Understanding the Benefits of Gaze Enhanced Visual Search

Pernilla Qvarfordt, Jacob T. Biehl, Gene Golovchinsky, Tony Dunningan
FX Palo Alto Laboratory, Inc.
3400 Hillview Bldg. 4, Palo Alto, CA 94304
{pernilla, biehl, gene, tonyd}@fxpal.com

Abstract

In certain applications such as radiology and imagery analysis, it is important to minimize errors. In this paper we evaluate a structured inspection method that uses eye tracking information as a feedback mechanism to the image inspector. Our two-phase method starts with a free viewing phase during which gaze data is collected. During the next phase, we either segment the image, mask previously seen areas of the image, or combine the two techniques, and repeat the search. We compare the different methods proposed for the second search phase by evaluating the inspection method using true positive and false negative rates, and subjective workload. Results show that gaze-blocked configurations reduced the subjective workload, and that gaze-blocking without segmentation showed the largest increase in true positive identifications and the largest decrease in false negative identifications of previously unseen objects.

CR Categories: H.5.2 [Information Interfaces and Presentation]: User Interfaces

Keywords: gaze-enhanced visual search, two-phase search, multiple targets.

1 Introduction

Human-based image analysis is a commonly performed task in many contemporary professions. For instance, radiologists and other medical professionals frequently examine medical images to diagnose and treat patients, airport security agents scan x-rays of luggage for prohibited items, and factory workers perform visual inspection of goods to assure quality. In these tasks, the examiner must combine knowledge of the domain and a high degree of mental concentration within a short amount of time to classify or interpret the images.

While there have been many advances both in the technology used to create images and in the training of human examiners, many image analysis tasks are still prone to significant error. For example, Goddard *et al.* [2001] showed that radiological image examination still has rates nearing 20% for clinically significant or major errors. Even with multiple examination (multiple radiologists investigating an image or set of images), the error rate is still high.

.

For instance, a study investigating the utility of triple examination showed error rates still at 11% [Markus et al. 1990]. Error rates go even higher when non-skilled examiners (*e.g.*, MDs other than Radiologists) are involved [Shaw et al. 1990]. A central recommendation from many of these studies is to employ a systematic methodology in how images are examined [Goddard et al. 2001]. That is, there is a need for a more formal, structured analysis methodology that ensures full examination of all anatomical components [Vock 1987].

Similar recommendations have been made for other critical task domains, such as transportation security. Many of the recommendations call for changing the current methodology from only requiring secondary screening on potential true positives to requiring mandatory multi-stage screening [Butler and Poole 2002, Fenga et al. 2008]. While false negative statistics are not openly published for the domain, such recommendations would not be made if there was not a meaningful need to reduce false negative error rates.

In this paper, we present a novel methodology for performing structured image analysis that is both systematic and adaptive to an examiner's search behavior. The goal of the methodology is to provide a more structured image examination process. Specifically, we propose a two-phase process that consists of an initial phase of free search, followed by a second phase designed to help the examiner to cover the whole image. This second phase can consist either of the image being divided into smaller sub-images (segments), of having previously viewed (fixated) segments blocked out, or of a combination of these techniques. This methodology was designed to increase the total coverage of examination while reducing redundant examination with the goal of reducing the overall number of false negatives, particularly false negatives that the examiner did not view or judge. Our proposed methodology is not domain dependent; we believe it can be applied to many domains, including security images, satellite images, maps, astronomical images, and scientific visualizations.

We evaluated the inspections techniques proposed for the structured viewing in a controlled laboratory study. The results shows that masking previously well inspected areas reduced the subjective workload. When the previous gaze is masked over the whole image, the increase of true positive identifications and reduction of false negative identifications are larger than in the other combinations. We believe these results can provide insights and design lessons for the construction of new visual search systems and techniques that improve search performance.

2 Related work

Research on visual search have been pursued in two directions, either to investigate the perceptual and cognitive processes underlying visual search, or to design procedures or tools for improving

the performance of professional image analysts. Our research falls in the latter category, but findings from perceptual and cognitive studies provide additional context.

2.1 Visual Search

Characteristics of visual search have been investigated extensively in cognitive psychology and perception. This line of research is focused on building models of perception and of higher-level cognitive processes. Although visual search studies are often laboratory studies removed from real world image search tasks, there are findings that carry over to the work of professional image analysts.

When searching for an item in an image, the eye scans the image to find the item. The character of these scanpaths has caused a long debate within psychology. Both early observations, *e.g.* [Norton and Stark 1971] and more recent studies, *e.g.* [Scinto, Pillalamarri, and Karsh 1986], have characterized the scanpaths as random. Although people tend to look at similar things in the image, people rarely follow the same scanpath, *i.e.*, they look at different parts of the image in different order.

However, scanpaths do not always appear random. Findlay and Brown [2006] showed that people can employ different strategies when searching through images with randomly placed targets. One such strategy, for example, is to follow the global external counter that the objects create. When the displays are arranged in a more structured ways, *e.g.* in a grid pattern, people tend to read the grid as they would read text [Gilchrist and Harvey 2006]. The practical implications from scanpath theories is that expert inspectors appear to develop structured scanpaths [Sadasivan et al. 2005]. There is, however, evidence from a luggage-screening experiment that performance gains appear to be related to changes in ability to recognize targets rather than to changes in scanpaths [Mccarley et al. 2004].

