An Eye Tracking Approach to Image Search Activities Using RSVP Display Techniques

Simone Corsato
Dip. di Informatica e Sistemistica
Università di Pavia
Via Ferrata, 1 - 27100 - Pavia - Italy
Phone +39 0382 985486
simocor@tele2.it

Mauro Mosconi
Dip. di Informatica e Sistemistica
Università di Pavia
Via Ferrata, 1 - 27100 - Pavia - Italy
Phone +39 0382 985486
mauro.mosconi@unipv.it

Marco Porta
Dip. di Informatica e Sistemistica
Università di Pavia
Via Ferrata, 1 - 27100 - Pavia - Italy
Phone +39 0382 985486
marco.porta@unipv.it

ABSTRACT
Rapid Serial Visual Presentation (RSVP) is now a well-established category of image display methods. In this paper we compare four RSVP techniques when applied to very large collections of images (thousands), in order to extract the highest quantity of items that match a textual description. We report on experiments with more than 30 testers, in which we exploit an eye tracking system to perform the selection of images, thus obtaining quantitative and qualitative data about the efficacy of each presentation mode with respect to this task. Our study aims at confirming the feasibility and convenience of an eye tracking approach for effective image selection in RSVP techniques, compared to the mouse-click "traditional" selection method, in view of a future where eye trackers might become nearly as common as LCD displays are now. We propose an interpretation of the experimental data and provide short considerations on technical issues.

Categories and Subject Descriptors
H.2.8 [**Database Management**]: Database Applications – *image databases*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *search process*; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces – *graphical user interfaces (GUI).*

General Terms
Performance, Experimentation, Human Factors.

Keywords
Image database, image presentation, image browsing, rapid serial visual presentation, eye tracking.

Conference'08, May 28-30, 2008, Napoli, Italy.

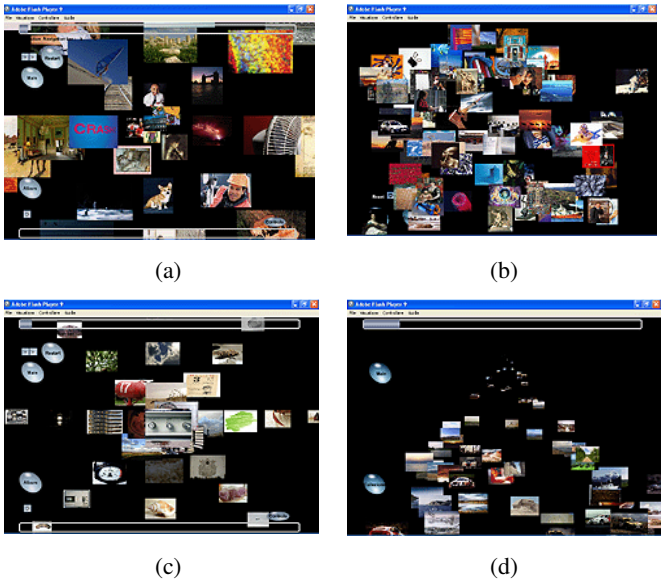
1. INTRODUCTION
Very often, we need to deal with large collections of images, and we want to select only some pictures according to certain criteria. For example, we may be interested in finding images with well-defined features, such as specific contents or technical properties; or, on the contrary, we may want to browse the database to search for something that only we can judge as suitable for our purposes. Quite common is also the case where the user simply desires to get some idea of the content of the picture database, like when rapidly riffling the pages of a book.

In the spatial domain, the most familiar visualization method is certainly the grid, in which pictures are arranged according to a matrix layout. In web pages and file folders, thumbnail images are usually displayed this way.

1.1 Rapid Serial Visual Presentation Modes
To achieve high search speeds, however, several dynamic visualization approaches have been proposed in the last years, among which those pertaining to the RSVP group deserve special attention. RSVP stems from Rapid Serial Visual Presentation and indicates a visualization mode where images are displayed in sequence, in the same location, for a short period of time (e.g. 100 milliseconds) [8]. A number of variants of RSVP have been proposed (also by our research group [4, 6]), which are more or less directly connected with it [1]. For instance, the *Floating* presentation mode is a time-dependent visualization technique in which small images appear about at the center of the screen and progressively enlarge, disappearing at the four sides (similarly to motorway signs which seem to move towards the driver). In our implementation (Figure 1a), to reduce image overlapping, pictures follow eight radial paths, and the angular distance between the directions of consecutive images is 135°. In the *Collage* display (Figure 1b), pictures appear very rapidly, in random positions, thus overlapping each other, like if being thrown onto a table. In the *Volcano* method (Figure 1c), images are "erupted" by the central "crater" of a virtual volcano, and slide down laterally along the virtual slopes, with a perspective effect; like in the Floating method, pictures follow eight radial paths. In the *Shot* display mode (Figure 1d), images, like "bullets", are "fired" by a virtual "gun" and progressively reach the lower part of the screen with a perspective effect.

