Journal of Digital Imaging

Evaluating Different Radiology Workstation Interaction Techniques with Radiologists and Laypersons

A. Moise,[1] M. S. Atkins,[1] and R. Rohling[2]

This paper presents a new methodology for evaluating radiology workstation interaction features, using lay subjects to perform a radiology look-alike task with artificial stimuli. We validated this methodology by evaluating two different workstation interaction techniques with two groups of subjects: laypersons and radiologists, using a set of artificial targets to simulate the reading of a diagnostic examination. Overall, the results from the two groups of subjects performing the same tasks were very similar. Both groups showed significantly faster response times using a new interaction technique, and the mouse clicks for both groups were very similar, showing that all the subjects mastered the style of interaction in a similar way. The errors made by both groups were comparable. These results show that it is possible to test new workstation interaction features using look-alike radiological tasks and inexperienced laypersons, and that the results do transfer to radiologists performing the same tasks.

KEY WORDS: User-interface evaluation, hanging protocols, radiologist productivity, user study subjects

Evaluating the design of a new radiology workstation interaction technique would conventionally have required a user study with many radiologists performing diagnoses on radiology workstations running new software to support the new interaction technique. We were concerned about the time and cost needed to execute such a study, so we took a different approach. We observed radiologists at work and then carefully designed a task and a set of stimuli that allowed us to simulate interpretation workflow, using a typical task: identifying anatomical abnormalities in a projection radiography chest reading scenario. The look-alike radiology task required identifying an artificial target in two images by performing a comparative visual search. The simulation was made possible by (1) abstracting the radiologist's task and (2) abstracting the basic workstation navigation functionality. We hypothesized that our new, scenario-oriented interaction technique attuned to the radiologist's interpretation task, called Stages (as it is based on the concept of staging[1]), would be faster than the traditional interaction technique using thumbnails, called Free User Interface (FUI), which is found in many current radiology workstations.[2] Stages relies on using stages to extend current hanging protocols for scenario-based interpretation.[3] Our hypothesis was that using workflow-oriented stages streamlines the radiological interpretation task, which leads to shorter completion times, less user–workstation interaction, and fewer errors.[4,5]

We then designed and conducted experiments to compare the new Stages interaction technique with a conventional thumbnail approach (FUI). We wished to perform a user study with laypersons to compare the interaction techniques. Using laypersons begs a question, however: are radiologists somehow better at searching images than other clinicians or laypersons? Because if radiologists are better at searching images than laypersons, then results from an experiment with

[1]*From the School of Computing Science, Simon Fraser University, Burnaby, BC, V5A 1S6, Canada.*

[2]*From the School of Electrical and Mechanical Engineering, The University of British Columbia, Vancouver, BC, V6T 1Z4, Canada.*

Correspondence to: M. S. Atkins, School of Computing Science, Simon Fraser University, Burnaby, BC, V5A 1S6, Canada; e-mail: stella@cs.sfu.ca

Online publication 19 April 2005
doi: 10.1007/s10278-004-2192-y

laypersons performing a visual search may not transfer to radiologists.

Two previous studies have compared radiologists and laypersons searching for hidden targets in complicated picture scenes.[6] One example of such tasks is finding Nina and Waldo in the *Where's Waldo* children's book. These tasks were similar to reading radiographs and searching for lesions because the targets of the search were embedded in complicated backgrounds that also had to be searched and interpreted in order to understand the scene. In that research, the radiologists were reported to spend more time searching images for targets and they fixated on the target much later in the search than laypersons did. The authors of that research conclude that lay subjects performed as well as radiologists in this limited art-testing experience. To support their conclusion that radiologists did not perform better than lay subjects, the authors referred to the theories of Osgood,[7] which relate the degree of transfer on the similarity of training and test situations, and Bass and Chiles,[8] which suggest that performance on perceptual tests has little correlation with diagnostic accuracy in detecting pulmonary nodules.

Inspired by these results we felt confident we could evaluate a radiology workstation interface with laypersons performing a radiology look-alike task, and the results would transfer to the radiologists performing the same task. To validate the new methodology reported in this paper we hypothesized that laypersons and radiologists would have similar performance and interactions.

We used two groups of subjects, novice laypersons and radiologists performing a visual search task for artificial targets to validate this hypothesis. Each subject performed the task using both interaction techniques. We hypothesized that both groups of subjects would make shorter completion times, less computer–user interaction, and fewer errors using workflow-oriented Stages, vs. the traditional thumbnail interface.

In the next section we describe our methods: the subjects, the task, and the interaction techniques. We then present the results for the response times, the mouse clicks, and errors. The section following has a discussion of the results. In the conclusion, we present a summary and plans for future work.

METHODS

Experimental Conditions and Subjects

For both interaction techniques, in each trial, the subjects were asked to find an abstract target on a gray background in the first study set of two images, and its evolution noted in a second study set of two images. The targets we designed are described in the next subsection, and the interaction techniques in the subsection following. Two groups of subjects performed the same experiments.* In the first experiment, a group of 20 university students were used as lay subjects. For the second experiment, a group of four radiology residents were used as subjects. Each subject performed two consecutive blocks of 15 trials, one for each interaction technique. The same 30 trials and the experimental design of these two experiments were identical, except for the fact that the order of trials for all experiments was randomized for the first experiment with laypersons, whereas the order of trials was identical for all four radiologist subjects, as this allowed for easier interpretation of the results. We counterbalanced for the interaction techniques, so that half the subjects in each group started with FUI and half started with Stages. We used a simplified version of a radiology workstation and a set of artificial targets to simulate the reading of a diagnostic examination, illustrated in Figure 1 for the Stages interface. However, the hardware and environment differed between the two experiments. For the first experiment with the 20 laypersons, we used a 17-in. Samsung LCD monitor, with a resolution of 1,280 × 1,024.[4,5] The experiment took place in our laboratory, a controlled environment buffered from distractions and noise. For the experiment with the four radiologists we had to use a room with a significant level of nearby traffic, and with a noticeable background noise. Although these conditions were not ideal, it did not seem to bother the radiologists—we assume they are quite used to focusing and working in similar conditions. The radiologists used a Dell Inspiron 8100 notebook with a 15-in. screen (1,400 × 1,050 native resolution) with the display resolution set at 1,280 × 1,024.**

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 Ref. 1. After learning about their task, the subjects were introduced to the application used during the experiment. Subjects were instructed to make a verbal statement, "target" or "no target" at the end of each trial, with the location of the target (if present) under the mouse cursor when the target was found. We videotaped and collected comparative information while subjects were

* This research was approved by the Simon Fraser University Office of Research Ethics and was conducted in 2003.

** Unfortunately, we could not use the identical monitor for both experimental groups, as a portable laptop was required for the radiologists, but the resolution was the same, and the gray-level contrast was similar for both experiments.

