Comprehension and Object Recognition Capabilities for Presentations of Simultaneous Video Key Frame Surrogates

Laura Slaughter[1], Ben Shneiderman[2], and Gary Marchionini[1]

[1]Digital Library Research Group
College of Library and Information Services
University of Maryland, College Park, MD 20742

[2]Department of Computer Science, Human Computer Interaction Laboratory
and Institute for Systems Research
University of Maryland, College Park, MD 20742

e-mail: lauras@wam.umd.edu, ben@cs.umd.edu, march@oriole.umd.edu

The demand for more efficient browsing of video data is expected to increase as greater access to this type of data becomes available. This experiment looked at one technique for displaying video data using key frame surrogates that are presented as a "slide show". Subjects viewed key frames for between one and four video clips simultaneously. Following this presentation, the subjects performed object recognition and gist comprehension tasks in order to determine human thresholds for divided attention between these multiple displays. It was our belief that subject performance would degrade as the number of slide shows shown simultaneously increased. For object recognition and gist comprehension tasks, a decrease in performance between the one slide show display and the two, three or four slide show displays was found. In the case of two or three video presentations, performance is about the same, and there remains adequate object recognition abilities and comprehension of the video clips. Performance drops off to unacceptable levels when four slide shows are displayed at once.

1 Introduction

It is anticipated that the digital libraries, world wide web contributions, and other information stores of the future will provide diverse collections of data. The amount of digitized video information included among these collections is expected to increase substantially due to the decreasing cost of storage space and improving communications technology. The growth in the number of video "documents" will necessitate the development of systems that allow users to search, browse and display these data formats in an effective and useful manner. A user may first conduct a search of the database, employing the use of query formulation in a goal oriented, systematic approach to information seeking. Then, the user may browse through the subset of video clips that were retrieved through the searching process. Browsing is the ability to allow the user to employ scanning, observing, navigating and monitoring strategies to find items to examine in closer detail (Marchionini, 1995). Specifically, the primary goal of the research presented in this paper is to contribute to the design of interfaces that support browsing, allowing rapid screening of video documents. In view of our expectations of the future, studies that investigate the methods of representing and displaying video data for retrieval and browsing should be conducted.

This investigation employed the use of video key frame surrogates that are displayed as a "slide show". Key frame surrogates used for this research consist of sets of salient still images extracted from three to five minute digital video clips, each set corresponding to a separate video clip. Video key frames are analogous to "abstracts" that are created for textual documents (O'Connor, 1991). In this experiment, key frames are extracted using a technique that automatically segments video clips according to a change in scene. It is maintained that these key frame surrogates characterize the content of videos (Kobla *et. al.*, 1996). Key frames can be viewed as a "slide show" in which frames are flipped through at a constant rate. In this study, a specific question about limits on the number of slide shows presented simultaneously is asked. One through four slide shows are shown followed by object identification and gist comprehension tasks that are used to determine human limitations. From the data described in this study we attempt to add to the list of necessary elements required in systems for video browsing.

One type of tool for creating and displaying video surrogates, the Video Streamer, has been developed at the MIT Media Laboratory (Elliot, 1993). The idea behind the Video Streamer is that although some aspects of video can be automatically parsed, more detailed representations of video content require user annotations. The video clips produced by the Video Streamer are depicted as three dimensional video blocks with the pictures stacked, representing the temporal attributes of the stream. Users may also be able to comprehend actions that take place in the clip from the edges of images seen on the sides of the block. The Streamer also includes a utility to automatically recognize cuts between shots. Using this feature, the user can save the top frame of each shot in the stream to create an overview.

The SWIM Project (Show What I Mean) is a video classification project developed for video parsing, indexing, and retrieval based on video content. The video parser segments video into individual shots and extracts key frames that best represent the content of each shot (Zhang *et. al.*, 1995). The key frames are used in a hierarchical browser that allows users to view video at two levels of granularity which are an overview level and a detail level. Key frames in the SWIM system are used to browse video shots following queries. The key frames are also used to formulate visual queries based on color or texture.

Christel *et. al.* (1997) describe several "video abstraction" methods implemented in the Informedia Digital Video Library Project at Carnegie Mellon University. These include a single image that represents a video clip, called a poster frame which is another term for key frame. The poster frames were found to be a faster method for the quick location of relevant videos than text lists of video titles. Filmstrips are a term used to describe a type of abstraction where sets of still images arranged in chronological order. Skims are compact representations that summarize video into short concatenations of the video and audio data. All of these abstractions can be considered possible design considerations for inclusion in multimedia digital libraries.

