 12 items correct on every trial.

4.4.4 Validity of the visual gist determination measure

As with several of the other measures, performance did vary with the source of the item. The items came from the target video (frames that were not seen in the surrogate; they were scored as correct if selected), from videos that were similar in style to the target video, and from videos that were different in style from the target video. For the measure of visual gist determination, those items from videos that were different in style were the easiest (90% correct). The items from the target video were the next easiest (70% correct). Those from videos that were similar to the target were the most difficult (53% correct). These differences were statistically significant (chi square (2df) = 223.9847, p<0.0001).

In the first study, this measure was not related to any of the other performance measures. However, the second study did reveal some relationships between visual gist determination performance and other measures. Scores on the visual gist determination and the full-text gist determination measures were weakly correlated (r=0.15, p=0.0392). (There was not a statistically-significant relationship with the multiple-choice gist determination measure.) It is likely that the inferences required to select the frames that "belong" in a video rely on an accurate understanding of that video's topical gist. The lack of relationship with the multiple-choice measure suggests that selecting a description of a video's gist is not relying on the same inferential processes.

Scores on the visual gist determination measure were also weakly correlated with textual object recognition scores (r=0.20, p=0.0064). Interestingly, performance on the visual gist determination measure was not related to performance on the graphical object recognition measure, even though the stimuli strongly resembled each other (only the task instructions differentiated them). The correlation between the visual gist task and the graphical object recognition task was 0.12 (p=0.5204) in the first study and 0.08 (p=0.5218) in the second study. This lack of relationship in participant performance suggests that the participants were able to distinguish the visual gist task from the object recognition tasks.

In summary, there is evidence of relationships between visual gist and two other measures of performance: gist determination (represented as free text generated by the study participants) and object recognition (with textual stimuli). These relationships, as well as the lack of relationships between visual gist and the other measures, should be further explored with additional studies, so that a clearer definition of this construct can be developed.

5 Discussion

5.1 Additional considerations in using the proposed measures

As these measures were implemented in two studies, the first in fall 2001 and the second in spring 2002, a variety of issues arose. While these issues were resolved for the purposes of the two studies conducted, they may be resolved differently within the context of a different study. The issues, discussed below, include the order in which the tasks should be assigned, the effects of user confidence on task performance, and the effects of video genre on task performance.

5.1.1 Order of the measures

If multiple measures are administered in one study, it is possible that they may interact with each other. For example, among the measures being considered here, it is likely that interacting with the recognition measures may improve a study participant's performance on the gist determination measures. For the meaures to be valid, they should be administered in an order that is least likely to bias the outcomes.

For the first study, the order in which subjects completed the performance tasks was based on each measure's importance to users' general interactions with video and on the amount of past research experience with the measures. First were the two gist determination measures, assumed to be the most important in relation to user interactions with video. Subjects were first asked to write a brief summary of the video, and then to select one correct gist statement from the five choices. Secondly, the three recognition measures were presented. Subjects were asked to select those frames, words or actions they remembered seeing in the video surrogate. The final measure was the visual gist task, the task with which there was the least prior research experience. Subjects were asked to select those frames they thought belonged in the video, even though they hadn't seen them in the surrogate.

In the first study, some problems were found in this order of arranging these tasks. The multiple-choice gist determination task, by providing alternative hypotheses about the gist of the video, could affect how subjects understand the gist of the video. This seems to correspond with van Dijk and Kintsch's model of discourse comprehension (1978, 1993). 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 appears to be used by video viewers. If this model applies to this set of measures, performance on any measures that follow the multiple-choice gist determination task could be affected by the subject's reading of the candidate gist statements as much as by the subject's viewing of the video surrogate.

For this reason, in the second study, the multiple-choice gist determination task was moved to the very end, after subjects finished all the other tasks. In spring 2002, with the re-ordered set of tasks, some subjects felt more confused when they did the multiple-choice gist determination task (last) because their hypothesis about the gist of the video was not totally consistent with any of the selections. This anecdotal evidence suggests that it was wise to move the multiple-choice gist determination task to be the last of the performance tasks.

