Improving Visual Search with Image Segmentation

Clifton Forlines[1,2], Ravin Balakrishnan[2]

[1]Mitsubishi Electric Research Labs
Cambridge, MA USA
forlines@merl.com

[2]University of Toronto
Toronto, ON Canada
ravin@dgp.toronto.edu

ABSTRACT

People's ability to accurately locate target objects in images is severely affected by the prevalence of the sought objects. This negative effect greatly impacts critical real world tasks, such as baggage screening and cell slide pathology, in which target objects are rare. We present three novel image presentation techniques that are designed to improve visual search. Our techniques rely on the images being broken into image segments, which are then recombined or displayed in novel ways. The techniques and their underlying design reasoning are described in detail, and three experiments are presented that provide initial evidence that these techniques lead to better search performance in a simulated cell slide pathology task.

Author Keywords

Visual search, image presentation, RSVP, segmentation

ACM Classification Keywords

H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION

Many critical tasks that require an operator to visually search images, such as medical image viewing, baggage screening at the airport, and satellite photograph analysis, all involve searching for uncommon targets. Luckily, conditions such as cervical cancer, severe blood disorders, and knives in carry-on luggage are rare; however, it is the very rarity of these conditions that makes searching for them so difficult. Recent research [27] has shown that an individual's ability to search images for a particular object degrades rapidly and severely when the prevalence of the sought objects becomes small.

For example, laboratory directors routinely report error rates of less than 2% for the common Papanicolaou smear slide ("pap smear") test for cervical cancer, which requires a human operator to search for certain visual qualities in a slide's cells. Renshaw [19] correctly points out these reported error rates are only meaningful when viewed relative to the incidence of a specific disease, not the overall caseload. Because the number of negative cases is much higher than positive cases, one could achieve a low overall error rate by simply reporting every test as negative.

False-negative (missing a target) error rates are believed to be much higher, as high as 20% [19]. Keenlyside et al. [13] rescreened Papanicolaou smear slides from over forty thousand women, and found false-negative error rates of 16% and 15% for low grade and high grade intraepithelial lesions respectively. Given the overwhelming evidence of the importance of accurate screening in reducing mortality [1], even small improvements to the false-negative error rates for these tests would have large consequences.

In this paper, we present three novel image presentation techniques designed to improve the visual search of images that can be segmented into pieces. These techniques rely on the images being broken into image segments, which are then recombined or displayed in novel ways. The three techniques and their underlying design reasoning are described in detail. This design rational, while rooted in established work, does not guaranty the techniques' usefulness in improving search. Thus, the three techniques are accompanied by three laboratory experiments that provide initial evidence that they lead to better search performance. These techniques are not intended to be comprehensive solution to the general problem of low prevalence visual search *per se*; rather, they are steps in the direction of improving people's ability to perform some difficult and important classes of visual search tasks.

RELATED WORK

Visual search has been studied extensively in the psychology literature for over a century (see Wolfe [25] for a good overview of this extensive literature). Many underlying theories have been presented as to the nature of the human visual system; however, no single model of the human visual system explains the variety of experimental results in the literature.

In a typical search experiment, participants are asked to look for a specific object within a stimulus image that may or may not contain the target and does contain a varying number of distracter objects. Normally, the target object is present in 50% of the stimulus images and the participant responds to each stimulus image by indicating whether or not they believe the target object is present. Participants are asked to respond as quickly and as accurately as possible, and the dependent variables measured are reaction time and

CHI 2009, April 4–9, 2009, Boston, Massachusetts, USA.

error rate. To achieve a better measurement of error rate, a variation of this method is used in which images are presented for a prescribed period of time before being removed from the display. This consistency in image presentation duration is believed to reduce differences among individuals who may have different thresholds for appropriate search time / accuracy tradeoffs.

For difficult searches, search time is linearly correlated with the number of distracter objects present in the image. Similarly, error rates are typically higher for images that contain a greater number of distracter objects. Overall error rates are typically around 10% as participants naturally slow down if they are making many errors and speed up when committing few errors.

Problems with Low Target Prevalence

Target object prevalence is the percentage of stimulus images in which a target item appears. In a recent study, Wolfe et al. [27] tested not only the typical target object prevalence of 50%, but also prevalences of 10% and 1%, pointing out that in many important real world tasks, such as baggage screening and x-ray analysis, targets of interest, such as knives and tumors, are very rare. Their somewhat disturbing results showed that error rate grew significantly as target prevalence dropped – from 7% error in the 50% prevalence trials, to 16% error for 10% prevalence, to 30% error for 1% prevalence. Wolfe et al. concluded that "if a target is rare, we rarely find it." While some steps have been made toward improving or discounting low-prevalence search [6,7,26], combating the negative effects of low target prevalence remains an open research issue.

Visual Search in the HCI Literature

Mello-Thoms, Nodine, and Kundel [15] conducted an eye tracking experiment in which radiologist searched mammograms for signs of cancer. They were motivated by the fact that 10-30% of cancers are missed by radiologists and are only found retrospectively. Their results included the interesting finding that missed tumors were often visually inspected by the radiologist, indicating that visual search was not the cause of the problem, but rather that decision making or some other perceptual process was.

In regards to modeling human behavior and performance in visual search tasks, Halverson and colleagues [10,11,12] are attempting to produce a unified visual search model for predicting search time in user interfaces. Early results indicate a close match between the model's predicted path and actual eye movements during hierarchical menu searching. While this model should be useful in the design of GUIs, it is unclear if it will help with the more general task of searching for targets in more general images.

Rapid Serial Visual Presentation

The human visual system is extremely adept at rapidly processing visual images [4]. While well understood in the field of psychology, there has been less research and fewer experiments in the field of human computer interaction. Spence gives an overview of research in the field of Rapid Serial Visual Presentation (RSVP) in terms of its use in HCI [23]. In general, RSVP techniques trade time for space when presenting a set of images, and differ mainly in their presentation and animation of images. Recent well publicized uses of RSVP in computer interfaces include Apple's "Coverflow" and Windows Vista's "Flip 3D".