There is also evidence that the characteristics of fixation change depending on judgments viewers make of the items in view. Pomplun, Reingold and Shen [2001] found that people who were asked to match two sets of images fixated for a longer time on matching objects compared to mismatched objects. Manning, Ethell and Donovan [2004] showed that when radiologists inspected chest x-rays, their gaze duration was half as long on lesions they did not report as a detected tumor as on lesions they reported. Another study [Nodine and Kundel 1987] reported longer gaze durations when giving a true positive judgment about a lesion compared to a false negative judgment. These results suggest that it might be possible to use gaze durations to provide feedback to the user of which parts of an image have been well inspected and which parts have not.

One interesting finding is that people have a higher error rate when targets are rare than when targets occure frequently [Wolfe et al. 2007]. A practical example of this is that when a radiologist screens mammograms for cancer, very few mammograms will show any signs of tumors. Research in low prevalence has suggested several ways to counter the effect, such as rewarding image searchers appropriately for finding targets [Navalpakkam, Kock and Perona 2009], and allowing people to go back and correct mistakes [Fleck and Mitroff 2007].

In most of perception studies that explore low prevalence of targets, participants are asked to find only one target. While the overall true positive rate maybe low, targets often do not exist in isolation. That is, there are likely to be many true positives in a single image, all of which need to be identified to correctly perform the search task. Examples of such recall-oriented activity include looking for suspected tumors in an X-ray, or for camflaged weapons in a satellite image.

2.2 Improving Visual Inspection

Visual inspection is important in many domains such as finding tumors in X-rays and identifying threats in airport luggage scans. Yet, visual inspection is error prone. Most attempts to lower the error has focused on training, see for example [Kollera, Drury, and Schwaninger, 2009]. Nickles, Melloy and Gramopadhye [2003] used dynamic tool for training inspectors to follow a specific scanning pattern during the inspection, but they could not show that the dynamic training was better than verbal instructions. Another study [Sadasivan et al. 2005] used gaze information from experts to create a visualization of a preferred scanpath for cargo bay inspection. This visualization was used to train novice inspectors and proved to be successful in improving performance.

Rather than focus solely on training, others have focused on supporting the inspection task. For instance, Haimson *et al.* [2004] added partitioning to a radar screen and thereby improved performance in finding targets. Forlines and Balakrishnan [2009] used image segmentation to separate objects in the image and to recompose the objects. Techniques for recomposing the image showed promise in reduced error rate, but longer search time was also found for two of the recomposition techniques. Image segmentation as an underlying technique may work well on some kinds of targets. Forlines and Balakrishnan focused on blood cell inspection; other type of images decomposition may not be as easy to decompose as cell slides used in their experiment.

As mentioned above, Nodine and Kundel [1987] found that the fixation times for false positive responses were lower than those of true positive when inspecting chest X-rays. They used this knowlege in a system which collected gaze information on where radiologists looked during a free form inspection. In a second phase, they asked the radiologists to re-inspect areas they had dwelled on for an extended time, but in which they did not find any incidence of tumors. In that work, participants were not asked to view areas they had not looked at previously, but only to re-evaluate already looked at areas.

3 Inspection Methodologies

In this work we compare four different inspection methodologies for visual search. These methodologies follow a two-phase inspection approach. In the first phase, searchers are allowed to view the image unconstrained, enabling them to gain an understanding of the image and its gross structure. In the second phase, the view of the image is restricted to create a more structured search. Below we describe each methodology studied and the rationale for its selection.

Structured Segmentation. Inspired by previous work that found performance benefits in systematic search of images [Sadasivan et al. 2005] and [Nickles et al 2003], this technique subdivides the image into a tiled grid and then displays the grid tiles serially (Figure 1). By reducing the viewable search area at any one time, this technique forces the user to focus attention on the currently-visible area. It also ensures that all parts of the display get equal attention and reduces distractions from other parts of the display. In our study, we divided each image into a three by three grid, keeping the area small enough to be able to quickly scan and mark during the time limit set for the experiment, and large enough to contain information when gaze-blocking was also used.