While also other RSVP variants have been devised, in the experiments we present in this paper we focused on the above-described techniques for two major reasons. Firstly, our main

interest is in presentation modes able to display many images at a time (so that very fast visualization rates can be achieved with large image collections), and the four display methods considered satisfy such requirement more than others. Secondly, unlike other techniques, such approaches have as a common trait the fact of being characterized by image spatial distributions that occupy most of the screen area, thus potentially requiring eye-intensive screen exploration from the user.



(a) (b)

(c) (d)

Figure 1. Floating (a), Collage (b), Volcano (c), and Shot (d) display modes

1.2 Eye Tracking for the Evaluation of RSVP modes

Eye tracking can undoubtedly be a valuable source of information in the study of image presentation modes. The observation of eye scanpaths can in fact provide hints about the design of display methods, as well as suggestions about new interaction modalities. This is the reason why in our experiments we have considered both common efficiency indicators for search tasks (e.g. the number of correct pictures found within a set in a defined time period) and eye tracking data, obtained through the use of an unobtrusive eye tracker.

In the last decade, the potential of eye tracking for assessing RSVP methods has begun to be exploited. One of the first investigations of this kind was [3], where four visualization techniques were considered, namely Carousel, Collage, Floating and Shelf. Experiments, conducted with two testers who were asked to search a pre-viewed target image using the above-quoted approaches, had the main purpose to find correlations between eye gaze data and the trajectories of pictures. The tests showed that while the four display modes do not present specific perceptual problems, there may be differences among them for what concerns the effort required from users: methods in which images move may, in fact, cause some strain of the visual system.

Another interesting work, described in [9], investigated the relation between the space and time domains in display methods. The study considered the task of detecting the presence or absence of a previously viewed picture within a collection, using three modes: Slide Show (where 64 images were displayed in sequence, in the same position, at regular time intervals), Static (which was a static grid of pictures) and Mixed (a combination of the two previous modes, where 2x2 grids of images were displayed in sequence, in the same position, at regular time intervals). Since the testers expressed a strong preference for the Mixed presentation approach, which seemed also to be the one less prone to errors, eye tracking was used to try to better understand such an outcome. In particular, the hypothesis was tested that users tend to fix their gazes at the center of the four images of the Mixed mode, thus reducing the eye exploration extent while getting a "quick-glance" understanding of the images being displayed.

A more recent study [1] exploited the Slide Show, Mixed and Static presentation modes along with three other RSVP display techniques, namely Diagonal (images move diagonally from the upper left to the lower right corner of the screen), Ring (pictures appear at the center of the screen, rotate around it, and then disappear through the upper edge) and Stream (images flow along a hyperbolic trajectory, starting from the lower right corner of the screen and disappearing in the opposite corner with a perspective effect). Three tasks were considered: searching a pre-viewed target image, searching an image described in detail and searching an image described in general terms. Three different presentation rates were used. The study took into account such parameters to obtain, for each display mode, data about recognition rate and accuracy, as well as other indirect measures.

Our study focuses on the identification of all the images matching a textual description. Such images have to be selected as soon as they are identified, without interrupting the normal presentation flow. We want the identification times to be decoupled from the selection times: all the images should be selectable with the same (minimal) motor effort. This can be achieved by means of an eye tracking approach.

1.3 A Scenario for an Eye Tracking Approach to RSVP

In view of a future where eye trackers might become as common as LCD displays are now, we desire to find clear evidences that an eye-driven approach, besides being more natural [5], could really speed up search activities within very large collections of images.