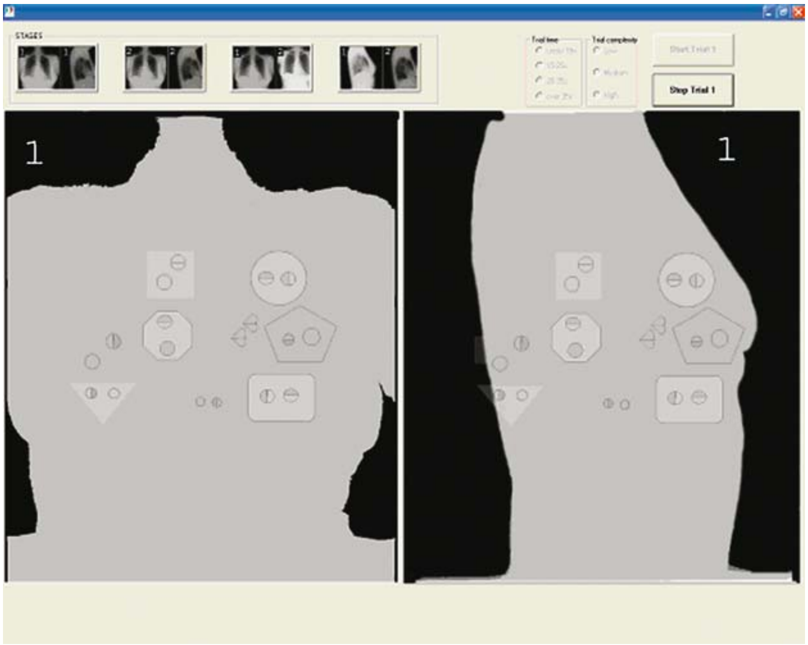


Fig 1. Screen layout from Stages. The images to be displayed are selected by clicking on the icons in the top left. A study with high complexity (many stimuli) is shown. Both images must be viewed to detect a target. The target (explained below) is in the heptagon at the bottom centre.

performing the task; the videotape and audio recording were used to identify if the correct target had been located.

Target Description

Target Design

The target is an item with two discs, of the same size, half-split along the same vertical or horizontal diameter, and half-shaded, as shown in Figure 2. Images may also contain distractors, taking forms such as unequally sized discs, hearts, or octagonal-sided discs. Identifying the target on a single image is too easy, so we increased the difficulty of the trials by allowing discrimination of a target from a distractor only by integrating the information from two related images displayed on a right and left viewport; that is, the target is incompletely revealed to the user due to partial occlusion.* The occlusion is simulated in our stimuli with the introduction of a “wild card,” which forces our subjects to register information between the two images of a study, in a comparative visual search. A wild card is used to represent the disc divider, an important characteristic feature of a target. A disc with a uniform fill can hide a disc divided either vertically or horizontally. The user must find on a related image the actual instantiation of a wild card. Depending on the orientation of the occluded disc divider, a wild card could either instantiate into a target, or into a distracter, as shown in Figures 3 and 4, respectively.

Registration is required for resolving the “wild card” into a target, a disc with the proper divider orientation. Therefore, for instantiation of a target, the discs on one image must be compared with the discs on the other image. Only the orientation of the divider is important. It does not matter which half of the disc is grayed-out. A third situation, also corresponding to a distracter, occurs when the wild card does not instantiate into a divided disc, illustrated in Figure 5. Note a potential target always has a wild card, so, for every potential target containing a wild card, complementary information from the two images of the same study must be viewed.

Target Evolution

To simulate the radiologist’s follow-up on a radiographic examination, we introduced a time dimension by presenting two instances, study 1 and study 2, of the same scene, corresponding to different time moments. 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. 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 image, respectively, from study 1. The target is free floating in the bottom left of each image. Figure 6(c) and (d) shows the two

* 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 tumor hidden behind a blood vessel.

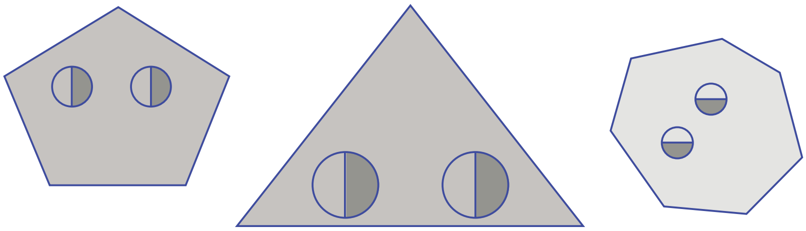


Fig 2. Typical targets: two spherical disks of the same size split in half in the same direction, either vertically or horizontally.

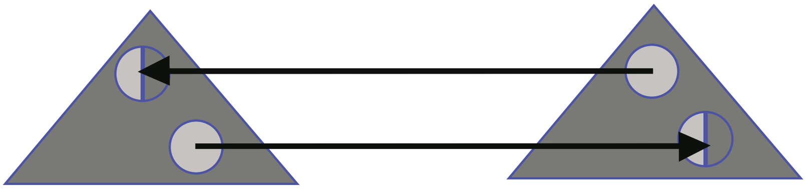


Fig 3. The target, with the same divider orientation, is incompletely presented on two different images.

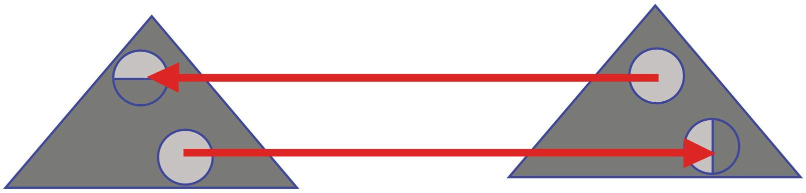


Fig 4. The wild card instantiates in a disc with incorrect divider orientation, so it is not a target.

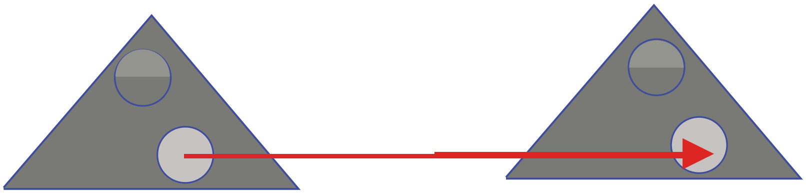


Fig 5. The two wild cards do not transform into a split disc, so it is a distracter.

