Running Head: Domain-specific expertise and initial scene processing

Worth a quick look? Initial scene previews can guide eye movements as a function of domain-specific expertise but can also have unforeseen costs

Damien Litchfield[1] *

Tim Donovan[2]

[1]Department of Psychology

Edge Hill University, UK

[2]Medical & Sport Sciences,

University of Cumbria, UK

---

*Corresponding author:

Department of Psychology, Edge Hill University, Ormskirk, L39 4QP, UK

Email: damien.litchfield@edgehill.ac.uk

Tel: +44 (0)1695 584085

Abstract

Rapid scene recognition is a global visual process we can all exploit to guide search. This ability is thought to underpin expertise in medical image perception yet there is no direct evidence that isolates the expertise-specific contribution of processing scene previews on subsequent eye movement performance. We used the flash-preview moving window paradigm (Castelhano & Henderson, 2007) to investigate this issue. Expert radiologists and novice observers underwent 2 experiments whereby participants either saw a 250ms scene preview or a mask before searching for a target. Observers looked for everyday objects from real-world scenes (Experiment 1), and searched for lung nodules from medical images (Experiment 2). Both expertise groups exploited the brief preview of the upcoming scene to more efficiently guide windowed search in Experiment 1, but there was only a weak effect of domain-specific expertise in Experiment 2, with experts showing small improvements in search metrics with scene previews. Expert diagnostic performance was better than novices in all conditions but was not contingent on seeing the scene preview, and scene preview actually impaired novice diagnostic performance. Experiment 3 required novice and experienced observers to search for a variety abnormalities from different medical images. Rather than maximising the expertise-specific advantage of processing scene previews, both novices and experienced radiographers were worse at detecting abnormalities with scene previews. We discuss how restricting access to the initial glimpse can be compensated for by subsequent search and discovery processing, but there can still be costs in integrating a fleeting glimpse of a medical scene.

Keywords: medical image perception, scene perception, eye movements, flash-preview moving window, expertise

Introduction

Detecting significant clinical findings from medical images is a key component of how expert practitioners make life-saving decisions (Beam, Krupinski, Kundel, Sickles, & Wagner, 2006; Field, 2014). Since medical image perception is a difficult task, even for expert radiologists, research over the last few decades has sought to understand what influences performance and what are the perceptual and cognitive reasons for why errors of up to 30% still persist (Krupinski, 2010). This body of research was largely based on the search for cancer from static chest radiographs and mammograms, but to address 21st century developments is now exploring a range of digital (Jaarsma, Jarodzka, Nap, Merrienboer, & Boshuizen, 2014; Krupinski et al., 2006) and volumetric imaging modalities (Bertram, Helle, Kaakinen, & Svedström, 2013; Drew et al., 2013b; Phillips et al., 2013). Nevertheless, at the heart of prevailing models of medical image perception (e.g., Nodine & Kundel, 1987; Kundel, Nodine, Conant, & Weinstein, 2007) is that within the first glimpse, the expert observer holistically processes the medical image and subsequently makes efficient search-related eye movements to potentially abnormal areas to support diagnostic decision-making.

The importance of the initial glimpse in relation to diagnostic accuracy was realized early on by Kundel and Nodine's (1975) tachistoscopic experiments, in which they found that experts could correctly detect 70% of abnormal images, even though such images were only presented for 200ms (Carmody, Nodine & Kundel, 1981; Evans, Georgian-Smith, Tambouret, Birdwell, & Wolfe, 2013; Mugglestone, Gale, Cowley & Wilson, 1995). The idea that holistic processing was integral to expert performance was also established by experiments that disrupted holistic processing, by requiring search through segmented (Carmody, Nodine, & Kundel, 1980) or rotated images (Oestmann, Greene, Bourgouin, Linetsky, & Llewellyn, 1993). Similarly, the efficiency in which expert observers search through medical images has been well documented, in that compared to less experienced

observers, experts are more likely to find abnormalities and do so faster and with fewer eye movements (Donovan & Litchfield, 2013; Kundel, Nodine, & Carmody, 1978; Kundel & La Follette, 1972; Kundel, Nodine, Krupinski, & Mello-Thoms, 2008; Manning, Ethell, Donovan, & Crawford, 2006). These enhancements are domain-specific in nature, as although expert radiologists may have better sensitivity in medical image discrimination tasks compared to novices (Sowden, Davies, & Roling, 2000), experts do not perform any better at general visual search tasks (Nodine & Krupinski, 1998).