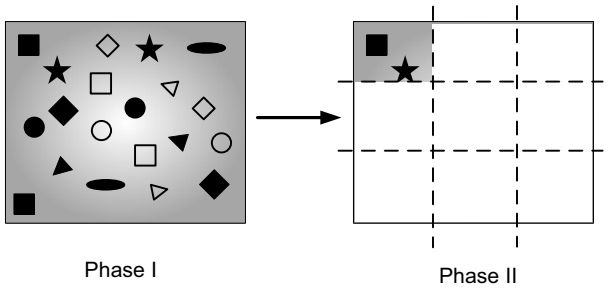

Figure 1: Structured segmentation

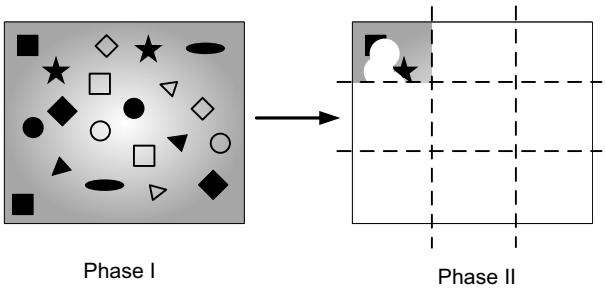

Figure 3: Structured segmentation with gaze-blocking

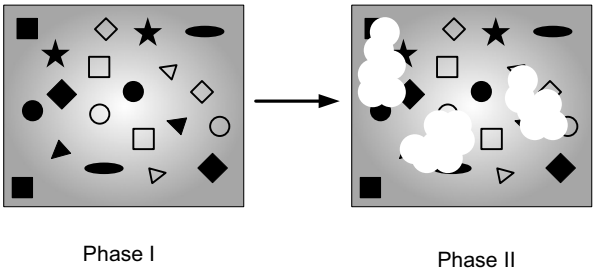

Figure 2: Gaze blocking

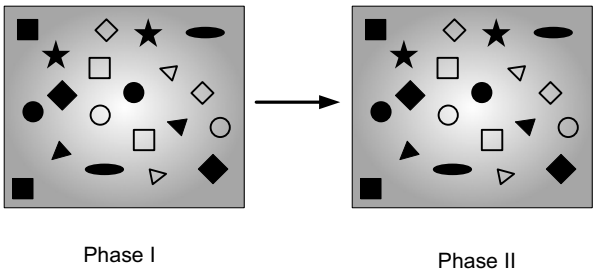

Figure 4: Full-image search

Gaze-Blocking. This methodology uses a searcher's gaze behavior from the first free-viewing search phase to determine where the searcher has previously inspected the image and blocks out those areas in the second phase to prevent unnecessary re-examination (Figure 2). The blocked-out areas are constructed from clusters of identified fixations from the first phase using a dispersion-based fixation detection algorithm [Salvucci and Goldberg 2004]. Fixations are clustered using a minimum-spanning tree algorithm based on their distance to other fixations. Only fixations clusters with a total gaze duration of more than 270 ms were blocked out. We based this threshold on 275 ms fixations for image search reported in [Rayner 1998], and based on our own experimentation. This threshold ensured that fixation clusters reflected some higher level of cognitive processes. Each fixation in the clusters is represented as a 40 pixel circle in our system, centered at the gaze location. This size was chosen as it represents the 1 visual degree of the fovea at the distance of 60 cm from the computer screen using 96 pixels per inch. Each blocked-out region was replaced with a white patch over the image.

Structured Segmentation and Gaze-Blocking. This technique combines the previous two methodologies. The image is re-presented in segments, each segment containing gaze blocks of previously inspected areas (Figure 3). This retains the advantages of structured search while eliminating redundant examination.

Full Image. In addition to these three methodologies, we also included a baseline condition in which the full image was re-presented to the searcher (Figure 4).

4 Experiment

We conducted an experiment to evaluate the inspection methodologies. We designed a controlled experiment to explore the effects of using both segmentation and occlusion of well-inspected areas.

4.1 Method

We used a 2x2 within-subjects design, where the independent variables were type of image display (image display), and visibility of fixation clusters (gaze display). Each of the two independent variable had two levels for *image display*: full image or segmented image, and for *gaze display*: with gaze-blocking or without gaze-blocking. This design gave us four conditions for displaying the image in phase 2: full image with previous gaze blocking (FB), segmented image with gaze-blocking (SB), full image without gaze-blocking (F), and segmented image without gaze-blocking (S). These conditions directly match the proposed inspection methodologies discussed in the previous section.

4.2 Participants

Seven men and one woman participated in the study. Participants were recruited using a broadcast email solicitation within our organization. Participants' age ranged between 30 and 60 years. All were screened before the study to make sure they had normal or corrected vision by contact lenses (4 participants wore contact lenses) and that the eye tracker could properly track their eyes.

4.3 Stimuli and Tasks

Careful consideration was given to the design of the task stimuli. Motivated by the fact that most inspection tasks are not only perceptual but also involve higher-level cognitive processes, we wanted to simulate a task that involved not only perception of targets, but also associated judgment. Furthermore, we did not want to require specific domain knowledge to avoid introducing domain-specific and participant-specific effects.

We designed two types of targets and distractors: symmetric and asymmetric (see Figure 5 shapes in black outline). For symmetric targets a mirror image of the target is indistinguishable from the target. Asymmetric targets on the other hand, have a mirror image that when rotated do not match the target (e.g. in Figure 5 first gray

shape to left of the assymetric target). Symmetric targets differed from distractors by the size of the insections (notches shown in Figure 5) in the basic shape. To identify symmetric targets, participants needed to judge details of the target. In Figure 5 for example, the participants need to judge if the circular insection was of the same size as the example target.To identify asymmetric targets, participants needed to perform mental rotation of potential targets.

Each target had one to three similarly looking distractors (close distractors, targets with gray outline in Figure 5). Targets were randomly assigned to locations where they would not overlap other targets or distractors. Targets and close distractors were not placed on segmentation boundries. Besides close distractors, stimulus images also include a random collection of distractors (random distractors). Targets and distractors were randomly rotated in steps of 30 degrees, starting a 0 degrees..