With respect to search, Wittenburg et al. [24] demonstrated a RSVP interface designed to help television viewers search for specific scenes when fast-forwarding/rewinding through recorded video. Their work included an experiment that demonstrated this interface resulted in faster search times as well as fewer errors over the traditional fast-forwarding/rewinding technique used for VCRs and DVDs.

Image Segmentation

The three techniques presented in this paper all rely on the original search image being segmented into its component pieces (Figure 1). The field of image segmentation is an active area of research, especially as it relates to medical image analysis. Many methods and techniques exist, including k-means clustering, histogram based techniques, edge detection, region growing, and so on. A good overview of image segmentation as it relates to medical image viewing is provided by Pham et al. [17].

Figure 1 shows the results of segmenting two cell slides using the Vincent-Soille watershed algorithm, as described by Roerdink & Meijster [20] and implemented in [2]. While we are not researchers in the field of image segmentation, we feel it is likely that experts in this field could perform a much better segmentation than the approaches we investigated. We believe this example demonstrates the feasibility of successfully segmenting these types of images for the purpose of enabling the interaction techniques described in the remainder of this paper.

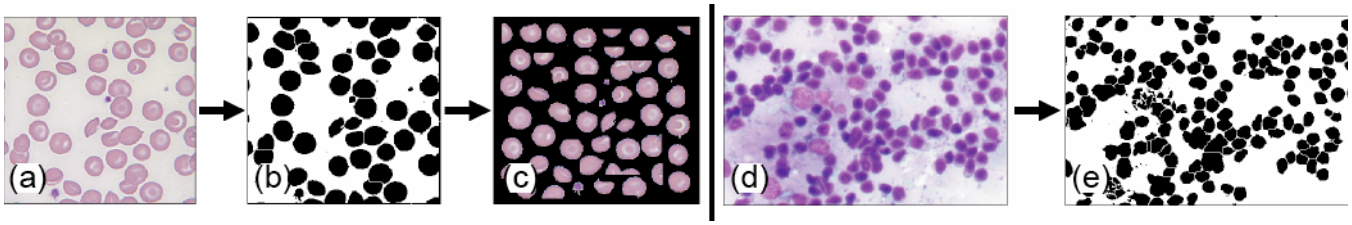

Figure 1: (a) A blood cell slide. (b) Results of Vincent-Soille watershed segmentation. (c) Components are broken apart and can be treated as separate images. (d) A more difficult image and (e) the resulting segmentation.

TECHNIQUE ONE: INCREASING TARGET PREVALENCE

Figure 2 shows an overview of our first technique. A set of images of a certain size having a certain target prevalence is segmented into a collection of individual pieces using the techniques previously described. These pieces are then randomly composited into larger, composite images with higher target prevalence that the user then searches through.

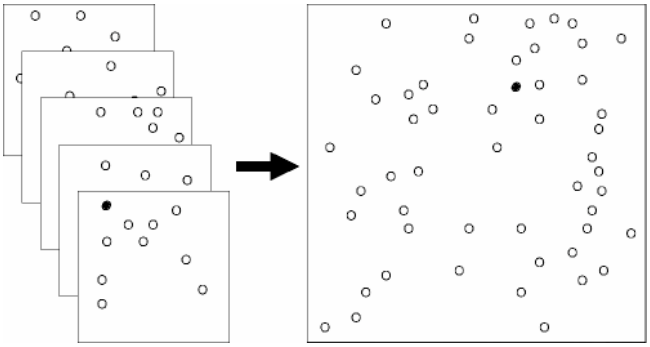

Figure 2: Image recombination into composite images with higher target prevalences.

The logic behind this approach is as follows. Define $p(T_I)$ as the probability of encountering the type of target T in an image I and $p(\check{T}_I)$ as the probability of not encountering T in I. The basic law of probability informs us that when N images are combined into a single composite image C, then the probability of not encountering T in C is equal to the probability of not finding T in I raised to N. In short, $p(\check{T}_C) = p(\check{T}_I)^N$. Because C either contains T or does not contain T, $p(T_C)+ p(\check{T}_C) = 1$. Thus:

$$p(T_C) = 1 - p(\hat{T}_I)^N$$

If $p(T_I)$ is unacceptably low and interferes with people's ability to accurately search through images, one can raise $p(T_C)$ through image recombination. Indeed, any desired target prevalence $p(T_C)$ that is less than 100% can be theoretically met by increasing N.

This tradeoff does not come without cost. With a larger number of distracter objects, the larger composite images will be more difficult to search than an original image. It is unclear how this tradeoff between task difficulty and target prevalence will affect peoples' ability to successfully search through image sets for target objects.

Why not skip the segmentation and random recomposition steps and instead simply tile the collection of low-prevalence images together to form a high-prevalence screen of images? There is some evidence that this type of tiling leads to higher error rates compared to the sequential viewing of each of the images alone [9]. Therefore, we chose to remove the appearance of tiled images in this technique through the randomized layout of the segmented pieces in the hopes that viewers would treat our composited images as a whole.

Experiment One

Given the tradeoff between task difficulty and target prevalence, we designed an experiment to measure how increasing target prevalence through image combination affects search accuracy.

12 individuals (4 female, 8 male, ages 20-52) from local colleges participated in our study. All had normal or corrected normal vision. Participants searched for vowels in images that contained collections of letters. The images were displayed on a desktop LCD monitor. There were two image sizes: 402x402 pixel original images and 900x900 pixel composite images. The area of the larger composite images was approximately five times the area of the smaller original images. Images contained a randomly placed collection of English letters rendered in one of three fonts, with original images containing 10 letters and composite images containing 50 letters. Participants were asked to search the images and indicate the presence or absence of vowels, with each image containing 1 or 0 vowels. To avoid confusion between consonants and vowels, there were no Y's in the images, nor were there any lowercase L's (which look like I's in some fonts).