We imagine a scenario where a graphic designer (the "user") deals with a collection of some thousands of images. According to certain criteria, he wants to rapidly reduce the number of images to a subset that can be reasonably managed later with more attention: for instance, he may want to pre-select all the pictures representing a cat. Quickness is here major concern: it is not important if some wrong images will be selected or if some appropriate ones will be missed, because a second, more accurate, selection will be later performed, based on other convenient criteria (for instance, to select a few images of lazy cats, suitable for a graphic project). This kind of research may be performed effectively also on small resolution images (as it happens on the web with stock photos).

Within our scenario, the user, after having spent few seconds for calibrating the eye tracking system, starts examining the rapid sequence of pictures displayed on the computer screen according to a RSVP mode. As soon as he identifies a proper picture, to select it, he just presses a key on the keyboard (or possibly acti-

vates a special sensor): the system marks the picture which corresponds best to the current user's gaze screen coordinates. After a session in which thousands of pictures have been displayed, the user can now concentrate in a small subset of few dozens pictures.

2. EXPERIMENTS

In the context of image search activities where the user is required to find pictures pertaining to well-defined categories within large databases, our hypothesis was that the described eye tracking approach for image selection in RSVP techniques could be feasible and convenient with respect to the mouse-click "traditional" selection method. Moreover, we expected to find significant differences in the efficacy of the chosen presentation modes, as suggested by researches mentioned in section 1.2, concerning similar tasks.

We tested our hypothesis by comparing the performances of a group of 31 students, all aged between 20 and 27. Each tester tried both the four considered dynamic RSVP methods (*Volcano, Floating, Collage, Shot*) with the eye tracking approach and a simple grid interface, with a point-and-click approach, which may be considered the present standard solution for this task. We stress again that the aim of this research was not to directly compare the four methods with the grid: a dynamic grid will be tested in future experiments.

Platform.

As an eye tracker, we used the Tobii 1750 [10], which integrates all its components (camera, infrared lighting, etc.) into a 17'' monitor. With an accuracy of 0.5 degrees and a relatively high freedom of movements, the system is ideal for simulating real-use settings, where it would be intolerable to constrain users too much in their activities. The device returns the *x* and *y* user's gaze screen coordinates, recorded by dedicated software 50 times a second. User interfaces for the tests were coded using Adobe Flash technology.

In our tests, as soon as a potential target picture was recognized in the four RSVP techniques, the user had to press any key on the keyboard. We didn't implement pictures selection entirely in real-time. Rather, we relied on data registered by the eye tracker, which include video files showing real-time scanpaths during the session and Microsoft Excel files reporting, among other information, fixation times and duration, as well as timestamps relative to 'key press' keyboard events. This way, it was possible to correlate 'a posteriori' images which were being looked at by the testers at a particular moment and their "conscious" selection action. In total, 155 clips were analyzed, each one lasting about 210 seconds.

Variables.

As it can be easily guessed, presentation speed in RSVP methods does influence the accuracy of search, as well as, of course, the total exploration time. However, in this preliminary phase of our activity we decided to limit the number of experiment variables, in favor of test sessions characterized by reasonable durations and well-defined comparable data. Before the actual tests, we carried out several pilot trials aimed at identifying the "optimal" presentation rates for each method, in terms of number of correct images found and subjective judgments about the chosen speeds ("too fast", "acceptable", etc.). We selected a presentation rate of a new picture every 105 milliseconds. This way, for each RSVP display mode, the presentation time for 2000 images was fixed to 3 minutes and 30 seconds.

Within the considered visualization techniques, pictures are displayed for different amounts of time, due to their different paths. They occupy different portions of the screen; they may overlap and they may progressively shrink or enlarge in different ways. Instead of devising sophisticated parameters to make the methods more directly comparable (such as, for instance, the integral of pictures area on the display time interval), we decided to separately "optimize" the different methods (in terms of picture sizes, paths and display time) during the preliminary phase in order to set the parameters of the tests. Figure 2 shows the average size (length of the diagonal, in pixels) and life-time (in seconds) of pictures as used in the experiments for the Floating, Volcano and Shot methods; for the Collage and the Grid methods the size is constant and the life-time is variable.