Figure 5: Examples of symmetric and asymmetric targets. The black shape is targets and the gray close distractors.

In addition, the task required target objects to match a particular color. To ensure color matching was non-trivial, we placed the targets and distractors against uniquely generated, non-uniform colored background and had distractors (both close and random) colored in the same color. For the targets, we used five different basic colors (blue, green, orange, purple, and red), and three shades of each color. Figure 6 shows an example of stimuli used with a mid-shade orange color as the target color.

Figure 6 further illustrates the combination of requirements and constraints of the study tasks. A legend at the top of the image describes the target shape and the color separately. In all, 24 images were generated. Four images included no targets matching the description in the legend. Twenty images, ten with symmetric targets and ten with asymmetric targets, were generated with a random number of targets ranging between 5-20. In addition each image had between 10-40 target shapes with wrong color, 35-40 close distractors with the target correct color, 10-40 close distractors with wrong color. Each stimulus image had 260-300 shapes in total. The difference between the total number of shapes and the targets plus close distractors were filled with random distractors in random colors. All shapes had approximately the same size, around 24 pixels high and wide.

4.4 Experimental Setup

A ViewSonic 18" CRT monitor set at 1024 x 768 pixels resolution was used to present task stimuli to the participants. A CRT was chosen over a LCD panel as it showed consistent colors independent of viewing angle. This was important because the task required participants to carefully judge colors. A Tobii X120 eye tracker was positioned in front of the monitor. Participants were seated approximately 60 cm from the monitor.

The Tobii X120 was configured to collect gaze data at a rate of 60 Hz. We found that this rate meet the input requirements of our system and provided more robust results compared to higher collection rates. The dispersion threshold for the fixation detection algorithm was set to 40 pixels. The minimum duration of a fixation was

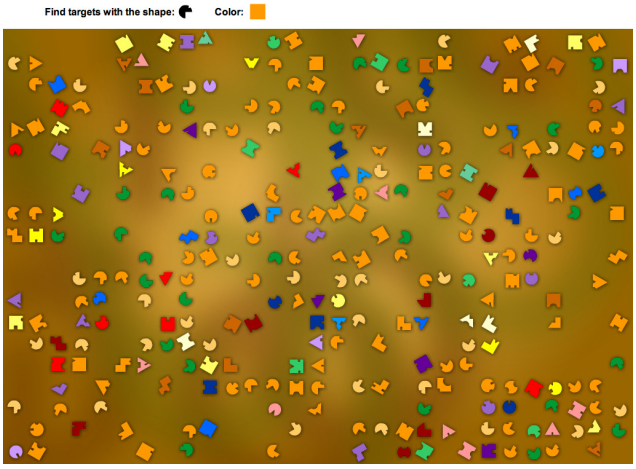

Figure 6: Example of stimuli image used in the study.

set to 100 ms. In addition, we only displayed fixation clusters in the gaze-block conditions where gaze duration was greater than 270 ms.

4.5 Procedure

Each participant's session took about one hour to complete. A researcher first demonstrated to the participant the basics of the experimental task by explaining and showing a sample search task. The participant was then asked to perform the experimental image search tasks. The experimental session was divided into four parts, one for each condition (baseline-full image, segmented, gaze-blocked, and combined segmented and gaze-blocked). Ordering of the conditions was counterbalanced using a 4x4 Latin Square.

In each condition, participants started the task by calibrating the eye tracker. Next, participants were shown a screen that described the current experimental condition. Pressing the space bar loaded the first stimulus image. The image would be displayed for 67.5 seconds (Phase 1). Next, the screen was blanked for three seconds. The image was then displayed according to one of the four conditions (Phase 2). The participant was given another 67.5 seconds to inspect the image and find targets. In conditions including segmentation of the image, each segment was shown for 7.5 seconds. Another three second break was given between successive tasks, for a total of 24 tasks per participant.

A total of six images were presented in each condition. The first image was a training image to familiarize participants with the presentation method and timing in the second phase. One image contained no targets. The no-target image and the training image were not included in the analysis. The order of images after the training image was randomized within each condition. Tasks were balanced over conditions.

Before repeating the search task for the remaining conditions, participants filled out a NASA TLX questionnaire [Hart and Staveland 1988] for assessing their subjective workload. After participants had gone through all four conditions, they ranked the rating scales, as specified in the final step in the NASA TLX procedure. Finally, the participants were asked a few debriefing questions about their experience of using the four conditions. They were also asked to rank the four conditions according to their preference.

4.6 Data Analysis

In all analysis, if nothing else is mentioned, we used a 2x2 full factorial repeated measurements ANOVA. We expected that participants' performance would not differ during the first phase, the free viewing phase. This assumption was confirmed on all dependent measures reported in the results section. Hence, we focused our analysis on the performance during the second phase. We looked at the True Positive (TP) rate, the False Positive (FP) rate and the False Negative (FN) rate. The false negative rate requires some additional elaboration because it consists of items viewed (gazed at) and not judged as matching, and also of items not viewed at all. Thus we calculated FN_V and FN_{NV} rates (where $FN_V + FN_{NV} = FN$). We calculated the overall performance in phase 2 and an adjusted performance based on visible targets not selected for TP and FN. The visible targets were the total number of targets present in the image minus the targets blocked or partially blocked by the gaze-blocking. The adjusted TP and FN hence reflect the number of targets in the second phase that the participants could see and judge in full.

