

Designing Better Radiology Workstations: Impact of Two User Interfaces on Interpretation Errors and User Satisfaction

Adrian Moise and M. Stella Atkins

This paper presents our solution for supporting radiologists' interpretation of digital images by automating image presentation during sequential interpretation steps. We extended current hanging protocols with support for "stages" which reflect the presentation of digital information required to complete a single step within a complex task. We demonstrated the benefits of staging in a user experiment with 20 lay subjects involved in a comparative visual search for targets, similar to a radiology task of identifying anatomical abnormalities. We designed a task and a set of stimuli that allowed us to simulate the interpretation workflow from a typical radiology scenario—reading a chest radiography exam when a prior study is also available. The simulation was enabled by abstracting both the radiologist's task and the basic workstation navigation functionality. The staged interface was significantly faster than the traditional user interface, provided a 37% reduction in the interpretation errors, and improved user satisfaction.

KEY WORDS: Radiology workstations, user interface, hanging protocols, interpretation errors, user satisfaction

INTRODUCTION

The radiologist has highly repetitive interpretation task, with stringent requirements of accuracy, confidence, and speed. Accessing the controls of current radiology workstations produces considerable disruption of the visual search,¹ which may lead to differences in the volume and type of information processed.

Hanging protocols (HPs) were developed to control the initial display of images for efficiently viewing films.¹⁻³ Film technicians would hang the films, but for soft-copy interpretation on computer monitors, the radiologist must select the hanging protocol. Strickland et al.^{4,5} presented default

modes for the display of images on PACS workstations, and a group from University of California at Los Angeles introduced structured display protocols that model the presentation of data according to the diagnostic task, represented using specific Unified Modelling Language™ diagrams.⁶ By automatically displaying images on multiple monitors, the requirement for users to interactively manipulate images is minimized. Still, the early implementations of HPs on PACS workstations were primitive, lacking user control over series placement, window/level settings, as well as the zoom factor and panning required. Even recent implementations of HPs fail to describe the scrolling interaction, the underlying layout algorithm, the complex screen layouts that cannot be represented as rows by columns, the series linking for synchronized navigation, or the display of flagged images only. Other challenges and limitations associated with current HPs are detailed in the work of Moise.¹ However, the major drawback with current HPs is that HPs provide only the initial presentation of images, which may be used solely to provide a gestalt on the study to be read. Furthermore, they do not handle other clinical information normally

From the Department of Computing Science, Simon Fraser University, 8888 University Drive, Burnaby, BC, V5A 1S6, Canada.

Correspondence to: M. Stella Atkins, Department of Computing Science, Simon Fraser University, 8888 University Drive, Burnaby, BC, V5A 1S6, Canada; e-mail: stella@cs.sfu.ca
Copyright © 2005 by SCAR (Society for Computer Applications in Radiology)

Online publication 25 January 2005
doi: 10.1007/s10278-004-1906-5

reviewed by radiologists, such as prior radiology reports and the patient's history.

This paper presents a solution for supporting radiologists' interpretation of digital images by automating image presentation during temporally sequential interpretation steps. Our solution supports scenario-based interpretation, which groups data temporally, according to the diagnostic process used by the radiologist.⁷ Scenario-based interpretation is especially suitable for complex studies, where screen estate is at premium. Starting from the question that needs to be answered (which was extracted from the diagnosis protocol)—what are the relevant data the radiologist requires for interpretation?—we then addressed how to best present the data on the current hardware platform.⁸ Our solution extends hanging protocols with support for "Stages," where a stage reflects the presentation of digital information required to complete a single step within a complex task. Stages provide context-sensitive navigation, enabling the gathering and filtering of information customized for users within a predefined domain.

We designed a task to test the hypothesis that the Stages approach would be faster, more accurate, and more satisfying to use than a traditional thumbnail-based approach and tested the hypothesis. A "typical" validation of our hypothesis would have required prohibitively expensive resources, including a fully functional radiology workstation as the test bed, radiologists as subjects, and real-life radiological images as stimuli. With the goal of performing inexpensive usability experiments related to radiology workstation design, we designed a new set of stimuli and adapted the experimental task in order to test our hypothesis using a basic workstation and novice users as subjects. We aim to transfer the results from our experiment to the radiological interpretation task. This is possible because we abstracted the radiologist's task and the basic workstation navigation functionality. The goal of the experiments was to analyse the performance of subjects engaged in a radiology look-alike visual search task for artificial targets located in four images under two different interaction techniques.