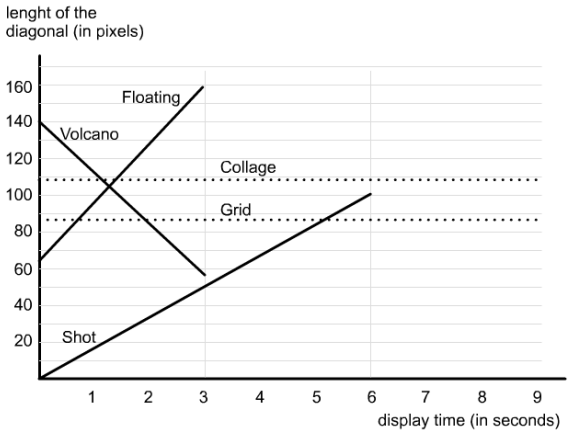


Figure 2. Average size (length of the diagonal, in pixels) and life-time (in seconds) of pictures used in the experiments for the different display modes

Experimental design.

After a short calibration procedure, necessary for the eye tracker to correctly understand where the specific tester is looking at, each user searched for target images using the four RSVP display techniques, plus the grid, in five different sessions.

Five different sets of 2000 images were employed. Each set contained 40 pictures pertaining to a specific target theme (namely cats, dogs, ships, planes and cars) and 1960 images with other content. Each presentation mode had of course a different set of pictures.

The testing order of display methods and their associated image sets and target themes varied among testers, so as to prevent results from being biased by learning effects, kind of target and possible user's mental fatigue. The total time necessary to introduce each participant to the experiment, explain the procedure, show examples and perform the five tests was about 50 minutes.

The purpose of the experiments was to find, in 3 minutes and 30 seconds, as many images as possible pertaining to a theme (40 out of 2000 in total). For the grid display, images were subdivided into 32 screens, arranged in 8x8 grids, and the user could move among them through 'next'/'previous' buttons.

At the end of each test session, users were also asked to express a subjective judgment about each method in terms of efficacy and fatigue, with values from 1 (lowest efficacy/fatigue) to 5 (highest

efficacy/fatigue). After the tests, data produced by the eye tracker were analyzed in detail, to extract both quantitative and qualitative data about the effectiveness of the display modes.

3. RESULTS AND DISCUSSION

Thanks to the devised approach, we have been able to compare the RSVP display methods according to "how easily" they allow image recognition (and selection), rather than on the basis of mouse clicks (which strongly depends on individual reaction times and hand/mouse spatial coordination ability). Main results are summarized in the following tables while the graphics in Figure 3 provide a clue about the distribution of the data.

Table 1. Average performances of testers by presentation method

	Floating	Volcano	Shot	Collage	Grid
num. of right pictures selected	29.51	28.00	19.29	19.35	15.58
num. of wrong pictures selected	1,87	2,06	2,35	1,64	0,74
wrong / right pictures (perc.)	7.12%	7.94%	16.72%	9.95%	6.25%

Table 2. Average eye movements for testers by presentation method (monitor resolution: 800x600)

	Floating	Volcano	Shot	Collage	Grid
scan path length per minute (in pixels)	13,462	14,810	25,783	36,349	31,868
aver. duration of fixations (msec)	274	307	182	187	181
aver. duration of saccades (msec)	46.5	50.1	60.0	56.2	59.3

Table 3. Average users feedback by presentation method. Subjective judgment in terms of efficacy and fatigue, with values from 1 (lowest efficacy/fatigue) to 5 (highest efficacy/fatigue)

	Floating	Volcano	Shot	Collage	Grid
efficacy (subjective)	3.97	3.70	3.26	2.42	2.16
fatigue (subjective)	2.52	2.71	3.13	3.90	1.74
efficacy / fatigue ratio	1,58	1,37	1,04	0,62	1,24

As can be noted from the previous tables, the results obtained do confirm our hypothesis according to which a selection method based on eye tracking is practicable and convenient compared to the state-of-the-art mouse-click approach. During the 3 minutes and 30 seconds allotted to each session, in the test with the grid users were able to freely control the presentation rate (screen change), but most of the time they did not succeed in examining all the images: on average, in fact, only 51,5% of the 2000 pictures was inspected, against 100% of RSVP methods. Of course, using the grid only few images were erroneously selected, since they were still. The ratio between wrong and correct images was very close to that of the Floating and Volcano techniques, which, however, allowed almost twice the number of images to be inspected. Also, users judged the "traditional" interface less effective than the others (although less tiring).

