How Fast Is Too Fast?
Evaluating Fast Forward Surrogates for Digital Video

Barbara M. Wildemuth, Gary Marchionini, Meng Yang, Gary Geisler,
Todd Wilkens, Anthony Hughes, & Richard Gruss
Interaction Design Laboratory
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599
+1 919 9663611
[wildem, march, yangm, geisg]@ils.unc.edu; [tpodd, hughes, gruss]@email.unc.edu

Abstract

To support effective browsing, interfaces to digital video libraries should include video surrogates (i.e., smaller objects that can stand in for the videos in the collection, analogous to abstracts standing in for documents). The current study investigated four variations (i.e., speeds) of one form of video surrogate: a fast forward created by selecting every Nth frame from the full video. In addition, it tested the validity of six measures of user performance when interacting with video surrogates. Forty-five study participants interacted with all four versions of the fast forward surrogate, and completed all six performance tasks with each. Surrogate speed affected performance on four of the measures: object recognition (graphical), action recognition, linguistic gist comprehension (full text), and visual gist comprehension. Based on these results, we recommend a fast forward default speed of 1:64 of the original video keyframes. In addition, users should control the choice of fast forward speed to adjust for content characteristics and personal preferences.

1. Introduction

The ability to create sophisticated digital video productions is now within the reach of anyone with a home computer, since technologies that support the capture, storage, and transmission of digitized video files are common marketplace items. Digital video cameras and cheap webcams are becoming household appliances. Inexpensive disk space allows consumers to store vast amounts of original or commercially produced video, and increasing bandwidth facilitates sharing these files over the internet. These hardware advances are in turn supported by basic software packages that aid in capturing, editing, and compressing the final digital video production. These technical developments stimulate traditional video enterprises—such as video rental or purchasing, stock footage clearinghouses and distance education purveyors—to move towards on-demand, asynchronous delivery of digital video via the internet. More importantly, they stimulate the incorporation of digital video materials into digital library collections. We believe that there is a crucial need for interfaces that will improve library users' access to digital video collections, and so are focusing our research on a combination of interfaces and surrogates that will support retrieval from digital video libraries.

The design of such interfaces should be rooted in a blend of (1) empirical evidence about how people interact with and understand video and (2) imaginative approaches to leveraging the digital medium. The Open Video Project aims to develop and maintain an open source digital video repository that serves as a testbed for video research, including user studies and evaluations of interface prototypes for digital video applications.

Our current emphasis is on user studies of specific surrogates that help people browse and select materials from libraries of digital videos. Video surrogates stand in for the videos in the collection, just as abstracts are surrogates that stand in for documents in a text-based library. We believe that these interfaces will be more effective if they leverage the digital video medium rather than simply mimic the analog interfaces of television and VCRs or the text-based interfaces of document collections. Thus, we are experimenting with a variety of surrogates using digital video materials, and evaluating them based on their ability to help users of digital

libraries make rapid, accurate decisions about the relevance of video materials.

This paper reports on a study of the use of fast forwards as one type of surrogate for digital video. This type of surrogate is worth investigating for two reasons. First, people are familiar with the concept of fast forward movement through a video because of their experience with their VCR's. This familiarity should increase the ease with which people interact with fast forward surrogates. Second, participants in our initial studies expressed the desire to see motion in the video surrogates. Other surrogates, such as poster frames or storyboards, use the images/frames present in the video, but the user loses any sense of movement. The primary question related to design of fast forward surrogates is the tradeoff between speed (i.e., faster speed will shorten the necessary viewing time) and understanding (i.e., if the surrogate is too "fast", it will not be useful in supporting accurate relevance judgments). This question is addressed by the study reported here. The results have import for designers of digital libraries that include digital video.

The current study also makes a methodological contribution. It presents a set of measures useful for evaluating the effectiveness of any type of video surrogate. These measures have been developed, revised, and extended in studies over several years and illustrate a multifaceted approach to assessing human understanding of complex media when using different surrogates.

2. Related work

In digital video libraries, the size of files and time to download and view each video make it particularly important to have meaningful metadata and surrogates that allow people to recognize or assess the pertinence of the full object. Titles, keywords, and other bibliographic metadata have traditionally been used in video retrieval, along with short textual descriptions that act as surrogates to help people assess relevance. In addition to these linguistic representations, the medium of video suggests that image-based surrogates may provide additional cues for people trying to assess the relevance of a particular video for a particular purpose.