One of the key principles of the holistic model (Kundel et al., 2007) is that the rapid initial holistic processing helps constrain search to suspicious areas in the image, and that with increasing expertise in medical image perception it is more likely that guidance towards abnormalities will be based on initial holistic processing. However, whilst there is supporting evidence for two distinct streams of information processing, 1) rapid initial holistic processing, and 2) slower processing relating to search and discovery (Kundel et al. 2008), how these two processes interact so as guide subsequent eye movements is not well understood. Thankfully, alongside this account of medical image perception, numerous psychological and computational models (e.g., Torralba, Oliva, Castelhano, & Henderson, 2006; Wolfe Evans, Võ, & Greene, 2011) have also been investigating global and local processing to address how observers are able to rapidly recognise the scene category, or 'scene-gist', and infer what objects would be in such scenes, where they are likely to be located, and how the initial glimpse of a scene guides real-world search (Biederman, Mezzanotte, & Rabinowitz, 1982; Greene & Oliva, 2009). Indeed, there is substantial overlap in the literature on scene perception and medical image perception (for a recent review see Drew, Evans Võ, Jacobson, & Wolfe, 2013a).

One of the main problems with the holistic model is that there is no direct evidence that isolates the specific contribution of the initial scene preview on subsequent eye movement

performance as a function of expertise. Instead, inferences are made based on how observers perform under tachistoscopic conditions where eye movements are prevented, or by analysing time-to-first fixation data. For example, data from several mammography studies showed that 57% of all cancers were fixated within 1 second, whereas the remaining cancers fixated in subsequent search (Kundel et al. 2008). However, we recently showed that with chest x-rays, only 33% of cancers were fixated within 1 second, whereas 56% of cancers were fixated within 2 seconds (Donovan & Litchfield, 2013). As a result, we cannot always equate the visual processing across imaging modalities. A more problematic issue with time-to-first fixation data is that it is obtained from eye tracking experiments under free viewing conditions, whereby the observer has constant access to the whole scene via peripheral vision, making it difficult to isolate the specific contribution of the initial scene preview on subsequent eye movement behaviour (Donovan & Litchfield, 2013). In the present study we make use of the recently developed gaze contingent 'flash-preview moving window' (FPMW) paradigm (Castelhano & Henderson, 2007) as it dissociates the initial scene representation from the ongoing scene representation obtained during search.

In the FPMW paradigm observers are shown a brief preview of a scene (or control) and then asked to search for a target object within the same scene whilst their peripheral vision is restricted to a small gaze-contingent window. Typically, target objects are detected faster with scene previews as eye movement metrics reflect greater efficiencies in initiating and executing windowed search, and this suggests that the initial representation of the scene can be retained in memory and used to plan subsequent eye movements (see also Hollingworth, 2009). These improvements in search are thought to be the product of the initially generated scene representation interacting with the target knowledge activated from viewing the presented target word (Hillstrom, Schloley, Liversedge, Benson, 2012), or picture (Castelhano & Heaven, 2010). Moreover, the scene preview benefit exists even when the

target object is not visible during the preview but only found through windowed search, thereby confirming the benefit of scene-context processing, irrespective of any additional local target processing that could occur when targets are actually present during previews (Castelhano & Henderson, 2007; Võ & Henderson, 2010).

So far the FPMW paradigm has helped reveal the time-course of the initial representation (Hillstrom, et al., 2012; Võ & Henderson, 2010), the individual differences associated with initial scene processing (Võ & Schneider, 2010), and the extent to which semantically consistent objects are processed within scenes (Castelhano & Heaven 2011; Võ & Henderson, 2011). Crucially, however, it is our experience with specific scenes and objects that allows initial scene processing to be exploited, for subsequent eye movements to be guided more effectively towards task relevant areas, and for decisions to be made faster as a function of expertise (Gegenfurtner, Lehtinen, & Säljö, 2011; Reingold & Sheridan, 2011). To our knowledge, no actual study has been conducted using the FPMW paradigm using experts and novices, and thereby directly confirming whether domain-specific knowledge contributes to the effective processing of the initial representation.

