CS-TR-3790

UMIACS-TR-97-40

CLIS-TR-97-06

A Study on Video Browsing Strategies

Wei Ding and Gary Marchionini
Digital Library Research Group
College of Library and Information Services
Human-Computer Interaction Laboratory
University of Maryland, College Park, Maryland 20742

Abstract: Due to the unique characteristics of video, traditional surrogates and control/browsing mechanisms that facilitate text-based information retrieval may not work sufficiently for video. In this paper, a video browsing interface prototype with key frames and fast play-back mechanisms was built and tested. Subjects performed two kinds of browsing-related tasks: object identification and video comprehension under different display speeds (1 fps, 4 fps, 8 fps, 12 fps and 16 fps). It was found that browsing the key frames between 8 to 12 fps could potentially define a functional limit in object identificationaccuracy. There was no significant performance difference found across display speeds tested. The results also showed that lower speeds were required for object identification than for video comprehension. How user performance was affected by individual characteristics such as age, gender, academic background and TV- or movie-watching habits, was investigated, but no significant difference was found due to the limit of sample size and other constraints.

1. Introduction

As a result of the rapid development of multimedia computing technologies and the improvement of networked information environment, the proportion of new data types such as video, audio and still images in digital library collections is continuously increasing. It is expected that video data will be as widely accessible to users as text. In text-based document retrieval systems, by browsing document surrogates (e.g. keywords, abstracts or portions of full text depending on the representation level) , users can rapidly get an overview of the information, filter out irrelevant objects, and further examine the relevant ones at a more detailed level. Due to the unique characteristics of video, traditional surrogates and control/browsing mechanisms that facilitates text-based information retrieval may not work sufficiently for video (Christel, 1995). Thus, more suitable surrogates and control mechanisms need to be explored for users to access video data in an efficient and effective way and best meet their information needs (Marchionini, 1996). In this paper, a video browsing interface prototype with key frames and fast play-back mechanisms was built and tested. Key frames are frequently used as video surrogates in current digital video libraries. Videos convey video signal (camera motion, scene changes, colors) and the audio signal (noises, silence, dialogue), the information is expressed both spatially and temporally. Not considering that the ambiguity of natural language may bias on video interpretation, sheer volume of video information is beyond accurate description simply via words. O'Connor argues in his series of research there is no systematic means of translating images into words(1985, 1986a, 1986b). He proposes using key-frames as video "abstracts", which are representative still pictures extracted from different scenes contained in each video sequence, based on video's physical and semantic properties. Supposedly, key frames and their temporal and content relationships can characterize the structure of the video (Kobla, Doermann & Rosenfield, 1996). When a video or video clip is stored as key frames, the storage space and data transferring time (from server to client) is greatly saved. More importantly, video segmenting and key frame extracting techniques with nearly 90% accuracy (England et. al., 1996), have been commonly applied in a number of systems of video indexing and retrieval (Kobla, Doermann & Rosenfield, 1996; Li et. al, 1996; Zhang, Low & Smoliar 1995; Chang et. al. 1994). Browsing is considered as a legitimate information-seeking strategy that augments analytical searching (Marchionini, 1995). By allowing users to gather overview information, monitor searching processes and discover/learn new information etc, browsing is especially helpful in an interactive information environment. From the perspective of human computer interaction, the information in digital libraries should not just be retrieved but should allow for rich interaction, so that users can tailor the information into effective and memorable renderings appropriate to their needs (Rao et. al. 1995). At present, there are two browsing techniques commonly applied in video database interfaces: sequential browsing and random browsing (Zhang, et. al. 1995; Yeung, Yeo & Liu, 1996). Sequential browsing takes place through a VCR-like interface with stop/start, fast-forward, reverse and pause/freeze controls, and the contents to be displayed could be key frames, video skims (Smith & Kanade, 1995; Wactlar et. al., 1996) or an entire video (clip). Random browsing takes place through a static hierarchical arrangement of key frames (Mills, Cohen & Wong, 1992). The VCR-like browsing interface with key frames provides two levels of granularity: overview and detail examination. Via the fast playback mechanism the overview granularity can be achieved by playing the key frames at a selected display speed; Via the mechanism " pause/free key frame", detailed granularity is provided - users can freeze a certain key frame of particular interest to examine the details of contents. As Wolf (1996) points out, the control mechanisms are analogous to thumbing through a book in the non-digital world: A reader can flip through book, and stop to read selected pages. Readers can gain an overview of the printed material in seconds by browsing: chapters are identified by typographical conventions, illustrations help identify relevant material, etc. Since the reader controls the rate at which the pages are flipped, he/she can slow down when getting near relevant material or skip past obviously unimportant pages. Although several primitive models of such video browsing interfaces have been implemented, little usability testing was conducted. In other words, it is still known whether the control mechanisms are suitable to user characteristics or meet user needs, and what aspects of the interface could be further improved. Specifically, under what circumstances do users need video fast playback? How fast can the key frames be played for user to get an overview sufficiently? It is likely that different speeds (key frame rates) are suitable for different tasks (overview vs. details examination) and different user groups (based on age, background, experience, and gender etc.). Once the best speed for each task is determined, the interface will be able to offer users with more control mechanisms (e.g., speed shift for different tasks) to facilitate their video access activities. This study aims at identifying the key factors that may affect user's performance in video access with the key-frame based VCR-like browsing interface. These factors will contribute to better interface designs for digital video searching and browsing libraries.

II. Experimental Design and Operational Definitions

This experiment could not fully implement the ideas of the proposed study due to time and other constraints. The focus of the experiment is, based on both user performance and their satisfaction, to investigate the best suitable fast playback speed for different video browsing-related tasks, or the boundaries of users' information processing abilities, such as the speed limit of video fast playback for different tasks. It is also to examine how user performance and satisfaction varies with their own characteristics, such as age, gender, background and experience.