Each participant viewed three types of images that differed in size and the prevalence of vowels. In the first set, a vowel was present in 10% of the original stimulus images and absent in the other 90% (low-prevalence). In the second set, a vowel was present in 50% of the original images (high-prevalence). In the third set, a vowel was present in 50% of the composite images (composite). Figure 3 shows target-present images from the low-prevalence, high-prevalence and composite conditions.

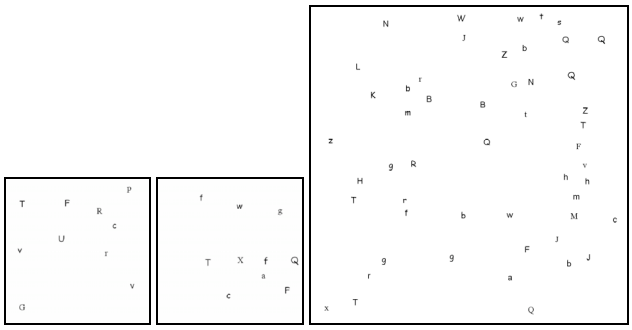

Figure 3: (Left pair) Original images used in the high and low-prevalence conditions. (Right) A composite image made from combining five original images.

Participants indicated whether or not a vowel was present in the stimulus image by using a small keypad with two keys, one to indicate target presence and the other target absence. After each answer, participants were given visual feedback as to their performance. When vowels were correctly located, they were highlighted in green for ½ second before the next image was shown. Similarly, when a participant correctly indicated the absence of vowels, the entire screen was briefly highlighted in green. When a participant missed a vowel, the vowel was highlighted in red, and the testing application penalized the participant by presenting this

negative feedback for 15 seconds. This pause was meant to reduce the chance that a participant would race through the trials with careless responses in order to quickly finish the experiment. False-positives (which are typically very rare) resulted in a similar 15 second pause with the entire screen highlighted in red. While participants were allowed to answer early, our testing application displayed original images onscreen for 2 seconds and composite images onscreen for 10 seconds. This consistency in image presentation duration is intended to normalize for differences among individuals who may have different thresholds for appropriate search times in this type of task.

Method

We used a within-participant, repeated measures design with *presentation technique* (low-prevalence, composite, and high-prevalence) and *target presence* (target-present and target-absent) as independent variables and *error rate* and *search time* as dependent variables. Because one of the main goals of the experiment was to study differences among target prevalence conditions, it was not possible to balance the experimental variables in the traditional manner. By definition, a low-prevalence condition will contain fewer target-present trials than a high-prevalence one. To address this issue, we averaged the results from the many repetitions of each presentation technique and used only the average error rate for each of the 6 *presentation technique / target-present* combinations in our analysis. In summary, our design was:

12 participants x
(105 high-prevalence, target-absent trials +
105 high-prevalence, target-present trials +
189 low-prevalence, target-absent trials +
21 low-prevalence, target-present trials +
21 composite, target-absent trials +
21 composite, target-present trials)
= 5,544 trials

Hypotheses

H1. Participants will commit fewer false-negative errors in the high-prevalence *presentation technique* than in the low-prevalence condition, as predicted by previous work.

H2. Participants will commit fewer errors in the composite *presentation technique* than in the low-prevalence set.

H3. Participants will commit fewer errors in the high-prevalence condition than in the composite condition.

Error Rate Analysis

53 trials in which the search time was more than three standard deviations from the mean for that trial's *presentation technique / target presence* combination were counted as outliers and removed. These outliers (1% of our data) occurred when participants took an unscheduled break, dropped the keypad, or were otherwise interrupted. A repeated measures ANOVA of all of our experimental data found no asymmetrical learning effects for the ordering of *presentation technique* among participants. Therefore, it is safe to continue analysis with a within-participant design.

As expected in a visual search experiment, *target presence* had a significant main effect on error rate ($F_{1,11} = 60.59$, $p < 0.001$). False-positives are generally rare, and the majority of errors in our experiment occurred during target-present trials (Figure 4). The remainder of the analysis in this section was performed on the target-present trials only.

Presentation technique had a significant main effect on error rate ($F_{2,22} = 12.30$, $p < 0.001$), with mean error rates of 17.3%, 12.5%, and 6.7% for low-prevalence, composite, and high-prevalence images respectively (Figure 4). A post-hoc pair-wise comparison shows a significant difference among all possible pairs of *presentation techniques*.

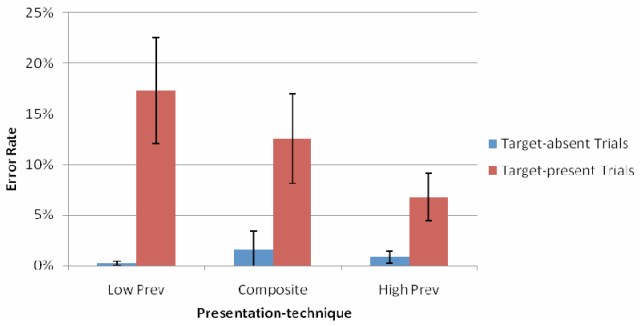

Figure 4: The significantly different error rates for the three *presentation techniques* for both *target presence* conditions. Bars represent 95% confidence intervals.

Search Time Analysis

Search time was measured as the time between the appearance of an image and the participant's pressing of a key. To compare the three *presentation techniques*, we first normalized the search times for the composite conditions. Because these images contained five times as many items, we divided the search time for these trials by five.

As is typical in a search experiment, there was a significant main effect from *target presence* on search time ($F_{1,11} = 155.25$, $p < 0.001$), (Figure 5). Target-present trials end when a target is located, while target-absent trials require the participant to pass a self-determined internal confidence threshold before deciding that the target is absent.