5.1.2 Effects of differences in confidence levels

During the fall 2001 study, a number of participants commented on their confidence in completing the performance tasks, particularly the object recognition tasks. 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.

To further investigate the possible effects of participant confidence, the spring 2002 study incorporated direct measurement of each subject's confidence on each selection. It was possible to include a measure of confidence with four of the six measures: the three recognition measures (textual object recognition, graphical object recognition, and action recogntion) and the visual gist determination measure. The confidence measure was incorporated in two ways. First, a sentence asking the subject to indicate their confidence was added to the instructions for each performance task. For example, the instructions for the object recognition task read, "Of the following pictures, which ones did you see in the video surrogate? For those you check, indicate how sure you are that your selection is correct. For those you do not check, indicate how sure you are that the object was not in the video surrogate you viewed." Second, a five-point rating scale (from "unsure" to "sure") was added for each selection. (See Figure 1 for an illustration of the confidence measure.) The researcher ensured that the subject rated his/her confidence in each selection (or non-selection) before they progressed to the next performance task.

The confidence data from the spring 2002 study was analyzed to determine whether confidence level was associated with the correctness of participant's response, the act of selecting (versus not selecting), characteristics of the surrogate (i.e., the speed of the fast forward surrogate, being evaluated in spring 2002), or characteristics of the video (i.e., whether the video was a documentary or more narrative in style).

For all four measures, participants were more confident when they were correct than when they were incorrect. The mean confidence ratings when the individual item was scored as either correct or incorrect are shown in Table 4. All four differences were statistically significant.

Table 4. Confidence ratings for correct and incorrect responses

	Confidence when item was correct	Confidence when item was incorrect	t	p
Object recognition (textual)	3.9	3.4	8.79	<0.0001
Object recognition (graphical)	3.9	3.3	12.60	<0.0001
Action recognition	3.8	3.1	9.26	<0.0001
Visual gist determination	3.9	3.5	8.73	<0.0001

For the two object recognition measures and the visual gist determination measure, confidence was also associated with whether the item was selected as being/belonging in the video surrogate viewed or not selected (indicating that the participant believed that the stimulus was *not* in the video surrogate viewed or *did not belong* in the video represented by the surrogate). The results of this analysis are provided in Table 5. For these three measures, t tests indicated that the differences were statistically significant. The difference in confidence levels for the action recognition measure was not statistically significant.

Table 5. Confidence ratings for selected and not-selected items

	Confidence when item was selected	Confidence when item was not selected	t	p
Object recognition (textual)	4.3	3.3	23.34	<0.0001
Object recognition (graphical)	3.9	3.7	6.29	<0.0001
Action recognition	3.8	3.7	0.91	0.3629
Visual gist determination	3.9	3.7	5.78	<0.0001

The primary research question addressed with the spring 2002 study was the effects of the speed of a fast forward surrogate on participants' performance on the six measures. Therefore, we also examined the effects of surrogate speed on confidence in responding to the items. For all four measures, there was a statistically-significant relationship between confidence and surrogate speed (p<0.001). In all four cases, participants were more confident of their responses when the surrogate speed was slower, but these differences were not large: less than a half point on the five-point confidence scale. The speed at which lower confidence levels began to appear (and to be statistically significant) varied by measure (as indicated by Bonferroni t tests). The mean confidence ratings at each speed are shown in Table 6.

Table 6. Confidence ratings at different surrogate speeds

	1:32	1:64	1:128	1:256
Object recognition (textual)	4.0	3.8	3.6	3.6
Object recognition (graphical)	3.9	3.9	3.8	3.6
Action recognition	3.8	3.7	3.5	3.6
Visual gist determination	3.9	3.8	3.7	3.7

Note: The surrogate speeds are based on the sampling rate of key frames used to create the surrogate. For example, 1:32 indicates that 1 of every 32 key frames was included in the surrogate, and so is considered the "slowest" of the fast forward surrogates.

