# Text or Pictures? An Eyetracking Study of How People View Digital Video Surrogates

Anthony Hughes, Todd Wilkens, Barbara M. Wildemuth, Gary Marchionini

Interaction Design Lab, School of Information and Library Science
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599-3360
Phone: (919)962-8366; Fax: (919)962-8071
ahughes@unc.edu; tpodd@email.unc.edu; [wildem, march]@ils.unc.edu

One important user-oriented facet of digital video retrieval research involves how to abstract and display digital video surrogates. This study reports on an investigation of digital video results pages that use textual and visual surrogates. Twelve subjects selected relevant video records from results lists containing titles, descriptions, and three keyframes for ten different search tasks. All subjects were eye-tracked to determine where, when, and how long they looked at text and image surrogates. Participants looked at and fixated on titles and descriptions statistically reliably more than on the images. Most people used the text as an anchor from which to make judgments about the search results and the images as confirmatory evidence for their selections. No differences were found whether the layout presented text or images in left to right order.

## 1 Introduction

Digital video is an active research area on many fronts, ranging from storage and transfer challenges to psychological studies of multichannel information processes. The Open Video Project (www.open-video.org) currently has more than 1800 video segments with a variety of automatically and manually generated metadata elements. The repository is used as a testbed for creating and testing highly interactive user interfaces for digital video (See [1] for an overview of the project and [2] for details regarding one of the previous user studies). These interfaces are crucially dependent on surrogates (metadata that 'stands for' the full object) that provide context and clarity during the retrieval process. These surrogates are displayed via a search results page containing objects that have been retrieved in response to a query. Before downloading a video, users rely on the system's representation of those videos to make relevance judgments. While textual surrogates are often revealing, because they must be read and decoded, they are often unable to characterize the rich amounts and types of data that can be transferred through visual media. It is assumed that utilizing visual surrogates will help users develop a more thorough understanding of the videos retrieved. In order to develop the most useful mix of textual and pictorial

representations of video objects, this study used eye tracking methods to explore the ways in which people interact with these media within the context of a video retrieval system interface.

## 2 Background

Eye tracking has been used to study people's processing of textual and pictorial information in a variety of contexts. Since the late 1880's there has been scientific interest in human eye movements and their ability to reveal cognitive strategies [3]. Of the many different patterns of eye movement, the saccade and point of fixation are most pertinent to a study of human computer interaction [4]. A saccade is a quick movement of the eye for the purposes of redefining the immediate visual field. During a saccadic interval, the eye doesn't collect information but is simply moving from one place to another [5]. A point of fixation is the period between saccades, typically ranging from 200-600ms[1], in which a stable visual point is held and information can be viewed [4,5,6]. An analysis of saccades and fixations can capture the wide range of visual perceptions and, presumably, cognitive processes people undergo when accessing information from a computer generated visual display [4,7].

A review of the literature provides few examples of research regarding people's eye movements as they integrated both textual and visual elements in an information seeking context. As noted by Rayner et al. [8], this may be due to the fact that the bulk of this research has probably been conducted by advertising agencies and the majority of it is not made readily available to the scientific community.

Faraday and Sutcliffe [9] studied the contact points between textual and pictorial elements. They suggested that a contact point "or co-reference between an image and a text" [9, pg. 29] should be carefully crafted to ensure that the maximum amount of encoded information is passed along to the viewer. Their findings revealed that, participants sought to link textual descriptions to visual depictions in a simple manner and if the link wasn't clear, participants often became confused as to how the two channels could be synthesized into a coherent whole. This work was based upon earlier research conducted by Hegarty [10,11]. As her participants viewed instructional diagrams comprised of both visual and textual directions, they constructed representations of the instructional material that were mainly text based. Across analyses, imagery was found to supplement an initial and lasting impression that was decoded from the textual material.