Keyframes [15] have been adopted by many digital video researchers as a basis for non-linguistic representations of the information content of a video object. There is a substantial body of work related to finding ways to segment video, extract keyframes or other features, and create indexes for the purposes of retrieval. There is less attention to creating user interfaces that support interactive search and browse capabilities. The Informedia project (www.informedia.cs.cmu.edu/) is perhaps the most comprehensive digital video effort that includes novel user interfaces and usability testing [1,2,24]. Their video skims are surrogates created from several kinds of features (transcripts, keyframes extracted with color and texture features, superimpositions, and other features such as face recognition). The Físchlár Project (www.cdvp.dcu.ie/) stores and provides access to video programming from broadcast TV. They have developed user interfaces that integrate several different types of surrogates to help users find video [10,23]. The ECHO project (pc-erato2.iei.pi.cnr.it/echo/) aims to provide access to large volumes of historical video in Europe. Their interfaces will support multilingual access. The CueVideo system (www.almaden.ibm.com/cs/cuevideo/) [19] extracts a variety of features as the basis for indexing (e.g., using speech to text analysis, image analysis, event analysis) and has been the basis for more specific user interface techniques, such as movieDNA [20], that provide visual patterns for where query images occur in lists of video segments. The SmartSkip interface [5] is one of the few interfaces that provide innovative fast forwards beyond the digital TV fast forwards. Their user study compared a standard skip interface and a fast forward interface with a user-controllable SmartSkip interface. They found that, although people found the SmartSkip interface more 'fun' to use, they performed better with the standard skip interface than with the other two interfaces on commercial skipping and weather finding tasks. These results parallel studies of slide shows and story boards [22] that demonstrate that, although people are able to perform effectively on retrieval tasks with very rapid slide shows, they strongly prefer the story board interfaces that give them more control but take more time to use.

The Open Video Project (www.open-video.org) began with efforts to provide digital video from sources like the Discovery Channel and the US Archives to middle school science and social studies teachers [12,21] and has been expanded to serve as an open source test bed for the research and educational communities. The repository points to about 2000 video segments (more than a half terabyte) and draws upon documentaries from many US government agencies, the Prelinger Collection in the Internet Archive, digitized films in the Library of Congress' American Memory collection, and videos from CMU's Informedia Project and the University of Maryland's Human Computer Interaction Laboratory. The MySQL metadata database is accessible from an interface that provides overviews and previews [6] and serves as the testbed for the surrogates developed and tested in the Interaction Design Laboratory at UNC-Chapel Hill.

Because different people may understand the same object differently, we aim to design a variety of surrogates and access mechanisms to support this

variability in human sense making. In addition, for any given surrogate or view, there will be variations in human abilities and experience that affect performance with those surrogates. Therefore, we also aim to establish effective ranges of use for those surrogates such as what speed ranges to provide on a slider bar mechanism for slide show surrogates. For example, in previous work [4] we investigated the relationship between speed of keyframe slide shows and performance on object identification and gist determination tasks. Slide shows allowed people to comprehend the video's gist at very high rates of speed (from 4kf/second to 16kf/second) with a predicted fall off in performance as speed increased. These performance effects were strongly moderated by an inverse relationship in user satisfaction—although participants' performance was relatively good at high rates, their preferences decreased at higher rates and they strongly preferred story board surrogates that require more time to view but give them control [9].

Additional studies have demonstrated the importance of linguistic cues in supporting understanding [4], the tradeoff between high performance possibilities and users' comfort levels, and the many influences that individual human characteristics (such as experience) and video content characteristics (such as genre, visual style, pace, and subject matter) have in determining overall user performance and satisfaction. Given the current early stage of mass popularization of digital video, it is important that researchers continue to devote attention to designing and testing interfaces to multimedia libraries that are both user-centered and take advantage of the particular characteristics of digital video.

To this end, the work reported here isolated and examined fast forward surrogates that go far beyond the capabilities of analog video. Home VCRs can support one or two fast forward capabilities but at very low speeds (2-4 times real-time speed) [8]. By contrast, fast forwards of digital video can be simulated at any rate by selecting/displaying each Nth frame. This type of fast forward surrogate[1] is created by sampling from the video frames (rather than speeding up the display of the frames), but the result for the person viewing the surrogate is the ability to speed through the video much faster while still being able to perceive the images. In pilot studies and in a previous study that compared slide show, storyboard, and fast forward surrogates [25], fast forwards constructed in this Nth frame fashion were judged to be effective and realistic by users. Thus, the fast forward surrogates used in this study are an approximation of what may be both technically possible and also useful from the human perceptual system point of view.

[1] Rather than using the awkward phase, 'Nth frame fast forward', we simply call these surrogates 'fast forwards'.

The goal of the study was to identify the fastest speed at which people could still gain an understanding of the video represented by the fast forward surrogate. While it was presumed that users should maintain the ability to change the speed of the fast forward surrogate, based on the characteristics of the situation or on their own preferences, we hoped to identify the speed that could be used as a "default" setting for video retrieval applications.

3. Methodology: assessing video browsing success

Because this paper aims to present an approach to assessing people's success in using surrogates to browse a digital video library, as well as the results of a study of one important class of surrogate, the methods section is presented in two parts. In this first part, the six measures of surrogate use are described; in the next section, we provide an overview of the study procedures.

