Rapid Serial Visual Presentation Techniques for Consumer Digital Video Devices

Kent Wittenburg*, Clifton Forlines*, Tom Lanning*, Alan Esenther*, Shigeo Harada, and Taizo Miyachi****

*Mitsubishi Electric Research Laboratories, Inc.
201 Broadway
Cambridge, MA 02139 USA
{wittenburg, lanning, forlines, esenther}@merl.com

**Mitsubishi Electric Corporation
Industrial Design Center
5-1-1 Ofuna, Kamakura, Kanagawa 247-8501 JAPAN
{sharada, miyachi}@idc.melco.co.jp

ABSTRACT

In this paper we propose a new model for a class of rapid serial visual presentation (RSVP) interfaces [16] in the context of consumer video devices. The basic spatial layout "explodes" a sequence of image frames into a 3D trail in order to provide more context for a spatial/temporal presentation. As the user plays forward or back, the trail advances or recedes while the image in the foreground focus position is replaced. The design is able to incorporate a variety of methods for analyzing or highlighting images in the trail. Our hypotheses are that users can navigate more quickly and precisely to points of interest when compared to conventional consumer-based browsing, channel flipping, or fast-forwarding techniques. We report on an experiment testing our hypotheses in which we found that subjects were more accurate but not faster in browsing to a target of interest in recorded television content with a TV remote.

Keywords

Rapid Serial Visual Presentation, RSVP, multimedia interfaces, video browsing, TV interfaces, consumer devices

INTRODUCTION

Most research in interfaces for video browsing, e.g., [2][3][4][8][10][13][14][17][19], has been done in the context of desktop systems for digital video and imagery, relevant to pc users but not necessarily to the mass of consumers who interact with digital image content through their cameras, video recorders, and television sets. The basic methods for consumers to visually browse recorded video, in particular, have changed little from its first inception in analog VCRs: the user manipulates direction and (perhaps) speed of a temporal image sequence in a fast-forwarding or reverse mode. Changes brought by the digital medium so far are generally restricted to offering indices, static page-based displays of image thumbnails and text, and a method for jumping to the content. Selecting television channels is equally limited to these two modes of interaction.

Recent digital video recording devices such as TiVo™ and Replay™ offer many interesting features that can fundamentally alter television watching behaviors. They enhance the features for skipping ahead by specified time increments when fast forwarding and (controversially) have offered automatic commercial detection and skipping features. However, the basic visual presentation and interaction paradigm for browsing is still the same: static indices or fast-forwarding in full-screen one frame at a time.

One of the published results that has investigated new visual browsing methods specifically for consumer video devices has examined the proposal to have users manipulate a scrolling presentation of key frame indices [9]. Users could alter the zoom level of the key frames, which were related to shot boundaries, as well as manipulate scrolling. Their results were encouraging--users were more satisfied with this interface than with the more conventional one. However, they were not able to show any differences in performance measures.

Rapid Serial Visual Presentation (RSVP) interfaces that explore trade-offs in temporal and spatial layout [6][7][16][18][19][20][21] offer one avenue to improve on the basic paradigms. These methods have the virtue of simplicity of controls, a major advantage to designers of consumer devices. The basic controls for such presentations require only methods for direction and variable speed, much like the controls necessary for fast-forwarding or reversing a video recorder [20].

Spence [16] provides an overview of work in RSVP interfaces. At one extreme of the RSVP methods is a temporal sequencing of single images where each successive image displaces the previous one, a paradigmatic case of video fast-forwarding or channel flipping. Spence calls this "keyhole mode," emphasizing the constricted view of the presentation in time. The more interesting techniques have combined some sort of spatial layout of images with the temporal sequencing. Spence mentions four variants: carousel mode, collage mode, floating mode, and shelf mode, shown in Figure 1. These all incorporate some form of spa-

UIST '03 Vancouver, BC, Canada

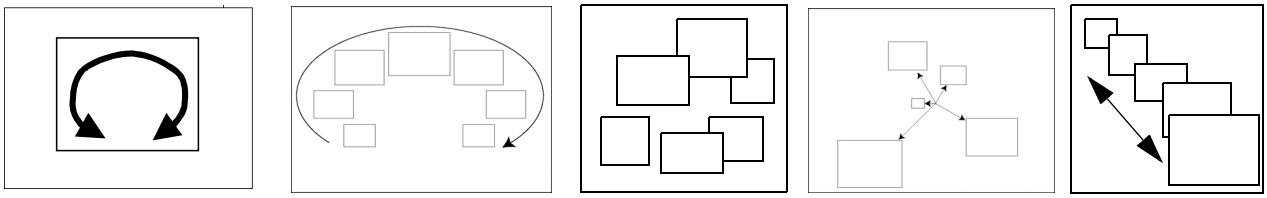

Figure 1: RSVP modes from left to right: keyhole, carousel, collage, floating, and shelf. The circular arrow within an image frame (as in keyhole mode) indicates replacement rather than movement.

tial/temporal layout of the image frames that add additional movement or displacement of the image content as the presentation proceeds. In three of these four modes (carousel, floating, and shelf), the images that are upcoming in the sequence are revealed in the background before moving to a more foreground position (or vice versa). In the collage mode, the images appear and disappear as the focus position cycles around the space [19].