The remaining research was generated by investigations of people's interactions with advertising materials. This work is mainly interested in what pieces of an advertisement are most likely to capture a user's interest. Work by Fox et. al. [12] and Lohse [13] are representative examples. A final study of particular interest was conducted by Rayner et. al. [8]. That work used magazine advertisements as source material for participants conducting searching related tasks. Participants were di-

[1] Due to the fact the fixation points are under voluntary control, fixation time can last as long as two or three seconds, depending on level of interest, but they generally fall within the 200-600 ms range.

rected to, for example, “decide which skin lotion and body wash to buy”, from a series of advertisements that contained both pictorial and textual elements. Results indicated that participants spent a majority of their time looking at the textual portions of the advertisement. While this was true, it was hypothesized that participants spent more time on the text due to the fact that images are more quickly and easily decoded than words. Their work also pointed to the fact that eye movement patterns are heavily affected by task. The current study is intended to address the interplay between text and images.

## 3 Study Methods

Twelve undergraduate students each completed ten search problems. As they browsed the results page for each search, their eye movements were tracked. They also completed a learning preference questionnaire and were debriefed at the completion of the study. The study methods are described in detail in this section.

The 12 subjects who participated in this study were all undergraduate students from UNC-CH. They came from a variety of departments, included 9 females and 3 males, and had a mean age of 20 (ranging from 18 to 24). All of the participants reported using computers on a daily basis and 4 of the 12 reported watching videos or films on a daily basis while the remainder (8) watched weekly. Half the participants reported searching for videos or films on a daily or weekly basis with the other half searching only occasionally or never. Each subject spent about two hours in the study and received $20 for participation.

### 3.1 Learning Preference, Layout Design and Lookzones

Because this study was asking users to view both pictures and text, it was hypothesized that verbal or visual learning preferences might affect results. Kirby, Moore and Schofield’s [14] Verbalizer Visualizer Questionnaire (VVQ) was used to assess these preferences. The VVQ contains three sections of ten questions each dealing with, respectively, verbal themes, visual themes and questions regarding dreams. The questions regarding dreams were dropped due to their lack of relevance to this study. Scores on the verbal and visual scales were analyzed directly and were also used to categorize participants as verbal, visual or balanced. Participants were categorized with verbal or visual preference if the difference between their verbal and visual scores were more than or equal to three. If their scores were less than or equal to two, they were categorized as balanced.

Because this study intended to assess how users interacted with a search results page that included text and pictures, it became clear that the placement of either text or picture on the page might play a large part in how the participant used them. It was hypothesized that, due to the left-to-right reading habits of Western culture, whatever information was placed on the left hand side might be used first. To account for this possibility, two versions of the search results page were designed, one with the textual metadata on the left and the visual metadata on the right and one in

which the placement was reversed. As users interacted with the search results pages, they were given either design one or design two depending on the protocol.

Because this study was also interested in how participants employed either visual or textual metadata, zones were defined for each instantiation of a search page, that delineated three areas-text, pictures and other. Areas that had titles and descriptions were assigned to the text zone, areas that had poster frames were assigned to the picture zone, and areas that included neither textual nor pictorial metadata (i.e. scrollbar, etc.) were defined as an "other" zone.

### 3.2 Study Procedures

The OpenVideo eye tracking study was conducted in the Interaction Design Lab at UNC-CH using an ASL 504 eye tracker and Flock of Birds head tracker. A separate piece of software called GazeTracker was used to correlate eye tracking data with data from the user workstation. A combination of the two was used as the basis for analysis and discussion of results. Based on a review of the literature [3,4,5,6,7], a threshold of 200 ms. was used to define fixations.

Each session included ten trials and in each trial the subjects were given a search stimulus and one search results page. Each page contained between 20-25 video segments. For example, while viewing the appropriate search results page, the participant was asked to find a video "That discusses the destruction earthquakes can do to buildings." The first eight stimuli were topical and the remaining two asked subjects to select a video they might like. Each participant interacted with both page designs and search tasks that were counterbalanced to control for order effects.