Using the FPMW paradigm we compare the performance and eye movement behaviour of expert radiologists and novice observers (psychology students) as they search for everyday objects from real-world scenes (Experiment 1), or lung nodules from chest x-ray images (Experiment 2). We also compare novice observers and experienced radiographers as they search through a variety of medical image types looking for different pathologies (Experiment 3). Across all experiments, the guiding rationale is that only when viewing familiar scenes would prior knowledge facilitate key processing decisions. Assuming that initial scene processing would be exploited based on domain-specific expertise (Donovan & Litchfield, 2013; Drew et al., 2013a; Kundel et al., 2007; Wolfe et al., 2011), it is expected that expert radiologists / experienced radiographers will be faster than novices at detecting

targets and show more efficient eye movement behaviour when presented with scene previews before search (for both real-world scenes and medical images). However, it is expected that novices will only benefit from scene previews when viewing scenes that they are highly familiar with (real-world scenes). Note that in all experiments the target will be visible during the preview. Whilst this means that scene guidance and local target processing could facilitate search within these scenes, these effects should be additive, and therefore lead to a stronger (if not purer) scene preview benefit compared to mask preview. It is our intention that if stronger scene preview effects can first be established then subsequent studies can further isolate the contribution of the respective processing mechanisms.

Experiment 1

Method

Participants

There were 28 participants consisting of 14 experts (13 male; mean age = 46.8 years) and 14 novices (9 male; mean age = 21.2.years). Experts were all board-certified consultant radiologists for the NHS with a minimum of 10yrs medical image perception experience, whereas the novices were all psychology undergraduates with no experience of medical images or the nodule detection task. All participants had normal or corrected-to-normal vision, and all participants completed Experiment 1 followed immediately by Experiment 2.

Stimuli and apparatus

For Experiment 1, the stimuli were 40 full-colour photographs of real-world scenes taken from the LabelMe database (Russell, Torralba, Murphy, & Freeman, 2008), a repository of copyright-free images. Half were indoor scenes (e.g., kitchens, offices, living rooms) and half were outdoor scenes (e.g., streets, parks, coastlines). All scenes were

presented on a 19-in. CRT monitor (1024 x 768 pixels, 120 Hz). The scenes subtended 24.24° x 18.18° of visual angle when viewed from 57 cm. Each scene contained a single unique target object with an average size of 3.58° x 3.55°. Across all scenes there was an equal probability of the target occurring on either the left or right side of the image. The mask preview was created in Adobe Photoshop and consisted of a random array of coloured pixels (Experiment 1) or greyscale pixels (Experiments 2 & 3).

Eye movements were recorded using an EyeLink 1000 desktop eye tracker (SR Research Ltd, Mississauga, Canada) and stimuli were presented via Experimental Builder software. Calibration points of eye position were only accepted if they had an average resolution less than 0.5° visual angle. Rectangular interest areas were created that best fit each target object. The key performance metrics were Response Time (RT), defined as the time from the onset of the windowed search screen until button press, and Accuracy (% targets correctly identified). To assess the efficiency of search as a function of preview and expertise, we examined the time until first target fixation (search latency) and number of fixations until first target fixation. We also examined the initial saccadic latency and initial saccadic amplitude of the first eye movement as these measures represent the first response relating to the rapid processing of the scene preview and the readiness to initiate search (Hillstrom, et al., 2012; Võ & Henderson, 2010; for a recent medical imaging equivalent, see Pietrzyk, McEntee, Evanoff, Brennan, & Mello-Thoms, 2014).

Procedure

Eye movements were calibrated using a 9 point calibration and validation. Participants were instructed that they would have up to 15 seconds to search for a target from a real-world scene under windowed viewing conditions, and that on some trials they may be shown a brief glimpse of the upcoming scene before commencing search. Figure 1 indicates the trial

sequence for each condition using the FPMW paradigm. Participants were presented with a black target word indicating the identity of the target object for 1500ms. Then a fixation cross for 200ms, and then either a scene preview or mask preview for 250ms, followed by a mask for 50ms. Following a second fixation cross for 400ms, windowed search of the scene began. A 2.5° radius window was used to restrict the field-of view during search and to detect a target participants pressed a gamepad button whilst directly fixating the target. By presenting the target word before the scene preview, we tried to control for the fact that in Experiment 2 the target object (lung nodule) was always known before scenes were presented. Four separate practice trials were presented beforehand to familiarise participants with the procedure followed by 2 blocks of 20 trials. Targets were only considered correctly identified if a fixation was within the target AOI during button press. Participants only saw each scene once and so scene/condition combinations were counterbalanced across participants. Trials were presented in a randomized order for each participant and the whole experiment took approximately 20min.