2.1. Research Questions:

Generally, the research questions are: for each task, under what playback speed can users gain best performance? What is the speed limit for different tasks? what user characteristics affect user's performance besides speeds? Since the research questions are broader than what could be done in this experiment, several related concepts needed to be narrowed down and operationally defined, such as tasks, user characteristics, performance, display speed, best display speed, and perceived best display speed. Two typical tasks are identified and to be tested: object identification and video comprehension. As mentioned above, detail examination and overview are primary cognitive activities often required in various information processing settings. For example, when presented with lots of search results, users tend to narrow the size of search results by quickly scanning the surrogates of the results (Fenichel, 1981) -to overview the gist of each item and filter out less relevant ones; When needed to identify the most suitable results among a number of similar ones, detail examination is often involved. An overview of the whole document can be obtained through scanning key paragraphs while details can be examined by slowing down the browsing speed and read some paragraphs word by word. If we assume key frames correspond to key paragraphs in text and the objects in each key frame serve as keywords dispersed among each paragraph, detail examination and overview are the same important browsing activities happened in video databases as in textual settings. In this experiment, the corresponding activities are specified as object identification and video comprehension respectively. User characteristics include age, gender, academic degrees and TV-watching habits in this experiment. Graded based on how well the participants finish the two tasks mentioned above, performance will be measured in score and in percentage. Besides absolute performance, acceptable performance will also be defined based on the experimental results in order to determine the speed limit for each task. Display speed is defined as the number of video key frames per second to be shown on screen. Best display speed is the speed at which best performance is obtained. Perceived best display speed is the best speed users perceive subjectively. Specifically, the research questions are as following:

* Research Question 1.

What are the best video display speed for object identification (OI) and video comprehension(VC)? Shown with two-dimensional coordinates, the performance for both OI and VC would decrease with the increasing of display speed based on common sense. However, it is not clear if the decreasing rate is constant across all the speeds or increases abruptly at a certain point (or in a range) in speed axis. It is hypothesized (H1) there exits a speed breakpoint or a proper range, beyond which user performance becomes unacceptable and independent of the speed variation. If it is true, the speed limit for each task would be found, and then the best display speed should be identified from between the baseline and the limit. The null hypothesis (Ho) is there is no significant difference in subjects' performance between the different display speeds for either OI or VC.

* Research Question 2.

Object identification (OI) and video comprehension(VC) are different cognitive processes. While VC needs global attention, OI involves focused attention. Thus, they could have different best suitable display speeds and speed limits. Based on an informal pretest at a very early stage of this study, OI was likely easier to be done at the same speed than for VC. So it is hypothesized (H1) at this point the best suitable speed and speed limit for object identification are higher than for video comprehension. The null hypothesis (Ho) is that at the same speed, the subjects will gain the same (performance) accuracy in both object recognition and video comprehension. Additionally, users' subjective estimation (their satisfaction or confidence) of speed may also vary with tasks, but it may not be exactly consistent with their performance. For instance, sometimes they may have higher satisfaction with a speed even though they don't gain good performance. As user's performance and their satisfaction are both important in user interface design, information on subjective estimation must not be neglected. Thus a related research question is whether users have the same speed expectation (perception) for different tasks. It is hypothesized that users' subjective estimation to comparable speeds (perceived speed) differ with tasks (H1), and the null hypothesis (Ho) is there is no significant difference between the subjective estimation to comparable speeds for either OI or VC.

* Research Question 3

User characteristics, are important factors to be considered in user-centered interface design. For some interface, academic background, age and gender often affect user performance to some extent. As far as this experiment concerned, vision ability is directly related to user performance, and human's visual perceptual abilities degrade with the growth of age. It is possible that the older subjects tend to gain poorer performance in both OI and VC. Besides, it is natural to be interested in gender difference, but there is no evidence to predict which direction the difference goes. Since the video-watching task requires human's adaptability of motion and visual representation, we would like to investigate whether there is a relationship between people's TV- or movie-watching habits and their video-watching performance in this experiment. It is hypothesized that the more time one spends on watching TV or movie during daily life, the more likely one gains better video-watching performance. In addition, the subject of academic degrees reflect subjects' professional background in a way, which may help them finish the task easier when the video theme match the subject's field. Finally, none of above variables are exclusively independent to each other. They are related to each other. The research hypothesis (H1) is variables of age, gender, academic degrees and TV-watching hour are significantly contributed to the variability of performance. H0 is none of the variables is significantly correlated to the performance.

2.2. Methodology

The basic idea of this experiment is to ask participants to watch pre-determined video clips at different display speeds and then finish the tasks assigned. Each video clip will be displayed by flashing the key frames one by one at a constant speed. By comparing their performance and subjective perception at each speed per task, best suitable speed and speed limit for different task are expected to be drawn. By associating participants' performance to their individual information, such as age, gender, background, possible correlation relationship between human performance and user characteristics might be seen.

2.2.1. Video Clips

The video clips used in this experiment were segmented and selected from the digitized (MPEG-1) Discovery Channel educational video resources. The video key frames were created by a color histogram-based segmenting and indexing technique developed at the Center for Automation Research at University of Maryland. Tests showed this technique gained nearly 90% accuracy in key frame extraction (Kobla, Doermann & Rosenfield). Six 3-5-minute video clips (one sample was for practice, and the other five for testing) were finally used in the experiment, each of which was composed of about 23-25 key frames. Video clip 1 is from Spirits of Rainforest, showing how researchers conduct studies on monkeys in a rain forest; Video clip 2 is from Flight Over Equator, showing scenes of Singapore's industrialization, its culture and people in daily life. Video clip 3 is from Spirits of Rainforest showing how a native American tribe makes a living in the jungle. Video 4 from The Revolutionary War showed enacted scenes of the Battle of Concord during that period. Video 5 from Space Shuttle was about the Apollo 11 astronauts' training and moon landing activities. The sample clip showed a researcher learning from a native American to identify medicinal plants. Supposedly all the video clips were under a same difficulty level to non-professionals (two former video clips used in the pre-test was replaced due to their unequal degree of difficulty and cultural unfamiliarity.)