Figure 1 (next page) depicts the general framework within which this study is situated. Our overall goal is to understand (and predict) various performance and preference outcomes. Four main classes of variables influence these outcomes: the user task/need, individual user characteristics, video characteristics, and characteristics of the surrogates that represent the full videos (some examples of each class are shown in Figure 1). Our focus in this paper is on the speed of the fast forward surrogate within the context of all the different types of tasks. The study also took into account the video's genre (documentary vs. narrative) and visual style (black and white vs. color), and the users' video experience and basic demographics (e.g., gender, age). These variables are boldfaced in the figure.

In contrast to text documents, interacting with video relies on multiple informational channels, e.g., sound and moving images. Therefore, in addition to the usual linguistic/ textual measures used to assess the success of people's interactions with textual objects, visually oriented measures were designed for use in our studies of video browsing. Our perspective is that people's interactions with video have multiple facets on multiple dimensions. One dimension is perceptual and includes facets such as text superimposed on the images (visual channel, linguistic encoding), aural representations (audio channel,

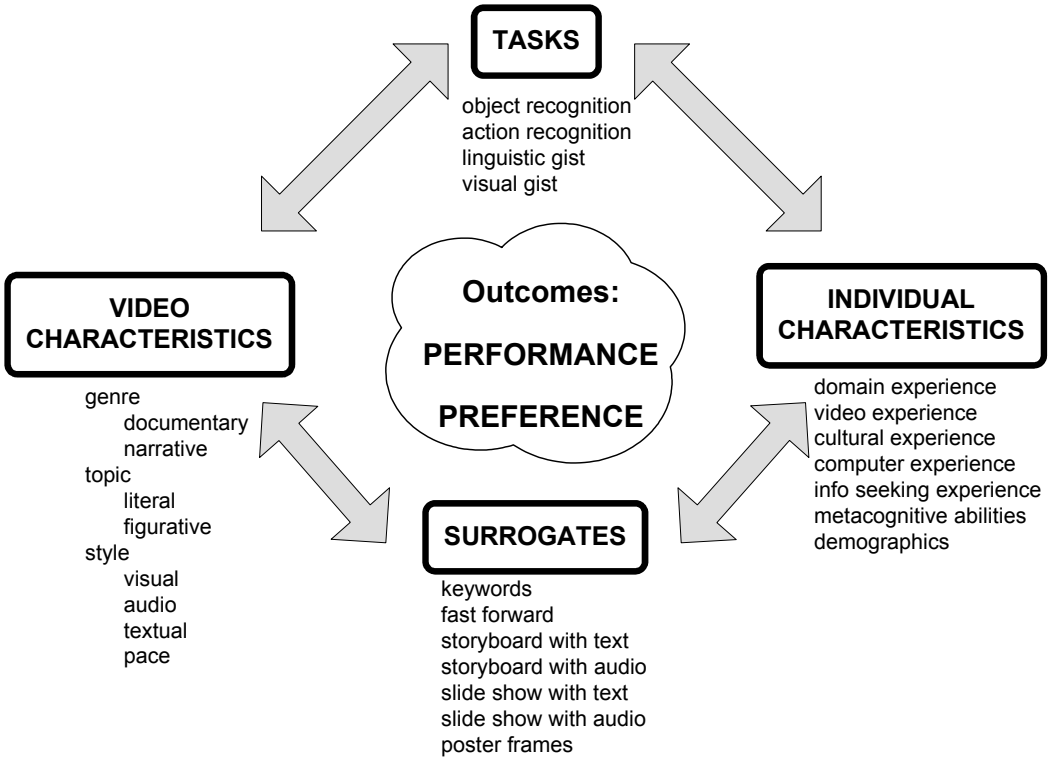

Figure 1. Video browsing assessment framework

linguistic encoding), non-verbal sound representation (audio channel, aural encoding), image representation (visual channel, graphical encoding), and motion representation (visual channel, temporal-graphical encoding). Another dimension is conceptual and includes facets such as the objects represented in video; the juxtapositions, actions, and interactions of these objects; and what these objects and actions, taken together, 'mean' to a viewer. Another dimension is pragmatic and includes facets related to the user's context such as current motivation, temporal and physical resources (e.g., how much time and what kinds of equipment, software, and authority they have), setting (e.g., work, home), and content facets such as the socio/cultural features inherent in the content[2]. Clearly, there are other dimensions and facets at play (e.g., Grodal's [7] theory of film strongly defines an emotional dimension of understanding), and many theories of how people process visual data (see Palmer [17] for a comprehensive treatment of vision science) and other sensory data.

In this and other studies, our aim is to instantiate some of these elements in well-defined tasks executable in studies of video browsing. For the present study, six tasks were defined:

- Object recognition (textual): Select objects seen in the surrogate from a list of nouns.
- Object recognition (graphical): Select objects seen in the surrogate from a set of still images.
- Action recognition: Select clips seen in the surrogate from a set of brief (2-3 second) clips.
- Linguistic gist comprehension (full text). Write a brief summary of the video represented by the surrogate.
- Linguistic gist comprehension (multiple choice): Select the best summary of the video from a set of five statements.
- Visual gist comprehension: Select objects that "belong" in the video represented by the surrogate, from a set of still images.