<< Insert Figure 1 about here >>

Results

All data were subjected to a 2 x 2 mixed measures ANOVA with preview (scene, mask) as a within-participant factor, and expertise (novice, expert) as a between participant factor. For all measures only trials with correct responses were analysed. With the exception of accuracy, we expected all response metrics would be more efficient with scene preview than mask preview, but that there would be no difference between the expertise groups on any of these responses. A summary of means can be seen in Table 1.

<< Insert Table 1 about here >>

Performance

Overall search accuracy averaged 88% (ranging from 60% to 100%) but did not differ as a function preview, $F(1, 26) = 2.62$, $p = .12$, $\eta^2 = .09$, or expertise, $F(1, 26) = .03$, $p = .87$, $\eta^2 < .01$, and there was no interaction, $F(1, 26) = 1.48$, $p = .24$, $\eta^2 = .05$. RTs averaged 3895ms across conditions and there was a main effect of preview, $F(1, 26) = 77.46$, $p < .001$, $\eta^2 = .74$. However, there was no main effect of expertise, $F(1, 26) = .34$, $p = .86$, $\eta^2 < .01$, and there was no interaction, $F(1, 26) = 0.60$, $p = .48$, $\eta^2 = .01$. For both groups of observers RTs were faster for scene preview than mask preview.

Search-related eye movements

Search latency averaged 2961ms across conditions and there was a main effect of preview, $F(1, 26) = 115.94$, $p < .001$, $\eta^2 = .81$. However, like the RT measures, there was no main effect of expertise, $F(1, 26) = .11$, $p = .74$, $\eta^2 < .01$, and no significant interaction, $F(1, 26) = 1.58$, $p = .22$, $\eta^2 = .01$. The number of fixations averaged 10.91 across conditions and there was a main effect of preview, $F(1, 26) = 126.95$, $p < .001$, $\eta^2 = .82$. However, there was no main effect of expertise, $F(1, 26) = .11$, $p = .74$, $\eta^2 < .01$, and no significant interaction, $F(1, 26) = 1.28$, $p = .72$, $\eta^2 = .01$. All these search metrics indicate that targets were identified faster and in fewer eye movements for scene preview than mask preview, but were not modulated in any way by expertise group.

First eye movement of search

The initial saccadic latency averaged 236ms across conditions and there was a main effect of preview, $F(1, 26) = 28.99$, $p < .001$, $\eta^2 = .53$, but no effect of expertise, $F(1, 26) = .15$, $p = .70$. $\eta^2 = .01$, and no interaction, $F(1, 26) = .08$, $p = .78$, $\eta^2 < .01$. The initial saccadic

amplitude averaged 2.03° visual angle across conditions and there was a main effect of preview, $F(1, 26) = 141.36$, $p < .001$, $\eta^2 = .84$. Once again however there no main effect of expertise, $F(1, 26) = 80$, $p = .38$, $\eta^2 = .03$, and no interaction, $F(1, 26) = .12$, $p = .73$, $\eta^2 < .01$. The first eye movement of search following a scene preview was both faster and of larger amplitude than following a mask preview, however, once again there was no difference across expertise groups.

Discussion

In line with previous research (e.g., Castelhano & Henderson, 2007; Võ & Henderson, 2010), all metrics indicated search for everyday objects from real-world scenes was more efficient when participants were presented with a scene preview than a mask preview. Following a 250ms glimpse of the upcoming scene, participants were quicker to initiate search and faster to fixate and identify the target. Importantly, there was no difference between the two expertise groups on any of the measures in Experiment 1. This is consistent with previous comparisons studies between experts and lay observers (Nodine & Krupinski, 1998), which show that experts in medical image perception do not demonstrate superior visual processing in search tasks outside their domain-specific expertise. Moreover, our results suggest novice observers were exploiting the initial glimpse of the scene in the exact same manner as expert radiologists. Experiment 2 attempts to isolate the expertise-dependent contribution of initial scene processing and eye guidance by requiring the same participants to this time search for lung nodules from chest-x-rays using the FPMW paradigm.