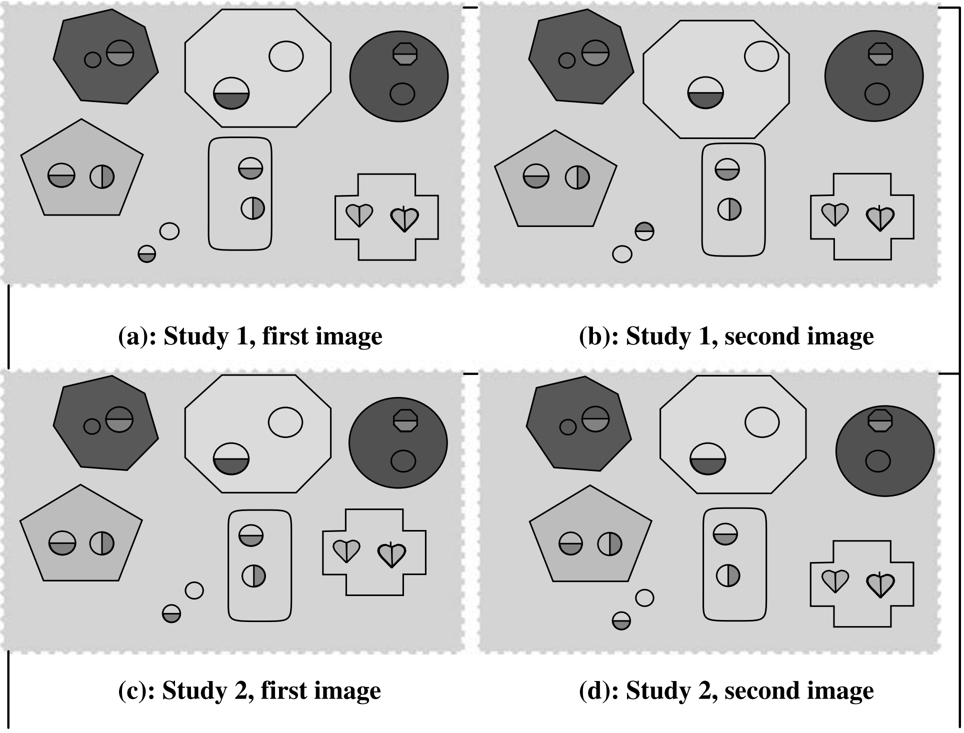
images from study 2. The target is no longer present in the second study.

Trial Complexity

The trials were designed to be of low, medium, and high complexity. 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 in the stimuli. Figure 1 shows an example of high-complexity stimuli, with the target present in the heptagon. Each image includes four potential targets and five distracters. Details on how stimuli were rated as low, medium, or high complexity are provided in Ref. 1.

Trial Outcome

We used the following notation convention for trial outcome: "0" means no target present in the study, "1" means a target was present. As each trial consisted of two studies,



(a): Study 1, first image

(b): Study 1, second image

(c): Study 2, first image

(d): Study 2, second image

Fig 6. (a,b) Two images of study 1; target is in bottom left. (c,d) Two images of study 2; no target, as discs are not resolved.

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.

Interaction Techniques

For both interaction techniques, the layout of the screen consisted of a left viewport, a right viewport, the controls used for image selection, and the controls used to start/stop the current trial, as illustrated in Figure 1 for Stages and in Figure 7 for FUI. Note the only difference between the two interfaces was in the functionality of the workstation controls at the top-left.

Only two images could be displayed simultaneously, so subjects had to interact with the system to see first the two images from the first study, and then the two images from the second study. In cases where 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 selection of the two images to be displayed on screen was done using the four thumbnail-size controls located in the top-left side of the screen. For the Stages interaction technique, each of the four controls corresponded to a predefined pair of images. A single click on one of the controls resulted in changing the images from both viewports at the same time. For example, by clicking on the leftmost thumbnail, the two images corresponding to the first study would be displayed on screen. For the FUI interaction technique, one toolbar containing the image thumbnails was used for the independent selection of the image displayed in each one of the two screen locations. As four distinct images could be displayed at each of the two screen locations, the user could create a total of 16 screen combinations. The four controls corresponded to the four images to be searched for targets. A two-step interaction was required to change the image in each viewport: first the user had to select the viewport (either left or right), and then the control corresponding to the image to be displayed in that viewport. Consequently, to change both images on screen, four clicks were required.*

Independent Variables

Our experiment had three independent variables: the interaction technique, the trial outcome, and the trial complexity. There were two conditions for the interaction technique: Stages and FUI. The following five trial outcome conditions were possible: "00," "01," "10," "11 same," "11 diff," where "11 same" meant that the target was the same size between study 1 and study 2, and "11 diff" meant that the target changed size between study 1 and study 2. In terms of the trial complexity, our

* The application remembered the last viewport selected on the screen. One could save the click used to reselect the last viewport. As there always was a selected viewport, one could change both images on the screen in only three clicks.

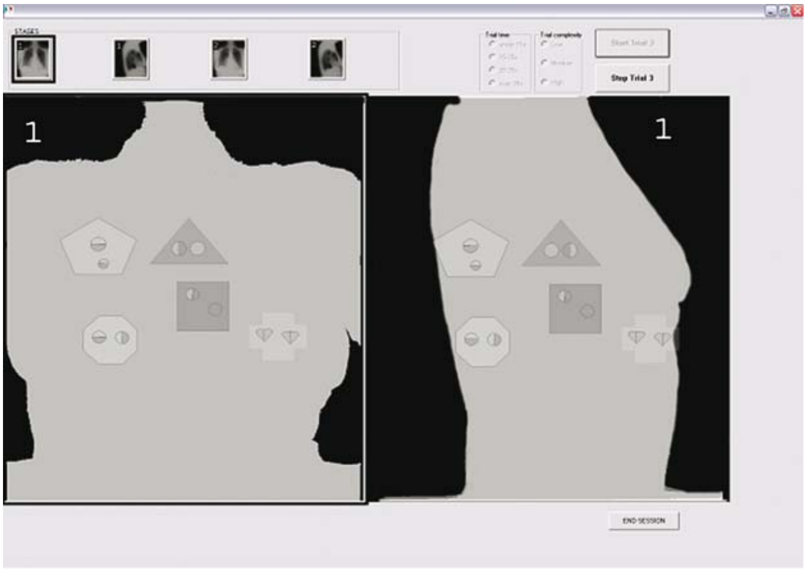


Fig 7. Screen layout from *FUI*. A study with low complexity is shown. The target is in the triangle.

trials had low-, medium-, or high-complexity stimuli images. Consequently, our experimental design had 2 × 5 × 3 = 30 distinct combinations. We associated one trial per combination (where a trial is a set of two studies, where each study has two images), for a total of 30 trials. The stimuli for the 30 trials were grouped into two disjoint blocks of 15 trials, one for each interaction technique.

Dependent Measures

Response Time

Individual trial completion times were recorded in a user-specific log file. The response time was measured from the moment the subject clicked on “Start trial” to have the stimuli displayed, until the subject clicked on the “Stop trial” and the stimuli were hidden.