Fig 1. Screen layout from *Stages*. The images to be displayed are selected by clicking on the icons in the top left. A study with high complexity (many stimuli) is shown. Both images must be viewed to detect a target. The target is in the cross.

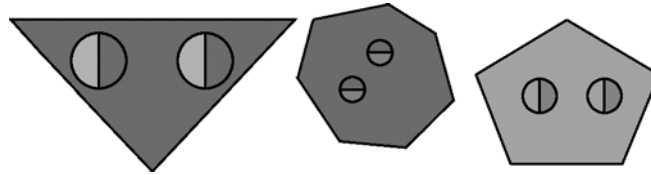


Fig 2. Typical targets: two spherical disks of the same size split in half in the same direction, either vertically or horizontally.

MATERIALS AND METHODS

Task Application and Interaction Techniques

The application provided the simultaneous display of two images that had to be viewed in a comparative visual search task for an abstract target (targets are described in a following subsection). More images had to be viewed in order to describe whether the target grew or shrank in size or appeared/disappeared. The subjects had to interact with the system to see the two images from the first study and then the two images from the second study, and in case a target was present in both studies, a comparison had to be made between one image from the first study and one image from the second study.

The new, scenario-oriented interaction technique is referred to as Stages because it is based on the concept of staging. We refer to the other interaction technique as free user interface (FUI) based on thumbnails of single images, which is found in many current radiology workstations. The screen layout for Stages is shown in Figure 1, where two images are displayed in side-by-side viewports. The only difference between the Stages interface and the FUI is in the four controls at the top left. For the Stages interaction technique, each of the four thumbnail controls corresponds to a predefined pair of images. A single click on one of the controls changes the images in both viewports at the same time. In FUI, each of the four controls holds one thumbnail that is used for the independent selection of the image displayed in each one of the two screen locations. Because four distinct images can be displayed at each of the two screen locations, the user can create a total of 16 screen combinations. For FUI, the four controls correspond to four images to be searched for targets. A two-step interaction is required to change the image in each viewport: first, the user has to select the viewport (either left or right) and then the control corresponding to the image they want displayed in that viewport. Consequently, to change both images on screen, four clicks are required.

Subjects

A group of 20 university students was used as subjects. Each subject performed two sets of 15 trials, one for each

interaction technique. The order of trials for all experiments was randomised. We used a 17-in. Samsung LCD monitor, with a resolution of $1,280 \times 1,024$. The experiment took place in our laboratory, a controlled environment buffered from distractions and noise.

Once the subjects were comfortably seated, instructions about the task were given using several training steps presented on the computer screen. Each training step was followed by a short practice session, where the subjects' understanding of the recently learned concepts was tested. Details are given in the work of Moise.¹ After learning about their task, the subjects were introduced to the application used during the experiment. They had to say aloud if they found a target and point the mouse to it.

Target Description

Target Design

In our experiments, the target is an item with two discs, of the same size, half-split along the same vertical or horizontal diameter, and half-shaded. Three examples of targets are presented in Figure 2.

Images also contained distracters, taking forms such as unequally sized discs or octagonal-sided discs. Identifying the target on a single image was too easy. Therefore, we increased the complexity of the trial by presenting the targets in such a way a subject must discriminate a target from a distracter solely by integrating the information from two related images, as shown in Figure 1. To achieve this, the target was incompletely revealed to the user due to partial occlusion. A similar occlusion occurs in radiology frequently due to anatomical structures shown as bright areas in the image, which overlay the lesion. Such is the case of a barely visible lung nodule hidden behind a rib on a chest CR or a liver tumour hidden behind a blood vessel. The occlusion was simulated in our stimuli with the introduction of a "wild card," which forced our subjects to register information between the two images of a study. A wild card was used to represent the disc divider, an important characteristic feature of a target. The disc divider was occluded by a disc with a uniform fill, which could hide a

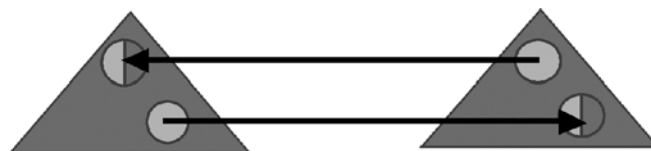


Fig 3. The target is incompletely presented on two different images.