Experiment 2

Method

Participants. These were all the same participants from Experiment 1

Stimuli and apparatus

A testbank of 60 chest x-ray images were used in Experiment 2, 36 images were abnormal and contained a single nodule located within the lung fields, and 24 images did not contain a nodule. Nodules were defined as discrete opacities in the lung field or mediastinum measuring between 5–30mm in diameter, and all nodules were histopathologically proven. The chest x-ray images subtended 22.48° x 20.44°, and lung nodules had an average size of 3.30° x 3.30°. We have successfully used this testbank in previous studies to establish expertise-related differences in search (Donovan & Litchfield, 2013), and how search-related eye movement behaviour can be used as learning cues for other observers (Litchfield, Ball, Donovan, Manning, & Crawford, 2010). For methodological reasons, medical image perception research typically adopts a 50% prevalence rate as this helps characterise observer performance, as it is important to not only detect targets, but to refrain from making false positive decisions on normal images. Since the primary interest of the present study is how quickly observers identified abnormalities to maximise the number of valid samples in the final analysis we adopted a prevalence rate of 60% in Experiment 2, and to be consistent with Experiment 1, we adopted a 100% prevalence rate in Experiment 3. It should be noted, however, that the prevalence rate in clinical settings can be substantially lower than this, and that low prevalence is a contributing factor as to why such targets are missed in medical image perception (Nakashima, Kobayashi, Maeda, Yoshikawa, & Yokosawa, 2013; see also Wolfe, Brunelli, Rubinstein, & Horowitz, 2013).

Procedure

Although the timings in which the medical images were shown were identical to Experiment 1, adapting the FPMW paradigm to a medical imaging task meant that there were two key differences in the experiment. One of the primary differences was that Experiment 2

included normal images that did not contain a target. Participants were told that if they did not identify a target within the 15 second maximum limit, then that trial would be coded as normal. As such, if observers finished searching the image within this time and believed the image was normal then they should allow the timer run out. This timeout feature was a key logistical constraint that has been used previously (e.g., Carmody et al., 1981), and it ensured that observers only had one response button to press if a nodule was detected and so were not making additional response compatibility judgements during this already demanding task.

The second key difference was that the search targets (lung nodules) were much more difficult to find in Experiment 2. Consistent with FPMW studies, the search targets in Experiment 1 were everyday objects that once fixated are easily recognised. In contrast, lung nodules are notoriously difficult to correctly identify, even in free viewing tasks. It is therefore standard practice in medical imaging tasks to obtain a confidence rating for each decision so that such information can be used in receiver operating characteristic (ROC) analysis to more accurately characterise performance (Chakraborty, & Berbaum, 2004; Green & Swets, 1966). Accordingly, once a participant had identified what they thought to be a nodule by pressing the gamepad button whilst looking at the suspected target, before proceeding to the next image participants were required to provide a 1-4 confidence rating (4 being highly confident) regarding their decision.

As the novice observers had no experience with medical images or lung nodules, before beginning the experiment two practice chest-rays containing a nodule were first shown to participants so that they understood what targets they were looking for. A further two separate practice trials (1 abnormal, 1 normal) were presented using the FPMW paradigm beforehand to help participants familiarise themselves with the modified procedure. Participants only saw each image once and conditions were counterbalanced across

participants. Trials were presented in 2 blocks of 30 trials in a randomised order and the whole experiment took approximately 30min.

Results

All data were subjected to a 2 x 2 mixed measures ANOVA with preview (scene, mask) as a within-participant factor, and expertise (novice, expert) as a between-participant factor. For all measures, analysis was restricted to only trials where targets were correctly detected. Since Experiment 2 involves a medical imaging task, we provide 2 analyses of diagnostic performance. To provide a clear comparison between Experiment 1 and previous FPMW studies, we first report accuracy levels based on the % of target nodules correctly identified (i.e., ignoring true/false negative decisions on normal images). We then report observer performance based on jackknife free-response ROC (JAFROC) analysis, which has been validated as a more sensitive measure of diagnostic decision-making than ROC (Chakraborty, & Berbaum, 2004). JAFROC was calculated using the freely available RJafroc software (http://www.devchakraborty.com) which uses the number of true positives and false positives observers report and their respective confidence ratings to produce a single figure of merit for each observer. This figure of merit represents the likelihood that a true positive will be given a higher confidence rating than a false positive. This single measure is superior to ROC because it simultaneously takes into account decision confidence and location information. Traditional ROC does not take into account location information but simply requires the observer to state whether the image contains an abnormality or not, without actually having to localize it. This can lead to problematic situations where an observer views an abnormal image, reports the image is abnormal, and so according to ROC the observer is making a correct decision (i.e., true positive). However, that observer could be stating the image is abnormal based on the wrong information (i.e., thinking normal anatomy is an abnormality)

and so ROC can overestimate observer performance. With JAFROC, a decision is only counted as a true positive if the observer correctly identified the location of the abnormality. Consistent with Experiment 1 and other FPMW studies, a correct response is determined by the participant directly fixating the suspected target and pressing the response button. JAFROC uses a chance level of .50 and we have previously shown that expert performance is usually represented by a figure-of-merit of .75 or above (Donovan & Litchfield, 2013; Litchfield et al., 2010). A summary of means can be seen in Table 2.