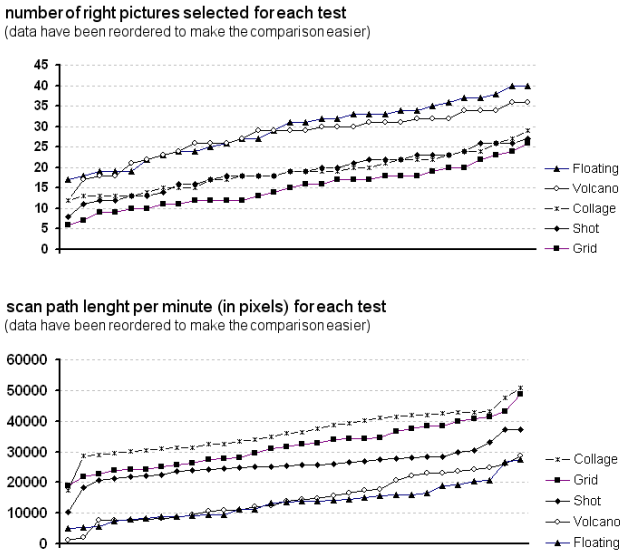


Figure 3. Number of right pictures selected and scan path length per minute for each test

To date, the study has demonstrated the viability of the eye tracking approach. We expect that also a display mode based on a dynamic grid (the *Tile* method described in [1]) could allow good performance using this new approach, due to the fact that pictures remain still. Our future experiments will just compare the Tile technique with the best among the four RSVP methods considered in this study.

According to our measured results and to the opinion of the testers, the Floating method has emerged as the most promising one, that is the most effective, efficient and satisfactory. Using the devised "visual selection", the performance of the Volcano mode is similar to that of the Floating technique, while in an interface based on mouse-click the selection of rapid-moving targets which get smaller and smaller would be rather demanding (Fitts' Law).

Data about eye-gaze behaviour confirm that the effort required from the observer is different in the various methods, and that such effort is related to the effectiveness of the methods themselves. The Floating and Volcano techniques, in fact, besides allowing better performances, are characterized by shorter average saccade times (even if this is only a rough figure) and shorter scan paths. Gazeplot and hotspot graphs generated by the eye tracking system prove that the Floating and Volcano modes are visually less "disorganized".

Currently, we are still examining the huge amount of data obtained to identify possible correlations among image size, motion speed and recognition time.

4. IMPLEMENTATION ISSUES

The implementation of an image search system based on eye tracking like the one proposed in this paper surely poses some technical issues.

The main problem is due to the difficulty of automatically identifying the image to be selected when another picture is very close to it, or even partially overlapped. The observation of key press

times within the recorded video clips has allowed us to correlate gaze positions to the actual "will" to select an image. A method which simply selects the image that is closer to the gaze center would be rather error-prone, as shown in Figure 4.

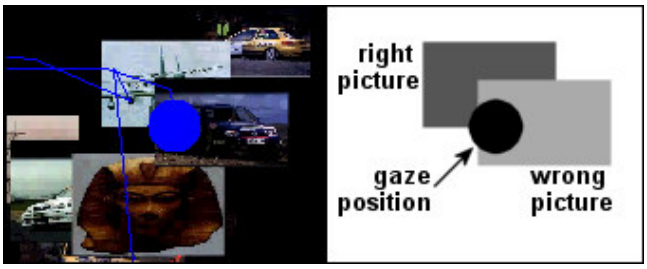


Figure 4. An example of image overlapping in the Shot display mode

However, the error probabilities could be reduced considering the user's eye-gaze behaviour. For instance, as shown in Figure 5, we have noted a tendency, more marked in some subjects, to anticipate the key press before the gaze is fully centered on a target picture.

Each column in Figure 5 shows the behaviour of a different tester while performing 5 selections (within each column, the different segments have been reordered by type). A black segment (type A) means that the corresponding keystroke occurred (more than 66 msec) before the gaze position was over the selected picture, which happens in about the 30% of the observed cases. Gray segments (type B) represent keystrokes which happen slightly prematurely (less than 66 msec before the "right" time, 9.7 %). Light grey segments (type C) represent "punctual" keystrokes (about 60%) . Less than 1% keystrokes were delayed (type D).