These tasks were selected for development because they closely relate to the tasks in which users engage when interacting with a library of digital videos [25]. The object recognition tasks are most closely related to the user goal of selecting particular frames from a video, just as the action recognition task is most closely related to the user goal of selecting a particular clip. For example, an elementary school teacher may be trying to locate an image or short clip illustrating the force of a hurricane; an effective surrogate will allow the teacher to recognize that such an image appears in the full video. The linguistic gist comprehension measures are most closely related to

[2] The socio/cultural 'meanings' parallel Panofsky's [18] iconographic level in his triarchic theory of of image understanding.

the users' ability to make relevance judgments concerning the video represented by the surrogate. If the user can accurately comprehend the gist of the full video by viewing only the surrogate, then we can conclude that the surrogate is useful in helping the user to select videos that are relevant to his or her current information need. The visual gist comprehension task is also related to making relevance judgments, but additionally incorporates stylistic considerations and so is most closely related to the users' desire to evaluate the movement or style in a video. For example, the user's information need may be for a modern-looking overview of the U.S. space program; an effective surrogate supports user judgments about these multiple facets of his or her information need. In summary, each of these measures is grounded in the real-world goals of users of digital video libraries.

Because these measures interact as people complete them, their sequencing is important. In the current study, participants were asked to write brief summaries (*linguistic gist comprehension, full text*) immediately after viewing each surrogate. Graphical and textual object recognition were the second and third tasks respectively. *Graphical object recognition* presented a set of 12 video frames with yes/no radio buttons. Half of the frames were from the stimulus video (i.e., they had been send in the surrogates) and half were not. Half the distractors (i.e., the incorrect frames) were selected from a portion of the video not included in the surrogate and half were selected from other videos. *Textual object recognition* presented a set of 12 words with yes/no radio buttons. Half the words were for objects included in the stimulus video and half were not. A mix of concrete objects (e.g., car) and abstract concepts (e.g., joy) were included among both the correct and distractor words. After the two object recognition tasks, the *action recognition* task was presented. The idea behind the action recognition task was to probe the roles that motion plays in video browsing. Here, participants were given six short clips (2-3 seconds each) and asked whether they had seen those clips in the surrogate (yes/no radio buttons). Participants could replay the clips if they wished (however, the original surrogate shown at the beginning of the tasks was not available for replay at any time in the session). Of the six clips, two were selected from the target video, two were from a video of a similar style, and two were from a video of a different style. Because none of the surrogates incorporated clips from the stimulus video, the participants would not have seen these clips before; however, they would have been exposed to individual frames from the two clips representing the target video. Next, the *visual gist comprehension* task was administered. Participants were given a set of twelve video frames with yes/no radio buttons. This time, they were asked to indicate whether the frames 'belonged' to the video represented by the surrogate they had seen at the beginning of the session. None of the frames had yet been seen by the study participants. Half the frames were selected from the same video but a different segment and half were selected from other videos. Finally, the *multiple choice linguistic comprehension* measure was administered. Subjects were given a set of five summary statements and asked to select the best. This measure was administered last so that the video summaries provided would not influence performance on the other measures.

4. The fast forward study methods

4.1. Participants

Study participants were recruited through the distribution of flyers on campus and especially in several classes related to video production, with the intention of recruiting study participants who would be interested in using a library of digital videos. The 45 subjects who participated in this study included 19 undergraduate students, 19 graduate students, 2 faculty members and 5 others. They came from a wide variety of departments, included 31 females and 14 males, and had a mean age of 26.1 (s.d.=7.9, ranging from 17 to 51 years old). Forty-four of the 45 subjects reported using computers on a daily basis and 32 of the 45 reported watching videos or films at least weekly; only 12 reported searching for videos or films on at least a weekly basis. The most common way to search for videos was online (32) followed by newspapers or magazines (10). Each subject spent about one hour in the study and received $10 for participation.

4.2. The videos

Four video segments were selected from the Open Video Project repository (www.open-video.org):

- '*Coney Island*' (1940, 9:19), a black & white documentary showing scenes of the amusement park;
- *'How Much Affection'* (1958, 19:48), a black & white educational film (narrative in style) exploring the boundaries of personal relationships;
- *'Iran'* (1954, 14:00), a color documentary on Iran in 1953; and
- *'On the Run'* (1956, 14:09), a color narrative about teenagers competing in the Mobilgas 'Safety Economy Run' in San Francisco.

A surrogate of a fifth video was used as a training example. The video was 'A Ride for Cinderella' (1937, 10:50), a cartoon advertising Chevrolets.