<< Insert Table 2 about here >>

Performance

Overall accuracy averaged just 58% (novices ranging from 22% to 82%; experts ranging from 44% to 88%). There was no main effect of preview, $F(1, 26) = 3.29$, $p = .08$, $\eta^2 = .11$, and there was no interaction, $F(1, 26) = .91$, $p = .35$, $\eta^2 = .03$. There was however, a main effect of expertise, $F(1, 26) = 21.95$, $p < .001$, $\eta^2 = .46$, with experts detecting more nodules (70%) than novice observers (46%). Moreover, performance assessed by JAFROC revealed a number of significant effects. There was a main effect of preview, $F(1, 26) = 6.40$, $p = .018$, $\eta^2 = .15$, a main effect of expertise, $F(1, 26) = .20.25$, $p < .001$, $\eta^2 = .44$, and a significant preview x expertise interaction, $F(1, 26) = 7.22$, $p = .012$, $\eta^2 = .19$, with experts better at detecting nodules than novice observers.

In unpacking the interaction, the simple main effect analyses revealed some rather surprising findings. Experts performed better than novices in both the scene preview condition, $F(1, 26) = 26.42$, $p < .001$, $\eta^2 = .50$, and mask preview condition, $F(1, 26) = 12.74$, $p < .001$, $\eta^2 = .34$. However, denying experts the initial glimpse of the image did not impact on their performance as there was no difference between the scene preview and mask

preview conditions, $F(1, 26) = .01$, $p = .91$, $\eta^2 < .01$. In contrast, novice observers actually performed better in the mask preview condition than the scene preview condition, $F(1, 26) = 13.61$, $p = .001$, $\eta^2 = .52$. As mentioned above, the JAFROC analysis is a much more sensitive measure of diagnostic performance than simple accuracy levels. We return to potential explanations of these findings in the discussion. Finally, RTs averaged 7291ms across conditions and there was no main effect of preview, $F(1, 26) = 2.85$, $p = .10$, $\eta^2 = .10$, no main effect of expertise, $F(1, 26) = .34$, $p = .86$, $\eta^2 = .06$, and there was no interaction, $F(1, 26) = 0.01$, $p = .92$, $\eta^2 < .01$.

Search-related eye movements

The overall search latency averaged 4499ms across conditions and there was a borderline main effect of preview, $F(1, 26) = 4.19$, $p = .051$, $\eta^2 = .13$. However, there was no main effect of expertise, $F(1, 26) = 1.40$, $p = .25$, $\eta^2 = .05$, and no significant interaction, $F(1, 26) = 1.02$, $p = .32$, $\eta^2 = .03$. The overall number of fixations averaged 15.28 across conditions and there was a main effect of preview, $F(1, 26) = 4.96$, $p = .035$, $\eta^2 = .16$, but no main effect of expertise, $F(1, 26) = .48$, $p = .50$, $\eta^2 = .02$, and no interaction, $F(1, 26) = .68$, $p = .44$, $\eta^2 = .02$. Contrary to our predictions, for novices as well as experts, the scene preview led to faster search latencies and fewer fixations compared to mask preview conditions.

First eye movement

Initial saccadic latency averaged 233ms across conditions but there was no main effect of preview, $F(1, 26) = 2.52$, $p = .12$, $\eta^2 = .09$, expertise, $F(1, 26) = .63$, $p = .44$. $\eta^2 = .02$, and no interaction, $F(1, 26) = .32$, $p = .57$, $\eta^2 = .01$. Likewise, the initial saccadic amplitude averaged 2.03° across conditions but there was no main effect of preview, $F(1, 26) = 1.56$, $p = .22$, $\eta^2 = .06$, expertise, $F(1, 26) = .79$, $p = .38$, $\eta^2 = .03$, and no interaction, $F(1, 26) = .04$,

p = .85, η^2 < .01. These measures indicated there was no difference in the speed or amplitude of the first eye movement of search following a scene preview compared to mask preview, or any influence of domain-specific expertise.