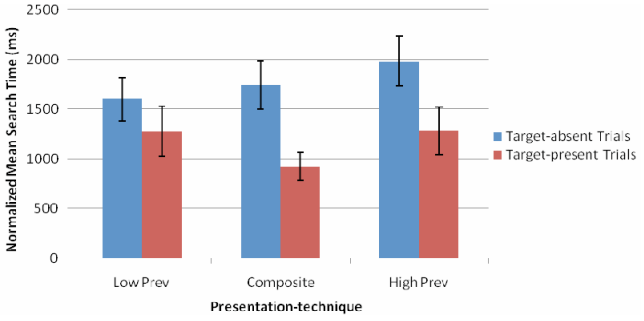

Figure 5: Normalized mean search times for the three *presentation techniques* for both *target presence* conditions. Bars represent 95% confidence intervals.

There was also a significant main effect on search time for *presentation technique* ($F_{2,22} = 6.57$, $p = 0.006$), as well as a significant interaction between *presentation technique* and *target presence* ($F_{2,22} = 13.60$, $p < 0.001$).

Discussion of Experiment One

Our results support *H1* and reinforce the previously published findings indicating that target prevalence significantly affects search accuracy, as shown by the significant difference between the low-prevalence and high-prevalence trials (with error rates of 17.3% and 6.7% respectively). Indeed, it seems that "when a target is rare, we rarely find it" [27]. While our participants searched images of letters looking for vowels, a false-negative error rate of over 17% is a cause for concern if these results are applied to highly consequential tasks such as cell slide pathology and baggage screening.

Our results indicate the negative effects of low-prevalence can be significantly mitigated by the recombination of multiple low-prevalence images into larger, higher-prevalence images. Although an individual composite image is more difficult to search than an individual low-prevalence image, the raising of target prevalence when the image sets are viewed as a whole resulted in a 28% decrease in false-negatives. Thus we accept *H2*.

It is important to remember that our participants viewed the same total number of targets and distracter objects in approximately the same time period in both the low-prevalence and composite *presentation technique* conditions. The only difference was the manner in which these targets were presented, and it appears that the presentation matters significantly.

Finally, our results also support *H3*. While we were able to improve upon our participants' ability to search through the low-prevalence image set by segmenting and then compositing the content of these images into a high-prevalence composite image, these composited high-prevalence images presented a more difficult task than simply searching through a similarly high-prevalence set of smaller images with fewer items in each image. This tradeoff is inherent in our technique, but is ultimately not troubling as our approach would never be applied to a high-prevalence search task.

While these benefits are encouraging, they have a cost. In terms of search time, the high-prevalence trials resulted in longer target-absent search times than the low-prevalence trials (Figure 5). It appears that when participants become use to targets being frequent, they take longer to convince themselves that they are absent from a particular image. This implicit user decision to tradeoff speed for accuracy, however, may be good for critical search tasks.

A second cost to consider is the increased monetary and physical foot-print cost of requiring a larger display size for this technique. This experiment investigated the raising of target prevalence from 10% to 50%, which (if keeping target density constant) requires a screen size that is roughly five times larger than needed for the low-prevalence images. For search involving very rare targets (1%-2% prevalence) a much larger display would be necessary to reach 50% target-prevalence. Put simply, practical limitations on display size will eventually limit the feasibility of reaching 50% for very rare targets. This limitation aside, there is evidence that even modest increases in target-prevalence (say from 2% to 10%) can result in significant decreases in errors; thus, system designers interested in utilizing this technique must balance increase in target-prevalence with increase in display size.

TECHNIQUE TWO: RE-LAYOUT

Figure 6 shows an example of our second technique. An original image with a scattered collection of objects is segmented using the previously described method. These pieces are then arranged in an orderly grid layout.

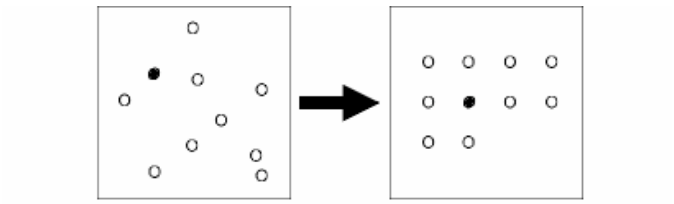

Figure 6: Image segmented and rearranged into an orderly grid layout.

The motivation behind this technique comes from previous research into gaze paths using eye trackers. There is some disagreement as to what is the "typical" gaze path one takes through a randomly ordered image (such as a slide of cells). Some researchers have observed that this path is itself random, with the image being investigated in an inefficient, seemingly random order [5,14,21]. These researchers have observed that searchers not only follow a random gaze path, but also often skip objects and revisit the same object multiple times when searching. Other researchers have observed that the gaze path one takes through a randomly ordered image is itself orderly [8]. By systematically scanning a randomly ordered image, one attempts to cover all portions of the image in a timely manner. Problems arise when the location of objects place them out of step with this orderly search. People often skip objects or visit the same objects on multiple occasions (although the later is rarer). One explanation for these competing theories is that the gaze path one takes through an image is heavily dependent on the qualities of the image and the task at hand.

Given either theory, we hope to encourage the user to take a more efficient path that minimizes traversal by presenting every image as an orderly collection. The orderly layout used in this technique should give users confidence that they have inspected every object in the image and allow them to visit each object only once (Figure 7).

Additionally, there is some evidence that an orderly layout of objects decreases search time. While Beck & Trafton [3] found no difference in search times between a random and orderly layout for target-present trials, they did observe a significant increase in search time for random layouts when the target was absent from the image. They hypothesize that in the random conditions participants took longer to convince themselves that the target was absent than they did in the orderly conditions. Others have studied random vs. orderly layouts for an image searching task and found the quickest searches with the shortest gaze path occurred with regularly spaced elliptical and matrix layouts [22].

While an orderly layout's effect on search time has been demonstrated previously, our interest is on search accuracy, specifically for rare targets. Thus, an evaluation of the effectiveness of this technique is needed.