The results of initial pilot experiments [6][20] have not yielded any evidence to date that these more sophisticated RSVP techniques are actually superior to the simple keyhole mode. Suggestive results so far are that users preferred a keyhole mode presentation of images in an e-commerce shopping task over a particular instance of collage mode [20] and that no advantage of carousel mode over keyhole mode was found in an image recognition task [6].

We suspect two factors may explain the results for the more sophisticated RSVP methods tested so far. First, as Spence mentions, presentations in which images move and/or appear and disappear in different positions in rapid succession naturally require more cognitive processing than keyhole mode presentations. From extensive psychology literature (an overview may be found in Coltheart [5]), we know that humans can process imagery presented extremely rapidly. Often only a quick glimpse is needed, a single fixation lasting only around 200 milliseconds, to extract content. Such rapid visual processing has been referred to as pre-attentive processing [12]. Keyhole mode is a condition much like many of the psychology experiments. Without moving their gaze position at all, users can extract a lot of information in a rapidly changing sequence of images. But if users have to change the position of their eyes as the presentation proceeds, they may lose some of the advantages of rapid presentation simply because they can't keep up. Second, we surmise that presentations that rely heavily on movement of images in a 2D plane (scrolling) are going to be more demanding to process than ones that move images forward or back in a depth dimension in a virtual 3D model. The basic psychology of human visual perception tells us that humans are wired to process imagery in a 3D world. In particular, rapid visual processing of approaching objects is a particularly important survival skill.

DeBruijn and Spence [7] have published comparative results on eye-tracking for some of the RSVP methods. They have characterized the eye-gaze patterns associated with certain modes. An observation about the behavior of one subject was revealing to us. They noted that in shelf mode, this subject seemed to focus only on the screen area in which new images would appear (before they moved off to a corner). Such behavior is indicative that the user was processing a particular image position as in keyhole mode (not shifting gaze). We surmise that the other images remained in the periphery to be attended to if necessary. Such a strategy seemed particularly interesting to us. In contrast, another subject followed each image as it moved along the track, shifting the eyes back each time to once again follow incoming images. These observations spurred us to consider whether support of keyhole mode processing shouldn't explicitly be a part of RSVP designs. Other questions arose. What tasks would keyhole mode attention patterns best support? What tasks would be supported by attending to the presentation of upcoming or already presented image sequences? We will discuss these issues in the context of tasks involved in video browsing in consumer devices later in this paper.

The result of these considerations and a good deal of prototyping has yielded our proposal for a basic paradigm in RSVP interfaces that we describe in the next section. It is a fusion of keyhole mode with methods for presenting and manipulating a neighboring 3D trail of images similar to shelf mode. We call our model Shift, Analyze, and Collect (SAC) in order to highlight three different aspects of the proposal. We then present several prototypes for consumer digital video devices. We describe a new interface for TV channel surfing and design variants for browsing and skipping recorded video and still images that may be saved on the hard disk drive of (future) TVs. We then describe an experiment we have run to test our hypotheses that our methods will be more precise and faster than conventional methods when browsing and skipping recorded video. We conclude with an indication of future work.

SHIFT, ANALYZE, AND COLLECT

The name "Shift, Analyze, and Collect" is for the following. *Shift* refers to movement along a stable 3D trajectory during advancement. *Analyze* is an indication of the role of system analysis that can be used to highlight or arrange the positions of the items in the 3D trail. *Collect* refers to the role of what we call a collector frame, which is the focus position that incorporates keyhole mode processing.

The method is generic to a sequence of any digital visual objects, although we will focus on images. Analyses of the

image sequence are designed to provide an "interest" function so that certain images drop out or pop up from image trail. In all cases we combine the spatial/temporal layouts with a common set of controls for continuous forward/back and adjustable rates of speed [20]. The visual effect of such controls is to advance the linear sequence of images towards or away from the main focus position close to the viewer.

Spatial layout

We propose a spatial layout of a sequence of images in a 3D space such that the trajectory formed as a line from center point to center point in the image sequence comprises a straight line or curve in which the position of the center point increases in the z dimension. The trails may represent either upcoming or just presented images, and a single design may incorporate one of each (upcoming and just presented). In two-dimensional coordinate systems, the depth dimension z is represented in the scale and position of the image. The effect for each trail is analogous to a roadway that may be straight or curved along which a series of signs appear in a spaced sequence. The viewer (driver) sees only a certain subsequence of the signs at any given point in time through the display (windshield)...

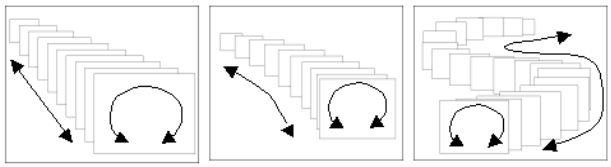

Figure 2: A few of the spatial/temporal layouts in the SAC method.

In the most basic case, all visual objects along a trajectory in a 3D space represent merely a strict ordering and are spaced evenly. However, the method provides for any spacing along a scale of choice. A spatial layout may reflect, say, temporal placement as well as any other semantically coherent relation that can be mapped onto values in a one-dimensional linear or nonlinear scale.

Traversing the sequence

The user is given controls governing the traversal of the sequence by manipulating forward/back direction and speed. As the user advances the sequence, the images appear to move closer (or further away) along a stable trajectory path. The analogy with signs along a highway is that the user can control the speed and direction of a car. Looking out the windshield, the signs will appear to move past at regular temporal and spatial intervals, appearing initially from far away and then moving closer. Or, if the car is moving in reverse and the driver remains facing forward, the signs will appear in view from over the shoulder and then regress into the distance. Whether the perceptual effect is that the viewpoint moves or that the trail of images moves is an interesting issue [18]. Interfaces that take the latter approach may avoid the lost in space problem recognized as an issue in many 3D interfaces.