There have been many mechanisms for representing video data that have been proposed (Elliot, 1993; Zhang *et. al.*, 1995; Christel *et. al.*, 1997). These mechanisms provide useful ideas that may contribute to the design of multimedia systems for browsing and retrieving video data. Included among these are methods for representing the length of the entire video, using key frames to browse different levels of granularity, and use of key frames for quick location of relevant videos. In this experiment, we

study a method of displaying key frames using a slide show format. Multiple slide shows can be an effective way to browse large numbers of video clips.

Using the slide show method of displaying key frames conserves screen space, however, the speed of the presentation must be optimal for human perception. A study by Ding (1996) provided exploratory data on human ability to perceive scrolling key frame surrogates at specific speeds. From her study, it was determined that the baseline speed of 1 frame per second (fps) allowed the best performance by users and that 12 fps could be a "speed breakpoint". It was postulated that beyond this speed, object identification performance will remain poor. Ding (1996) further concluded that eight fps is an acceptable speed for the purpose of object identification and gist extraction. Another interesting finding was that slower speeds were required when completing object identification tasks than for the comprehension tasks given. Ding suggests that higher speeds can be used when a basic understanding of the content is desired, while lower speeds are necessary for identifying individual objects. The current study displays key frames at the baseline speed of 1 fps. This speed for key frame presentation in slide shows was used to allow the best performance by subjects for object recognition and gist comprehension. Furthermore, it was anticipated that the addition of a slide show will be perceived by subjects as increasing the speed of the key frames. For example, if there are four simultaneous slide shows, speed perception will be at least 4 times 1 fps giving a perceived speed of 4 fps. Even at a perceived speed of 4 fps, Ding's results show that human abilities for object recognition and gist comprehension are at an acceptable level.

The design of systems for browsing video data should focus on human perceptual and cognitive abilities rather than system characteristics (Shneiderman, 1998). The fundamental problem with a simultaneous presentation of these key frame sets is that humans have limited abilities to divide attention between stimuli, all of which need to be processed. It is often difficult to maintain several things in working memory. These limits of divided attention correspond to the limited cognitive abilities of humans to time-share performance between two tasks.

There are two theories of divided attention, single resource theory and multiple resource theory, both of which can be used to define elements of this experiment. Single resource theory (Kahneman, 1973) proposes that there is a single undifferentiated pool of resources available to all tasks and mental activities. As task demands increase, either by harder tasks or more tasks, the supply of resources increases until the increase is insufficient to compensate for the demand, at which time, performance declines. Multiple resource theory (Wickens, 1992) states that instead of one pool of resources, there are several different dichotomous dimensions of resources. These dimensions are stages, perceptual modalities, and processing codes. Perceptual modalities is of primary interest for this experiment. In the perceptual modalities dimension, dual tasks that are cross-modal, that is, are split between two different sources (auditory, visual) produce better performance than tasks that are inter-modal, which split tasks between one source (two visual inputs). (Wickens, Sandry and Vidulick, 1983). As in our experiment, because it is inter-modal, placing the video clip screens far apart may cause difficulty due to the additional visual scanning between them. Due to this, the screens are placed beside each other forming a square to allow for easier visual scanning. Due to the inter-modal nature of the task, and the limitations of human resources to divide attention, the theory leads us to believe that users will

experience difficulty dividing attention between the simultaneous video displays. What we expect to learn is at what point resources are no longer available for viewing multiple video screens. It is hoped that this experiment will shed light not only on optimal system design for browsing video data but will elucidate the psychological aspects as well.

2 Experiment

The primary goal of this research is to explore an aspect of video surrogate display that may be useful for browsing video data. User needs for quickly locating a particular video clip among a multitude of video data within a digital library may be met through an interface that allows simultaneous display of these surrogates. Before implementing an interface for browsing video data that allows concurrent slide shows, one must first understand human limitations and abilities to process the information. The research questions posed in this experiment were designed so that human aptitudes for viewing multiple slide shows can be carefully assessed.

2.1 Hypotheses

Research Question 1

The first research question addressed the primary interest in completing this study. We wished to determine the threshold for human performance. This overall research question asked if performance is dependent on the number of concurrent video clips presented. Our hypothesis states that subjects should have a better understanding and ability to recognize objects for a clip in conditions with fewer slide shows. This is believed to be the case simply because the subjects must divide their attention simultaneously for several videos at once. The addition of a slide show uses up more mental resources, making it more difficult for the user to get the gist of the story and focus on content for each video.