After completing the ten searches, participants were debriefed with such questions as: Which of the two designs did you prefer and why? Name two or three strengths of the design you prefer. Name two or three weaknesses. Did the text give you different information than the pictures? If so, could you describe the difference?

In analyzing the data, we first asked how participants utilized text, pictures or other zones as they browsed the results pages. In addition, the effects of the layout, the search task, and the participants' learning preferences were investigated. These results were analyzed with a variety of tools (e.g., GazeTracker, SAS, MS Excel) and methods, including analysis of variance and contingency tables.

## 4 Results

After summarizing participants' use of textual and pictorial zones in the search results page, this section discusses the variables affecting the study results.

### 4.1 People's Use of Textual and Pictorial Representations

The primary question to be addressed in this study was how people use textual and pictorial representations of video objects. A summary of the results addressing this

question is presented in Table 1. It is clear that participants looked at the textual surrogates much longer than the visual surrogates. The mean time that participants spent looking at text, for the study as a whole, averaged more than 22 seconds longer, per search, than time spent looking at pictures—a difference that was statistically significant ($t=14.08$, $p<0.0001$). The differences in mean duration of fixation time over text, pictures or other was found to be statistically significant ($F=170.86$ with 2, 326 df, $p<0.0001$) and Duncan's multiple range test indicated that each mean was different from the other two. Thus, subjects consistently spent much more time decoding the text. The number of fixations appeared to be similarly focused on text, with participants spending 75% of their fixations over text, 18% of their fixations over pictures and less than 6% of their fixations over other sections of the results page. These differences were also statistically significant ($F$ (2, 178)$=181.06$, $p<0.0001$). Interestingly, participants' first fixations were distributed approximately equally between the text (65) and pictures (54).

**Table 1.** Summary of performance for all participants over all search tasks

| | Text | | Pictures | | Other | |
|---|---|---|---|---|---|---|
| | Mean | s.d. | Mean | s.d. | Mean | s.d. |
| Mean time (in seconds) | 29.78 | 16.2 | 6.89 | 6.2 | | |
| Mean # of fixations | 56.44 | 11.4 | 14.52 | 3.9 | 4.7 | 2.4 |
| Mean duration of fixation time | 18.9 | 11.4 | 4.60 | 3.9 | 2.0 | 2.4 |
| Area of first fixation (frequency) | 65 | | 54 | | 1 | |

### 4.2 Effects of Layout, Tasks and Learning Preference

The effects of the placement of the pictures and the text in the interface was evaluated in two ways: amount of time in each zone and the zone which was examined first. The layout did not affect the total fixation time in each region, nor the number of fixations in each region. The mean time per fixation in the text zone was affected by the layout, with slightly more time spent gazing at text when the text was on the left (.34 seconds versus .31 seconds, $p=0.0018$). The zone examined first also varied with the layout (Fisher's exact test, $p=0.0017$). When the pictures were on the left, participants predominately fixated over text first and when the text was on the left, participants predominately fixated over pictures[2] first.

When looking at differences across search tasks, three results were of interest regarding participants' usage of the textual metadata. The total amount of time spent looking at text was affected by the task ($F=5.19$ with 9, 107 df, $p<.0001$; post hoc tests indicated a gradual change in this variable across the ten search tasks). Regardless of

[2] This may simply be due to the fact that participants began each scan on the middle section of the page and as the designs were switched, either text or pictures occupied the middle portion of the page.

counterbalancing plans, some tasks required more time for completion than others. A similar pattern was detected when examining the amount of time spent in fixations in the text zone ($F=3.74$ with 9, 110df, $p<.0004$). Frequency of fixations in the text zone also varied across tasks ($F=4.53$ with 9, 110 df, $p<.0001$). No statistically significant effects of task was found for the time spent in the picture zone, fixation time spent in the picture zone, or number of fixations in the picture zone.