Collector frame

In the mode of operation we have described above, as an image gets closer, it will eventually move past the field of view and disappear. This is the standard metaphor of what would happen if a driver in a vehicle passes signs along a roadway, looking straight ahead. We propose a variant on this basic mode that has the advantage of maintaining continuity across related visual objects (say, frames within a video) and providing a focal point for the users so that they do not have to shift their eyes to process a sequence at high speed. We use a collector frame, which is a visual container that exists at a z position closest to the viewer along the main trajectory. As the images move closer in the sequence they eventually approach the position of the collector frame. Instead of moving past, they replace the last image that was formerly in the view. Our method thus combines the advantages of conventional video players (with fast forward and reverse) with a layout that affords being able to look ahead (or behind) the focal point in the sequence of image frames or visual objects. The frames in the frontmost position in Figure 2 are collector frames, the circular arrow representing replacement rather than movement during advancement.

Analysis (clustering/segmentation) of the sequence

Having a sequence of images "exploded" into a third dimensional trail has advantages in providing a spatial context for a temporal presentation. It can provide an aid to visual memory unavailable in keyhole mode and can help users find and target a frame of interest. However, analysis of the sequence suggests other possibilities. A useful feature in video browsing is to select key frames at varying intervals that might be related to a zoom factor [9]. The relative temporal relationships can be maintained in the spacing. In other applications we will describe shortly, a user might initiate a query that would select a subset of the images. For instance, in channel surfing, a user might select a category such as sports or a favorite channel list. Here we speculate that there may be advantages in maintaining the original relative positions in the overall channel sequence instead of just replacing the overall sequence with a new sequence. Other ideas for analysis are suggested in [21], a paper in which we consider using SAC for visual data mining on conventional desktop hardware.

APPLICATIONS IN CONSUMER VIDEO DEVICES

Here we discuss our experience with applications of the SAC paradigm for consumer electronics devices and some of our explorations in the overall design space. The dimensions of the design space include path layout, spacing of items within the path, rendering of selected items, and variations in temporal "movement." From a historical perspective, we began this project with a new idea for TV channel surfing. We followed with fast-forwarding and reverse in recorded digital video and then integrated it with methods for video summarization (key frame selection). In parallel we prototyped television-based browsers for digital photo collections.

Figure 3: Screen shot of channel surfing prototype. Shown are images sampled offline for each of the channels with a second tuner.

Channel Surfing

One of the problems that has arrived with digital TV and satellite broadcasts is a delay during channel switching caused by the digital tuner. This delay is frustrating for seasoned channel surfers, who seem to like to blaze through channels sometimes for its own sake. (Other members of the family, of course, are annoyed to no end with such behavior.) Whatever one's feelings about channel surfing, we decided to build a prototype of what it would be like using a SAC interface. A screenshot is shown in Figure 3.

What is shown are images representing the current content of each of the available channels. There are at least two models for how the application could acquire such images. In one scenario, the TV would be equipped with a second tuner and sufficient memory to grab and save images coming live to the set. This tuner would be constantly cycling in the background to update its cache. The rate of refresh for the images would be an important factor in the application's effectiveness and would be related to the speed of tuning and the number of channels. In another scenario, representative images could be supplied by the content providers as an extension to electronic programming guides currently used in the industry. Although one may not be able to discern this from the black and white version of the figure, the visual quality of this prototype is high, the more so since it is capable of rendering at high definition TV resolution. Metadata in the form of the channel's call numbers and logo are rendered on a semi-transparent field on the same plane as the key frame content. As the user advances through the channels with a remote control, the images in each plane smoothly animate forwards or back. The collector frame in the front can be turned on or off in our prototype so that potential users and producers can get a feel for its effect. When the collector frame is off, images continue toward the user and "disappear" once they reach the camera plane. We experimented with numerous designs for the trajectory of the 3D trail. The curved trajectory shown in Figure 3 is one that worked well. Informal reactions to this demo have been extremely encouraging. There is something visually seductive about viewing images moving in 3D in this way. We believe that consumers would find such an alternative to channel surfing novel and pleasurable.

Further experiments with this prototype included variations in temporal layout. In one case, we tried methods involving nonlinear speed. The setting for this prototype was user control over viewing a subset of the total number of channels. The issue was traversal of the channel list when a subset has been selected. Our hypothesis was that there might be advantages to maintaining the global list in view even when browsing a subset. Users still might catch something unexpected as well as maintain a spatial sense of the overall order. So we provided a means for selecting subsets of the channels (favorites or specific categories) and then traversing through the subset so that each target in the subset would take constant time for transition. For example, suppose the user has selected a subset of frames as indicated in Figure 4 through a menu or button operation. The user manipulates the direction and speed controls in the usual way, moving forward or backwards, faster or slower, in a global sense. However, the rate of traversal depends on distances from targeted frames so that transitions from one target frame the next takes a constant time. This has the

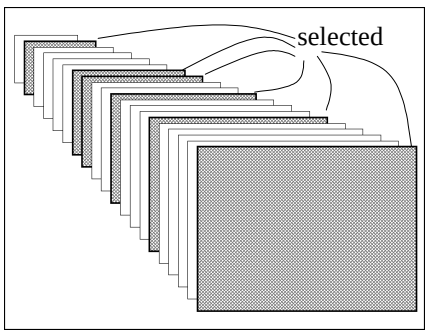

Figure 4: When traversing a subset of image frames, time is constant between each selection.

perceived effect that the speed of passing images varies according to how far away the current target frame is from the next target frame. We also experimented with variations that would show only target frames in the collector frame.