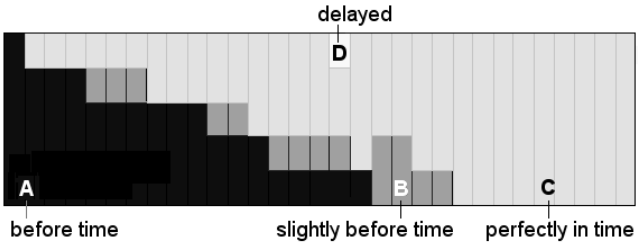


Figure 5. User's behavior: keystrokes occur before the eye gaze is centered on the target image in 1/3 of cases

Thus, while solutions without overlaps seem more promising, there are optimizations which may help to reduce ambiguities in display techniques that imply potential image conflicts.

5. CONCLUSIONS

In this paper we have considered the problem of searching well-defined target images within very large image databases. Eye tracking has been used to compare four RSVP display methods each other and with the "traditional" grid layout. Through tests involving 31 users, we have collected a huge amount of data, which we have now started to analyze. For example, we have discovered that the Floating and Volcano techniques are better than the Collage and Shot modes, in terms of number of correct images found, length of the eye path and user judgment. Indeed, the inspiring motivation of our study was the conviction that, for image selection, an eye tracking approach can be more efficient than the usual point-and-click mouse-based solution. The results of our investigations do confirm our hypothesis.

6. ACKNOWLEDGMENTS

This work has been supported by funds from the Italian FIRB project "Software and Communication Platforms for High-Performance Collaborative Grid" (grant RBIN043TKY).

7. REFERENCES

[1] Cooper, K., De Bruijn, O., Spence, R., and Witkowski, M. 2006. A Comparison of Static and Moving Presentation Modes for Image Collections. In *Proc. of AVI 2006*, Venice, Italy, May 23-26.

[2] De Bruijn, O., and Spence, R. 2000. Rapid Serial Visual Presentation: A space-time trade-off in information presentation. In *Proc. of AVI 2000*, Palermo, Italy, May 23-26, 189-192.

[3] De Bruijn, O., and Spence, R. 2002. Patterns of Eye Gaze during Rapid Serial Visual Presentation. In *Proc. of AVI 2002*, Trento, Italy, May 22-24, 209-214.

[4] Demontis, G., Mosconi, M., and Porta, M. 2003. Experimental Interfaces for Visual Browsing of Large Collections of Images. In *Proc. of HCI International 2003 (HCI '03)*, Crete, Greece, June 22-27.

[5] Oyekoya, O. K., and Stentiford, F. W. M. 2004. Eye Tracking as a New Interface for Image Retrieval. *BT Technology Journal* (Springer), Vol. 22, N. 3 / July, 161-169.

[6] Porta, M. 2006. Browsing Large Collections of Images through Unconventional Visualization Techniques. In *Proc. of AVI 2006*, Venice, Italy, 23-26 May.

[7] Simonin, J., Kieffer, S., and Carbonell, N. 2005. Effects of Display Layout on Gaze Activity During Visual Search. In *Proc. of INTERACT 2005 (13th International Conference on Human-Computer Interaction)*, Rome, Italy, September 12-16, 1054-1057.

[8] Spence, R. 2002. Rapid, Serial and Visual: a presentation technique with potential. *Information Visualization*, 1, 1, 13-19.

[9] Spence, R., Witkowski, M., Fawcett, C., Craft, B., and De Bruijn, O. 2004. Image Presentation in Space and Time: Errors, Preferences and Eye-gaze Activity. In *Proc. of AVI 2004*, Gallipoli (LE), Italy, May 25-28, 141-148.

[10] Tobii Technology AB 2003. Tobii 1750 Eye-tracker (Release B), November '03.

[11] Wittenburg, K., Chiyoda, C., Heinrichs, M., and Lanning, T. 2000. Browsing Through Rapid-Fire Imaging: Requirements and Industry Initiatives. In *Proc. of Electronic Imaging 2000*, San Jose, CA, USA, January 23-28, 48-56.

[12] Wittenburg, K., Forlines, C., Lanning, T., Esenther, A., Harada, S., and Miyachi, T. 2003. Rapid Serial Visual Presentation Techniques for Consumer Digital Video Devices. In *Proc. of 16th ACM Symposium on User Interface Software and Technology (UIST '03)*, Vancouver, Canada, November 2-5, 115-124.