Discussion

One of the first issues to note is that for both groups of observers, the accuracy in detecting targets in this task was much lower than Experiment 1. Nevertheless, as one would expect, novices performed much worse than experts in detecting nodules from chest x-rays images. In addition, our JAFROC analysis confirmed that across all conditions, experts (M = .64) outperformed novices (M = .53) but also that novice observers actually performed worse in the scene preview condition (M = .51) compared to the mask preview condition (M = .55). We have shown in previous research (e.g., Donovan & Litchfield, 2013) that novices searching for lung nodules are more likely to fixate regions in the image that contain nodule-like distractors, but which are in fact normal anatomy (e.g., the hilar and mediastinum). Since rapidly distinguishing normal anatomy from pathology is a hallmark of expertise, one potential explanation for why novices in Experiment 2 performed worse in the scene preview condition is that whilst encoding the initial scene representation, novices may have been biased towards these distractors regions, which experts with years of experience would have learned to attenuate. Indeed, novices made significantly more false positives in scene preview (M = 9.57) than mask preview (M = 7.43), $t(13) = 5.30$, $p < .001$, Cohen's d = .38) which would have contributed to the lower JAFROC figure-of-merit. In contrast, expert observers showed no such difference in false positives between scene preview (M = 4.07) and mask preview (M = 4.21), $t(13) = -.20$, p = .85 Cohen's d = .04). This difference in false positive rates for scene preview confirms that the performance impairment in scene preview by

novices was not because they thought abnormal images were normal and therefore gave up search, but rather that they mistook normal features for pathology.

One of the key aims of this study was to establish the expertise-based contribution of initial scene processing on diagnostic and search performance. In applying the FPMW paradigm to medical imaging, we were surprised to find that experts showed no advantage in diagnostic performance (either in accuracy or JAFROC) in the scene preview condition. Providing an initial glimpse of the scene did not appear to contribute to expert performance, and as discussed above, actually reduced novice performance. Moreover, in the mask preview condition performance can only be attributed to slower processing relating to search and discovery, and not rapid initial holistic processing (Kundel et al. 2008). The fact that experts were better than novices in the mask preview condition but showed no greater advantage with scene previews indicates that the importance of search and discovery should not be underestimated as a marker of expert diagnostic performance.

Previous research has used eye-tracking measures such as time-to-first-fixation as an indirect measure of rapid holistic processing (Kundel et al. 2008). The FPMW paradigm provides a more rigorous manipulation of the initial scene preview on subsequent eye movement performance. As shown in Table 2, we found that both novices and experts fixated nodules faster (M = 4530ms, M = 3956ms) and in fewer eye movements (M = 15.04, M = 13.64) when provided with an initial glimpse of the upcoming medical image. This suggests that whilst there was not a diagnostic advantage of scene previews, there was a search advantage for both observer groups. However, this facilitation of search-related eye movements was a much smaller effect than that observed in Experiment 1. For example, the search latency effect sizes for the scene preview benefit in Experiment 1 was η^2 = .81, but only η^2 = .13 in Experiment 2. Indeed, only search latency and number of fixations showed significant improvements, whereas we found no modulation of the speed or amplitude of the

first eye movement of search that typically accompanies such scene preview benefits (Castelhano & Henderson, 2007; Hillstrom, et al., 2012; Pietrzyk et al., 2014).

Given the weak nature of this effect in the medical image perception task, we also examined expert and novice search performance in isolation. A scene preview did not enable novices to fixate nodules quicker, (M_{diff} = -260ms), $t(13) = -.77$, $p = .457$, Cohen's $d = .27$), or in fewer eye movements(M_{diff} = -1.21), $t(13) = -1.15$, $p = .272$, Cohen's $d = .31$) than the mask preview condition. Whereas for experts, the effect of scene preview was approaching significance for search latency (M_{diff} = -765ms), $t(13) = -2.08$, $p = .058$, Cohen's $d = .78$) and in the number of eye movements made (M_{diff} = -2.55), $t(13) = -1.93$, $p = 075$, Cohen's $d = .80$). According to Cohen (1988) these effects of scene preview for experts could be considered as medium to large. However, to put these effect sizes in the context of Experiment 1 using the same individual analysis, this showed that the scene preview effects for real-world scenes were at least twice as large and in this case did enable novices to fixate real-world targets quicker (M_{diff} = -1620ms), $t(13) = -8.57$, $p < .001$, Cohen's $d = .1.74$) and with fewer eye movements(M_{diff} = -5.52), $t(13) = -9.04$, $p < .001$, Cohen's $d = 2.55$), and there were similarly strong effects for experts regarding search latencies (M_{diff} = -1282ms), $t(13) = -6.68$, $p < .001$, Cohen's $d = 1.97$) and number of fixations (M_{diff} = -4.51), $t(13) = -6.96$, $p < .001$, Cohen's $d = 1.70$) .Taken together, this suggests that if domain-specific expertise in medical image perception is modulating how the initial scene is processed, its effect above and beyond our shared expertise in initial scene processing is weak at best.