The above example of variable rate with constant time is just one example of issues in temporal layout in RSVP interfaces. This design space has barely begun to be explored. We should further comment that the scheme described above is one of many that might be explored for presenting subsets of selected channels. One might present just the subset without including the global context, and of course one could explore alternative orderings of the subset rather than maintaining the original ordering. There are trade-offs evident in these design variants such as maintaining visual continuity vs. offering a more specialized mode suited to the task.

Browsing and Skipping Recorded Video

Our second application area for consumer video devices was browsing and skipping of recorded digital video. This has relevance to consumer televisions, VCRs, and DVD recorders, and also to digital video cameras. We have primarily considered scenarios using TVs as the display device controlled by a remote outfitted with a jog dial or similar continuous speed controller. In the experiment section that follows we discuss one scenario in detail, namely, our experimental setting that used purely temporal sampling for acquiring key frames. Figure 5 illustrates another of our design explorations, which was to integrate automatic video summarization techniques using key frame selection. Our focus was on how to visually present key frames that are the output of an analysis process. Collaborators from our lab have developed techniques that use motion and audio analysis for this purpose [22]. Other key frame selection techniques, of which there are many, e.g., [1][23], would also be compatible.

Figure 5: A video browsing interface in which key frames are drawn with a white border and are brought to the front of the drawing order.

Figure 5 includes another type of path layout for SAC interfaces along with one design choice for highlighting key frames. The path shows not only frames that precede the focus collector frame in time but also those that come after. This sort of "V-formation" was also used in one of the variants of shelf-mode [16]. For revealing the position of key frames, this design brings the key frames to the front of the drawing order so that they are entirely visible rather than being obscured by images earlier in the temporal order. We have also experimented with simply dropping the nonselected frames out of the sequence while retaining the relative positions of the selected frames.

Digital Photo Collections

SAC interfaces are also well suited for digital photo collections. We expect that many home entertainment systems of

Figure 6: Browser for digital photos on a TV.

the future will be a repository not only of personal and commercial recorded video but also of still photo collections. The same sort of interface would be usable there. One key advantage of this form of RSVP over the usual designs that include pages of thumbnails is again the simplicity of controls. Instead of having to include separate controls for page flipping, scrolling, and/or selecting, the SAC style can integrate the functions into one simple linear stream in which the selected item is in the collector frame. This one-dimensional stream for presentation purposes has also been exploited in a very different setting for document browsing in desktop interfaces [11].

EXPERIMENT

In this section, we will discuss an experiment that we conducted in order to test SAC against the traditional presentation of images for users fast-forwarding through recorded video on an NTSC television monitor using a remote control. Our hypotheses were that users would be more accurately able to reach a desired point in a recorded video with SAC than with the traditional fast-forwarding technique and that they would be able to reach the desired point faster.

Two Fast-Forward Interfaces

For the purpose of this study, we based one interface on a VCR with variable speed fast-forward and the other on a simple version of Shift, Analyze, and Collect. We built an application that played QuickTime™ movies and fast-forwarded through them using a television for a display and a commercially available remote for control.

Variable Speed Fast-Forward

The first interface was designed to be familiar to anyone who has used a VCR with variable speed fast-forward. When advancing through recorded video, the interface pre-

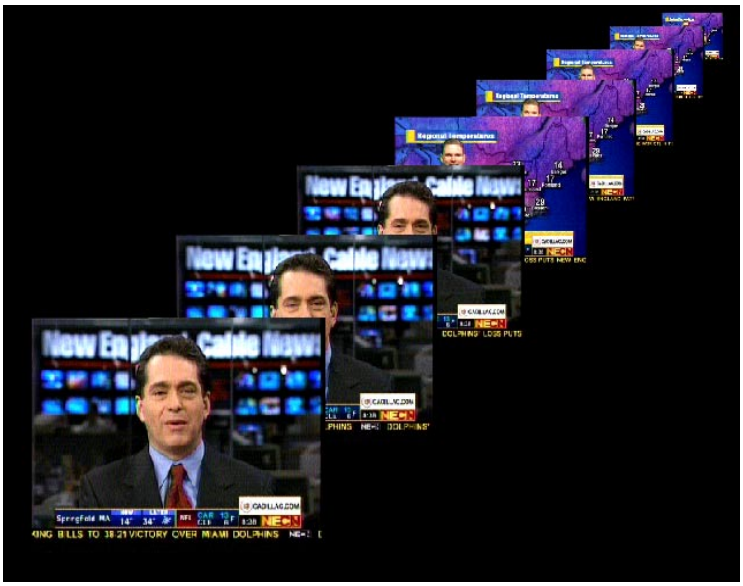

Figure 7: The SAC display in the experimental condition. A segment change is visible at the half-way point through an overall shift in visual composition.

sents frames from that video sampled at 1 frame(s) per second (fps). The frames are presented at a rate between 4 and 11 fps, so that the user is able to advance through the content at 4 to 11 times real-time. A sampling rate of 1 fps was chosen to represent what could easily be pulled out of a digital video stream using I-frames. The upper bound of 11 times real-time was chosen based on the maximum rate that a consumer DVD player could read. Many variations of the sampling and presentation rate could be used, and a further exploration of the interaction and trade-offs between them is left for future study.