H1: As the number of slide shows increases, performance will decrease on the object recognition and comprehension tasks.

Research Question 2

The experiment was repeated twice by each subject in order to determine how previous viewing may affect user performance. Each subject viewed the same slide shows following the first trial. They then repeated the exact same object recognition and comprehension tasks that they did in the first trial. We hypothesized that after viewing video key frames a second time, the subjects would improve their accuracy for identifying objects and would have a greater understanding of the gist of the video clip.

H2: Viewing the slide shows a second time will increase performance on the object recognition and comprehension tasks.

Research Question 3

A questionnaire measuring user perceptions was completed by each subject. It was our belief that user ability to obtain the gist of the video and identify objects within the key frames would not be dependent on perceptions of the speed and number of slide shows.

H3: User perceptions of the difficulty of viewing multiple slide shows will not affect object recognition and comprehension performance.

2.2 Subjects

Twenty-eight undergraduate students (20 males, 8 females) enrolled in introductory psychology at University of Maryland, College Park participated in this experiment. Subjects took part in the investigation in order to fulfill a course requirement. The subjects voluntarily chose this experiment, it is unknown why the number of males who signed up is so much greater than the number of females.

2.3 Materials

The video key frames used in the experiment were segmented from digitized MPEG video clips. These were extracted from 3-5 minute video clips from the following Discovery Channel educational programs: Spirits of Rainforest, Flight Over the Equator, The Space Program and The Revolutionary War. All the video programs were the same difficulty level for comprehension of content. They were meant as learning tools for novices and students (Baltimore Learning Communities Project, http://www.learn.umd.edu). Five 3-5 minute video clips were chosen so that the short segment conveyed a meaningful story. Video Clip 1 (Spirits of Rainforest) showed how researchers conduct studies on monkeys in a rain forest. Specifically, it showed how the researchers tagged monkeys for tracking. Video Clip 2 (Flight Over Equator) showed scenes of Singapore's industrialization, its culture and people in daily life. Video Clip 3 (Spirits of Rainforest) showed how a native American tribe makes a living in the jungle. Video Clip 4 (The Revolutionary War) showed enacted scenes of the Battle of Concord during the Revolutionary War. Video Clip 5 (The Space Program) was about the Apollo 11 astronauts training and moon landing activities.

This study used an automated procedure for key frame extraction. The key frames were created by a color histogram-based segmenting technique developed in the Center for Automation Research at University of Maryland (CFAR) that creates key frame shots based on scene changes. This technique has been shown to be useful for browsing, indexing and retrieval (Kobla *et. al.*, 1996). Although selected key frames were chosen from the computer-generated output in order to eliminate duplicate pictures and "fuzzy" images, the key frames used in this study were more from an automated procedure than human selection of "shots". Computer-generated extraction methods are used in this experiment for generalization of results to computerized extractions. This is due to the fact that surrogates generated from large archival databases of video data will most likely be done through the use of automated procedures. There were five slide shows of the key frames (one for each video clip), each of these was composed of eighteen individual key frame shots. For the two through four multiples, the slide shows flipped through the key frame shots synchronously. All slide shows were displayed at 1 frame per second.

The experiment took place in the Academic Information Technology Services (AITs) Teaching Theater at the University of Maryland, College Park campus. The teaching theater is equipped with 25 Gateway(TM) Pentium computers running the Windows 95 operating system. Subjects viewed the video key frames using Netscape version 3.0. Monitors used in the experiment displayed 256 colors at 800x600 pixels

resolution. A HTML/Javascript file in Netscape controlled the look of the interface and the rate at which the key frames were shown (1 fps) to the subject. HTML files were used to administer the object recognition, sentence writing, comprehension tasks and the user perception questionnaire. The file displaying the video key frames was placed on the hard drive of the machine so that the speed for the key frames would not be affected by speed of the server. All other experimental files were placed on a WWW server.

The interface used for displaying the video key frames is shown in Figure 1. All subjects were given a practice trial prior to the actual experiment. In the practice trial, the subjects were shown video clip 5 in the upper left corner of the video display interface. The videos were placed in a square-shaped format to consolidate the videos into a single space. This minimized the need for subjects viewing three and four multiple slide shows to move their heads and eyes in order to get an overall view of all the videos at once. In the three slide show condition, the third was added to the bottom right hand corner with the expectation that subjects would move their eyes in a clockwise motion while viewing the screen, enhancing visual scanning.