We analyzed patterns of gaze on targets, and found that participants often fixated multiple times on a target. For this reason we analyzed total gaze duration on the target rather than the fixation durations. This method of analysis is also more consistent with how fixation clusters were calculated.

5 Results

5.1 Performance and Error

In any inspection task, it is important to have as many true positives (TPs), correctly identifying a true target, as possible. In our study the participants found on average 86% (SD=10.7) of all targets after both phases across all conditions. In phase 2, they improved their performance on average 31% (SD=20.2) across all conditions. See Table 1 for a summary of TP and TP adjusted for visible targets (TP_{adj}) for all conditions. We did not find any main effect on TP or $TP_{abj}$for the independent variables *image display* and *gaze display*, or any effect of gaze blocking or image segmentation. However, when testing the improvement of TP from phase 1 to phase 2 adjusted for visible targets, we found a borderline significant interaction between gaze blocking and segmentation ($F(1,7)=4.589$, $p=0.069$). As Figure 7 shows, the combination of gaze-blocking and no segments gave a higher improvement in adjusted TP rate (49%) compared to all other combination of con-

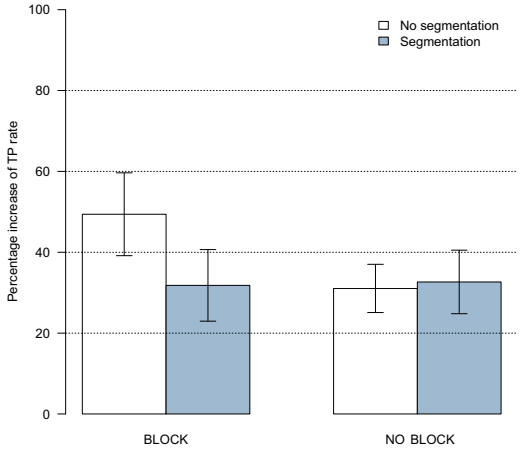

Figure 7: Average improvement of TP rate adjusted for visible targets. Error bars indicate ±1 standard error

Condition	TP	TP_{adj}	FN	FN_{adj}
Overall	86%	88%	14%	11%
Block	83%	88%	17%	11%
No Block	88%	88%	12%	12%
Segmentation	85%	87%	15%	13%
Full Image	86%	90%	14%	10%
Block + seg	83%	86%	14%	10%
Block + full image	83%	90%	14%	7%
No Block + seg	87%	87%	11%	11%
No Block + full image	90%	90%	8%	8%

Table 1: TP, TP_{adj}, FN, FN_{adj} rates at end of phase 2.

False positives classifications (FP) were rare in our experiment. Across all conditions, only 6% of the total number of selections were FPs (SD=7.6). No significant difference was found between conditions.

The false negative rate (FN), incorrectly identifying a true target as false, was 15% (SD=10.9) across all conditions. During the second phase, participants reduced FNs by, on average, 59% (SD=21.7). We did not find any main effects for *gaze display* and i*mage display* variables.

For many image inspection tasks, reducing the FN rate is important. In this work, we are particularly interested in reducing the FN_{NV} rate. We compared the number of FN_{NV} outcomes in phase 1 with the three different outcomes after phase 2: FN_{NV} (still not viewed), FN_V (viewed but not selected) and TP (viewed and selected). Table 1 shows ratios for these three categories for all conditions. As the table illustrates, we again did not find any main effects, but we found a significant interaction for the FN_{NV} in phase 1 transitioning to TPs durring phase 2 ($F(1, 7)= 6.321$, $p<0.05$). In line with the previously found interaction, the combination of gaze-blocked and no segmentation (FB) gave more TP from previously not seen targets, an average of 82% compared to the overall mean of 67%.

5.2 Revisitations of Targets and Image Search Strategy

Previous literature has shown that people revisit targets and that revisitation may change the person's judgement or classification [Nodine and Kundle 1987]. In our study, the gaze-blocking condi-

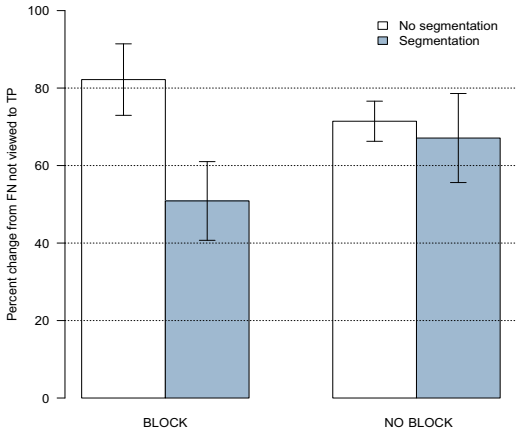

Figure 8: Average transition of FN not viewed in phase 1 to TP during phase 2. Error bars indicate ± 1 standard error.