2.2.2. Display Speeds and Perceived Speeds

Display speed will vary from 1 frame per second (fps), 4 fps, 8 fps, 12 fps to 16 fps. Based on early studies (Potter & Levy, 1969) and current research (Healey et. al., 1996), 1 fps will serve as a baseline rate and 16 fps as an upper limit which is higher than the commonly-accepted recognition limit (for preattentative visual estimation), 105 milliseconds (about 9.5 fps). Perceived speed is defined as users' subjective estimation to the actual display speeds. With a 1-7 scale (from too slow to too fast) , users can estimate each speed by selecting a proper scale. In this case, 4 will be treated as perfect - not too slow and not too fast.

2.2.3. Tasks: Object Identification and Video Comprehension:

Object identification and video comprehension will be tested in this experiment. The former will consist of identifying objects from videos watched by subjects; the latter will request subjects to select the best of four statements summarizing the video. For object identification, cued-recall with a check list was employed instead of free-recall because of the advantages in quantitative analysis with controlled vocabulary. Among 20 objects in each list, half appeared in the video with the other half as distractors. Kept in an alphabetical order, lists were carefully created for face validity. To maintain the objects at the same specificity and difficulty level as much as possible, and terms could not be "too specific" or "too broad." Derivable parts from a certain object, such as face, arms, and clothing were not listed if the whole object (e.g., man/woman) appeared in the video. Only objects that were reasonably visible (at least visible at the lowest speed, 1 fps) were selected. The ones that could be misleading were not put on the list. For example, the background of one of the key frames looked both like snow and desert, and no other clue could help make the judgment. Therefore, neither desert nor snow was selected in the list. Finally, a person never exposed to any of the videos was asked to guess which objects should be in the list without watching the video. When the probability of being picked up for the distractors was close to that for the real objects, it was assumed that the list was workable. Otherwise, more work would be needed to further revise it. For video comprehension, subjects were asked to both write down the gist of each video, and select the best answer from four choices. While the write-up directly reflected the subjects' comprehension status of each video (even though there might exist individual differences in verbal expression), the multiple choice questions greatly simplified the variety in responses and data analysis. The creation of multiple choices was one of the most difficult parts in the whole design. Two principles were applied: first to maximize the distinction between the choices, and minimize the use of additional prior knowledge about the videos. As long as the video was watched carefully at baseline 1 fps, the answer should be obvious enough on average; the second principle was, if the first one makes the distinction too obvious, cautiously create less distinctive choices which require some prior knowledge to identify the answer. The multiple choice sentences had to be revised several times. First, each video clip was summarized based on its verbal indexing data (available in video database for the Baltimore Learning Community project at http://www.learn.umd.edu). Then distractors were created based on the write-ups by the 6 subjects of the pre-test and the variability results at the baseline. Further modification and polishing were made based on the feedback from several other people involved in this project.

2.2.4. The Interface:

The interface for this experiment (figure 1) was developed with JavaScript and HTML under Netscape Navigator 3.0 (at http://www.glue.umd.edu/~weid/movie/Viewer.html). Six video clips, including the sample for practice, were used in the experiment. As mentioned above, each video clip was composed of 23-25 key frames, which were 72 dpi GIF still images. The five display speeds are available in a listbox (in the upper part of the screen). To the right of the listbox, there is a control button labeled "Click here to play". In the lower left part of the screen, is a list of video clips. When a video is selected, it will be ready to be presented.

Figure 1. The Interface for the Experiment

[Image]

Clicking on the button "Click here to play" , the selected clip will be shown one key frames after another in the display area (the lower right part of the screen) with the selected speed. Between video shows, the display area is covered with a mask, consisting of random lines and dots. Although the interface is freely accessible on the world wide web, the display effect of key frames more or less varies with the computer platform used. Based on our experience, the best display effect was gained on a Power Mac 8500 with a 15- inch monitor. To maintain a consistent environment for all the subjects, the same computer was used throughout the whole experiment.

2.2.5. Subjects:

Twenty University of Maryland graduate and undergraduate students participated in this study voluntarily. Excluding the six for the pretest, fourteen (3 males, 11 females) participants went through the experiment formally. Their age ranged from 20 to 60 years old. Due to time and other limitations, only the students accessible to the experimenter were recruited. Methodologically, the size of the sample was too small, and the sampling was not random enough. As we will see in Results section, some of the data analyses were unable to be fully implemented as planned due to the small sample size.

2.2.6. Pretest:

To further polish the design and ensure its face validity and operational feasibility, a pretest was conducted with 6 (3 male and 3 female graduate students of library and information science) participants before the formal experiment. The early version of object lists was used for object identification, and as video comprehension task, the participants were asked to write down the gist of each video clip watched (without multiple-choice questions). Based on the pre-test results, some of the distractors in object lists that were never picked by the subjects were replaced with supposedly better ones. Two of the 6 video clips were replaced with new ones (video 1 and 3 as mentioned in Section 3.2.1) in order to maintain all the video clips under the same difficulty level (Note this was not quantitatively done in this study). Some of the video gist results from the pre-test participants were adopted as the statements for the multiple-choice questions in the formal experiment.