Mouse Clicks per Trial

Used to measure the number of user interaction steps required for the interpretation of each trial.

Interpretation Errors

The interpretation accuracy of each trial was assessed by video and audio analysis. The video footage captured both the screen content and the subject’s verbal interpretation. Our subjects were instructed to be as accurate as possible, so a correct diagnosis was their primary task. Completing each trial in the shortest possible time interval was a secondary requirement.

RESULTS

Results are presented in three sections, for response times (RT), mouse clicks, and errors. The average RT is presented for both interaction techniques for both experimental groups, then the average RT is further subdivided for different trial complexity and trial outcomes, as well as the ordered RT for both interaction techniques.

Response Time

Response Time per Interaction Technique

Table 1 shows the average response times for both interaction techniques for both experimental groups.

We used the Normal Q–Q Plot to determine whether the distribution of the RT variable matches the normal distribution. The RT distribution did

Table 1. Response time (seconds) per trial for both interaction techniques (IT) for both experimental groups

Experimental group	IT	Average response time (SD) per trial (s)
Second user experiment—4 radiologists	Stages	15.3 (5.2)
	FUI	16.8 (4.5)
First user experiment—20 laypersons	Stages	17.0 (6.4)
	FUI	19.7 (6.6)

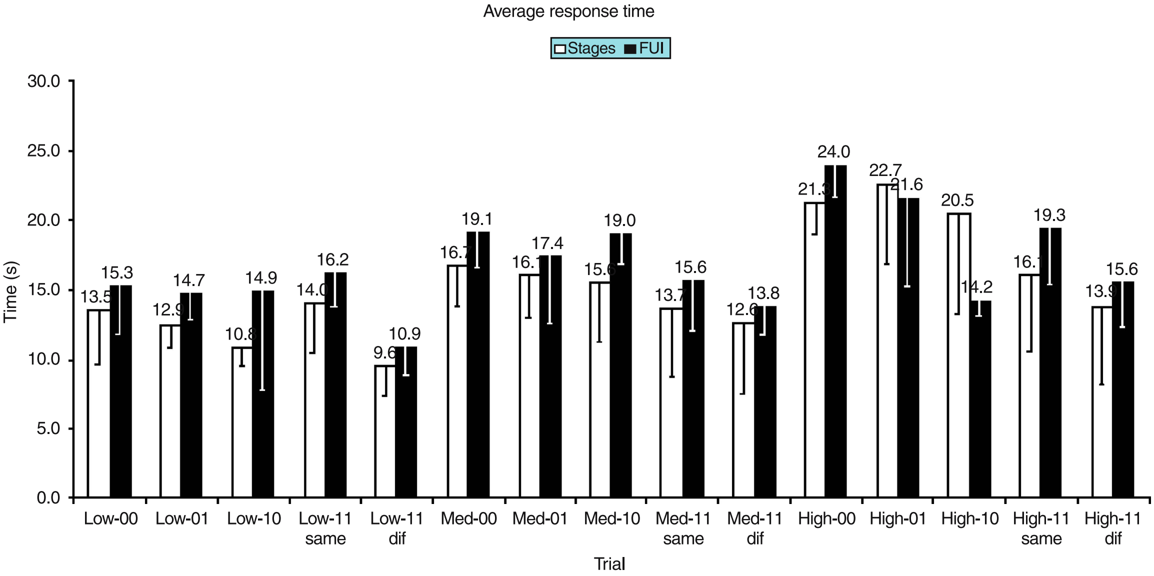


Fig 8. Average response time (seconds) for both interaction techniques for each outcome and trial complexity, for the four radiologists.

closely match the normal distribution for both experiments, as the points of the Q–Q plot clustered around a straight line, so we used the RT values for significance tests. A General Linear Model (ANOVA) on the RT showed a significant ($p = 0.04$) 9% reduction in the response time for Stages compared with FUI for the four radiologists, and a significant ($p < 0.001$) 14% reduction in the response time for Stages compared with FUI for the 20 laypersons.

Response Time per Trial

A breakdown of the average response time for each trial, by outcome and trial complexity, for the four radiologists is shown in Figure 8. Error bars on all the figures represent 1 SD. Note the complexity of the trials increases with increasing trial number: trials 1–5 correspond to low-complexity stimuli, trials 6–10 correspond to medium-complexity stimuli, and trials 11–15 correspond to high-complexity stimuli.*

Response Time vs. Trial Complexity

The average response time vs. trial complexity for the four radiologists for both interaction techniques is shown in Figure 9. Complexity has a significant effect ($p < 0.001$) on RT for the radiologists. Figure 10 shows the average response time vs. trial complexity for laypersons; however, complexity does not have a significant effect ($p = 0.319$) on RT for the laypersons.

Response Time vs. Trial Outcome

The response time vs. trial outcome for the four radiologists is shown in Figure 11. Trial Outcome has a significant effect ($p < 0.001$) on response time.

Figure 12 shows the effect of trial outcome on the response time for laypersons. The trial's outcome had also a significant effect ($p < 0.001$) on RT for the 20 laypersons.

Differential Learning Effects on Response Time

Figure 13 shows the response time vs. the interaction technique in order, for the radiologists. The two radiologists who performed Stages first

* Note that a comparable plot for the 20 laypersons was not possible as their trial ordering was randomized and less readily obtained.

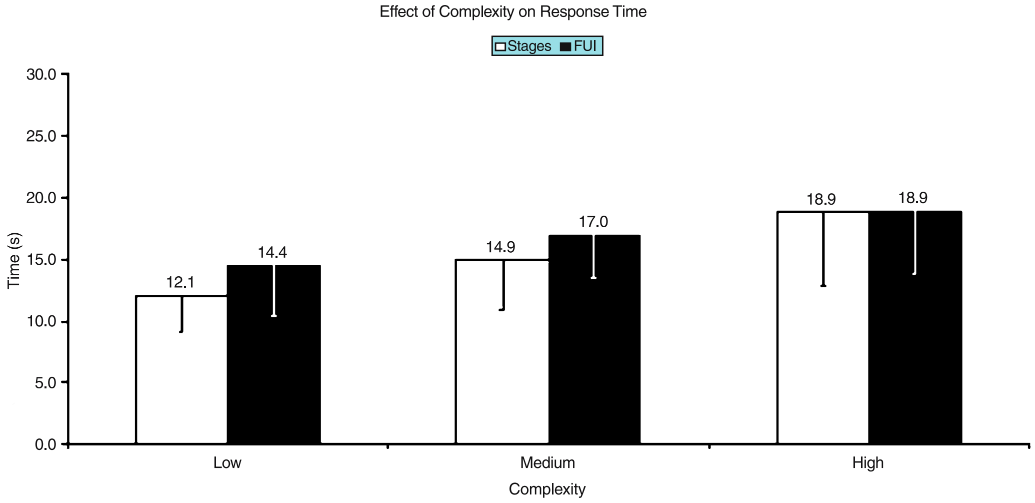


Fig 9. Average response time (seconds) vs. trial complexity for both interaction techniques, for the four radiologists. Response time increases significantly with increasing complexity ($p < 0.001$).

took an average of 17.6 s/trial while they were learning and speeded up slightly, to 17.0 s/trial for the FUI later. The other two radiologists who performed FUI first took 16.5 s/trial and speeded up to 13.0 s/trial later. This shows a significant differential learning effect for radiologists (p = 0.003) in that the subgroup starting with FUI had a steeper learning curve than those starting with Stages.