Shift, Analyze, and Collect

The design space for SAC is huge, and each of the variations can potentially have an effect on the users' performance. For the purpose of this study, we decided upon a simple layout consisting of eight frames of video arranged linearly and appearing to recede into the distance. A picture of the testing application in the SAC condition is shown in Figure 7. During fast-forwarding, frames would appear at the upper right and proceed to the lower left. Upcoming segment changes (shifts to ads, for instance) would often be revealed in the overall visual properties of sets of images that would span several frame positions. In Figure 7, one can see the weather forecasting segment of a newscast is upcoming after the current anchor desk shot.

To reduce confounding variables, the testing application contained no animation between frames. Images appeared to jump directly to the next position in the layout. The images used in this condition were sampled at 1 fps from the video, and presented at 4 to 11 fps allowing the user to advance through the content at 4 to 11 times real-time. The sampled frames used in each of the two conditions were identical so that neither group had access to more video information.

Testing Content

QuickTime video was digitized directly from television broadcasts, so the clips accurately reflected the type of video users could expect to encounter. Other studies have shown the relationship between the genres of program and the performance of subjects browsing those programs [9][14]. Our aim was to include a reasonable sampling of a variety of content types. The content included clips from sports, game shows, cartoons, news broadcasts, and cooking shows.

Method

Fifteen subjects participated in this experiment. A few were interns and employees within our company and not directly paid for this experiment, others from outside the company were paid $20 for about 45 minutes of their time. Of the fifteen subjects, 10 were male and 5 were female. Their ages ranged from 21 to 53 years old. All but one described themselves as regular television watchers, and all had experience fast-forwarding through recorded video using a VCR.

Subjects were asked to perform simple fast-forwarding tasks using both interfaces (SAC and traditional) on 2 sets of video data, each set having 7 clips. The order that the sets were presented and the order that the techniques were used were balanced to control for interfaces and content.

Hardware Setup

Our goal was to implement the evaluation of SAC in a manner that closely resembled normal television watching. We wanted our subjects to say that they were "watching TV" rather than "using a computer." Toward this end, the test was controlled solely with a commercially available Mitsubishi remote control that included a jog dial. An IR receiver was placed discreetly next to the television so that the user could manipulate the remote by pointing it at the screen. The setup is shown in Figure 8.

The IR receiver was connected to a 2GHz Windows 2000 based PC with 1024 MB of RAM, 64 MB video memory, and 40 GB of disk space. The computer was configured to display video at analog 640x480 resolution through the 27" television. The computer also played sound through the television's internal speakers. Subjects sat about 5 feet from the television screen and used the jog dial of the remote control to operate both conditions.

Figure 8: View of the experimental setup.

Software Setup

The testing application was written using Java with the QuickTime for Java plug-in. Still images were generated by sampling each video clip at one-second intervals. Both testing conditions (SAC and traditional) used the same video files and the same collection of still image files.

The application not only responded to the commands sent by the user from the remote control, but also automatically recorded all of the signals sent from the remote control to the computer in a log file. These logs provided a quantitative description of all of the users' control actions during the experiment. To begin fast-forwarding through a section of video, the user simply turned the jog-dial clockwise. Turning the dial farther resulted in a faster rate of traversal. The rate of traversal was adjustable between four and eleven times real-time in both the SAC and traditional conditions. To stop fast-forwarding, the user simply let go of the dial, which would re-center itself. For the purpose of this study, rewinding was not implemented.

Procedure

Each session started with instructions on how to use the remote control and how to use both fast-forwarding interfaces. A warm-up video was loaded into the application and the user would watch a short bit of an academy award-winning movie. The subject was then instructed to practice fast-forwarding through portions of this movie using each technique. For the SAC technique, the experimenter pointed out the features of the SAC layout to the subject -- features such as the fact that each image on the screen was an upcoming frame from the video they were watching or that they could control the speed of fast-forwarding using either technique by adjusting the jog wheel of the remote control. The subjects were asked to "Play around with both techniques until you feel comfortable with both."

The subject was then asked to watch each of the 14 video clips in turn. The experimenter controlled the starting of each clip. Each clip was preceded by written on-screen instructions asking the user to watch some portions of the clip and fast-forward through other portions. Example instructions include "Please watch this program and fast-forward through the commercials" and "Please fast-forward to the start of the fireworks show." The subjects were instructed that speed and accuracy were both important in reaching the desired location in the video.

The first 7 clips were fast-forwarded through using either the SAC or traditional techniques depending on which group the subject was in. The remaining 7 clips were fast-forwarded through using the opposite technique. There was a short break between sets. The application recorded both the speed and accuracy with which the user reached the desired location as well as time-stamped signals received from the remote control. For each set of clips, the first 3 were practice clips meant to reduce learning effects. Only the last 4 contributed to the results.

At the end of the session, most users gave their impressions and ideas about the tools they just used.

Results

Our hypotheses were that users would more accurately reach a desired point in a recorded video with SAC than with the traditional fast-forwarding technique and that they would be able to reach the desired point faster. While the data collected supports the first hypothesis, the second is less clear.