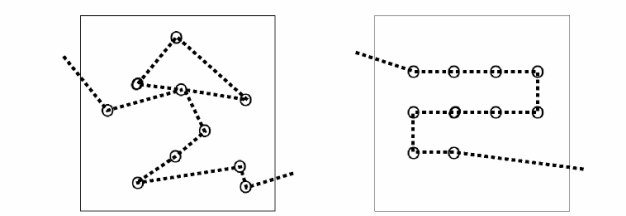

Figure 7: Possible gaze path through images with a (left) Random and (right) Ordered layout. We hypothesize that an Ordered layout will reduce the total traversal distance, as well as remove the revisiting of potential targets in the image.

Experiment Two

We recruited 12 individuals who did not take part in Experiment One from local colleges and universities to participate in our study. The 7 female and 5 male participants' ages ranged from 21 to 45 years old, and all had normal or corrected normal vision.

Because our desire was to compare the results from experiment two with those from experiment one, much of the study's design is identical. Participants again search for vowels in images that contained collections of letters. The display configuration and input keypad from experiment one were used. The 402x402 pixel images contained an orderly positioned collection of 10 English letters rendered in one of three fonts. Participants were asked to search the images and indicate the presence or absence of vowels, with each image containing one or zero vowels. Each participant viewed two sets of images that differed in the prevalence of vowels, a 10% set (low-prevalence) and a 50% set (high-prevalence). As in experiment one, we again provided visual feedback after each answer and penalized participants when they committed an error.

Method

We used a repeated measures design with *target prevalence* (low-prevalence and high-prevalence) and *target presence* (target-present and target-absent) as within-participant independent variables, *image layout* (random and ordered) as a between-participant independent variable (where the random layout data came from experiment one), and *error rate* and *search time* as dependent variables. In summary, our design was:

12 random layout experiment one participants x
(105 high-prevalence, target-absent trials +
105 high-prevalence, target-present trials +
189 low-prevalence, target-absent trials +
21 low-prevalence, target-present trials) +
12 ordered layout experiment two participants x
(105 high-prevalence, target-absent trials +
105 high-prevalence, target-present trials +
189 low-prevalence, target-absent trials +
21 low-prevalence, target-present trials)
= 10,080 trials

Hypothesis

H4. Participants in the ordered *image layout* condition will commit fewer errors than those in the random *image layout* condition in the low-prevalence image sets.

Error Rate Analysis

80 trials in which the search time was more than three standard deviations from the mean for that trial's *target prevalence / target presence* combination were counted as outliers and removed from our analysis. These outliers occurred when participants took an unscheduled break, dropped the keypad, or were otherwise interrupted and represent 1.5% of our data. As expected in a visual search experiment, *target presence* had a significant main effect on error rate ($F_{1,22} = 83.25$, $p < 0.001$). False-positives were again very rare, and the vast majority of errors occurred during target-present trials. The remainder of the analysis in this section was performed on the target-present trials only.

Target prevalence had a significant main effect on error rate ($F_{1,22} = 23.31$, $p < 0.001$), with error rates of 14.8% and 7.5% for low-prevalence and high-prevalence images respectively. While the between-participant variable *image layout* did not have a significant main effect on error rates ($F_{1,22} = 0.57$, $p = 0.46$), it did have a significant interaction with *target prevalence* ($F_{1,22} = 4.76$, $p = 0.04$). Figure 8 shows the mean error rates for each *target prevalence / image layout* combination. In the low-prevalence conditions, the random and ordered layouts had mean error rates of 17.3% and 12.2% respectively. This indicates that an ordered layout is beneficial for low-prevalence tasks, but may incur a penalty for high-prevalence tasks.

Search Time Analysis

People immediately stop searching when a target is found, so it was not surprising that target-present trials were on average shorter than target-absent trials and thus *target presence* had a significant main effect on search time ($F_{1,22} = 79.79$, $p < 0.001$). No other significant effects or interactions were observed relative to search time.

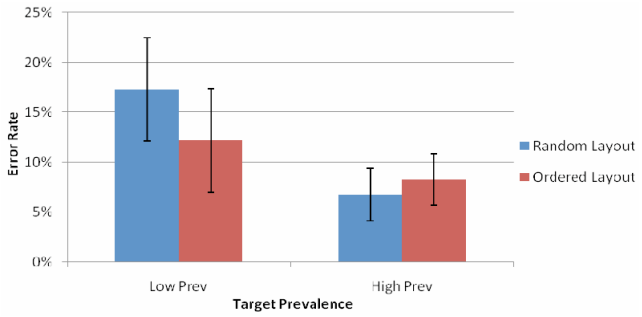

Figure 8: The interaction between target prevalence and image layout on error. An ordered layout helps users find rare targets, but incurs a penalty for high-prevalence tasks.

Discussion of Experiment Two

Once again, *target prevalence* was shown to significantly affect error rates, with infrequent targets resulting in larger error rates. In regards to H4, our participants committed fewer errors when viewing ordered images than when viewing random images in the low-prevalence conditions (Figure 8). This drop represents about a 30% improvement in the false-negative error rate. Additionally, this improvement seems to have occurred without cost in terms of search time, with our participants taking similar times to search through the random and ordered layout images.

TECHNIQUE THREE: SPACE/TIME TRADEOFF

When searching an image, a large fraction of time is spent moving one's eyes around the image and fixating them on potential targets. Humans can only fixate on a target every 250ms. The psychology literature provides ample evidence that humans are able to rapidly process visual images [4], with only a 200ms glimpse of an image being enough for recognition and other real world tasks. Because some types of image processing occur faster than gaze direction, eye movement becomes the limiting factor in many visual searching tasks (Figure 9). Our third presentation technique takes advantage of this characteristic by rapidly presenting the component pieces of an image over time in the same screen location. A user does not have to move their eyes, and can process many image pieces quickly.