Similarly, Figure 14 shows the response time vs. the interaction technique in order, for the laypersons. This figure shows a similar significant learning effect ($p < 0.001$) in that the 10 laypersons who performed Stages first took an average of 19.8 s/trial and took 20.1 s/trial with FUI later, whereas the other 10 laypersons who performed FUI first took 19.3 s/trial and speeded up to 14.1 s/trial later.

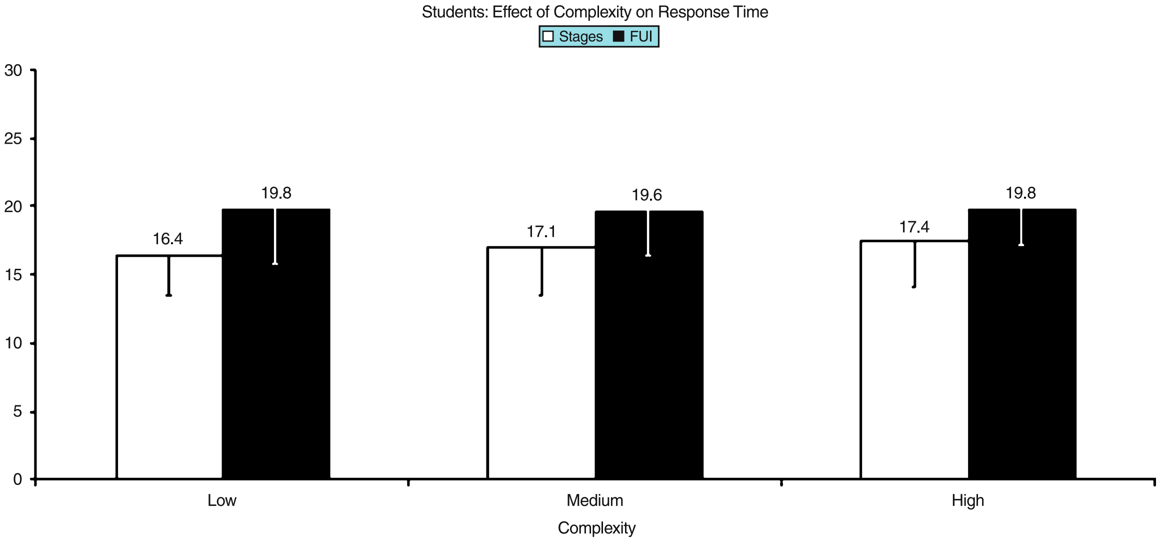


Fig 10. Average response time (seconds) vs. trial complexity for both interaction techniques, for the 20 laypersons. Response time does not increase significantly with increasing complexity (p = 0.319).

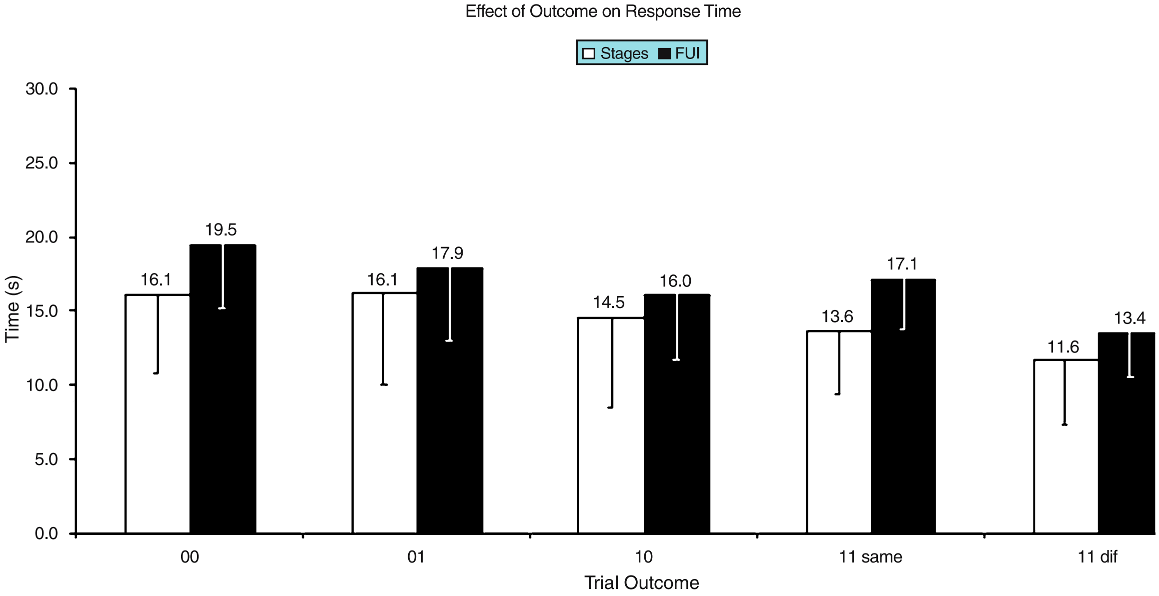


Fig 11. Average response time (seconds) vs. trial outcome for both interaction techniques, for the four radiologists. Response time depends significantly on the trial outcome (*p* < 0.001).

Mouse Clicks

Table 2 shows the average number of mouse clicks to complete a block of 15 trials for both interaction techniques for both experimental groups. As expected, the interaction technique had a significant effect ($p < 0.001$) on the number of clicks required, for both experimental groups.