2.2.7. Experimental Procedure:

Each subject watched 5 different video clips each at a different speed. Before the experiment, after the consent form was signed, subjects were asked to fill out a questionnaire about their age, gender, academic degrees, and how much time they spent on average watching TV, video and movies (see Appendix 3 for the questionnaire before the experiment). Then a practice session with a sample video and sample tasks was given to ensure that the subjects understood the tasks and the procedure, and to help them get used to watching videos at different speeds, especially the high speeds. To avoid bias on different speed settings, the video display sequence and speed were randomized for each subject by a small computer program. (see Appendix 1 for a computer-created random sequence list for a sample of 30 subjects. In this experiment, the first 14 of the sequences were followed by the subjects). Within one minute before watching each video clip, the subject was first presented with a list of 20 objects (only nouns). The list was printed on a sheet of paper. After watching the video, subjects were asked to complete three kinds of activities. (1) To check off the objects in the video on the list of 20 objects as quickly as possible; (2) To write down the gist of the video in 1 to 2 sentences, and give the display speed a subjective rating on a 1 to 7 (from too slow to too fast) scale for both of the identification and comprehension tasks; (3) In another separate page of paper, to select one statement of four presented which best represented the gist of the video. If the subject was not sure about the answer, they should make a guess. These procedures were repeated for each video. (See Appendix 2 for the object lists and the multiple choice questions for all the videos). After all five videos, a brief interview was conducted to hear suggestions and opinions about this experiment from the subjects.

III. Results:

This study is preliminary and limited to many constraints. For example, the sample size (only 14 subjects) was small, and the sampling was not ideally random and representative. Therefore, the results only serve for testing purpose, and the conclusions are not very reliable for real decisions. Before going into the specific data analyses, it is necessary to explain and define the measures and related terminology.

3.1. The Measures and Terminology:

For the purpose of data analysis, we need to find the relationship between user performance and video display speed, between actual and perceived speed, and between user performance and user characteristics. To be specific, the performance will be measured based on the accuracy scores and accuracy percentage participants gained in each task. For object identification (OI), there were two kinds of scores and percentages were used: score A (identification score) for correctly identifying objects in the video and score B (un-identification score) for correctly not identifying objects not in the video. The total accuracy score is the sum of A and B. As mentioned earlier, for each video there are twenty items in a list, ten of which appear in the video with the other 10 as distractors. If one picked 8 of the 10 items in the video, s/he would gain 8 points (score A) for correct identification, and the percentage (percentage A) is (8/10=) 80%; at the same time, if s/he picked 3 of the 10 items not in the video, then s/he would gain another (10 - 3 =) 7 points (score B) for correct un-identification, with percentage B as 70%. So her/his total accuracy score would be (8 + 7 =) 15 points with a total accuracy percentage of (15/20=) 75%. Under random conditions, the accuracy probability is 50%. Furthermore, using this scoring system, accuracy probability is 50% if all 20 items are identified or none are. For video comprehension (VC), only the multiple-choice question has been analyzed so far. For each question, 1 point was given for a correct selection, 0 point otherwise, with 100% and 0% accuracy percentage respectively.

3.2. Data Analysis Plan:

To test the hypotheses of the research questions, a series of statistical analyses were planned to be done. Research question 1 is to find the best display speeds for OI and VC, and then to identify the corresponding speed limits. The independent variable (IV) is display speed, and the dependent variable(DV) is accuracy score (an equivalent of accuracy percentage). With a one-way ANOVA first conducted, if the result of ANOVA is significant, a multiple t-test will be applied to see what speeds are divided into same or different sub-groups based on their mean differences. With a significant ANOVA, the speed (or speed subgroup) with the best performance mean will be the best display speed, and the speed limit will be identified accordingly. Otherwise, that would mean there is no significant difference in subjects' performance between the different display speeds, and Ho cannot be rejected. Research question 2 is to find out if same performance can be obtained at same speed for different tasks, such as OI and VC. The accuracy performance is DV, and task (OI or VC) is IV. A t-test is needed to compare if the means of the performances for OI and VC are significantly different. If significant, the null hypothesis will be rejected, otherwise it will hold. Similarly, for the related research question–if there is significant difference between the subjective estimation to a same speed for OI and VC (perceived speed as IV, and tasks as DV) – if the t-test is significant, it will be evident that users expect higher speeds for one task than for the other. Research question 3 is to test the relationship between individual user characteristics (age, gender, and watching habits as IV) and their performances (DV). Ideally, analyses of correlation coefficient should be conducted to test if the correlation between each characteristics and the corresponding performance is significantly high. And then a proper calculation of a regression on all the user characteristics would be drawn. However, limited to the small size of the sample (14 subjects), such a statistical procedure could not be done at this point. Brief qualitative analyses will apply instead. During the design of this study, it has been assumed all the video clips used in the experiment are at the same difficulty level, and subjects don't need much professional knowledge involved to do the tasks. However, this assumption turned out not being exactly held based on the interviews with the subjects and their performances. As a third-party variable, it would affect most of the results. Therefore, the differences in the video clips will be stated in Data Analysis Section later on.

3.3. Data Analysis:

3.3.1. Relationship between Performance and Speed for Object Identification:

For object identification, a one-way ANOVA indicates significant performance differences between the display speeds, and the data were divided into three homogeneous subsets under a student multiple t-test: group 4 and group 5 (12 fps and 16 fps), group 3 and group 2 (8 fps and 4 fps), and group1 (1 fps). Note the performances in different speed subsets are significantly different from each other, and homogeneous for the ones within a same subset. In other words, 1 fps resulted in the best performance, 4 fps and 8 fps resulted in better performance than the highest speeds, but were not significantly different from each other. Also there was no significant difference between 12 fps and 16 fps, at which the accuracy was only around 60%, just a little bit higher than random probability (see figure 2a and 2b on next page). Analysis of Variance (ANOVA) (Variable SCORE by Variable SPEED) Sum of Mean F F Source D.F. Squares Squares Ratio Prob. Between Groups 4 140.4857 35.1214 12.3543 .0000 Within Groups 65 184.7857 2.8429 Total 69 325.2714