Fig 4. The wild card instantiates in a disc with incorrect divider orientation, so it is not a target.

disc divided either vertically or horizontally. The user must find on a related image the actual instantiation of a wildcard. Depending on the orientation of the occluded disc divider, a wild card could either instantiate into a target, as shown in Figure 3, or into a distracter, as shown in Figure 4. Registration is required for solving the “wild card” into a target. This is called a comparative visual search.

Only the orientation of the divider is important. It does not matter which half of the disc (e.g., top or bottom for a horizontal divider) is grayed out. A third situation, also corresponding to a distracter, occurs when the wild card does not instantiate into a divided disc, as illustrated in Figure 5. Note that a potential target always had a wild card, so, for every potential target containing a wild card, subjects had to register complementary information from the two images of the same study.

Target Complexity

For each trial, the complexity was rated according to the presence of a contour around the target, the target’s contrast compared to the background, and the number of distracters and potential targets. Figure 1 shows an example of low complexity stimuli, with the target present in the cross. Details on how stimuli were rated as low, medium, or high complexity are provided in the work of Moise.¹

Target Evolution

To simulate the radiologist’s follow-up on a radiographic examination, we introduced a time dimension by presenting to our subjects two instances of the same scene, corresponding to different time moments. Hence, we asked our subjects to detect the target from the two images in study 1 and then track the evolution of the target in time. Therefore, each trial consisted of two studies, where each study had two images. The two images of study 1 were presented first, and the two images of study 2 had to be viewed next to detect the evolution in size of any target seen in study 1. An example of stimuli used in the two studies of a trial is presented in Figure 6. Figure 6(a) and (b) shows the first and second images, respectively, from the first study. The target

is free floating in the bottom left of each image. Figure 6(c) and (d) shows the two images from the second study. The target is no longer present in the second study.

Trial Outcome

We used the following notation convention for trial outcome: “0” means no target present in the study and “1” means a target was present. Because each trial consisted of two studies, an outcome of “01” means “no target in the first study, target in the second study.” Hence, in the example trial shown in Figure 6, the outcome is “10” as the target was present in the first study but was not present in the second study. If the target was present in both studies and did not change size, the outcome is represented as “11.” If the target changed size, the outcome is represented by “11c.”

Procedure

Each subject performed two consecutive blocks of 15 trials, one block for each interaction technique. In each trial, a target consisting of equal-sized discs split in half had to be located in the first study set of two images and its evolution noted in a second study set of two images. The same 30 trials were performed by each subject. Inside each block, the order in which the trials were presented to the subjects was randomised. The interpretation accuracy of each trial was assessed by video analysis, and the user satisfaction was recorded with a questionnaire.

RESULTS

Response Time

The interaction technique had a significant effect on response time (in a generalized linear model ANOVA for the 20 lay subjects, $p < 0.001$). The average response time was 17.0 and 19.7 s for Stages and FUI, respectively. Hence, Stages



Fig 5. The two wild cards do not transform into a split disc, so it is a distracter.