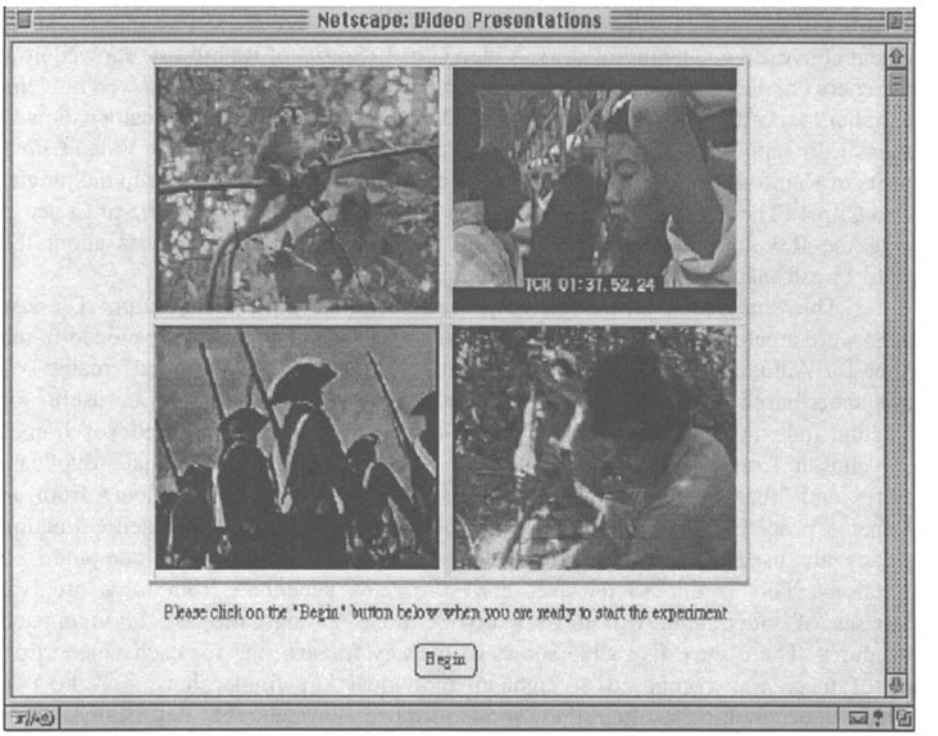

Fig. 1. Experiment interface showing four simultaneously displayed key frames. Pictures shown are Copyright ©Discovery Channel , Inc. All rights reserved.

2.4 Experimental Tasks and Data Preparation

For the object recognition tasks, there were an equal number of distracter objects and objects that actually appeared in the key frames. The list was developed by equalizing the probability of selecting false positives with the true items listed. In other

words, the distracter objects were designed to fit things expected to be the key frames that might be a part of the story of the video clip. For example, although a horse might be expected to be present in a battle during the Revolutionary War, it is never shown in the video key frames presented to the subjects. In the design of the object lists, face validity was an important aspect and terms were chosen so that they were at the same specificity and difficulty level (Ding, 1996). Performance on object recognition tasks was measured based on accuracy scores and accuracy percentage. This is calculated based on two scores. The first is the number of objects that were correctly identified by the subjects. The second is the number of distracter objects that were not identified by the subject. These two are added together and divided by 20 (total objects in the list) to obtain the subjects' overall accuracy percentage. The number of wrong items checked was also recorded and percentages obtained.

The sentence writing task simply asked the subjects to write down in one or two sentences what they believed was the gist of the video. A content analysis was performed on the sentences provided by the subjects. Keywords and phrases were derived from the sentences and placed into one of the following categories: People, Objects, Actions/Concepts, and Places. The keywords and phrases were taken directly from what was written in the sentences. However, some keywords or phrases varied slightly from subject to subject, but the main idea remained the same. In this case, a single keyword/phrase was chosen by the experimenter and used as the description.

A single multiple choice question was presented for each video clip to measure comprehension of the video clip's meaning. The comprehension questions were designed using two principles. The first was to maximize the distinction between choices. The second was to minimize additional prior knowledge about the videos (Ding, 1996). For each subject, it was indicated whether or not the answer for the multiple choice question was correct or incorrect. The percentage of correct answers was obtained for conditions 2, 3, and 4.