There is no significant effect of Trial Outcome correlating mouse clicks and response time, although most mouse clicks were recorded for the outcome (11 same) where subjects performed

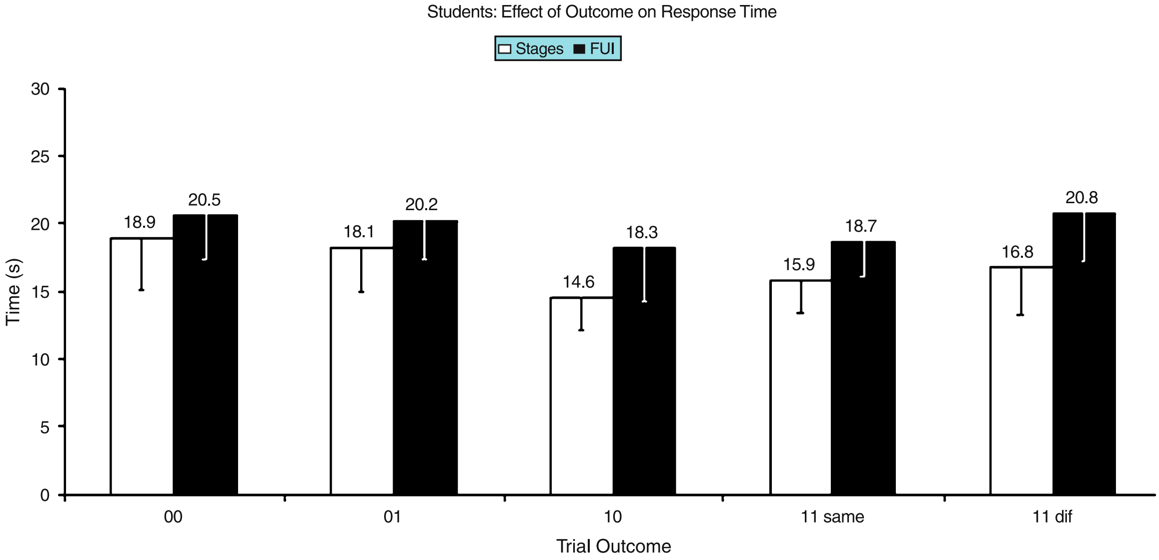


Fig 12. Average response time (seconds) vs. trial outcome for both interaction techniques, for the 20 laypersons. Response time depends significantly on the trial outcome (*p* < 0.001).

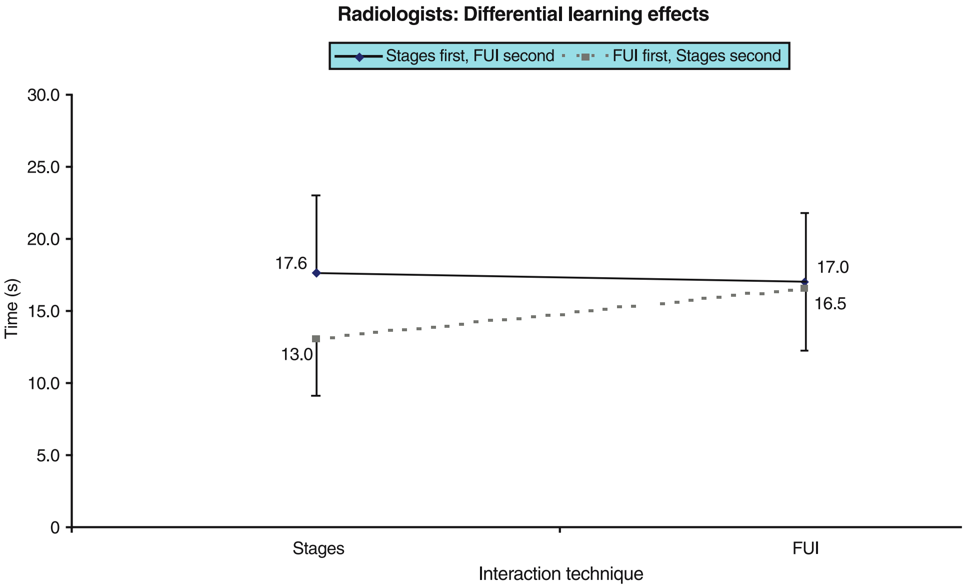


Fig 13. Response time (seconds) vs. Interaction Technique for learning both interaction techniques, for the four radiologists. Upper line corresponds to the group that performed Stages first; lower line corresponds to the group that performed FUI first. The two subgroups show a significant differential learning effect, with a much faster response time for the group switching to Stages after starting with FUI.

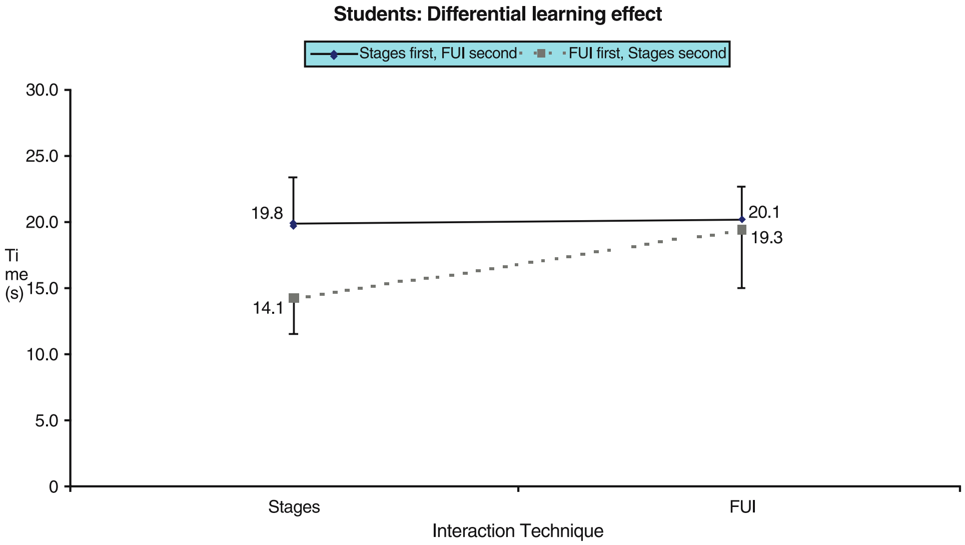


Fig 14. Response time (seconds) vs. Interaction Technique for learning both interaction techniques, for the laypersons. Upper line corresponds to the group that performed Stages first; lower line corresponds to the group that performed FUI first. The two subgroups show a significant differential learning effect, with a much faster response time for the group switching to Stages after starting with FUI.

additional clicks to check for slight changes in the target's size.

Interpretation Errors

Table 3 shows the number of errors made vs. trial complexity, for both groups and for both interaction techniques. Laypersons' errors were broadly analyzed in Ref. 5. Detailed error analysis of the four radiologists' errors follows.

Errors vs. Trial Outcome for the Four Radiologists

For FUI, the seven errors were distributed as: 3 in outcome 10, 1 in outcome 11 same, and 3 in 11 diff.

For Stages, the six errors were distributed as: 1 in outcome 00, 2 in outcome 11 same, and 3 in 11diff.

Types of Errors for the Four Radiologists

We have defined four types of errors, and the number of errors is presented in the format [Stages, FUI] beside each type of error. Because of difficulties with the extensive video/audio analysis, this comparable analysis was not performed for the laypersons.