4.3. The fast forward surrogates

For each video segment, the four fast forward surrogates were created from the full video (MPEG-1 format). As noted above, fast forward surrogates for digital videos, created by sampling every Nth frame, could be produced at any speed by varying the value of N. The research team created and reviewed surrogates over a wide range of speeds, eventually selecting 32, 64, 128, and 256 as values for N in the current investigation. Thus, a full video of 18,000 frames would take approximately 10 minutes to view at the standard speed of 30 fps. A surrogate for it, created with N=32, would display 562 frames (every 32nd frame), taking about 19 seconds to view; whereas a surrogate for the same video at N=256 would take about 2 seconds to view. Thus the speeds of these four surrogates, compared to their original video speed, were 1:32, 1:64, 1:128 and 1:256. The surrogate for the training video was at 1:32 only.

4.4. Procedure

The study was conducted in the Interaction Design Lab (IDL) at the University of North Carolina at Chapel Hill. Each individual session was videotaped and the transcripts analyzed. Subjects first signed the consent form and filled out questionnaires about their experience and background. The session included five trials (including one practice trial with a surrogate at N=32) and in each trial the subjects were asked to watch one fast-forward surrogate and complete six tasks/measures. The four videos and four fast forward rates were counter-balanced so that each video/surrogate speed combination was approximately equally represented. After completing the six tasks/measures described above, each subject was debriefed with questions such as: What would you say are two strengths of this video surrogate? Did this surrogate have any strengths related to any of the tasks you had to perform? What would you say are two weaknesses of this video surrogate? Did this surrogate have any weaknesses related to any of the tasks you had to perform? Do you have any suggestions for improving this surrogate?

4.5. Data analysis

The surrogates and measures were presented through a web front end that piped all responses to a MySQL database. The responses were then analyzed through correlation analysis, analysis of variance, or Fisher's exact test. For the full-text linguistic gist comprehension task, an 8-point scoring scheme was devised and two team members scored the 180 responses independently. There was a .76 correlation (Pearson's *r*) between the raters' scores, and the mean of the two scores was used in further statistical analyses.

5. Results

In general, study participants were able to perform the tasks successfully with these four surrogates (see Table 1), scoring at above the midpoint on all tasks except the two linguistic gist comprehension tasks.

Table 1. Summary of performance

	Max. possible score	Mean	s.d.	Actual Min/Max
Object recognition (textual)	12	8.6	1.35	5/11
Object recognition (graphical)	12	9.7	1.65	5/12
Action recognition	6	4.5	0.93	2/6
Linguistic gist comprehension (full text)	8	2.9	1.72	0/8
Linguistic gist comprehension (multiple choice)	100%	46%		
Visual gist comprehension	12	8.4	1.41	5/12

The speed of the surrogate had a statistically significant effect on four of the tasks (see Figure 2, next page): object recognition (graphical) ($F=3.81$ with 3df, $p=0.0112$), action recognition ($F=3.62$ with 3df, $p=0.0143$), linguistic gist comprehension (full text) ($F=10.77$ with 3df, $p<0.0001$), and visual gist comprehension ($F=3.88$ with 3df, $p=0.0102$). Across all these tasks, as the "speed" of the surrogate increased, performance decreased. However, the point at which the performance difference became statistically significant was tested with Duncan's multiple range test and was found to vary from task to task, as noted in Figure 2.

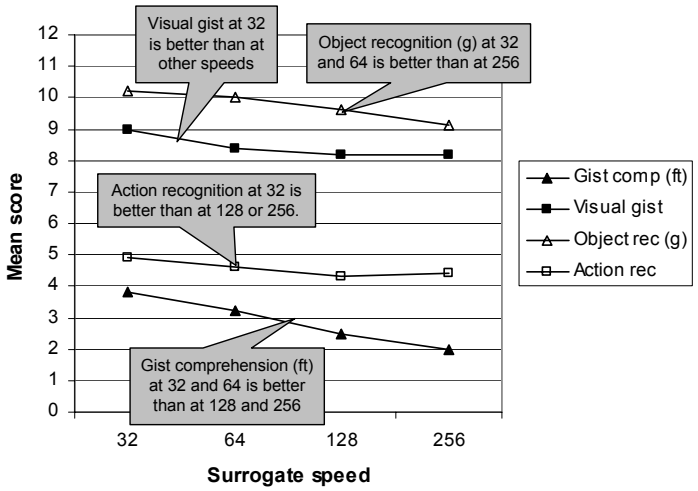

Figure 2. Effects of surrogate speed on performance

Performance on some tasks was also affected by the video with which the participant was interacting (see Table 2). Specifically, video characteristics affected object recognition (textual) ($F=11.94$ with 3 df, $p<0.0001$), object recognition (graphical) ($F=36.56$ with 3df, $p<0.0001$), gist comprehension (full text) ($F=3.15$ with 3df, $p=0.0263$), gist comprehension (multiple choice) (Fisher's exact test, $p<0.0001$), and visual gist comprehension ($F=3.64$ with 3df, $p=0.0139$). Duncan's multiple range test was used to investigate these differences further. Performance on object recognition (textual) was higher on 'Coney Island' and 'On the Run' than on the other two videos. Performance on object recognition (graphical) was highest on 'How Much Affection', followed by 'Iran', followed by the other two segments. Linguistic gist comprehension (full text) was higher on 'How Much Affection' than on 'Coney Island' or 'On the Run'. While no post hoc tests could be run, it would appear that linguistic gist comprehension (multiple choice) was highest on 'Iran', followed by 'Coney Island', followed by the other two videos. Visual gist performance was higher on 'On the Run' than on 'How Much Affection' and 'Iran'. In addition to the main effects of surrogate speed and video, the interactions between these two variables were investigated. They were significant only for action recognition.