The user perception questionnaire measured subject's perception of the speed that the key frame shots were shown and their perception of the number of slide shows presented at one time. User perceptions were measured using a likert scale (1 through 7) with the adjectives listed next to the question anchored above the left and right ends of the scale. Note that the "number of videos" questions are not asked for condition 1 because there was only one slide show. Listed below are the questions that were asked. The subjects were instructed to circle the number that corresponds to their perception of the multiple slide shows.

About object identification please evaluate....
1. The speed that the videos were presented was: too slow 1 2 3 4 5 6 7 too fast
2. The number of videos presented was: imperceptible 1 2 3 4 5 6 7 perceivable

About video comprehension please evaluate....
3. The speed that the videos were presented was: too slow 1 2 3 4 5 6 7 too fast
4. The number of videos presented was: imperceptible 1 2 3 4 5 6 7 perceivable

2.5 Procedures

A series of HTML/Javascript files was used to administer the experiment and gather the data collected. Subjects were first shown a "Welcome" screen that gave

instructions for the experiment. This was read to the subjects. Subjects were also told that it was important to complete all tasks in the experiment, including sentence writing. The subjects were told that they would complete tasks based on what they saw in the key frame shots, and that it was important they pay close attention to the slide shows presented. Each subject was randomly assigned to one of four groups. (Groups 1-4 which correspond to the number of video key frame surrogates presented.) Subjects were presented with a practice trial and then completed the practice object recognition and comprehension tasks. The subject viewed the video surrogate slide show(s) in the experimental condition assigned. The subject then completed the object recognition and comprehension tasks. It is important to note that the multiple choice comprehension question was displayed to subjects after they wrote sentences describing what was believed to be the meaning of the videos. This eliminated any bias that would have been introduced otherwise. Lastly, the subject received a questionnaire in order to evaluate the subject perception of the speed of key frame presentation and number of slide shows. The subject then completed the experiment a second time (excluding the practice) using the same slide shows.

3 Results

The dependent variables in this experiment were a) object recognition for the first and second time subjects completed the experiment b) multiple choice comprehension for the first and second time subjects completed the experiment c) content analysis for sentences written by subjects describing the gist of the video clips and d) evaluation scales for the number and speed of videos (1-7 Likert scales).

3.1 Object Recognition, Multiple Choice Comprehension Questions

Using the Kruskal-Wallis non-parametric test, a significant difference was found between conditions 1 through 4 for object recognition after watching the key frames one time through. $H(3)=15.96$, $p<0.001$. The rank information is listed in Table 1. The mean rank for condition 1 is listed as the highest indicating that condition 1 performed better on the object recognition tasks than conditions 2, 3, and 4. Conditions 2 and 3, having almost the same mean rank, both perform better than condition 4.

For object recognition data for the second time through the experiment, the Kruskal-Wallis test again showed a significant difference between the four conditions. $H(3)=12.74$, $p<0.005$. The rank information is listed in Table 1. The mean rank for condition 1 is listed as the highest indicating that condition 1 performed better on the object recognition tasks than conditions 2, 3, and 4. Conditions 2 and 3, having almost the same mean rank, both perform better than condition 4. Results are almost identical in performance for the first time through the experiment.

There were no significant differences found between conditions for the multiple choice comprehension questions (first time through). Kruskal-Wallis revealed $H(3)=3.88$, $p<0.25$. The rank information is listed in Table 1. The mean rank for all conditions appear approximately the same indicating that subjects in all conditions performed equally well on the comprehension questions.

There were also no significant differences found between conditions for the multiple choice comprehension questions (second time through). Kruskal-Wallis revealed $H(3)=.591$, $p<0.90$. The rank information is listed in Table 1. Results are almost identical in performance for the first time through the experiment. The mean

rank for all conditions are approximately the same indicating that subjects in all of the conditions performed equally well on the comprehension questions.

Table 1. Mean Ranks for Object Recognition and Multiple Choice Comprehension Tasks. For each condition, n=7.