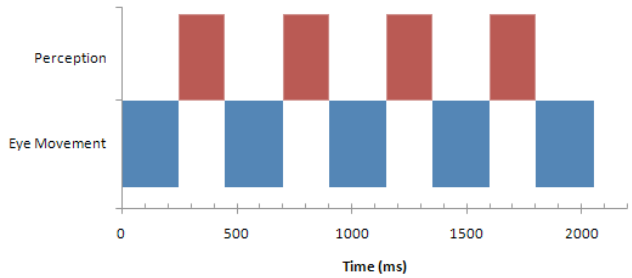

Figure 9: Significant time is spent moving one's eyes among targets when searching an image. RSVP techniques exploit this by presenting objects in the same location over time.

Figure 10 shows an overview of this technique. The original image is first segmented into its component pieces. Rather than recombining these pieces as in the previous two techniques, we keep them separated and present them sequentially to the user. Individual segments are small enough to be perceived *at a glance*, and the operator's eye can remain focused on a fixed location on the display. Similar to the re-layout technique described previously, this RSVP technique ensures that every item in the original image will be visited and that none will be skipped over as the user's eyes move from item to item around the display.

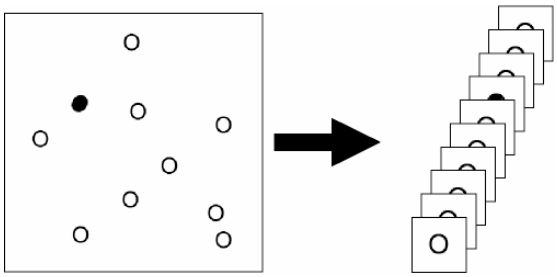

Figure 10: In the RSVP technique, images are segmented into their component pieces, which are themselves sequentially presented to the user in the same screen location.

One well known characteristic of RSVP presentation is known as "attentional blink" [18]. When viewing a rapid presentation of images, people often fail to observe a target when it is presented shortly after another target. The RSVP technique described in this section is susceptible to the attentional blink phenomenon; however, this is not necessarily a problem. Once the first target in the series had been identified, the entire set could be labeled as positive and the original images examined in detail.

Another danger lies in presentation pacing. Because the presentation of the image segments is system paced, visual distractions or physical blinking could interfere with search. It is also possible that the rapid presentation of images will simply overwhelm the user and cognitively interfere with search. Some RSVP systems provide the user with an interface for controlling the rate of presentation; however, the management of these controls may itself interfere with the cognitive facets of search. Given these dangers, it is unclear if this type of presentation will lead to lower error rates than searching the original images as a whole.

Experiment Three

The same 12 individuals who participated in experiment two took part in experiment three. With the goal of comparing the results of experiments one and three, we took the images from experiment one and segmented them into their component pieces. These individual letters were then displayed sequentially and participants again indicated the presence or absence of vowels.

At the start of each trial, a black cross indicated the region of the screen that the participant should look at. After a 2 second pause, the RSVP presentation of the 10 letters began. So that we could compare the results of this experiment to previous ones, participants had the same 2 seconds to view the letters that they had in experiment one

– this resulted in each letter being displayed on the screen for 200ms (Figure 11). At the end of the presentation, the screen became blank and participants entered their answer with the keypad used in experiment one. The same visual feedback and timing penalties encouraged accuracy and discouraged the careless racing through of the experimental trials. Each participant viewed a low-prevalence set of letters using the RSVP presentation technique.

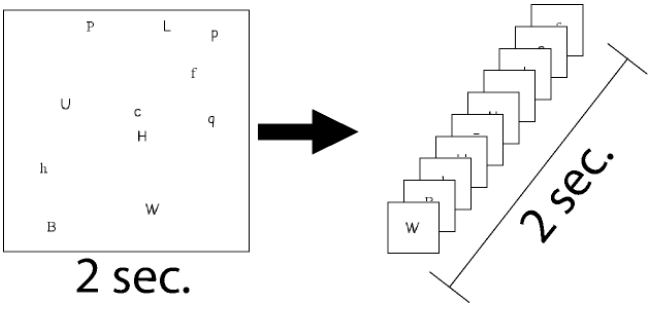

Figure 11: (left) Participants in experiment one viewed images such as this one for 2 seconds. (right) In experiment three, participants viewed the same collections of letters over the same 2 second period using an RSVP keyhole technique.

Method

We used a repeated measures design with *target presence* (target-present and target-absent) as a within-participant independent variable, *image layout* (random and RSVP) as a between-participant independent variable (where the random layout data came from experiment one) and *error rate* and *search time* as dependent variables. In summary, our design was:

12 random layout experiment one participants x
(189 low-prevalence, target-absent trials +
21 low-prevalence, target-present trials) +
12 RSVP layout Experiment Three participants x
(189 low-prevalence, target-absent trials +
21 low-prevalence, target-present trials)
= 5040 trials

Hypotheses

H5. Participants will commit fewer errors in the RSVP *image layout* condition than in the random *image layout* condition.

Error Rate Analysis

31 trials with search times of more than 3 standard deviations from the mean for that trial's *target presence* were treated as outliers and removed from our analysis. This was 1.2% of our data. Once again *target presence* had a significant main effect on error rate ($F_{1,22}$ = 70.93, $p < 0.001$), with false-positives being extremely rare. The remainder of the analysis in this section was performed on the target-present trials only.

The between-participant variable *image layout* had a significant main effect on error rate ($F_{1,22}$ = 13.29, $p = 0.001$). On average participants committed fewer errors when searching with the RSVP layout than with the random layout (with 17.3% and 7.1% error rates for random layout and RSVP layout respectively).

Search Time Analysis

Again, search time was recorded as the time between the presentation of the first letter and the participant's pressing of a key. There was a significant difference between the mean search times for each *layout technique* ($F_{1,22}$ = 54.46, $p < 0.001$), with mean search times of 1.29s and 1.98s for random layout and RSVP layout respectively. *Target presence* had both a significant main effect on search time ($F_{1,22}$ = 222.20, $p < 0.001$) and a significant interaction with *image layout* ($F_{1,22}$ = 18.04, $p < 0.001$). Figure 12 illustrates.