Multiple Range Tests: Student-Newman-Keuls test with significance level .050 (*) Indicates significant differences which are shown in the lower triangle

G G G G G r r r r r p p p p p 4 5 3 2 1 Mean SPEED 12.4286 Grp 4 12.7857 Grp 5 14.3571 Grp 3 * * 14.9286 Grp 2 * * 16.2857 Grp 1 * * * *

Obviously, H0 was rejected, and the best performance was gained at baseline 1fps --best display speed. More importantly, even though the performance decreased with an increase in display speed in general, there was an abrupt performance drop between 8 fps and 12 fps, and no significant difference between the performance at 12 fps and 16 fps. If we assume that 60% (comparing to the random probability 50%) is the lowest acceptable identification accuracy (or performance), there could exist a speed breakpoint close to 12 fps (as circled in the figures). However, if the acceptable performance could be even lower than 60%, then the speed breakpoint might be located even beyond 16 fps. More details, such as the variation of score B, could be identified in the figures. It was reasonable that score B kept high across the speed. At the lower speeds, subjects tended to check the objects they saw with confidence. At higher speeds, if they did not randomly pick up objects, the probability to pick distractors could also be very low. It was even difficult to see the objects that were really in the video, let alone the distractors that did not exist at all.

Figure 2a. Identification Performance by Average Score [Image] Figure 2b. Identification Performance by Accuracy Percentage [Image]

3.3.2. Relationship between Performance and Speed for Video Comprehension:

Analysis of Variance (Variable SCORE By Variable VIDEO)

Sum of Mean F F Source D.F. Squares Squares Ratio Prob. Between Groups 4 17.9429 4.4857 .9803 .4245 Within Groups 65 297.4286 4.5758 Total 69 315.3714 For video comprehension task, the one-way ANOVA above showed there was no significant difference between the performance across the speeds tested. In figure 3, comprehension performance did not show any pattern without much variety across the speeds. Two of the speeds gained around 70% accuracy (as circled the figure), and the other three gained around 60% accuracy. Therefore, the null hypothesis could not be rejected. Probably, only one multiple-choice

Figure 3. Different Performances in OI and VC [Image]

question could not fully reflect the comprehension performance. Thus, a speed breakpoint could not be found based on our data at this point.

3.3.3. Performance Differences in Object Identification and Video Comprehension:

Also in figure 3, performance in object identification was a little higher than in video comprehension. A t-test showed there was no significant difference(p=.171) between the performance in OI and VC. Interestingly, even though comprehension performance was lower than object identification, there was consistently higher subjective estimation to comparable speeds for OI than for VC (see figure 4 on next page. In the y-axis, 4 indicates optimal in perceived speed, 7 means too fast, and 1, too slow). A t-test showed significant difference between the subjective speed estimation for OI and VC (p=.001). Therefore, the null hypothesis was rejected, which preliminarily confirmed our assumption: that different tasks requires different display speeds. Meanwhile, it resulted in the contrary of one of the aspects of the research hypothesis: based on subjective estimation, it is easier to get the video gist than to identify all the objects (details) at comparable speeds, which likely implied that the speed limit for gist comprehension could be even higher than 16 fps. However, VC defined in this experiment was just roughly knowing rather than completely understanding. More reasonable definition and related tasks for video comprehension need to be worked out in the future.

Figure 4. Subjective Speed Estimation for Different Tasks [Image]

3.3.4. Performance Analysis by Video:

As mentioned in section Data Analysis Plan, the video difficulty level could not be well controlled. To have an estimation how much this affects our results, the data was analyzed by video clips tested (performances as DV, and video clip as IV). An ANOVA was conducted against the performances for OI and VC respectively. No significant difference was found between performance in OI, and significant in VC (see figure 5). In other words, the difficulty level did not affect the performance in OI much, on the contrary, it was very likely a factor in the VC performance. For VC performance, the difference between the video clips could be explained by the difficulty level and subjects' familiarity of the video (please refer to section 3.2.1 in page 4). Video 5 gained the highest performance 79% while video 4 gained the lowest, only around 43% . Video 5 was about the Apollo 11 astronauts training and moon landing, which was very familiar to the subjects. Video 4 showed scenes in the Battle of Concord during the Revolutionary War. It was much easier for the subjects to identify the Revolutionary War than to determine the exact battle, which was required in the multiple choice question. Video 1 gained the second highest performance (71%) and video 3, the second lowest (50%). Both of them showed people's activities in a rain forest, but the former was about researchers' studying on monkeys and the latter was about how a native American tribe makes a living in the jungle, which was less familiar than the former to the subjects.

Figure 5. Performance Analysis by Video [Image]

Based on a brief qualitative analysis above, it is likely that difficulty level and prior knowledge somewhat affect participants' performance in different tasks, but quantitative evidence was not obtained yet. Proper measures should be developed in later studies.

3.3.5. Relationship between Performance and User Characteristics:

User characteristics in this study are age, gender, subject of academic degrees and TV-watching hours per week. Limited to the sample size, a series of calculation of correlation coefficient and regression coefficient cannot be applied to the data. Here, only very descriptive analyses were conducted to gain a rough estimation on those relationships. In general, as shown in figure 6, the individual differences in object identification were in a regularly symmetric distribution (close to normal distribution). Also, the distribution of comprehension differences was basically symmetric with a few exceptions, which partially validated the data analyses. Probably, the distribution would tend to be normal if there were a larger number of subjects.

Figure 6. [Image]

Performance vs. Age: Figure 7 shows the age distribution of the sample. The subjects were divided into four groups by age. Group A: 20-24 years old; Group B: 25-30 years old; Group C: 31-40 years old; and Group D: 41-60 years old.