- Search errors in study 1, such as missing a target (a false negative) or taking a distracter as a target (a false positive) [2, 1]
- Search errors in study 2, such as missing a target (a false negative) or taking a distracter as a target (a false positive) [4, 4]. We chose to separate type 1 and type 2 errors due to effects such as satisfaction of search, and also because when a target was found in the first study, it acted as a hint on where to search for in the second study.
- Usability errors, such as making the diagnosis by looking at the wrong pair of images [0, 0]
- Evolution errors, meaning the target's evolution in size was incorrectly assessed [0, 2]

Table 2. Mouse clicks per trial for both interaction techniques (IT), for both groups

Experimental group	IT	Average number (SD) of mouse clicks/trial
Second user experiment—4 radiologists	Stages	2.3 (0.3)
	FUI	6.3 (0.9)
First user experiment—20 laypersons	Stages	2.3 (0.3)
	FUI	6.3 (0.5)

Table 3. Interpretation errors versus trial complexity for both interaction techniques (IT), for both groups

Experimental group	IT	Total errors	Low complexity	Medium complexity	High complexity
Second user experiment—4 radiologists	Stages	6	0	5	1
	FUI	7	2	4	1
First user experiment—20 laypersons	Stages	17			
	FUI	40			

Subject 4 made three of the six errors with Stages, all errors in the second study of the trial:

- (Trial Outcome 00, high complexity): the subject spends a great deal of time on this study and ends up pointing to a false target in study 2 (we label this a decision-making error);
- (Trial Outcome 01, high complexity): based on the mouse-pointing activity, we label this a pattern-recognition error;
- (Trial Outcome 01, low complexity): probably a pattern-recognition error.

DISCUSSION

Response Time

Table 1 shows that both the radiologists and the laypersons were significantly faster using Stages rather than the FUI. The decrease in RT between FUI and Stages was 9% for the radiologists and 14% for the laypersons. We believe the radiologists were more familiar with the idea of a comparative visual search, so they were less affected than the laypersons by the interaction technique they had to use. Also, the radiologists were on average a little faster than the laypersons, which can be explained by the fact that radiologists are more familiar with visual search tasks.

Figure 8 shows that for the four radiologists, Stages produced on average faster response times than FUI for 13 out of the 15 trials. The exceptions were trials 12 and 13. In trial 12, a high-complexity image with no target in either study, FUI is just slightly faster than Stages, mainly because subject 1 spent an excessively long time—31.2 s on this trial using Stages, whereas for the equivalent trial another subject spent only 15.6 s when using FUI. In trial 13, also corresponding to high-complexity stimuli, with the target only present in study 2, FUI is much faster than Stages. The reason is that subject 1 again spent a long time, 26.5 s, to complete this trial with Stages. However, trial 13 was the first overall trial for subject 1 and the slow response time was likely due to the learning effect (see below under the subsection "Differential Learning Effect on Response Time"). Subject 3 also spent a long time on trial 13 using Stages; the subject initially misdiagnosed the second study, but then reviewed study 1 and study 2 again in order to produce the correct diagnosis, generating a total of eight mouse clicks instead of the usual two clicks (see also the discussion under the subsection "Mouse Clicks").

We believe the faster response times using Stages are due mainly to the fact that the cognitive workload is reduced in Stages compared with the FUI. With FUI, some of the target information is "flushed" from the short-term memory in order to "load" the four point-and-click steps required in FUI.

Response Time vs. Trial Complexity

Figure 9 summarizes the RT vs. complexity of the stimuli images for the four radiologists, and Figure 10 summarizes the same data for the 20 laypersons. As expected, for the radiologists, the RT steadily increases with increasing complexity of the images, and the effect is significant ($p < 0.001$). For laypersons, the effect of RT with increasing image complexity is not significant ($p < 0.319$), although a trend is noticeable. This may be because the laypersons were less familiar with the visual search problem, so they took longer on even the simpler, less complex images.

Comparing the RT for Stages and FUI against image complexity, we note for the radiologists, for low- and medium-complexity images, the RT is significantly faster for the Stages interaction technique rather than the FUI technique, whereas for the high-complexity images, the RT is similar for both techniques. This may be explained by the fact that for high-complexity images, the time to perform the visual search overwhelms the effect of the interaction technique.

Response Time vs. Trial Outcome

Figure 11 shows the RT for the radiologists against the trial outcome for both interaction techniques, and Figure 12 shows the same data for the laypersons. Both groups show a significant effect of trial outcome on RT, where the longest trials had an outcome of "00," meaning no target was present in either study. This occurs because users had to perform a full search on both study 1 and study 2. Also, for both experimental groups, outcomes where a target was present in the first study were the fastest to diagnose, because an exhaustive search on study 2 was no longer required. For the radiologists, the response time was the smallest when both study 1 and study 2 had a target, and the target changed size (11 diff). Figure 12 shows slightly different effects for the laypersons, who spent longer trying to assess the size changes in trials with outcomes of "11." This timing difference may have arisen because the radiologists were trained to keep their eyes on abnormalities in images and detect size changes between studies of the same patient. In these trials, the change in target size in the stimuli was quite obvious—in study 2, the target was 1.5 times bigger (or smaller) than in study 1, and many of the radiologists observed the change in size of the target without having to review the images of study 1. However, when the target, present in both studies, kept the same size (11 same), the radiologists occasionally performed additional comparisons to check for subtle changes in target size, which explains the slight increase in the response times. In contrast, the laypersons nearly always returned to view the images of study 1 after detecting a "11" outcome; they had not been trained to look for changes in size of targets.

Differential Learning Effect on Response Time

Each subject performed two sessions (two blocks of 15 trials): one session with Stages, and the other one with FUI. In the first session each subject became familiar with the main task—performing a visual search for targets in two studies. Figure 13 shows significant differential learning effects for radiologists, and the same significant effects are seen in Figure 14 for laypersons. For both experimental groups, the subgroup that started with FUI showed a big speed increase in going to Stages (i.e., a lower RT). For example, the two radiologists who started block 1 trials with FUI (right side of the graph) performed with an average RT of 16.5 s. In their second block with Stages, these radiologists had a drop of 3.5 s in their average response time per trial. These results show that the subjects starting with FUI enjoyed a learning effect and speeded up considerably when switching to Stages. However, those who started with Stages took almost the same RT when they went to FUI for their second block of trials.

We believe the knowledge about the main task acquired in the first session transferred to the second session, but to a different degree, depending on which interaction technique was used first. Our interpretation of this result is that the extra difficulty brought to the task by FUI offsets the RT saving produced by learning; it appears that the use of Stages is more important than the learning effect for reducing RT. In fact, for laypersons, the RT actually increases as they make the transition from Stages to FUI (although it is not significant). This means the learning effect is obscured by the fact that FUI increases the difficulty for the subject.