Effects of participant characteristics on performance were also investigated. Sex of participant was not related to performance. Age effects were investigated by comparing the performance of those over 25 (n=96 observations) with those under 25 (n=84 observations), splitting the sample at the median age. There were no age effects except for action recognition ($t=2.32$ with 178 df, $p=0.0214$), where the older participants performed better (mean score of 4.7 versus 4.4). A parallel finding was associated with participant status ($F=4.54$ with 2 df, $p=0.0119$); undergraduate students (mean score = 4.3) did not perform as well on the action recognition task as the graduate students and faculty/other participants (mean scores = 4.7 and 4.8, respectively). In addition, the frequency with which participants searched for videos was weakly related to linguistic gist comprehension (full text) (Spearman's rho = 0.16, $p=0.0275$), and object recognition (textual) (Spearman's rho = 0.15, $p=0.0430$).

Table 2. Mean performance scores, by video

	Iran	Coney Island	On the Run	How Much Affection
*Gist comprehension (full text) (max=8)	3.2	2.5	2.5	3.3
*Gist comprehension (multiple choice)	89%	49%	24%	22%
*Visual gist (max=12)	8.0	8.4	9.0	8.3
*Object recognition (textual) (max=12)	7.9	9.2	9.1	8.3
*Object recognition (graphical) (max=12)	10.1	8.9	8.6	11.2
Action recognition (max=6)	4.8	4.5	4.6	4.3

Asterisk indicates statistically significant differences, by video

6. Discussion

We began with the question, how fast is too fast? Participants in this study were able to perform well on a variety of tasks, regardless of speed. Increased surrogate speed had negative effects on performance on four measures: object recognition (graphical), action recognition, full text linguistic gist comprehension, and visual gist comprehension. For these four measures,

participants performed better at the two slowest surrogate speeds. From these results, we conclude that designers should create fast forward surrogates that include at least 1/64th of the video frames, to be sure that user performance is adequate. In similar studies (e.g., Ding's [4] study of the use of slide show video surrogates and Öquist's [16] study of the rapid serial visual presentation of text such as), it has been found that, even when study participants performed relatively well, they were not pleased with the experience. Therefore, the selection of 1:64 as a recommended speed for the fast forward surrogate was intended to be conservative, supporting good performance and user satisfaction with the experience of using the surrogate.

On two of the measures, object recognition (textual) and multiple choice linguistic gist comprehension, performance was not affected by the speed of the surrogate. On these two measures, mean performance was adequate: 8.6 out of 12 correct on object recognition and 46% correct on this measure of gist comprehension. Therefore, we can conclude that performance is adequate, even at the highest speeds tested in this study. If designers set a default speed of 1:64, users should be able to perform these tasks well.

In a previous study [25], participants raised the notion that an important component of video materials is the motion perceived in viewing them. Thus, surrogates that simulate this motion (even in high speed) can help people comprehend the video's gist more completely. A variation of the 'how fast?' question for this implementation of fast forward surrogate is how many frames can be removed before reaching the point at which the surrogate is perceived as a slide show of discrete images, a single image, or none at all? From this point of view, selecting an N in the 50-100 range works well for many genres of video. Picking every 60th frame, for example, means that the user sees a frame from each 2 seconds of real-time video (playing at 30 fps) and this is enough to give a sense of motion for many videos without too many jarring visual disconnects between frames. For high-action video, N may need to be adjusted downward and for low-action video N could be adjusted upward.

Thus, we recommend a default speed of 1/64 of the video's frames based upon our empirical results but also strongly suggest that mechanisms for user control of display speed be included in a video browsing interface. Users may want to adjust the rate depending on video genre, the kind of task at hand, or personal preferences. For a number of tasks, study participants were able to obtain high performance when viewing as few as 1/256 of the video's frames, and there was certainly individual variation in the speed at which people could perform adequately. This variation in optimal speed will likely be influenced by the task the user brings to the browsing session, the past experience and knowledge of the user, and the setting of the browsing session. It may also be affected by the augmentation of the fast forward with other metadata or representations of the video. For example, the surrogate may be augmented with audio keywords, or the viewing of the surrogate may be preceded by the viewing of metadata describing the video. The current study eliminated any augmentation of the fast forward surrogate in order to isolate the effects of speed on performance, but any real-world digital library would not be constrained in this way and would likely include audio keywords or other metadata in representing the videos in the collection. The addition of these other sources of information will most likely improve user performance with the fast forward surrogate.