- **There is a significant difference between the accuracy of the two techniques.** For each video clip for every user, the testing program recorded the distance in time between when the user stopped fast-forwarding and the ideal place to stop fast-forwarding (such as the first frame of a television program after a set of commercials). The average distance was calculated for each of the 8 performance measuring video clips. The SAC group had significantly lower errors than the traditional group (on average 6.87 vs. 9.18, $t(13) = -2.183$, $p = 0.023$). The average errors for each clip using both methods are shown in Figure 9.

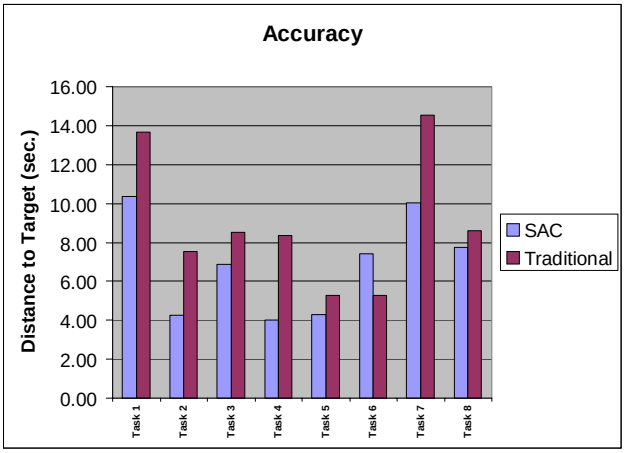

Figure 9: Accuracy rates, in seconds, for the two test conditions.

- **There is no significant difference between the speed of the two techniques.** For each video clip for every user, the testing program recorded the time taken to fast-forward to the desired location in the video clip. The average time to fast forward was calculated for each of the 8 performance measuring video clips. There was no significant difference found between the two groups (with means of 55.1 vs. 53.2, $t(13) = 0.22$, $p = .826$). The average task completion time for each clip using both methods is shown in Figure 10. Because the clips for each task were different lengths, the differences in task completion times between tasks is not relevant, only the difference between the two conditions within each task. It is notable that while there is no significant difference between the two techniques, users seem to perform slightly slower using SAC than the traditional technique.

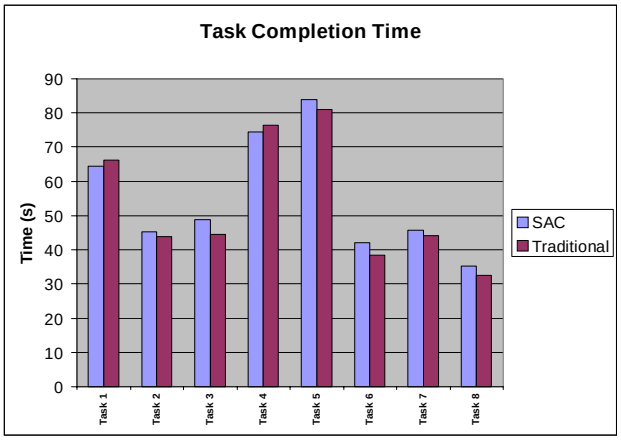

Figure 10: Average completion times for the two test conditions.

While not part of our initial hypotheses, the following finding came out of additional analysis of the collected data.

- **There is a significant difference between the variability of the rate with which users fast-forward through video using these two techniques.** The remote control jog dial allowed users to fast forward through the recorded video at variable rates (4-11 times normal speed). The testing application recorded each change in rate that the subjects initiated, and the total changes in rate were counted for each of the 8 performance measuring video clips. Users varied the rate at which they fast forwarded through the recorded video significantly more in the SAC group than in the traditional group (on average 12.9 changes in rate per clip vs. 9.8 changes, $t(13) = 2.56$, $p=0.037$). The average number of rate changes for each clip using both techniques is shown in Figure 11. While this comparison was not part of our original hypothesis, it seems that users take better advantage of the rate variability that the jog dial allows while using SAC than while using the traditional technique. This finding could explain the slightly slower task completion times of the SAC technique.

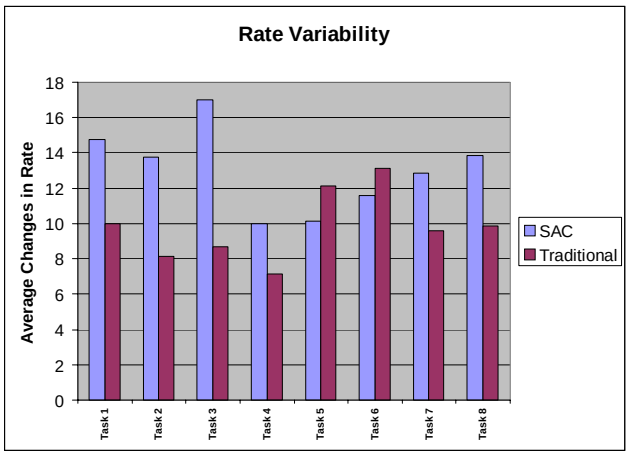

Figure 11: Comparison of the number of rate changes across the two conditions.

Discussion

The findings of this preliminary study were encouraging; however, as is often the case, many questions arose.

Because the instructions were written on screen, subjects were asked to perform a visual search for a written description of a position in the video. The findings may have differed if the subjects were given a visual cue to look for rather than a written description. Similarly, subjects were tested using clips they had never seen before. While we can guess that performance would increase using both methods for familiar video clips, we cannot say whether or not the differences between the techniques would remain.