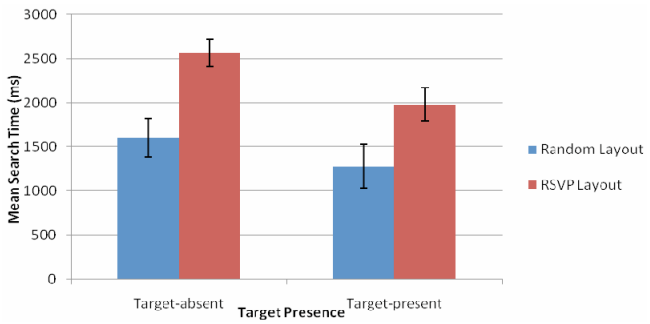

Figure 12: Participants took longer to search images in the RSVP *image layout* condition than in the Random *image layout* condition.

Discussion of Experiment Three

In regards to H5, our participants performed better using the RSVP *image layout* than they did using the random *image layout*. By simply changing the presentation of the items in the images, we observed a nearly 60% reduction in false-negative errors in this difficult, low-prevalence image set.

This improvement in error rate was not without cost. Participants took longer to search when using the RSVP *image layout* than with the Random *image layout* (Figure 12). This difference was not unexpected for the target-absent condition. During the RSVP trials, a participant had to watch the presentation of every letter in order to safely determine that there were no targets present. In essence, the system controlled the pacing as the participants had to wait for the RSVP presentation to finish before answering. The mean search time of about 2.5s for the RSVP, target-absent trials equates to the 2s of RSVP presentation plus 500ms of reation time to input one's answer. In the random *image layout* trials, participants were free to answer at any time within the alloted 2 seconds per image, and did so after some internal confidence threshold was passed.

GENERAL DISCUSSION

All three of the techniques presented in this paper lead to improved search performance, with reductions of false-negative error rates of 28%, 30%, and nearly 60% for Techniques One, Two, and Three respectively. This

provides strong evidence that the means of presenting the segmented components of an image can greatly affect the viewer's ability to find targets within images. While the motivation behind the design of each of these techniques is rooted in previously published literature from the fields of HCI and psychology, these motivations only provide a hypothesis as to why the observed differences occurred. To gain a better understanding of these techniques, further experimentation is needed to understand the underlying mechanisms as to *why* the observed improvements occur.

Image segmentation is at the core of our three search techniques, and any difficulties in the segmentation algorithm will likely result in a negative effect on search performance. One can partially addresses this problem by identifying borderline cases in the image segmentation process and handling them differently. For example, two objects A and B that the algorithm could not separate with a high-degree of probability would be displayed separately as A and B as well as together as AB. While this adds to the number of objects that must be investigated, it should compensate for many image segmentation errors.

In the following two sections, we discuss threats to the *internal* and *external validity* of the three experiments described in this paper. It is our hope that these comments help the reader interpret the results presented in the previous sections in a contextually appropriate manner.

Threats to Internal Validity

Participants in Experiments One and Two both searched two sets of 402x402 pixel images; however, Experiment One participants also searched a set of 900x900 pixel composite images. Experiment One participants effectively received extra "practice" images in the form of the composite images that may have either improved their performance on the smaller images or worsened it due to fatigue. How large is this danger? There was no effect in Experiment One from the order of presentation of the three experimental conditions, indicating that performing the composite image searches before the original image searches did not affect search performance. Additionally, there was no evidence that fatigue was a major factor in any of the experiments. These two factors match one's intuitive consideration of this danger: because the task was searching through English letters, it would be difficult to argue that extra practice would greatly affect an activity that native English speakers have been completing since childhood.

A similar issue exists in the design of Experiment Three. The participants in Experiment Three performed the RSVP search after completing the low and high-prevalence orderly search conditions for Experiment Two. There is again a danger that completing these activities would somehow change their performance relative to those participants in Experiment One. Again, it would be difficult to argue that participants became significantly better at identifying vowels over the course of an hour-long experiment; however, the danger does exist. In terms of the effect of this extra "practice" on the overall results and conclusions of the study, the effect size of the RSVP technique studied in Experiment Three was very large (nearly 60%), so it is unlikely that any confound that did exist would alter the relative ranking of normal vs. RSVP search or the overall conclusions of Experiment Three.

With only 12 participants per group, individual differences among the participants will likely manifest themselves in terms of task performance. A more powerful within-subject design in which participants completed both the orderly and random layout searches would result in a more powerful statistical test. While a limited number of participants is always a cause for concern, the statistical model employed in the analysis of Experiment Two takes the number of participants (24) into consideration when determining significance. Thus, the results can safely be interpreted as differences between conditions, not among subjects.

Threats to External Validity

While cell slide pathology is one example of a search activity that is amenable to the type of segmentation and display presented in this paper, some types of search may be negatively affected by the removal of context from each segment. Context has been shown to impact search performance in a number of search tasks [16], as viewers are directed toward the target by other objects in the scene. Future work is needed to investigate the tradeoff between the negative effects of removing target context and the positive effects of the approach presented in this paper.

As with any highly abstracted experimental task, there is a danger that the task used in our experiments may differ significantly from real work activities that involve image search. Such concerns about the *external validity* of our experiments are slightly assuaged by the very similar error rates observed in Experiment One and previously published results. Wolfe et al. [27] reported error rates of 16% and 7% for the 10% and 50% prevalence conditions in their experiment that included a simulated baggage screening task. These are very similar to our measured error rates of 17.3% and 6.7% for our 10% and 50% prevalence simulated cell side pathology tasks, and this similarity suggests a high level of *experimental reliability* and a more universal relationship between target-prevalence and error rate that may extend to many search activities.