Condition	FN_V	FN_{NV}	TP
Overall	19%	13%	68%
Block	19%	14%	66%
No Block	19%	11%	69%
Segmentation	26%	15%	59%
Full Image	12%	11%	77%
Block + seg	29%	20%	51%
Block + full image	10%	8%	82%
No Block + seg	24%	9%	67%
No Block + full image	15%	13%	71%

Table 2: Ratio of FN_{NV} from phase 1 to FN_V, FN_{NV}, or TP at end of phase 2.

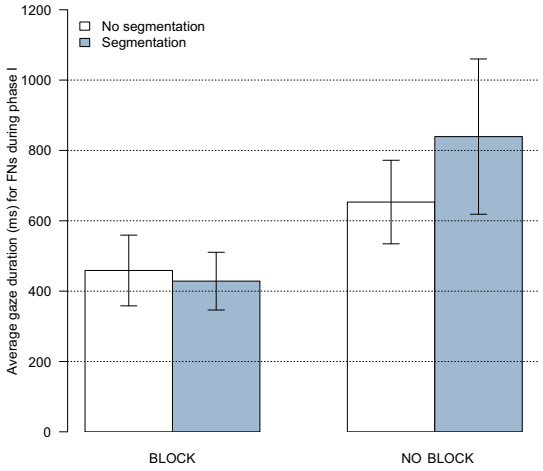

Figure 9: Average duration on FN during phase 1. Error bars indicate ± 1 Standard Error

tions make areas previously visited non-viewable if they are well inspected. The goal is to guide examiners to focus on finding new targets. If people frequently revisit targets and change their mind, this might negatively effect the performance of the gaze-blocked level. For this reason, we investigated how frequently participants revisited visible targets and how often they changed judgments from FN to TP. We found that 51% (SD=23.1) of all new TP in phase 2 had been viewed in phase 1. This difference was not significant, $F(1, 7)=.222$, *ns*. There was no main effects for the two independent variables.

This result shows that adding a second inspection phase is beneficial since people improve performance by re-examining potential targets. Further, it appears that the gaze-blocking did not significantly affect the number of targets being re-examined. The number of FN_V blocked in phase 2 was small: only three of eight participants experienced any blocking of FN_Vs, and of these, two participants had FN_Vs blocked in both gaze-blocking conditions, where on average 17% of FN_V were blocked.

When examining the gaze duration in phase 1 of viewed FN targets which the participants changed to TP after re-examination, we saw a difference in gaze duration between the two levels in the gaze display variable (Figure 9). On average the gaze duration in the gaze-blocking conditions was 443 ms (SD=251.2), while in the non-blocking conditions it was 746 ms (SD=493.7). This difference was very close to significant ($F(1, 7)=5.574$, $p=.0503$). This

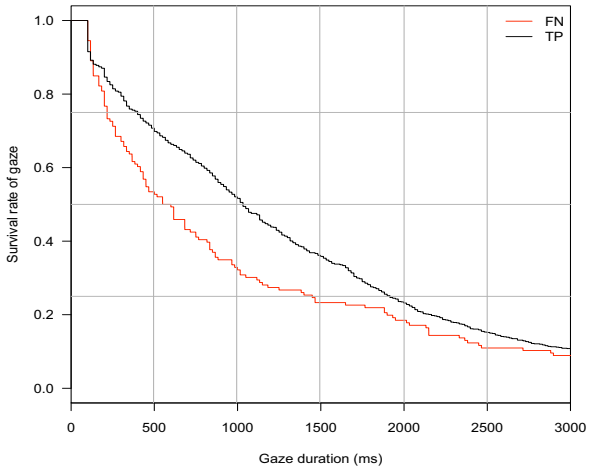

Figure 10: Survival rate of gaze duration on TP and FN targets during phase 1.

analysis indicates that participants likely used different strategies for searching for targets in different inspection conditions. During the first phase of the gaze-blocking conditions, participants likely searched for easily identifiable targets and used the second phase to identify targets that required a higher degree of examination to distinguish targets from distractors. For example, one participant indicated in the debriefing interview that he would examine less dense clusters of objects first, and then move on to examine more dense object clusters. Another participant stated he would try to identify objects "that jumped out at [him]" first.

5.3 Subjective Workload and Preference

Although the differences between the different conditions for the task load index (TLX) were small in magnitude, we found a significant main effect for the *gaze display* variable ($F(1, 7)=6.905$, $p<0.05$). When gaze-blocking was present, participants experienced a significantly lower workload (mean=63.6) compared to it was not present (mean=67.5). This correspond to 6% reduction in task load. We did not find any effect due to segmentation.

Participants were asked to rank their preferred methodology from 1 (most preferred) to 4 (least preferred). Differences among methodologies were small and no one methodology was overwhelming preferred across all participants. The highest overall rated methodology was no segmentation and no gaze-blocking (F: summed score: 17), followed by no segmentation and gaze-blocking (FB: summed score: 18). However, when we correlated the preference scores with the measured workload, we found a significant correlation ($r=0.45$, $t(30)=2.785$, $p<0.01$). That is, participants tended to prefer conditions where they reported lowest subjective workload.