Figure 7.[Image] Figure 8. [Image]

Figure 8 shows that Group A gained the best, and Group D gained the worst performance in both identification and comprehension. Group C had slightly better performance than Group B. With limited numbers of subjects, it is not possible to conclude that the performance gets poorer with the growth of age. However, the difference between the youngest group (A) and the oldest group (D) suggests that age might be an important factor, with consideration in user interface design for video display. Performance vs. Gender/ Subject of Academic Degrees: Besides the small sample size, the subjects were not evenly distributed by gender (3 males and 11 females). No analysis was done to compare the performance difference by gender at this point. Due to the large variety of the subjects of academic degrees, they have not been properly coded. More work needs to be done to test the correlation between performance and academic subjects. Performance vs. TV-Watching Hours: In the questionnaire, the subjects were asked about how much time per week they spent on watching TV, video or movie. Supposedly, the more people watch TV, the better they can get used to the dynamic representation mode and catch the information efficiently. However no correlation between the amount of TV watching and the performance was found either in OI or VC. On the contrary, it seemed that one of the best performers (Subject #11) did not watch TV much, while the poorest performer (Subject #9) did not watch TV at all.

Figure 9. [Image]

In summary, this study had many constraints. Because the number of subjects was limited, no analysis was done relating performance and gender and academic background; No stronger relationship was found between age and performance. Other speeds (e.g., the speeds between 8 fps and 12 fps, or even higher than 16 fps) were not tested. Finally, that two tasks (overview and detail examination) were performed at the same time, and subjects were not allowed to freeze particular key frames might have caused results somewhat distorted.

IV. Discussion

As a preliminary study, the design of the experiment needs to be reviewed. Although efforts were made to improve the reliability of the tasks, some problems still remained unsolved. Unlike most psychological studies that isolate each cognitive task separately, the object identification and video comprehension tasks involved multiple tasks and multiple cognitive processes that were unable to be isolated and simplified during the video display. According to some subjects, the two tasks conflicted with each other. While identification required focused attention to each key frame, comprehension needed synthetic processing of all the information, including the background information of each frame (Boyce et. al., 1989), the sequence of frames and the change of the repeatedly occurring objects, etc. Furthermore, since the video stimuli used came from the real world, they were not comparably informative or representative. The objects were more or less different in their sizes, frequencies, resolution, brightness, locations in the picture (Luck et. al., 1996), with different background complexity, different specificity and vividness (Marks, 1973); The objects in object lists might not be selected at a regular interval from the frame sequences; The objects shown later in the frame sequences were more likely to be memorized than the ones shown earlier (Dale, 1973), and so on. All these uncontrolled or not well-controlled variables may contaminate the data and results. Additionally, for the comprehension task, a single multiple-choice question may not sufficiently reflect subjects' real understanding status, not mentioning the uneven difficulty level and accuracy of all the multiple-choice questions. Since this experiment is exploratory and limited in many respects, it is impossible to draw conclusions formally. However, some of the results are encouraging and show potential for further exploration. The results can be summarized as following:

* Identification performance decreased with an increase in the display speed. Performance at 12 fps and 16 fps was significantly different from the others, but homogeneous with each other. Therefore around 12 fps there could be a speed breakpoint, beyond which identification performance would stay poor, independent of speed.
* Lower speed was required to finish object identification task than to video comprehension. The speed limit (both actual speed and perceived speed) for video (rough) comprehension may be higher than that for object identification.
* Identification performance may be relatively independent of video content, while comprehension performance demonstrated dependence on prior subject knowledge of the video.
* Both identification and comprehension performance were inversely correlated with age.
* There was no significant correlation found between performance and TV watching hours.
* A video interface could be developed to incorporate high speeds of playback for rapid comprehension and overview, while detail examination is less likely supported (could be achieved by freezing particular frames).

V. Further Work

Further work includes: (1) Improvement of the experimental design and (2) comparison and integration of different browsing strategies. Besides increasing the number of subjects and videos, other possible improvements involve:

* Testing additional display speeds between 8 fps to 16 fps for object identification, and the speeds around 16 fps for video comprehension. If 12 fps is an acceptable speed for comprehension with key frames, and they are picked up from every five frames, the display speed with the entire video clip would reach 12X5=60 fps while still supporting comprehension.

* Adding up more control mechanisms, trying other tasks with more videos and using better measurement instruments. For example, fully implement or improve the VCR-like interface with forward, reverse and freeze, and let users decide the speed of fast playback. Subjects then could perform other tasks, such as freely browsing video database and solving various problems. Thus, the suitability and use frequency of different control mechanisms would be studied at a more natural information seeking atmosphere when subjects are given freedom to choose the best control mechanism based on their own needs, such as playing more than one video simultaneously, watching a video as many times, or freezing a frame for examination. A measure of performance could be both the number of correct answers and response time. and

* Taking different approaches to data collection. In this study, the

data was collected manually. The subject answered questions with paper and pen and was timed by the experimenter. Although they were asked to look at the object list within one minute before watching the video to check the list as quickly as possible afterwards, the time used by the subjects was not equivalent – some subject finished watching the object list less than one minute, and some checked off the list quicker than others. By automating the procedure, all the subjects will be exposed to the stimuli for the same amount of time, and answer questions on the computer. Data can then be gathered and graded consistently and automatically.

Empirical data about the pros and cons of each browsing/control mechanism will lead to a better understanding of the strengths and weakness of various browsing control mechanisms for different tasks. Video random browsing technique provides users with random access to any point in a given video. The top level representative frames serve as a table of contents for the video (clip) so that users can go directly to the sub-unit of the video to see all the key frames in detail. However, the user has to go through the hierarchy by clicking, which may take longer than fast playback. If we can build the VCR-like mechanisms into each key frames at the top levels, then users would be able to rapidly access the video starting from any point. They can also use static display for detail examination and comparison, and use fast playback to gain an overview.