Given the results presented in this paper, we feel that there is sufficient evidence to justify further experimentation in which the participant population and tasks better match real work conditions. While we attempted to simulate some of the motivations that practitioners feel, this is impossible to do completely in the lab. No experiment can simulate the pressure one feels when making consequential decisions about patient care or airline safety. Similarly, one cannot reproduce the pressure a pathologist must feel to complete a diagnosis with an ever increasing number of patients and tests, or the pressure that a baggage screener feels when facing a long line of frustrated customers at the airport.

CONCLUSION

We have presented three novel image presentation techniques designed to improve visual search of segmentable images and provided initial evidence that these techniques improve people's ability to accurately search for target items. These techniques are similar to one another in that they all use the component segments of original images in novel ways to improve visual search performance. While these three design variations were motivated by an understanding of search behavior and previously published results, they are by no means the only techniques in this design space. It is quite likely that there are other presentation methods in this space that will lead to better visual search performance than viewing the original images themselves. It is our hope that this paper might inspire the interested reader to contemplate this general approach when attempting to design user interfaces that support people performing this important class of visual search tasks.

REFERENCES

1. Adami, H., Pontén, J., Sparén, P., Bergström, R., Gustafsson, L., & Friberg, L. (1994). Survival trend after invasive cervical cancer diagnosis in Sweden before and after cytologic screening. 1960-1984. *Cancer*, 73(1). pp 140-147.
2. AForge.NET. http://code.google.com/p/aforge/
3. Beck, M., & Trafton, J. G. (2007). Local spatial layout consistency affects strategies but not memory during Visual Search. *Journal of Vision*, 7(9):1060.
4. Coltheart, V. (Ed.) (1999). *Fleeting memories: Cognition of brief visual stimuli*. MIT Press.
5. Ellis, S. R., & Stark, L. (1986). Statistical dependency in visual scanning. *Human Factors*, 28(4), 421-438.
6. Fleck, M. S., & Mitroff, S. R. (2007). Rare targets are rarely missed in correctable search. *Psychological Science*, 18, 943-947.
7. Forlines, C., Shen, C., Wigdor, D., & Balakrishnan, R. (2006). Exploring the effects of group size and display configuration on visual search. In *Proc. of Computer Supported Cooperative Work*. pp. 11-20.
8. Gilchrist, I. and Harvey, M. (2006). Evidence for a systematic component within scan paths in visual search. *Visual Cognition*, 14(4):704-715.
9. Groopman, J. 2008. *How Doctors Think - Chapter 8: The Eye of the Beholder*. Houghton Mifflin Company, New York. ISBN-13:978-0-618-61003-7.
10. Halverson, T. & Hornof, A. J. (2007). A minimal model for predicting visual search in human computer interaction. In *Proc. of ACM CHI Conference on Human Factors in Computing Systems*. pp. 431-434.
11. Halverson, T. (2006). Integrating models of human computer visual interaction. In *Extended Abstracts of the ACM CHI Conference on Human Factors in Computing Systems*. pp. 1747-1750.
12. Hornof, A. J., & Halverson, T. (2003). Cognitive strategies and eye movements for searching hierarchical computer displays. In *Proc. of ACM CHI Conference on Human Factors in Computing System*. pp. 249-256.
13. Keenlyside, R. A., Collins, C. L., Hancock, J. S., Gagnon, M. C., Cohn, R. D., Menoff, A. L., Dodd, L. G., Kurtycz, D. F., Hearn, T. L., & Baker, Jr. E. L. (1999). Do proficiency test results correlate with the work performance of screeners who screen Papanicolaou smears? *American Journal of Clinical Pathology*, 112(6). pp. 769-776.
14. Krendel, E. S., & Wodinsky, J. (1960). Search in an unstructured visual Feld. *Journal of the Optical Society of America*, 50, 562-568.
15. Mello-Thoms, C., Nodine, C. F., & Kundel, H. L. (2002). What attracts the eye to the location of missed and reported breast cancers? *ACM ETRA Symposium on Eye Tracking Research & Applications*. pp. 111-117.
16. Pomplun M. (2006) Saccadic selectivity in complex visual search displays. *Vision Research.* Jun;46(12):1886-1900.
17. Pham, D., Xu, C., & Prince, J. (2000). Current methods in medical image segmentation. *Annual Review of Biomedical Engineering*. pp. 15-37.
18. Raymond, J., Shapiro, K., Arnell, K. (1992). Temporary suppression of visual processing in an RSVP task: an attentional blink?. *Journal of Experimental psychology. Human perception and performance*, 18 (3): 84960.
19. Renshaw, A. (2001). Measuring and Reporting Errors in Surgical Pathology. *American Journal of Clinical Pathology*, 115. pp 338-341.
20. Roerdink, J., & Meijster, A. (2001). The Watershed transform: definitions, algorithms and parallelization strategies. *Fundamenta Informaticae* 41. pp. 187-228.
21. Scinto, L., Pillalamarri, R., & Karsh, R. (1986). Cognitive strategies for visual search. *Acta Psychologica*, 62, 263-292.
22. Simonin, J., Kieffer, S., & Carbonell, N. (2005). Effects of display layout on gaze activity during visual search. In *Proc. of INTERACT Conference*. pp. 1054-1058.
23. Spence, R. (2002) Rapid, serial and visual: a presentation technique with potential. *Information Visualization*, 1 (1). Pp. 13-19.
24. Wittenburg, K., Forlines, C., Lanning, T., Esenther, A., Harada, S. & Miyachi, T. (2003). Rapid serial visual presentation techniques for consumer digital video devices. In *Proc. of ACM UIST*. pp. 115-124.
25. Wolfe, J. M. (1998). Visual search. In H. Pashler (Ed.), *Attention*, East Sussex, UK: Psychology Press.
26. 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 Psychology: General.*
27. Wolfe, J. M., Horowitz, T. S. & Kenner, N. (2005) Rare items often missed in visual searches. *Nature*, 435. pp. 439-440.