5.4 Fixation Patterns on Targets

For a gaze-blocking to be feasible, gaze durations on targets need to be distinguishable. When designing gaze-blocking, we relied on observations that fixations on TPs were longer than those on FNs [Nodine and Kundle 1987] in the domain of radiology. We examined gaze durations on targets, and found that in our domain too TPs were gazed at longer than FNs. The average duration of FNs were 661ms (SD=628.5), and for TPs 1386 ms (SD=1032.5). This difference was significant ($F(1,7)=16.533$, $p<0.01$). This result shows that the duration of gaze on targets differed depending the participant's interpretation of the target.

As illustrated in Figure 10, there is a definite pattern in the data which indicates TP, on average, are dwelled on longer by examiners compared to FN. For instance, at 550 ms, the participants had stopped looking at 50% of FNs, but only stopped looking at 32% of the TPs. The 50% gaze duration survival rate for TPs was at 1032 ms. On the other hand, at gaze duration of 1049 ms, 70% of the FN were shorter. Although both TPs and FNs has a large spread, these numbers reaffirm that it should be possible to use gaze duration as an implicit relevance feedback mechanism to support image analysis.

We also found a significant difference in fixation duration between phase 1 and phase 2 ($F(1, 7)=71.597$, $p<0.001$). During phase 2, participants' fixation duration was on average 813 ms (SD=749) and in phase 2 the duration was on average 1235 ms (SD=1037). There was no difference between the conditions. This result is further evidence that the participants shifted strategy between the two phases.

6 Discussion

Intuitively, masking off where you have looked during an image search task is both a bad idea and a compelling idea. When viewed parts of the image are blocked out, the searcher loses context and it is not possible to correct previous decisions. On the other hand, in a complex image it can be hard to identify all occurrences, and even to remember what you looked at already. In addition, when the task involves not just perceptual processes but higher level cognitive processes, fatigue sets in quickly due to the demanding nature of the task, resulting decreased performance. We described four different inspection methods using combinations of gaze-blocking and image segmentation techniques. In the study, we took an extreme approach to understand limitations of the methods. The performance results from our study show two things: masking an image based on where searchers looked previously did not negatively impact performance, and masking gazed-at areas has potential.

The results that gaze blocking reduced participants' workload is an important finding. Image search can be a demanding task and if the workload can be reduced, performance can be improved. We believe it is important to find methods supporting image analysts that also reduce their workload.

In terms of performance, not surprisingly the gaze-blocking condition did on average somewhat worse in true positive rate, in particular true positive rate not adjusted for visible targets. This difference was not significant, and hence could just as well as have been caused by other factors than those manipulated. However, when adjusted for targets blocked by the gaze mask, the gaze-block in combination with whole image showed a larger increase in the TP rate and a larger reduction in FN during phase 2 than the other conditions. This result indicates that allowing participants to focus on areas not previously viewed helped them find and select more of the remaining targets compared to the other three inspection methods. This result is promising for future work.

Segmenting the image did not seem to affect the performance, expect in combination with gaze block. The SB condition had the overall lowest performance. Interestingly, when the transition rate from FN_{NV} to TP were highest in gaze block on full image (FB), it was also lowest in SB. It appears that restricing the viewable area too much has negative affects on performance. Future research is needed to explain this phenonemon.

Our results showed that slightly less than half of the performance gain when previous gaze was blocked in phase 2 came from previously non-viewed objects. This is a promising result, in particular for situations where image analysis is done under time pressure. When time is unlimited, more targets may be viewed during the first phase. An interesting research issue to look at in the future is how people can optimize their performance using a two phase inspection method and what trade-offs the image analyst faces when previous gaze is blocked during a second phase.

Revisitation of previously looked at targets is another important factor for the performance increase during phase 2. Although gaze blocking masked previously well-inspected areas, we could not find a significant impact of gaze-blocking during the second phase in terms of revisitations. In our study, a low percentage of the previously viewed and un-selected targets were blocked out, and in particular, in gaze-blocking and full-image conditions, participants made up for that performance loss by finding previously-unseen targets. This is a promising result, since one of our goals was to lower the number FNs during the second phase.

Another interesting finding in our study is that participants adjusted their behavior during the gaze-blocking condition. During the first phase, they adopted the strategy of scanning the image for the easiest to find targets. In the next phase, they would focus on targets that were harder to find and harder to interpret. It is possible that this strategy contributed to lowered workload in the gaze-blocking condition. This result is intriguing since it suggests more research on how to design image search techniques that can promote a more efficient use of time and mental resources.

Gaze patterns collected during study confirm that it is possible to give indications to image searchers of which parts of an image they have not inspected in enough detail. As also observed in other studies, we found that on average, participants spent significantly longer looking at targets they ended up selecting (TPs) than those they discarded as distractors (FNs). Because of the large individual difference in gaze duration on TP targets and FN targets, thresholds for gaze-blocking algorithms may need to be set dynamically based on each user's behavior and preference, or by using machine learning approaches. In systems using manual pointing to for marking selections, users' selections can also be used to get gaze durations for TPs. In this study, we only looked at the sum of all fixations on the target. Other gaze characteristics, such as number of fixations on target, duration between fixations, etc. can potentially be informative. The image search task may also influence what method the system uses to determine which parts of the image would be helpful for the user to look at in more detail.

Our study does not describe a realistic system; it merely tests the limits and the feasibility of using eye tracking to support image search by masking viewed areas. A realistic system would need ways to make correction, which in the study was only possible on non-masked areas. A realistic system would also need to allow image analysts to control the pace, segmentation, and masking of viewed areas. However, results from this study are encouraging, indicating that eye tracking can provide useful support during image search.