Among these 12 participants, the mean score on the VVQ verbal scale was 5.8 (s.d. = 2.5) and the mean on the visual scale was 8.5 (s.d. = 2.2). When these scores were used to categorize the participants, six were picture-oriented and six were balanced in their preferences; none were text-oriented. Further analysis revealed that the balanced group spent more time looking at text than did the picture-oriented group (Fisher's exact test, $p<.0017$).

## 5 Discussion

It seems clear from the data that participants looked at and fixated on titles and descriptions far more than on the pictures in the displayed results.

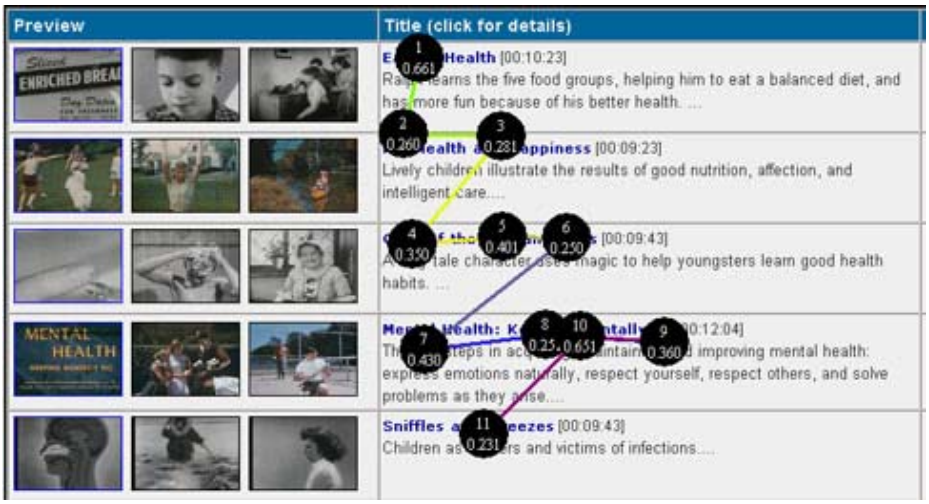


**Fig. 2.** Example Text Scanning Pattern. (black circles are areas of fixation with number of fixation and time of fixation noted)

Even more important, it seemed that most people used the text as an anchor from which to make judgments about the search results. In the analyses of scan paths (see sample in figure 2) and interview responses, it was clear that participants felt most comfortable searching for and finding videos with the textual metadata. While the data is consistent with this interpretation, there are some alternative possible explanations. As they searched, participants did spend more actual time and fixations over the text, but this could be attributable to the fact that information encoded in text takes longer to decode than information encoded in a picture [8]. It could also be

attributed to the fact that a title and description probably carry more topical information about a video than the three images included in this interface. Were people spending less time looking at and fixating on the pictures because the images offered information that was more quickly accessed or because the images offered less information or different kinds of information? Further study is needed to answer this question, but in this preliminary study, it seems clear that participants were thoroughly proficient at using text as their primary tool for selecting relevant videos.

Even so, the visuals were not ignored completely. Some participants spent upwards of 25-30% of their time looking at and fixating on the pictures. The pattern that began to emerge was one of scanning the text for a possible video candidate and then confirming or rejecting that choice with the pictures. See figure 3 for an example screenshot of the moment a participant found the video that they subsequently chose. Note the fixation numbers begin over the text, with 2-6 falling over the description, and then move to the images, where 7-9 are located, and then back to the text.


**Fig. 3.** Example of Text Scanning with Image Confirmation

It is interesting to note that two participants reported using the images as the main access point for their search, but subsequent analysis of their scan path didn't reveal this to be true. Regardless, participants consistently noted that they liked the pictures and felt that they were necessary to their searches. Even though the visuals weren't utilized as much temporally, they were obviously adding some value to the search process. Our study was not designed to reveal what this added value might be, but some exploratory conclusions can be made.