An interesting observation arose from watching subjects practice both techniques with the warm-up video at the beginning of the experiment. Several subjects commented at the end of the session that the transition between scenes in the motion picture warm-up clip were easier to distinguish than scenes from the television broadcasts. In hindsight, it would have been better to have the subjects practice with television content for the television trials; however, this has led to the interesting question of what differences might arise from content as different as television, motion pictures, home videos, etc., while using SAC.

While our study probed both accuracy and task completion time, it is not entirely clear which measure is the "most important." Furthermore, other measurements such as attention or user satisfaction might be more important for some tasks. Several participants suggested at the end of the study that SAC was the "cooler," "more exciting," or "sexier" of the two interfaces, but how this would affect the success of a consumer device and how much of the novelty of SAC would wear off over time are questions left unanswered.

There are other limitations to the this study. First, while the study was run using a real TV and remote control, the setting was different from normal television watching. The experiments took place in the experimenter's office with the experimenter in the room. The clips were artificially short (three to five minutes each) and were chosen randomly from a broadcast schedule. Allowing subjects to choose the programming to record and review, as well as letting them view it in their typical television watching setting, would be an improvement. Furthermore, there was some confusion among subjects as to what point in a video clip constituted the desired location. For example, when fast-forwarding through a commercial break, some users stopped immediately after the last commercial while others continued to fast-forward through the following lead-in sequence. Finally, users' performance while using SAC continued to improve as they performed more tasks. A better study would have included more clips so that one could compare the plateaus of performance.

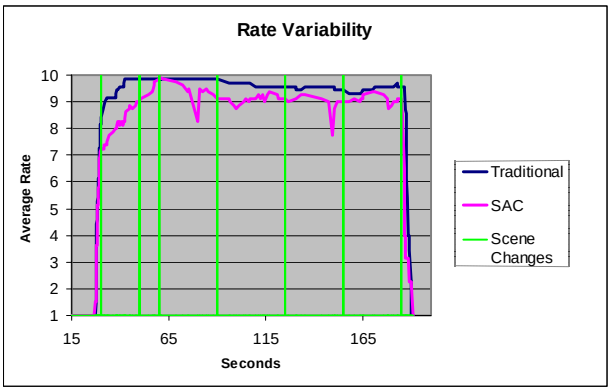

Figure 12: A graph comparing fast-forwarding rates of each condition where vertical lines represent scene changes.

Despite these cautions, we are able to report a significant positive difference in task performance using a novel interface compared to an extremely familiar one. Such results are not common in the literature of user interface systems and technologies. Our guess as to why users were able to be more accurate with the SAC interface than with the conventional one has to do with the ability to be able to better predict an upcoming visual target. This hypothesis is supported by looking at the patterns of a comparative sample of remote control usage shown in Figure 12. Note the downward spikes representing a slowing of speed for the SAC condition. They all happen just before a scene change, represented by the vertical lines. We speculate that the users noted a scene change coming up and thus slowed down to discern if the scene change was actually the target or not. In the traditional condition, users could not look ahead and thus did not slow down before scene changes.

The properties of the SAC presentation that afford this behavior have to do with the fact that scene or shot changes are often visually evident as a change in overall color and scene composition across a series of individual key frames in the trail. Consider again the example in Figure 7, in which a target scene would be visible as a segment (several images) of the upcoming trail that differs from the immediately preceding segment. Both segments are of course visible at the same time in a single frame of the overall SAC animation. In keyhole mode, which corresponds to the traditional fast-forward condition, users must integrate a series of single frames over time to form a mental image of a scene. It would seem that visual processing could be aided by presenting a series of single visual frames in which the boundary change is evident. The boundary would of course move through the path of the SAC presentation over time. Recognizing such a scene shift would not require detailed processing of any single sampled video key frame, but rather the processing in a more global sense of visual properties across a trail of key frames. We speculate that such a mental task, since it can be done in the global presentation (itself a single image), may be done pre-attentively [12]. Eye-tracking studies duplicating the experimental conditions we have discussed here could lend evidence to the accuracy of our speculations.

CONCLUSION

This paper has extended earlier work in RSVP interfaces to propose a principled paradigm for RSVP that integrates keyhole mode within a larger temporal/spatial context. We call the model Shift, Analyze, and Collect (SAC). We then presented some prototypes of applications for consumer digital video devices, indicating some of the many dimensions in the design space for SAC interfaces that are just beginning to be explored. We followed up with an experiment to test our hypotheses that SAC interfaces will yield better performance for users engaged in video browsing and skipping tasks. Our findings were that users were indeed more accurate in finding the target of a common video browsing task compared to a standard video fast-forwarding interface. We also noted an unanticipated finding that users made more use of variable speed control with the SAC interface than with the conventional one. In our discussion, we speculated that SAC style interfaces performed better because each frame of an animated SAC sequence in this task incorporates global properties of a series of individually sampled video frames. In conventional interfaces, users have to integrate the individual video frames over time to understand scene shifts. In the SAC interface, scene shifts are evident as visual properties of single SAC frames, and thus users can take advantage of rapid preattentive image processing to glean this information. There are many more experiments that could and should be done to confirm these hypotheses and to explore the relationship of SAC interfaces to user tasks more thoroughly.

As for practical concerns, we have shown that the model of RSVP interfaces we are proposing can be accommodated without radical changes to input devices currently available to home video consumers. Our hope is that we may see the introduction of such interfaces to the marketplace in the not too distant future.