Finally, the interaction between confidence and video characteristics was investigated. Two of the videos used in the study were narrative in structure and two were documentaries. This characteristic of the stimulus video did not appear to have a consistent effect on participants' confidence in responding to the performance measures. No statistically significant difference in confidence ratings were found for the action recognition measures (F (1, 1360) = 0.16, p=0.6913) or the visual gist determination measures (F (1, 2710) = 2.64, p=0.1045). There were statistically significant effects for the object recognition measures, but the results were mixed. For the textual object recognition measure, participants' confidence was higher on documentary videos, while for the graphical object recognition measure, confidence was higher on narrative videos (in both cases, p=0.0001). These mean confidence ratings are shown in Table 7.

Table 7. Confidence ratings for narrative versus documentary videos

	Confidence on narrative videos	Confidence on documentary videos	t	p
Object recognition (textual)	3.6	3.8	3.94	0.0001
Object recognition (graphical)	3.9	3.7	3.84	0.0001
Action recognition	3.7	3.6	0.40	0.6913
Visual gist determination	3.8	3.7	1.62	0.1043

The implications of differences in study participants' confidence for the validity of the proposed measures of performance must be considered. The fact that participants were more confident when they were correct suggests that their confidence is appropriate (i.e., they are not inappropriately over- or under-confident). The relationship between confidence and whether an item was selected (versus not selected) suggests that our study participants were fairly conservative in their approach to these measures; if they were not confident about selecting an item, they tended to not select it. Neither of these relationships would appear to have negative effects on the validity of the performance measures evaluated.

The particular characteristics of the surrogates/videos being examined in the spring 2002 study were also investigated for their effects on participants' confidence in responding to the performance measures. The fact that confidence dropped as the surrogate speed increased suggests that, even when performance levels can be maintained, it might be worthwhile to select surrogate speeds with which users are more comfortable, i.e., with which they are more confident in their ability to remember or understand objects and actions seen in the surrogate. The lack of a consistent effect of video style on confidence ratings suggests that differences in video style are not having a negative effect on the validity of these measures. However, since two of the measures were affected by video style, further investigations of this effect are warranted.

5.1.3 Effects of differences in video structure and style

In addition to the possible effects of video characteristics on users' confidence in using video surrogates, researchers should be concerned about the direct effects of video characteristics on performance with video surrogates. In the fall 2001 study, participants commented on video structure (documentary vs. narrative) and style (color vs. black and white) and their effects on perceptions of video content. Subjects preferred color films to black and white ones, since they thought that color would help them to differentiate objects in the surrogate and they could remember more details. Additionally, they thought they could determine what a documentary film was about much easier than a narrative film. Their explanation for this difference was that they needed to construct a story for a narrative film and this process was difficult if only the surrogate had been viewed, while for documentary films they didn't need as many details.

During the fall 2001 study, the structure of the video was not manipulated. Therefore, the spring 2002 study followed up on the questions raised concerning the effects of video structure and style on task performance. For the spring 2002 study, four films were selected: two color and two black and white; two narrative and two documentary. Each subject performed all six tasks for each of these four videos. Only on the multiple-choice gist determination measure was there a statistically-significant effect of video structure: 69% of the responses on documentary videos were correct, while only 23% of the responses on narrative videos were correct (chi-square = 37.5829, $p<0.0001$). The style of the stimulus videos (color versus black and white) affected performance on three of the measures.

Performance was higher with color videos on multiple-choice gist determination and action recognition, and higher with black and white videos on graphical object recognition. On the multiple-choice gist determination measure, participants got 61% of the items correct when the stimulus video was in color, and only 39% correct when the stimulus video was in black and white (chi-square = 8.0710, p<0.0045). On the action recognition measure, the mean score was 4.7 when the stimulus was in color and only 4.4 when the stimulus was in black and white (F (1, 178) = 4.09, p=0.0447). On the graphical object recognition measure, the effects of color were reversed: the mean score on black and white videos was 10.0, while it was only 9.4 on color videos (F (1, 178) = 7.09, p=0.0085).