	Mean Rank			
	Object Recognition (first time through)	Object Recognition (second time through)	Multiple Choice Comprehension (first time through)	Multiple Choice Comprehension (second time through)
1 slide show	24.1	22.0	10.6	12.8
2 slide shows	14.1	14.5	19.1	14.1
3 slide shows	12.9	15.1	14.1	15.5
4 slide shows	6.9	6.4	14.1	15.7

3.2 Sentence Analysis

Subjects were asked to write about what they believed was the gist of the videos. A summary of the sentence descriptions written by subjects was constructed for a content analysis. The contents of the sentences were broken into four different categories of responses, each about a different aspect of the videos that was mentioned by subjects: Gist, People, Objects, and Places. Keywords and phrases were extracted from the sentences and identical responses were grouped together under a single keyword/phrase using subjects' actual comments. The keywords/phrases that subjects wrote were listed and it is possible for a subject to have included multiple entries for each category. For example, the subject could list two descriptions for the gist of a video clip story, "research on monkeys" and "observing monkeys".

The primary analysis made from the sentence data concerned the subjects' ability to correctly identify the central idea of the video clip from the slide show. The Gist category formed the basis for this examination. The number of correct concepts described by subjects was counted for each experimental condition, across all slide shows shown to them and divided by the total number of concepts written about. When the subject wrote that they did not know what the original video clip was about, or were unsure of it, these comments were grouped as "main idea unclear". Table 2 lists the results of this analysis.

The People, Objects, and Places categories were useful for getting a sense of what the subjects thought they saw in the slide shows. For the most part, subjects wrote correct responses about the objects, types of people, and places shown in the slide shows. An examination of subject differences can be seen using Table 3 to compare experimental conditions (number of slide shows seen) for the video clip 1 slide show.

Even though subjects completed sentences twice through the experiment, only sentences from the first part of the experiment are analyzed due to the fact that in the second part subjects either wrote sentences based on the fact that they recently saw the multiple choice comprehension question or wrote "same as before".

Table 2. Gist Comprehension: The percentage of correct concepts identified by the subjects, and the number of times comments were made by the subjects that they were not sure of the gist of the video clip. For each Slide Show condition, the number of subjects is seven. (n = 7).

Condition	% Correct	# Times Main Idea Unclear
1 Slide Show	87.5	0
2 Slide Shows	73	2
3 Slide Shows	44	4
4 Slide Shows	28	5

Table 3. Object, Places, and People List for Video Clip 1 Slide Show: The number of responses made by subjects for each description listed. "C" and "I" indicate Correct/Incorrect responses made by subjects. For each slide show condition, the number of subjects is seven. (n = 7).

People		1 Slide Show	2 Slide Shows	3 Slide Shows	4 Slide Shows
man	C	5	2	1	0
people	C	2	1	3	2
scientists	C	2	1	0	0
woman	C	3	0	0	0
forest rangers	I	0	1	0	0
doctors	I	0	0	0	1
Places		1 Slide Show	2 Slide Shows	3 Slide Shows	4 Slide Shows
jungle/ rainforest	C	3	4	3	1
"not in U.S."	C	0	0	1	0
zoo	I	0	1	0	0
Objects		1 Slide Show	2 Slide Shows	3 Slide Shows	4 Slide Shows
bananas	C	4	0	2	1
cage	C	3	0	2	1
food	C	1	0	0	1

3.3 User Perceptions

Figure 2 gives an overview of user perceptions. The mean values for each question asked in the evaluation of the interface show that for a single slide show (condition 1), subjects rated the speed of the video key frames for the object recognition task as neither too fast nor too slow. Subjects in conditions 2, 3, and 4 rated the speed closer towards "too fast". Even though all the video key frames are displayed at the same rate, subjects perceive the slide shows as going "too fast" in the conditions where two, three and four are shown simultaneously.

Results for the speed of videos for the comprehension task are similar to those found for the speed question for object recognition. Subjects in condition 1 perceived the speed of video key frames as neither too fast nor too slow. Subjects in conditions 2, 3, and 4 rated the speed closer towards "too fast". Even though all the video key frames are displayed at the same rate, subjects perceive the slide shows as moving "too fast" in the conditions where two, three and four are shown simultaneously.

For both the object recognition tasks and the comprehension tasks, as the number of slide shows shown increased, subjects rated the "number of videos" more towards the "imperceptible" end of the scale.