Pictures were consistently used to confirm video choices that were made with the text. The added visual metadata seemed to make people more comfortable with their choices. One participant said pictures give you an "an actual visual object to look at." This makes sense in a medium (i.e., video) where visual information is one of two main signals (sound being the other). People want to "see" what the video looks like, not just read what it's about. It also makes sense that the visuals were communicating certain amounts and types of information that the text couldn't encode, such as the "feel of the film", what the characters looked like, and the colors used in the film. It was also noted by more than one participant that pictures can give you a quick overview of the film without having to read the text. To hazard a contrast between the two sources of metadata, text seemed to transmit information regarding what the video was about and the visuals seemed to transmit information regarding what the video was like.

### 5.1 Discussion of Layout, Tasks and Learning Preference

An assessment of the two layout designs used in this study reveals some interesting similarities and differences. First, the layouts did not radically affect searching patterns. For the most part, participants utilized a personal searching style across both designs. Regardless of where the text was placed it was utilized as the main information source. It is uncertain why subject's average length of fixation was slightly increased when the textual metadata was presented on the left hand side and not the right. It is possible that peoples preference for reading from left to right was more easily accommodated in this situation and therefore they fixated, on average, for slightly longer. It also seems possible that, while a statistically significant difference was found, the actual temporal difference of three one hundredths of a second was simply due to small sample size. As to the finding that participants' first fixations were different for the two layouts, it is uncertain whether this difference is caused by the layout, or simply due to people's tendency to start in the middle of the page. This tendency, alone, would account for the study results, because in one design, text was in the center of the page and in the other, pictures were in the center. It is possible that people weren't being influenced by the layout at all.

In any study that involves search tasks, the ways in which those tasks may have affected performance must be evaluated. The ten different tasks fall into two main categories. In the first eight tasks participants had fairly specific targets (e.g., find a video with nurses), while the final two tasks were much more open ended (e.g., find a video you like). It was hypothesized at the start of the study that the search type might influence visual behavior, but no evidence for such an influence was found. For all ten tasks, people made their initial decisions based on text and confirmed those decisions with pictures. A study of the transcriptions does reveal that most participants reported utilizing the visual metadata more for searches nine and ten. It seems likely that, when presented with a more open choice, people made relevance decisions that included the more "subjective" information presented in the images. Further study regarding the characteristics of search tasks and their effects on searching performance will need to be conducted to verify this preliminary finding.

The statistically-significant effects of individual tasks on performance must also be interpreted. Certainly some tasks took longer than others and we can surmise that those tasks were either harder to accomplish or had some inherent randomness that forced participants to search longer. Why this is true is not easily discerned from the data. Perhaps a larger sample size or a different set of search tasks specifically designed to investigate this issue would produce more interpretable results.

The Verbalizer Visualizer Questionnaire scores revealed an interesting result. The group that had a balanced score, meaning that they had no preference for either visual or verbal learning, spent more time looking at the text than the visual group. The most likely explanation for this result is that the balanced learners had a stronger preference for text than the visual group, and so spent more time with it. It is also possible that balanced learners couldn't decode the information presented in the text as quickly as the visual learners could, but to make this assumption would require more research.

## 6 Conclusion

This study explored the role of text and image representations of the video objects listed on a search results page. A small number of study participants were asked to react to a specific set of search results pages in relation to ten specific search tasks. Thus, the conclusions from the study should be generalized only with caution.

This research began with an overall question regarding the inclusion of visual metadata in a search results page. Would this help or hinder users as they tried to find videos? Will people use images if they are included with other textual metadata while they search for videos? Based upon the evidence presented in this study, the answer seems to be positive. Participants actively utilize visual metadata and include it in the process of making relevance judgments. Text was certainly utilized to a higher degree, but the images were frequently accessed by participants as well, especially to confirm judgments.