7 Conclusion

In this paper, we proposed a two-phased approach to image search and investigated the performance, workload, and user preference trade-offs of four different second phase image re-presentation methodologies. Results from a controlled laboratory study showed that in a two-phase approach the gaze-blocking treatment, which

masks well examined regions of the image in the second phase of inspection, significantly reduced overall subjective workload. Gaze block on a full image increased the transition of previously unseen false negatives to true positives, and was nearly as preferred as free-form search. Further, our results also suggest that segmentation and gaze-blocking likely encourage searchers to adopt an more structured search strategy in *both* phases of examination. Results from our study can have broad impact on the design of future techniques and systems that aid in visual search.

8 Acknowledgments

We thank all our participants for their time and effort, and Lynn Wilcox and Larry Rowe for supporting this research.

9 References

BUTLER, V. AND POOLE, R.W. 2002. Rethinking Checked-Baggage Screening. *Reason Public Policy Institute*; July. 25p. Policy Study No. 297.

FENGA, Q., HANDE, S., AND KAPURB, K.C. 2008. Designing Airport Checked-Baggage-Screening Strategies Considering System Capability and Reliability. R*eliability Engineering & System Safety.* 94(2) 618-627.

FLECK., M.S. AND MITROFF, S. 2007. Rare targets are rarely missed in correctable search. Psychological Science, 18(11), 943-947.

FORLINES, C., BALAKRISHNAN, R. 2009. Improving visual search with image segmentation. In *Proceedings of CHI'2009*, Pp. 1093-1102.

GILCHRIST, I. D. AND HARVEY, M. 2006. Evidence for a systematic component within scan paths in visual search. *Visual Cognition*, 14, 704-71

GODDARD, P., LESLIE, A., JONES, A., WAKELEY, C., AND KABALA, J. 2001. Error in radiology. *British Journal Radiology* 74, 949-951.

HAIMSON, C, BOTHELL, D., DOUGLASS, S.A., ANDERSON, J.R. 2004. Partitioning Visual Displays Aids Task-Directed Visual Search. *Human Factors*, 46(3), 551-566

HART S.G., STAVELAND L.E. 1988. Development of a NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research." In Hancock PS, Meshkati N, (eds). *Human Mental Workload*. Amsterdam: North-Holland: pp. 139–83.

KOLLERA, S.M., DRURY, C.G., AND SCHWANINGER, A. 2009. Change of search time and non-search time in X-ray baggage screening due to training. Ergonomics, 52(6), 644–656.

MANNING, D.J., ETHELL, S.C. AND DONOVAN, T. 2004. Detection or decision errors? Missed lung cancer from posteranterior chest radiograph. *British Journal of Radiology*, 77, 231-235.

MCCARLEY, J.S, KRAMER, A.F., WICKENS, C.D., VIDONI, E.D. AND BOOT, W. R. 2004. Visual Skills in Airport-Security Screening. *Psychological Science*, 15(5), 302-306.

NAVALPAKKAM, V., KOCH, C., AND PERONA, P. 2009. Homo economicus in visual search. *Journal of Vision,* 9(1):31, 1-16.

NODINE, C. F. AND KUNDEL, H. L. 1987. Using eye movements to study visual search and to improve tumor detection. *RadioGraphics* 7(6) 1241-1250.

NORTON, D., AND STARK, L. 1971. Eye movements and visual perception. *Scientific American*, 224, 34-43.

MARKUS, J.B., SOMERS, S., O'MALLEY, B.P., AND STEVENSON, G.W. 1990. Double-contrast barium enema studies: effect of multiple reading on perception error. *Radiology* 175, 155-156.

POMPLUN, M., REINGOLD, E.M. AND SHEN, J. 2001. Investigating the visual span in comparative search: the effect of task difficulty and divided attention. *Cognition*, 8, B57-B67.

RAYNER, K. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3), 372-422.

SADASIVAN, S., GREENSTEIN, J. S., GRAMOPADHYE, A.K., DUCHOWSKI. 2005. Use of eye movement as feedforward training for a synthetic aircraft inspection task. In *Proceedings of CHI'2005*, 141-149.

SANTELLA, A. AND DECARLO, D. 2004. Robust clustering of eye movement recordings for quantification of visual interest. In *Proceedings of ETRA'04*. Pp27-34

SCINTO, L. F., PILLALAMARRI, R., KARSH, R. 1986. Cognitive strategies for visual search. *Acta Psychologica.* 62(3), 263-292.

SHAW, N., HENDRY, M., AND EDEN, O. 1990. Inter-observer variation in interpretation of chest X-rays. *Scott Med J* 35, 140-141.

VOCK, P. 1987. Frequent errors in the interpretation of chest x-ray films. *Schweiz Med Wochenschr* 117, 928-932.

WOLFE, J.M., HOROWITZ, T.S., VAN WERT, M.J., KENNER, N.M., PLACE, S.S., KIBBI, N. 2007. Low target prevalence is a stubborn source of errors in visual search tasks. *Journal of Experimental Pshychology: General.* 136(6), 623-638.