These very mixed results do not allow us to draw any strong conclusions about the effects of video characteristics on performance with video surrogates. Because each study is likely to involve only a few videos as stimuli, it is critical that they be selected with attention to their structure (narrative versus documentary) and style (color versus black and white). It is also possible that other video characteristics (such as whether they treat a topic literally or figuratively, the amount and style of audio content they include, or the pace of scene changes) might affect performance on the six proposed measures. It is recommended that all future studies of interactions with videos include the effects of video characteristics among the research questions to be investigated.

5.2 Limitations of the proposed measures

While we will continue to use these measures to evaluate the effectiveness of different video surrogates and encourage other researchers to adopt them as well, they do have some limitations. First, they are performance measures, and do not take into account users' preferences. The development of attitudinal and satisfaction measures appropriate for use in studies of video retrieval is a necessary parallel activity that has not yet been addressed. Second, there were still some interactions between these measures. During our studies, the participants commented that they learned what the videos were about while completing the measures. Paying particular attention to the order of the measures can diminish the interaction effects, but it is likely that interactions will occur whenever multiple measures are used in a single study.

6 Conclusion

This paper describes several user performance measures that may be useful in evaluating different video surrogates, presents some initial data concerning their implementation, and discussed their validity. These measures represent 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 models of the cognitive process by which viewers perceive and understand images and videos (Greisdorf & O'Connor, 2002). The three object/action recognition tasks require that the user recognize objects or actions that occur in the video surrogates viewed; this type of task is associated with the users' needs to select video frames or clips for re-use. The gist determination and visual gist tasks require that the user infer an overall understanding of the video from viewing only the surrogate; these tasks are associated with the users' needs to select videos from a collection 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.

7 References

Bann, S. (1996). Meaning/interpretation. In Nelson, R. S., & Shiff, R. (eds.), *Critical Terms for Art History*. Chicago: University of Chicago Press, 87-100.

Christel, M. G., Winkler, D. B., & Taylor, C. R. (1997). Improving access to a digital video library. *Human-Computer Interaction, INTERACT '97: IFIP TC13 International Conference on Human-Computer Interaction (Sydney, Australia, July 14-18, 1997)*, 524-531.

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

Clark, J. M. (1987). Understanding pictures and words: comment on Potter, Kroll, Yachzel, Carpenter, and Sherman (1986). *Journal of Experiental Psychology: General, 116*(3), 307-309.

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

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 accessed May 14, 2003.

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.

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.

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

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. *Proceedings of ACM Multimedia 1999,* 489-498.

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

Marchionini, G., & Geisler, G. (2002). The Open Video Digital Library. *D-Lib Magazine, 8*(12). http://www.dlib.org/dlib/december02/marchionini/12marchionini.html. Last accessed June 20, 2003.

Nielsen, J. (1993). *Usability Engineering*. Boston: Academic Press Professional.

O'Connor, B. (1985). Access to moving image documents: background concepts and proposals for surrogates for film and video works. *Journal of Documentation, 41*, 209-220.

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.

Potter, M. C., & Kroll, J. F. (1987). Conceptual representation of pictures and words: reply to Clark. *Journal of Experimental Psychology: General, 116*(3), 310-311.

Potter, M. C., & Levy, E. I. (1969). Recognition memory for a rapid sequence of pictures. *Journal of Experimental Psychology, 81*(1), 10-15.

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: Proceedings of the First European Conference (EDSL '97, Pisa, Italy),* 41-54.

Smith, M., & Kanade, T. (1998). Video skimming and characterization through the combination of image and language understanding. *Proceedings of the 1998 IEEE Workshop on Content-based Access of Image and Video Databases (Bombay, India, January 1998),* 61-70.

Tse, T., Marchionini, G., Ding, W., Slaughter, L., & Komlodi, A. (1998). Dynamic key frame presentation techniques for augmenting video browsing. *Proceedings 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. Paper accepted for presentation at the Joint Conference on Digital Libraries, 2003.