Performance on several of the measures was affected by the video used as a stimulus. Differences in the videos were related to the participants' ability to recognize objects from the surrogate (both graphical and textual) and to identify frames that "belonged" in the stimulus video (i.e., the visual gist measure). In addition, the effects of differences in the videos interacted with surrogate speed in relation to action recognition performance. It is unclear which characteristics of the stimulus videos are the sources of these effects. The videos were selected to represent both narrative and documentary styles (two of each) and both color and black-and-white (two of each); the observed effects were not related to these video characteristics. Other possibilities identified as relevant video characteristics in our research framework (see Figure 1) include the rate of scene change, homogeneity of content [25] or other topical or stylistic features. These video characteristics should be investigated through additional analyses of data from the current study, as well as through additional studies.

Reliable, valid measures of user performance in video browsing are needed in order to make progress in this line of research. The six measures employed in this study are a good starting point for such efforts. They represent multiple facets of video browsing behavior: some more conceptual (gist comprehension and visual gist), some more perceptual (object and action recognition); some text-based (object recognition (textual), linguistic gist comprehension) and some image-based (object recognition (graphical), action recognition, visual gist). Further analyses of the measures' reliability are currently being conducted. While they already have some face validity, further analyses of their measurement validity will need to be based on a stronger theoretical understanding of video browsing behaviors. In particular, their applicability in studies of interactions with videos of additional genres, e.g., news broadcasts, should be investigated. We encourage other researchers to employ

these measures in their studies and test their psychometric qualities within a variety of video browsing contexts.

In addition, studies of users' interactions with interfaces that provide access to digital video libraries should incorporate measures of user satisfaction. While no such measure was incorporated in the current study, user comments concerning their reactions to the surrogates were systematically collected. These comments will form the foundation on which a valid measure of user satisfaction with video browsing interfaces can (and will) be developed for use in future studies.

7. Conclusion

As one of the early studies on people's use of fast forward surrogates, the results of this study must be evaluated in terms of the necessary limitations of the study design. The design was intended to isolate the effects of surrogate speed, and so could not take into account the effects of potential interactions with other surrogate features if implemented in context. For example, no audio was provided with the fast forward surrogates investigated here; yet, it is likely that a fully-functioning digital library interface would incorporate audio (such as the audio keywords investigated in a previous study [25]). Similarly, these surrogates were viewed in isolation; in a fully-functioning interface, it is likely that users would have viewed additional metadata (e.g., video title or poster frame) before accessing the fast forward surrogates. As such, the fast forward surrogate speed of 1:64 is probably a conservative estimate of the speed at which people can perform well with such surrogates augmented with audio or other metadata.

Nonetheless, for Nth frame fast forwards, we plan to adopt 64 for N as the default setting for fast forward surrogates implemented on the Open Video site.[3] In addition, we will provide control mechanisms that will give users control over the speed of the fast forward display. We are particularly interested in seeing the effects of this design decision as this class of surrogate is implemented within the context of a fully-functioning system (incorporating alternative surrogates and control mechanisms).

We are also interested in the relationship between users' ability to perform with high-speed fast forward surrogates and their satisfaction with that interaction. We are convinced that there is a performance-satisfaction tradeoff—although people may be able to perform accurately at high speeds, they seem willing to exchange some performance benefits for surrogates that are comfortable and satisfying. While the discrepancies between users' performance and their satisfaction have long been an issue in relation to usability [13], there is recent renewed interest in the affective dimensions of people's interactions with computer-based tools [3,14]. Our future studies will incorporate a measure of user satisfaction, thus explicitly taking into account the "user experience" as people interact with digital video surrogates and the mechanisms that control them.

What is clear from this work is that creating effective digital library interfaces that support video browsing and retrieval will demand a range of user control mechanisms and underlying representations for video. Making sense of video content is a complex cognitive act, depending on multiple facets and cues. Interfaces that aid people in making sense of video based on surrogates must aim to provide a rich mix of these facets and cues and to place them under user control. Designers of digital library interfaces are advised to consider providing such a mix in their implementations.

[3] For videos less than 10 minutes in duration, an N of 64 does not produce enough frames to create a fast forward surrogate of useful length. We plan to use an N of 32 for shorter videos.

8. Acknowledgments

We thank the participants in the study. This work is supported by National Science Foundation (NSF) Grant IIS 0099638.

9. References

[1] M. Christel, A. Hauptmann, A. Warmack, and S. Crosby, "Adjustable filmstrips and skims as abstractions for a digital video library", *IEEE Advances in Digital Libraries Conference, (Baltimore, MD, May, 1999),* pp. 19-21.