Many claims have been made about the value of non-textual cues in supporting video retrieval. This study demonstrated that although text dominates how people make sense of retrieval sets, images add confirmatory value and are strongly liked. In future work, we will continue to investigate how textual surrogates and visual surrogates interact to benefit video retrieval performance. A second iteration of the above study could investigate further variations in the design of the results page by including only text, only images or by comparing manually selected surrogates with randomly selected surrogates. It would also be worthwhile to investigate the amount and variety of information transfer through images and text as they relate to the source video. This work could be further extended through a study of search tasks and their effects on subject performance. It is anticipated that results from this line of research will support the architecture and design of future digital video retrieval interfaces.

## 7 Acknowledgements

This research was supported by grant NSF IIS-0099638 from the National Science Foundation.

## References

1. Wildemuth, B., Marchionini, G., Wilkens, T., Yang, M., Geisler, G., Fowler, B., Hughes, A., & Mu, X. (2002) Alternative surrogates for video objects in a digital library: users' perspectives on their relative usability. In *Proceedings of the European Conference on Digital Libraries*. Rome, September 16-18, 2002. pp. 493-507.
2. Marchionini, G. & Geisler, G. (2002). The Open Video Digital Library. dLib Magazine, 8(12). http://www.dlib.org/dlib/december02/marchionini/12marchionini.html

3. Just, M. & Carpenter A. (1976) Eye Fixations and Cognitive Processes. *Cognitive Psychology,* 8. pp. 441-480.
4. Jacob, R. (1990) What You Look at is What you Get: Eye Movement-Based Interaction Techniques. *ACM Transaction on Information Systems*, 9(3). pp. 152-169.
5. Salvucci, D. & Goldberg, H. (2000). Identifying Fixations and Saccades in Eye-Tracking Protocols. *Proceedings of the Symposium on Eye Tracking Research and Applications 2000.* 2000, Palm Beach Gardens, Florida, US. pp. 71-78.
6. Velichkovsky, B., Dornhoefer, S., Pannash, S. & Unema, P. (2000) Visual Fixations and Level of Attentional Processing. *Proceedings of the Symposium on Eye Tracking Research and Applications 2000.* 2000, Palm Beach Gardens, Florida, US. pp. 79-85.
7. Peiz, J., Canosa, R. & Babcock, J. (2000) Extended Tasks Elicit Complex Eye Movement Patterns. *Proceedings of the Symposium on Eye Tracking Research and Applications 2000.* 2000, Palm Beach Gardens, Florida, US. pp. 37-43.
8. Rayner, K., Rotello, C., Stewart, A., Keir, J., & Duffy, S. (2001) Integrating text and pictorial information: eye movements when looking at print advertisements. *Journal of Experimental Psychology: Applied*, 7(3). pp. 219-226.
9. Faraday, P. & Sutcliffe, A. (1998) Making Contact Points between Text and Images. *Proceedings of the sixth ACM International Conference on Multimedia*. Bristol, UK. pp. 29-37.
10. Hegarty, M. (1992) The Mechanics of Comprehension and Comprehension of Mechanics. In *Eye Movements and Visual Cognition: Scene Perception and Reading*. ed. Rayner, K. New York: Springer-Verlag. pp. 428-448.
11. Hegarty, M. (1992) Mental Animation: Inferring Motion from Static displays of Mechanical Systems. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 18. pp. 1084-1102.
12. Fox, R., Krugman, D., Fletcher, J. & Fischer, P. (1998) Adolescents' Attention to Beer and Cigarette Print Ads and Associated Product Warnings. *Journal of Advertising*, 27. pp. 57-68.
13. Lohse, G. (1997) Consumer Eye Movement Patterns on Yellow Page Advertising. *Journal of Advertising*, 26. pp. 61-73.
14. Kirby, J., Moore, P., & Schofield, N. (1988) Verbal and Visual Learning Styles. *Contemporary Educational Psychology*, 13. pp. 169-184.