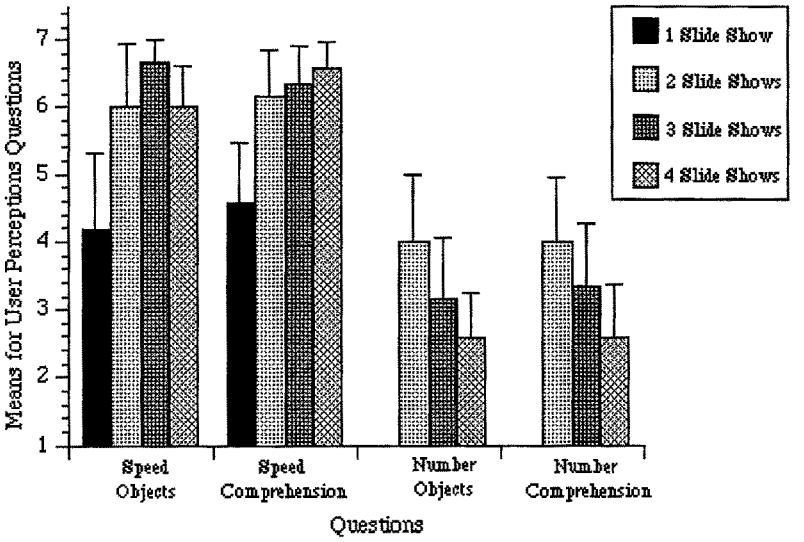

Fig. 2. Results of evaluation questions. +1 Standard deviation.

4 Discussion

The main objective of this study is to define elements that may be used in building user-oriented interfaces that employ key frame surrogates for browsing video data. Results of this study show that use of simultaneous viewing of slide shows are a possible design consideration, and may be used for both browsing for specific objects and for gist comprehension of the original video.

One interesting finding from this study was that subjects who viewed two and three simultaneously displayed slide shows, had almost the same performance on the object recognition task. It was expected that these groups would have lower accuracy scores than subjects who only viewed one slide show, and that both these conditions would be able to identify objects with much greater accuracy than when four are displayed at once. From the especially high performance by the subjects viewing three multiple slide shows, it may be postulated that human abilities for object recognition are acceptable for up to three slide shows at once. Two and three slide shows may be presented with a loss in performance beyond the baseline of one slide show. However, since performance degrades so dramatically when four videos are displayed at once, it can be assumed that attention resources are insufficient to compensate for the demand required to identify objects in this case.

The low accuracy scores for the object recognition task obtained by subjects viewing four slide shows indicate that either distracters are being selected more often or that subjects are not identifying many objects. In the first case, it can be said that subjects are relying on schemas to identify objects in the film. A schema is a knowledge

framework that is a way people organize information about various concepts or events (Ellis and Hunt, 1989). The subjects synthesize all of the objects and scenes shown in the key frames in order to build a schema for what they believe to be the story behind the video. The subjects check this schema in the object identification task to decide which fit this "story" or framework. In the second case, subjects simply cannot attend to all videos at once well enough to identify objects in the videos or grasp enough meaning in order to construct a schema. The sentence content analysis indicated that probably both are occurring as the number of simultaneous slide shows increases, subjects are less able to identify the meaning of the video and of those that do identify a meaning, it is an incorrect and incomplete analysis of the actual video content.

Interestingly, after viewing the video key frames a second time, subjects did not improve on the object recognition task. Because of their closeness to the results found after one time through the video it might be assumed that this is the best human performance that can be achieved for each type of display (1 through 4 simultaneously). We may be able to say that the maximum amount of attention resources (Kahneman, 1973) were allocated for each task.

The analysis of the multiple choice gist comprehension questions used for this study was not statistically significant. One reason for this result is that the comprehension question, given as multiple choice, may not be a good measure of subjects ability to "get the gist" of the video. Instead, subjects are choosing from several predetermined selections which does not indicate what they believe is the meaning of the video clip. The fact that after the second time through no differences were found between the groups confirms that the questions produced equally poor performance from all conditions. Looking at the sentence content analysis, however, depicts a different picture of subject understanding which may further indicate that these questions are not a good measure of gist comprehension. Instead, we will use the content analysis of the sentences written to provide more insight to the subject's comprehension of the videos.

As the number of slide shows displayed simultaneously increases, the percentage of replies that correctly identified the essence of the video clips decreased and the number of statements by subjects indicating confusion rose. When attention is divided between so many slide shows, subjects may have constructed erroneous schemas relating to the gist of the video. Looking at the content analysis for people, place and objects categories, the decrease in the number and quality of responses by subjects as the number of slide shows increases supports this theory. The lack of these descriptions corresponds to subjects inability to formulate the story behind the key frames, hence the poor explanations of the gist of the video clip. This indicates a large amount of attention resources available for gist comprehension of the video clips may be "used up" as each slide show is added.