REFERENCES

1. Aoki., H., Shimotsuji, S., & Hori, O. (1996) A shot classification method of selecting effective key-frames for video browsing. In Proceedings of ACM Multimedia '96 (November, Boston MA,USA), ACM, pp. 1-10.
2. Boreczky, J., Girgensohn, A., Golovchinsky, G., & Uchihashi, S. (2000) An interactive comic book presentation for exploring video. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (The Hague, Netherlands), ACM, pp. 185-192.
3. Christel, M. G., Winkler, D. B., & Taylor, C. R. (1997) Multimedia abstractions for a digital video library, Proceedings of the second ACM international conference on Digital Libraries (July 23-26, Philadelphia PA, USA), ACM, pp.21-29.
4. Christel, M. G., Hauptmann, A. G., Wactlar, H. D., & Tobun, D. N. (2002) Collages as dynamic summaries for news video. In Proceedings of the ACM Multimedia '02 (December, Juan-les-Pins, FRANCE), ACM, pp. 561-569.
5. Coltheart, V. (Ed.) (1999) Fleeting Memories: Cognition of Brief Visual Stimuli. MIT Press, Cambridge MA, USA.
6. De Bruijn, O., & Spence, R. (2000) Rapid serial visual presentation: a space-time trade-off in information presentation. In Proceedings of the Working Conference on Advanced Visual Interfaces (AVI 2000) (May, Palermo, ITALY), ACM, pp. 189-192.
7. De Bruijn, O., & Spence, R. (2002) Patterns of eye gaze during Rapid Serial Visual Presentation. In Proceedings of Advanced Visual Interfaces (AVI 2002) (May, Trento, Italy), ACM, pp. 209-217.
8. Ding, W., Marchionini, G., & Soergel, D. (1999) Multimodal surrogates for video browsing. In Proceedings of the fourth ACM Conference on Digital Libraries (August, Berkeley CA, USA), ACM, pp. 85-93.
9. Drucker, S. M., Glatzer, A., De Mar, S., & Wong, C. (2002) SmartSkip: Consumer level browsing and skipping of digital video content. In Proceedings of CHI 2002 (April, Minneapolis MN, USA), ACM, pp. 219-226.
10. Elliot, E. (1993) Watch, grab, arrange, see: thinking with motion images via streams and collages. MSVS thesis, Massachusetts Institute of Technology, Cambridge MA, USA.
11. Freeman, E., & Gelernter, D. (1996) Lifestreams: a storage model for personal data. ACM SIGMOD Record, 25, 1, 80-86.
12. Healey, C. G., Booth, K. S., & Enns, J. T. (1996) High-speed visual estimation using preattentive processing. ACM Transactions on Computer Human Interaction (TOCHI), 3, 2, 107-135.
13. Lee, H. Smeaton, A., Berrut, C., Murphy, N. Marlow, S, & O'Connor, N.E. (2000) Implementation and analysis of several keyframe-based browsing interfaces to digital video. In Proceedings of Research and Advanced Technology for Digital Libraries, 4th European Conference, ECDL 2000 (September, Lisbon, PORTUGAL), LNCS 1923, Springer, pp. 206-218.
14. Li, F. C., Gupta, A., Sanocki, E., He, L., & Rui, Y. (2000) Browsing digital video. In Proceedings of CHI 2000 (April, The Hague, Netherlands), ACM, pp. 169-176.
15. Spence, R. (2001) Information Visualization. ACM Press & Addison-Wesley, Harlow, England.
16. Spence R. (2002) Rapid, serial and visual: a presentation technique with potential. Information Visualization, 1, 1, 13-19.
17. Tse, T., Marchionini, G., Ding, W., Slaughter, W., & Komlodi, A. (1998) Dynamic Key Frame Presentation Techniques for Augmenting Video Browsing. In Proceedings Advanced Visual Interfaces (May, L'Aquila, ITALY), ACM, pp. 185-194.
18. Wittenburg, K., Ali-Ahmad, W., LaLiberte, D., & Lanning, T. (1998) Rapid-Fire image previews for information navigation. In Proceedings of Advanced Visual Interfaces (May, L'Aquila, ITALY), ACM, pp. 76-82.
19. Wittenburg, K., Nicol, J., Paschetto, J., & Martin, C. (1999) Browsing with dynamic key frame collages in web-based entertainment video services. In Proceedings of IEEE International Conference on Multimedia Computing and Systems (June, Florence, ITALY), IEEE, Vol. 2, pp. 913-918.
20. Wittenburg, K., Chiyoda, C., Heinrichs, M., & Lanning, T. (2000) Browsing through rapid-fire imaging: requirements and industry initiatives. In Proceedings of Electronic Imaging '2000: Internet Imaging (January, San Jose CA, USA), SPIE, pp. 48-56.
21. Wittenburg, K., Lanning, T., Forlines, C., & Esenther, A. (2003) Rapid serial visual presentation techniques for visualizing a third data dimension. In Proceedings of HCI International 2003 (June, Crete, GREECE), Lawrence Erlbaum, Vol. 4, pp. 810-814.
22. Xiong, Z., Radhakrishnan, R., & Divakaran, A. (2003) Generation of sports highlights using motion activity in combination with a common audio feature extraction framework. In Proceedings of IEEE International Conference on Image Processing (ICIP) (September, Barcelona SPAIN), IEEE, to appear.
23. Zhang, H., Kankanhalli, A., & Smoliar, S. (1993) Automatic partitioning of full-motion video. Multimedia Systems 1, 1, 10-28.