References:

Boyce, S.J.; Pollastsek, A. & Rayner, K.(1989). Object Identification Depends on the Background Information. Journal of Experimental Psychology: Human perception and Performance. 15(3), 556-566.

Chang, S.; Anastassiou, D.; Eleftheriadis, A.; Meng, J.; Peak, S.; Pejhan, S. & Smith, J. (1994). Development of Advanced Image/Video Servers in a Video on Demand Testbed. Proceedings of the IEEE Visual Signal Processing & Communications Workshop (VSPC).

Christel, M. G. (1995). Addressing the Contents of Video in a Digital Library. Electronic Proceedings of the ACM Workshop on Effective Abstractions in Multimedia (San Francisco: California, November 4, 1995). at http://www.cs.tufts.edu/~isabel/christel/ christel.html

Dale, H. C. (1973). Short-Term Memory for Visual Information. British Journal of Psychology, 64(1), 1-8.

England, P.; Allen, R.B.; Sullivan, M.; Heybey, A.; Bianchi, M. & Dailianas A. (1996). I/Browse: The Bellcore Video Library Toolkit. SPIE, (Jan. 1996). 1-11.

Fenichel, C. H. (1981). Online Searching: Measures that Discriminate among Users with Different Types of Experiences. Journal of American Society for Information Science, 32, 23-32.

Healey, C.G.; Booth, K. S. & Enns, J. T.(1996). High-Speed Visual Estimation Using Pre-attentative Processing. ACM Transactions on Computer-Human Interaction, 3(2), 107-135.

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

Maryland

Li, W.; Gauch, S.; Gauch, J. & Pua, K.M. (1996). VISION: A Digital Video Library. Proceedings of the 1st ACM International Conference on Digital Libraries (Bethesda: Maryland, March 20-23, 1996). 19-27.

Luck, S. J.; Hillyard, S. A. & Mouloua, M. (1996). Mechanisms of Visual-Spatial Attention: Resource Allocation or Uncertainty Reduction? Journal of Experimental Psychology: Human Perception and Performance. 22(3), 725-737.

Marchionini, G. et. al.(1996). Interface Design Parameters for Effective Browsing: Surrogates and Control Mechanism for Video and Statistical Information. NSF-STIMULATE proposal.

Marchionini, G. (1995). Information seeking in electronic environments: Cambridge University Press.

Marks, D. F. (1973). Visual Imagery Differences in the Recall of Pictures. British Journal of Psychology, 64(1), 17-24.

Mills, M.; Cohen, J. and Wong, Y. (1992). A Magnifier Tool for Video Data. CHI '92 Conference Proceedings of Human Factors in Computing Systems. (Monterey: CA., May 1992). 93-98.

O'Connor, B. C. (1985). Access to Moving Image Documents: Background Concepts and Proposals For Surrogates for Films. Journal of Documentation. 41(4), 209-220.

O'Connor, B. C. (1986a). Moving Image-Based Serial Publications. Serials Review. Summer & Fall, 19-24.

O'Connor, B. C. (1986b). Representation and the Utility of Moving Image Documents. Proceedings of 49th American Society of Information Science, v. 23, 237-243.

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.

Potter, M.C. & Levy, E. I. (1969) Recognition Memory for a Rapid Sequence of Pictures. Journal of Experimental Psychology, 81(1), 10-15. Cited from the NSF grant proposal by Marchionini et. al.

Rao, R., Pedersen, J., Hearst, M., Mackinlay, J., Card, S., Masinter, L., Halvorsen, P.-K., and Robertson, G. (1995). Rich Interaction in the Digital Video Library. Communications of the ACM, 38(4), 29-39.

Smith, M. A.; & Kanade, T. (1995). Video Skimming for Quick Browsing Based on Audio and Image Characterization. Technical Report, CMU-CS-95-186.

Wactlar, H.; Kanade, T. & Stevens, S. (1996). Intelligent Access to Digital Video: Informedia Project. IEEE Computer, May, 46-52. Wolf, W. (1996). Key frames selection by motion analysis. Paper presented at the IEEE ICASSP '96.

Yeung, M., Yeo, B.-L., & Liu, B. (1996). Extracting story units from long programs for video browsing and navigation. Proceedings of the International Conference on Multimedia Computing and Systems.

Zhang, H., Low, C. Y., & Smoliar, S. (1995). Video Parsing and Browsing Using Compressed Data. Multimedia Tools and Applications(1), 89-111. Zhang, H. J., Low, C. Y., Smoliar, S., & Wu, J. H. (1995). Video parsing, retrieval and browsing: an integrated and content-basedsed solution. Proceedings of ACM Multimedia '95 at http://www iss.nus.sg/RND/MS/Projects/vc/vidorigin.html.

Appendix A: The Random List for Sequences of Video and Speed (for a sample of 30 subjects) * V means video, and S means speed * V1-- 1 fps, V2 -- 4 fps, V3 --8 fps, V4 -- 12 fps, V5 -- 16 fps