The results of the sentence content analysis described above gave a different picture for the use of multiple shows than the object recognition data. While performance with two and three multiple slide shows for the object recognition were found to be similar, subjects decreased in their accuracy considerably between two and three when describing the gist of the video clips. The difference may be explained as two different thresholds, one for object recognition and another for gist comprehension, however, further study would confirm this interpretation.

The Ding(1996) study found that slower speeds were required for identifying individual objects and gist comprehension. One contradiction that was observed in the data for this experiment related subject perception of the number of slide shows and their actual abilities on the tasks given. This was found by comparing the user perception results to those of the object recognition task. In the questionnaire data, subjects viewing two, three and four slide shows perceived the videos as very much "faster" than those viewing only one. Adding up the number of slide shows seems to increase the perceived speed, however, subjects still performed better on the "faster" two and three slide show conditions than those viewing four. From this we may deduce that although there is the perception of "faster" speeds, it does not affect actual performance in dividing attention between the slide show displays.

5 Conclusions

The results of this study provide guidelines for building tools for browsing video data. We conclude by summarizing the main findings from this research that may be used as a starting point either for future research or as an outline for system designers. For object recognition and gist comprehension tasks, it is possible to identify objects when there are one through three concurrent slide shows. It is also possible for users to give high performance correctly describing the gist of the video clips when two simultaneous slide shows are displayed. However, as the number of slide shows increases, users are less able to use the key frame images to synthesize the story of the video clips. Performance on all tasks degrades dramatically at four. Lastly, user abilities to perform object recognition and gist comprehension of video clips is not dependent on their perceptions of slide show speed.

5.1 Suggestions for Future Researchers

Further work in this area is necessary to provide data on a number of related topics. A few of these questions are summarized below.

1. Do certain types of videos require specific types of surrogates to represent them?
2. Are key frame surrogates best suited to object recognition and comprehension tasks or are there others in which this type of representation provides the most information to users?
3. What are the effects of combining both the number and speed (Ding, 1996) of these video surrogate displays?
4. How do increased/decreased video compression rates affect performance on object recognition and comprehension tasks?

Acknowledgments

The authors would like to thank Wei Ding for the use of materials for the experiment. We thank Tony Tse for his invaluable comments. We also thank Ian Robinson and Navin Sharma for their help with testing subjects and sorting through the data. Lastly, we thank Ellen Yu Borkowski for use of the AITS teaching theater. Without the generous help of these individuals, this investigation would not have been possible.

6 References

Christel, M.G., Winkler, D.B., and Taylor, C.R., (1997). Multimedia Abstractions for a Digital Library. In Proceedings of the ACM Digital Libraries '97 Conference. Philadelphia, PA, July 1997.

Ding, (1996). A Study on Video Browsing Strategies. Technical Report CS-TR-3790 UMIACS-TR-97-40 CLIS-TR-97-06 , University of Maryland at College Park

Ellis, H.C. and Hunt, R.R., (1989). Fundamentals of Human Memory and Cognition. Wm. C. Brown Publishers. Dubuque, Iowa.

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

Kahneman, D.(1973). Attention and effort. Englewood Cliffs, NJ: Prentice Hall.

Kobla, V., Doermann, D., & Rosenfeld, A. (1996). Compressed domain video segmentation. Technical Report CAR-TR-839 CS-TR-3688, University of Maryland at College Park

LeDoux, J.E. and Hirst, W., eds. (1986). Mind and brain: Dialogues in cognitive neuroscience. Cambridge: Cambridge University Press.

Marchionini, (1995). Information seeking in electronic environments. Cambridge Series on Human Computer Interaction 9. Cambridge University Press. New York

O'Connor, B. C. (1991). Selecting Key Frames of Moving Image Documents: a Digital Environment for Analysis and Navigation. Microcomputers for Information Management. 8(2), 119-133.

Shneiderman, B. (1998). Designing the User Interface; Strategies for Effective Human-Computer Interaction, Third Edition. Addison-Wesley, Reading, MA.

Wickens, C.D. (1992). Engineering psychology and human performance. Second Edition. New York: HarperCollins.

Wickens, C.D., Sandry, D., & Vidulich, M. (1983). Compatability and resource competition between modalities of input, output, central processing. Human Factors, 25, 227-248.

Zhang, H.J., Wu, J.H., Low, C.Y. and Smoliar, S.W., (1995). A Video Parsing, Retrieval and Browsing System In Proc. of ACM Multimedia '95 San Francisco, Nov.7-9, 1995.