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

[3] A. Dillon, "Beyond usability: process, outcome and affect in human computer interactions", paper presented as the Lazerow Lecture, Faculty of Information Studies, University of Toronto, 2001.

[4] W. Ding, G. Marchionini, and D. Soergel, "Multimodal surrogates for video browsing". *Proceedings of Digital Libraries '99. the Fourth Annual ACM Conference on Digital Libraries (Berkeley, CA, August 11-14, 1999)*, pp. 85-93.

[5] S. Drucker, A. Glatzer, S. DeMar, and C. Wong, "SmartSkip: Consumer level browsing and skipping of

digital video content", *Proceedings of CHI '02: Human Factors in Computing Systems (Minneapolis, April 20-25, 2002)*, pp. 219-226.

[6] G. Geisler, G. Marchionini, M. Nelson, R. Spinks, and M. Yang, "Interface concepts for the Open Video Project", *ASIST 2001: Proceedings.of the 64th ASIST Annual Meeting (Washington, DC, Nov. 3-8, 2001), Volume 38,* pp. 58-75.

[7] T. Grodal, *Moving Pictures --- A New Theory of Film Genres, Feelings, and Cognition.* Oxford: Clarendon Press, 1997.

[8] P. Hoff, *Consumer Electronics for Engineers.* Cambridge: Cambridge University Press, 1998.

[9] A. Komlodi and G. Marchionini, "Key frame preview techniques for video browsing", *Proceedings of ACM Digital Libraries '98 (Pittsburgh, PA, June 24-27,1998),* pp. 118-125.

[10] H. Lee and A. Smeaton, A. "Designing the user interface for the Físchlár digital video library", *Journal of Digital Information, 2*(4), 2002. http://jodi.ecs.soton.ac.uk/Articles/v02/i04/Lee/

[11] R. Lienhart, S. Pfeiffer, and W. Effelsberg, „Video abstracting", *Communications of the ACM*, *40*(12), 1997, pp. 54-62.

[12] G. Marchionini, V. Nolet, H. Williams, W. Ding, J. Beale, A. Rose, A. Gordon, E. Enomoto, and L. Harbinson, "Content + connectivity => community: digital resources for a learning community", *Proceedings of ACM Digital Libraries '97 (Philadelphia, PA: July 23-26, 1997),* pp. 212-220.

[13] J. Nielsen, and J. Levy, "Measuring usability: preference vs. performance", *Communications of the ACM, 37*(4), 1994, pp. 66-75.

[14] D. A. Norman, "Emotion & design: attractive things work better", *ACM Interactions, 9*(4), 2002, pp. 36-42.

[15] B. O'Connor, "Access to moving image documents: background concepts and proposals for surrogates for film and video works", *Journal of Documentation, 41*(4), 1985, pp. 209-220.

[16] G. Öquist, Adaptive rapid serial visual presentation, Masters' thesis, Dept. of Linguistics, Uppsala University, 2001. http://stp.ling.uu.se/~matsd/thesis/arch/2001-009.pdf.

[17] S. Palmer, *Vision Science: Photons to Phenomenology*. Cambridge, MA: MIT Press, 1999.

[18] E. Panofsky, *Meaning in the Visual Arts: Papers In and On Art History.* Garden City, NY: Doubleday, 1955.

[19] D. Ponceleon, A. Amir, S. Srinivasan, T. Syeda-Mahmood, and D. Petkovic, "CueVideo: Automated multimedia indexing and retrieval", *ACM Multimedia '99 (Orlando, FL, Oct. 1999),* p. 199.

[20] D. Ponceleon, and A. Dieberger, "Hierarchical brushing in a collection of video data", *HICSS'34 (Hawaii International Conference on Systems Science), MiniTrack on Video in the office, Maui, HI, January 2001*.

[21] A. Rose, W. Ding, G. Marchionini, J. Beale, Jr., and V. Nolet, "Building an electronic learning community: from design to implementation", *CHI Conference Proceedings (Los Angeles, April 18-23, 1998),* pp. 203-210.

[22] T. Tse, G. Marchionini, W. Ding, L. Slaughter, and A. Komlodi, A., "Dynamic keyframe presentation techniques for augmenting video browsing", *Proceedings of AVI '98: Advanced Visual Interfaces (L' Aquila, Italy, May 25-27, 1998)*, pp. 185-194.

[23] The user-interface development of Físchlár digital video system, 1999. http://www.computing.dcu.ie/~hlee/ProgressHtml/Progress.html

[24] H. Wactlar, S. Stevens, M. Smith, and T. Kanade, "Intelligent access to digital video: the InforMedia Project", *IEEE Computer, 29*(5), 1996, pp. 46-52.

[25] B. Wildemuth, G.. Marchionini, T. Wilkens, M. Yang, G. Geisler, B. Fowler, A. Hughes, and X. Mu, "Alternative surrogates for video objects in a digital library: users' perspectives on their relative usability", *Proceedings of the European Conference on Digital Libraries (Rome, September 16-18, 2002)*, in press.