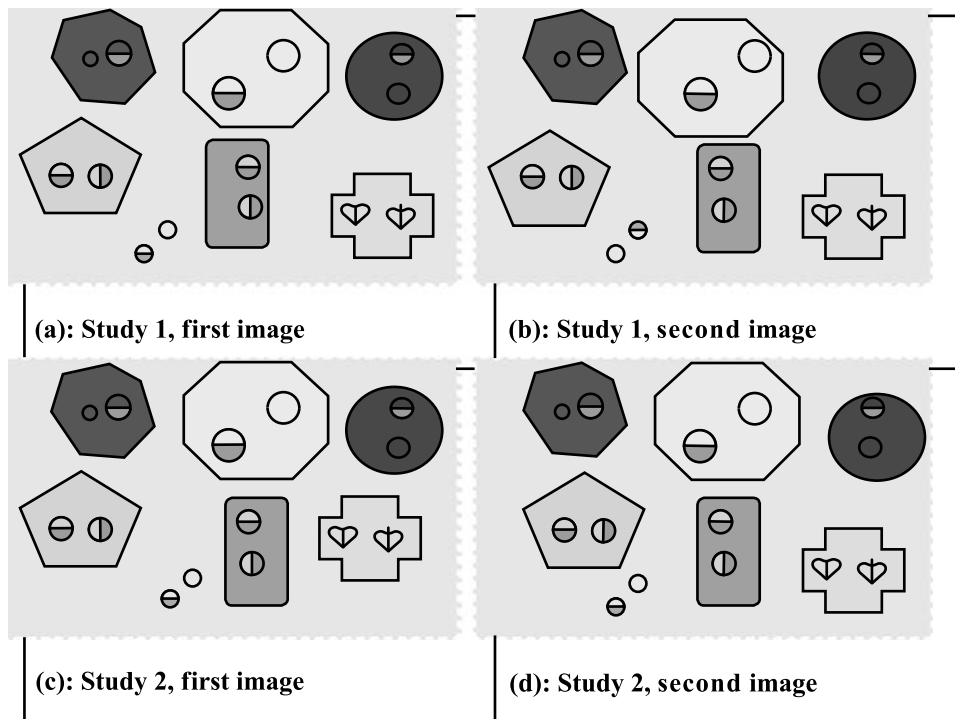


Fig 6. (a) First image of the first study. (b) Second image of the first study; target is in bottom left. (c) First image of the second study. (d) Second image of the second study; no target, as discs are not resolved.

reduced the response time on average by 14%. Details of the response time performance are given in the work of Moise and Atkins.⁹

Interpretation Errors

Our subjects, as their primary task, were instructed to be as accurate as possible in their diagnosis. Completing each trial in the shortest possible time interval was a secondary requirement. Therefore, our hypothesis made no references to the distribution of errors between the two interaction techniques: we traded time for accuracy. Our 20 novice subjects made a total of 27 errors when FUI was used and only 17 errors with “Stages.” There were three types of errors, and their number is presented in the format [“Stages”, FUI]: search errors, such as missing a target or taking a distracter as a target [17, 16]; usability errors, such as making the diagnosis by looking at the wrong pair of images [0, 9]; and evolution errors, which means the target’s evolution in size was incorrectly assessed [0, 2].

We also analysed the errors by the outcome condition for the trial, presented in Figure 7. Recall that the outcome condition 11c stands for the trials where a target was present in both study 1 and study 2, and it changed in size.

Subjective Ratings

We used two user satisfaction questionnaires to record subjective ratings for the two interaction techniques. The first questionnaire, called System Usability Scale, is a standardized usability questionnaire introduced by Digital Equipment Corporation in 1986. Subjects gave a higher usability rating to “Stages” (average score 82) than to FUI (average score 74), which shows a significant difference ($p = 0.03$).

We designed the second questionnaire to include 19 frequently asked usability questions, with answers selected on a scale ranging from 1 (strongly disagree) to 7 (strongly agree). A significant difference ($p = 0.03$) was again recorded for this questionnaire, as subjects again gave higher usability ratings to “Stages” (average score 5.85) than to FUI (average score 5.4).

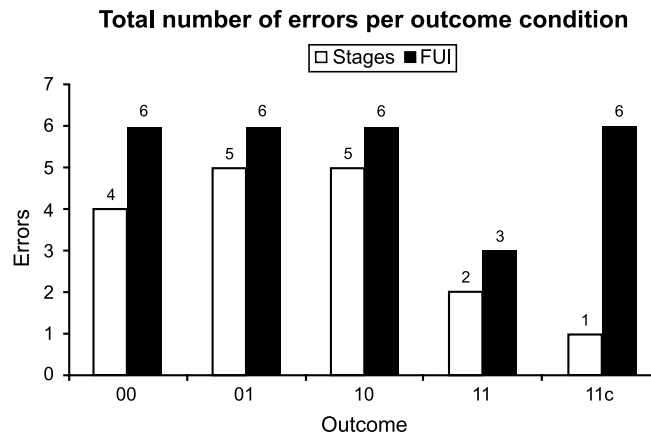


Fig 7. Number of errors per outcome condition.