Mouse Clicks

Table 2 shows that both groups employed a similar number of mouse clicks, with significantly more mouse clicks for the FUI than for the Stages interaction technique. Furthermore, as already noted, both groups showed significantly faster response times using the Stages interaction technique than using the FUI and had similar learning effects. These results indicate that all the subjects mastered the style of interaction in a similar way.

However, the extra time with FUI is far longer than the time needed to perform the extra few mouse clicks. Therefore, we believe the faster response times using Stages are not only due to the fact that there are fewer mouse clicks per trial, but also are due to the reduced cognitive workload in Stages, and to the fact some of the information was "flushed" from the short-term memory in order to "load" the four point-and-click steps required in FUI.

There is no significant effect of trial outcome correlating mouse clicks and response time; most mouse clicks were recorded for the outcome (11 same) where subjects performed additional clicks to check for slight changes in the target's size. As already noted, the most time-consuming trials were recorded for outcome (00) where subjects had to perform a complete search for targets in both study 1 and study 2, but this case had fewer mouse clicks than for outcome (11 same).

We hypothesized that the improvement in time achieved with our new interaction technique would be more than just the time needed to perform the extra clicks with the traditional interaction technique; these results validate this hypothesis. We conclude that extended visual search has a higher impact on the response time than the extra mouse clicks.

Interpretation Errors

Our hypothesis made no references to the distribution of errors between the two interaction techniques: we traded time for accuracy. However, both experimental groups of subjects generated more interpretation errors with FUI than with Stages, as seen in Table 3. Overall, the radiologists' accuracy is more consistent with the two interaction techniques, and the layperson's error rate is more influenced by the interaction technique, showing their lack of experience. However, the radiologists made slightly more errors with Stages than recorded by the laypersons. We believe this slightly higher number of errors is due to the different, less-than-perfect (background noise, traffic, and slightly smaller screen size) experimental conditions for the radiologists. Errors per trial complexity showed that the radiologists made fewer errors in the highly complex trials, probably because the image complexity forced them to view these images for a longer time, and they were therefore more careful in these situations. Errors per trial outcome showed that the radiologists made most errors (total 9) when there was a target in both studies (outcome 11).

This implies one target was missed in seven cases, as there were only two evolution errors (when the target's evolution in size was incorrectly assessed). These results likely confirm the findings from earlier research which identified three causes for false-negative errors: faulty visual search, faulty pattern recognition, and faulty decision making.[9–11] Eye-gaze data may be able to reveal why so many false-negative calls were made.

Implications of Results

When compared with our user-centered Stages interaction technique, the traditional thumbnail interaction technique not only adds a few extra clicks to the interaction, but also reduces productivity by causing disruptions of the scenario analysis, particularly disruptions of visual search. Using Stages not only has the potential of saving the time the user spends interacting with the workstation, but it can also alleviate the radiologist's workload and cognitive overhead caused by the absence of a skilled technician to handle the initial presentation of images prior to review by the radiologist, and by workstation manipulation tasks and workstation constraints not found on film. Consequently, we believe Stages is a better interaction technique, as it allows the users to optimize their workflow and to increase their productivity as they become accustomed with the main task: the visual search for targets. The main limitation is that these results are obtained from only four radiologists and may not extend to larger numbers.

CONCLUSIONS

Summary

We presented a hypothesis that it is possible to design radiology look-alike tasks to test new workstation design features on stripped-down workstations using inexperienced laypersons, rather than conducting more costly user studies involving radiologists. We validated this hypothesis by evaluating two different interaction techniques with two groups of subjects: laypersons and radiologists. We hypothesized that laypersons and radiologists would have similar performance and interactions in that using workflow-oriented Stages would streamline the radiologic interpretation task, which would lead to shorter completion times, less user interaction, and fewer errors.

Overall, the results from the two groups were very similar. Both groups show benefits in using the new "Stages" interaction technique over the traditional thumbnail-based interface. These results show that it is possible to design radiology look-alike tasks to test new workstation design features on stripped-down workstations using inexperienced laypersons, rather than conducting more costly user studies involving radiologists. The results for laypersons do transfer to radiologists, implying our experimental design using a set of high-quality stimuli and a carefully designed look-alike interpretation task did abstract the radiologist's task well.

To achieve this we took an interdisciplinary approach to identify key components of the radiologist's task. Our main contribution is pioneering inexpensive usability studies by designing a radiology look-alike task that can be performed by nonradiologist subjects. We proved this experiment can give us insight on the benefits of using user-centered interaction techniques for radiology softcopy reading even by employing naïve subjects. We then confirmed the results from the experiment with laypersons in a follow-up experiment using radiologists.

Future Work

We plan to analyze the eye-gaze patterns of the radiologists to determine where they are looking while they are performing the tasks, in particular to determine causes for errors and to establish how much time is spent viewing the workstation controls. We also plan to evaluate other measures for optimizing the user interface, for example, how to introduce new tools to avoid visual distraction. It is also hoped that work will soon start with vendors and radiologists to build staged hanging protocols for other radiological scenarios such as abdominal CT exams.

REFERENCES

1. Moise A: Designing better user interfaces for radiology workstations. Computing Science PhD thesis, Simon Fraser University, Burnaby, 2003

2. Moise A, Atkins MS: Design requirements for radiology workstations. J Digit Imaging 17(2):92–99, 2004

3. Moise A, Atkins MS: Workflow oriented hanging protocols for radiology workstation. Proc-SPIE 4685:189–199, 2002

4. Moise A, Atkins MS: Interaction techniques for radiology workstations: impact on users' productivity. Proc-SPIE 5371:16–22, 2004

5. Moise A, Atkins MS: Designing better radiology workstations: impact of two user interfaces on interpretation errors and user satisfaction. Presented at SCAR 2004, Vancouver, and to appear in Journal of Digital Imaging, 2004

6. Nodine CF, Krupinsk EA: Perceptual skill, radiology expertise, and visual test performance with NINA and WALDO. Acad Radiol 5(9):603–612, 1998

7. Osgood CE: Method and theory in experimental psychology. New York, NY: Oxford University Press, 1956, p 532

8. Bass JC, Chiles C: Visual skill: correlation with detection of solitary pulmonary nodules. Invest Radiol 25:994–998, 1990

9. Kundel HL, Nodine CF, et. al: Visual scanning, pattern recognition and decision-making in pulmonary nodule detection. Invest Radiol 13(3):175–181, 1978

10. Krupinski EA, Roehrig H: Influence of monitor luminance and tone scale on observers' search and dwell patterns. Proc-SPIE 3663:151–156, 1999

11. Berbaum KS, Franken EA, et. al: Role of faulty decision making in the satisfaction of search effect in chest radiography. Acad Radiol 8:304–314, 2000