V5 S2 V3 S3 V1 S1 V4 S4 V2 S5 V4 S2 V1 S3 V5 S4 V2 S1 V3 S5 V1 S5 V2 S4 V5
S3 V4 S1 V3 S2 V3 S4 V1 S2 V2 S3 V4 S5 V5 S1 V1 S4 V2 S2 V3 S1 V4 S3 V5 S5
V5 S1 V3 S3 V4 S2 V1 S5 V2 S4 V3 S2 V1 S1 V2 S3 V4 S5 V5 S4 V2 S1 V3 S5 V1
S2 V5 S3 V4 S4 V4 S3 V1 S4 V3 S1 V5 S5 V2 S2 V1 S3 V2 S5 V3 S4 V4 S1 V5 S2
V5 S5 V1 S3 V3 S1 V2 S4 V4 S2 V4 S4 V3 S2 V1 S5 V5 S3 V2 S1 V3 S4 V5 S1 V1
S2 V2 S5 V4 S3 V1 S4 V5 S2 V2 S3 V3 S5 V4 S1 V1 S1 V2 S2 V3 S3 V4 S5 V5 S4
V3 S3 V1 S5 V4 S2 V5 S4 V2 S1 V3 S5 V5 S2 V4 S3 V2 S4 V1 S1 V2 S5 V5 S3 V1
S2 V3 S4 V4 S1 V5 S5 V4 S4 V1 S3 V2 S2 V3 S1 V1 S4 V2 S3 V3 S2 V4 S5 V5 S1
V4 S1 V2 S2 V1 S4 V3 S3 V5 S5 V3 S4 V4 S5 V1 S3 V2 S1 V5 S2 V2 S3 V4 S4 V5
S1 V1 S5 V3 S2 V2 S4 V5 S3 V4 S2 V1 S1 V3 S5 V1 S2 V2 S5 V3 S1 V4 S3 V5 S4
V1 S1 V3 S5 V5 S4 V4 S3 V2 S2 V2 S4 V1 S2 V4 S1 V3 S3 V5 S5 V4 S4 V1 S5 V5
S1 V2 S3 V3 S2 V1 S4 V3 S1 V2 S5 V4 S2 V5 S3 V1 S3 V2 S1 V3 S4 V4 S5 V5 S2

Appendix 2: Task 1 -- Object Identification

Video 1 Objects

__ banana __ bird __ binoculars __ cage __ camera __ coconuts __ flashlight __ flowers __ glasses __ gorilla __ man __ monkey __ peanuts __ collar __ suitcase __ syringe __ tarp __ trees __ watch __ woman

Video 2 Objects

__ advertisement __ Buddha __ children __ church __ construction site __ hair salon __ harbor __ flute player __ kindergarten __ palace __ police __ restaurants __ shopping center __ skyscrapers __ subway __ television __ theater __ traffic lights __ trucks __ wet cement

Video 3 Objects

__ bicycle __ bridge __ bow __ cabin __ child __ canoe __ crab __ duck __ fish __ gourd __ ladder __ loom __ monkey __ palm __ pulley __ raincoat __ river __ shrimp __ spear __ woman

Video 4 Objects

__ barns __ boat __ bridge __ bricks __ cannon __ carriage __ corn __ dead body __ dogs __ drum __ glasses __ grassland __ horse __ moon __ musket __ smoke __ soldiers __ trees __ window __ woman

Video 5 Objects

__ airport __ astronauts __ control center __ craters __ earth __ footprint

__ flag __ helmet __ ice __ instruments __ laboratory __ lightning __ lunar module __ meteorite __ parachute __ rats __ rocket __ red rocks __ ship __ skylab

Appendix 2: Task 2 -- Video Comprehension

Video 1

Please select the statement that best describes the gist of this video clip.

1. Poachers capture monkeys to get their fur.

2. People tag monkeys in order to track them.

3. A veterinarian operates on a monkey that had fallen ill.

4. Two people in a jungle safari photograph monkeys.

Video 2

Please select the statement that best describes the gist of this video clip.

1. Pollution is rapidly increasing in Singapore due to industrialization and commercialization.

2. Singapore is a modern society that maintains its traditional culture.

3. Singapore is a great place to shop.

4. Working mothers in Singapore are experiencing problems in juggling their many activities.

Video 3

Please select the statement that best describes the gist of this video clip.

1. The Machiguenga are building a bridge across the river.

2. The Machiguenga have developed special fishing techniques.

3. A Non-Machiguenga boy is adopted by the tribe.

4. Machiguenga families make a living using jungle resources.

Video 4

Please select the statement that best describes the gist of this video clip.

1. The American rebels defeat the British in a battle during the Revolutionary War.

2. The British aided the American colonists during the French and Indian War.

3. The British destroyed the Americanís supplies and ammunition in a battle during the Revolutionary War.

4. No British and American lives were lost during this battle of the Revolutionary War.

Video 5 Please select the statement that best describes the gist of this video clip.

1. Astronauts did a space walk.

2. Astronauts conducted experiments in space and landed on the Moon.

3. Astronauts trained and then landed on the Moon.

4. Astronauts released a new communication satellite.

Appendix 3: Questionnaire before the Experiment

Subject #: ____________ Age: ____________ Gender: Male ______ Female _______

1. What is the subject of your undergraduate degree? _____________________

What is the subject of your graduate degree? _____________________

What is the subject of your Ph D degree? _____________________

2. On the average, how much time do you spend watching TV, video and movies per week? ____ hours ______ minutes

Appendix 4 : Task of Speed Evaluation *

After each video was watched, the subject was asked to evaluate the specific speed. Subject # ______________ Please Evaluate the Display Speed.

I. About object identification 1.

1 frame per second too slow too fast 1 2 3 4 5 6 7 2. 4 frame per second too slow too fast 1 2 3 4 5 6 7 3. 8 frame per second too slow too fast 1 2 3 4 5 6 7 4. 12 frame per second too slow too fast 1 2 3 4 5 6 7 5. 16 frame per second too slow too fast 1 2 3 4 5 6 7

II. About video comprehension 1.

1 frame per second too slow too fast 1 2 3 4 5 6 7 2. 4 frame per second too slow too fast 1 2 3 4 5 6 7 3. 8 frame per second too slow too fast 1 2 3 4 5 6 7 4. 12 frame per second too slow too fast 1 2 3 4 5 6 7 5. 16 frame per second too slow too fast 1 2 3 4 5 6 7

Please briefly describe what is the best aspect and the worst aspect of this kind of video display. (this was done in the end of the experiment)