Subjective Comments

Most subjects preferred to use “Stages” and described it as “straightforward,” “normal interaction,” “easier to use,” “easier to learn and operate,” and “less work.” Free user interface was criticized as “annoying” and “mental and physical workload” due to the extra clicks it required, the extra flexibility not being useful for the given task.

When asked which interaction technique they would prefer if they had to do the same task again, only two of the 20 subjects chose FUI. However, the performance of these two subjects was much better for “Stages,” with fewer interpretation errors and shorter response times.

DISCUSSION

In our experiment, both Stages and FUI gave rise to almost equal number of type 1 errors, mostly in the form of false negative or false positive target recognition. Just two errors occurred due to incorrect interpretation of the target’s evolution in time using FUI. No such error occurred with “Stages.” More interestingly, nine errors occurred using FUI due to a fault-prone interaction technique: the subjects meant to analyse the images from the second study, but they were not examining the appropriate images. Such error is not technically possible when using “Stages.” It is worth noting that this effect occurred despite the fact that each image had a

clear label (1 or 2) depending on which study it belonged to.

According to our results, the most difficult outcome condition occurred when a target, present in both studies, changed its size. We based our belief on the results reflected by dependent measures such as response time and number of errors. Under this outcome condition, the differences between “Stages” and FUI were most visible: the errors were one and six for “Stages” and FUI, respectively. This led us to believe that the benefits of a good interaction technique, such as “Stages,” were more likely to be visible under heavy user cognitive overload. This is the typical situation of a highly repetitive task that has stringent requirements of accuracy and speed. The user thinks he or she learned the sequence of interactions—four point-and-click steps in our case: click on the left viewport, click on PA2 thumbnail, click on the right viewport, and click on LAT2 thumbnail (relies on this sequence to provide the appropriate data for interpretation). This is a major hazard in highly repetitive tasks: even when the information of image label/study acquisition time is clearly displayed on the screen to prevent such errors, this information becomes irrelevant to the subject as they rely on an interaction pattern to produce the correct images on the screen. Bringing the second study for interpretation using FUI is more error prone not only because more steps are involved but also because the order of the steps is critical.

However, our subjects did not have enough practice to become expert users, so the four-step

interaction sequence did not migrate into their “muscle” memory. This is also the case with many radiologists; they do not spend enough time with one workstation to advance from the level of beginner/intermediate user to the level of expert/power user—the reason is that they either read studies in various offices with different software products or because the interaction in new software releases is not consistent with the previous version. There are known situations of radiologists reporting an abnormality on one side when it should have been reported on the other side. The problem is usually caused by the fact that most radiologists expect to see the right side of a patient displayed on the left side of the screen, and they sometimes rely on the software to flip the image accordingly and do not check the patient’s orientation using the markers displayed on the images.

FUTURE WORK

We have started to validate these results with local physicians. Initial comments from medical professionals and initial results from radiologist

fellows suggest that our results from nonexpert users can be transferred successfully to radiological softcopy interpretation tasks.

REFERENCES

1. Moise A: Designing Better User Interfaces for Radiology Workstations. Simon Fraser University, PhD thesis. <ftp://fas.sfu.ca/pub/cs/theses/2003/AdrianMoisePhD.pdf>, 2003
2. Arenson RL, Chakraborty DP, et al: The digital imaging workstation. *Radiology* 176:303–315, 1990
3. Lou SL, Huang HK, et al: Workstation design. Image manipulation, image set handling, and display issues. *Radiol Clin North Am* 34(3):525–544, 1996
4. Strickland NH, Allison DJ: Default display arrangements of images on PACS monitors. *Br J Radiol* 68:252–260, 1995
5. Strickland NH, Allison DJ, et al: Design for the optimal arrangement of magnetic resonance images on PACS monitors. *Proc SPIE* 3031:432–439, 1997
6. Valentino DJ, Wei J, et al: Standardization of hanging protocols using the unified modelling language. *Proc SPIE*, 2002 4319:319–327, 2001
7. Moise A, Atkins MS: Design requirements for radiology workstations. *J Digit Imaging* 17(2):92–99, June 2004
8. Moise A, Atkins MS: Workflow oriented hanging protocols for radiology workstation. *Proc SPIE* 4685:189–199, 2002
9. Moise, A, Atkins MS: Interaction techniques for radiology workstations: impact on users’ productivity. *Proc SPIE* 5371:16–22, 2004