A possible reason for these weak effects could be because the same image type and target type was searched repeatedly throughout Experiment 2. The scene preview benefit has so far been demonstrated as a robust effect that diminishes after the first few fixations (Hillstrom et al., 2012). However, like Experiment 1, in all FPMW studies different scenes and targets are used across trials and this maximises the advantage of the scene preview. In

contrast, observers in Experiment 2 were repeatedly accessing broadly the same scene guidance and target knowledge and so this could have minimized the scene preview advantage. Even with a mask preview, observers still knew they would always be searching through chest x-ray images for lung nodules and could have exploited that single scene-gist and target template information to help guide search. Because novices could also take advantage of these repeated search conditions, experts may have been prevented from showing their faster processing of these domain-specific scenes and outmatch novice search. With enough trials, repeated search for the same target within the same scene will decrease search times, but change the task, even within the same scene, and that search benefit is lost (Võ & Wolfe, 2012). If we were to use a range of medical image types and different abnormalities as search targets this should once again maximise the advantage of the scene preview, but specifically for the experts, as it is they that should be faster at recognising the scene-gist and accessing the appropriate target knowledge of where to look. Experiment 3 was designed to directly address this issue by varying the medical image type and pathology type across trials. In addition, to be more consistent with Experiment 1 where robust scene preview effects were found, we adopted a 100% prevalence rate so participants only needed to detect the target on each image, and no longer had to make any normal decisions or confidence ratings.

Experiment 3

Method

Participants

We recruited 22 novices (9 male; mean age = 20.9 years) and 19 experienced radiographers (8 male; mean age = 32.7 years). Radiographers were all trained in detection of pathology and had a minimum of 3yrs experience (mean experience = 9.5 years) and included

reporting radiographers for the NHS. We have previously shown that on nodule detection tasks (e.g., Donovan & Litchfield, 2013; Manning et al., 2006) and skeletal fracture tasks (Donovan, Manning, Phillips, Higham, & Crawford, 2005) experienced radiographers demonstrate comparable detection performance to radiologists, with both groups typically finding skeletal images easier than chest x-rays.

Stimuli and apparatus

The testbank consisted of 100 abnormal images from 3 different imaging modalities: 30 chest x-ray images, 20 single axial slice CT or MRI brain images (half each), and 50 skeletal digital x-ray images. All images contained a single discrete pathology and were clinical cases that had previously been reported by a consultant radiologist. The 30 chest images were randomly selected from the 36 abnormal chest images used in Experiment 2, and as such, the pathology was still a single nodule located within the lung fields. The pathology for the 20 brain images were all brain haemorrhages or tumours, whereas the pathology for the 50 skeletal images were all bone fractures. All images were presented on the same monitor as the previous experiments and subtended 22.44° x 23.27° for chest images, 18.07° x 19.09° for brain images, and 27.46° x 31.44° for skeletal images, and the abnormalities were all of comparable size (lung nodules: 3.31° x 3.33°, brain haemorrhages/tumours: 3.86° x 4.17°, fractures: 3.89° x 3.40°). Since we only analyse the eye movement data of correct (i.e., true positive) decisions our 100 % prevalence rate should maximise the number of valid samples in the final analysis and ensure we have sufficient power to detect scene preview effects without being confounded by fatigue effects (Krupinski, Berbaum, Caldwell, Schartz, & Kim, 2010).

Procedure

Participants were given practice examples of the 3 image types and the pathology they would be searching for and were told that there would be a single pathology on every image. All timings of how images were shown were identical to Experiment 2. A critical difference was that rather than searching for the same target (lung nodule) across the same image type (chest), participants searched for pathology on the given image. The target word (pathology) was presented before the scene preview and this generic word was chosen as it did not indicate the upcoming image type. Presentation of image type (chest, brain, skeletal) and preview (scene, mask) conditions were randomised across 4 blocks of 25 trials. Participants saw each image once and preview was counterbalanced. Trials were terminated either by button press or timed out after 15s. Note that as there were no normal (i.e., target-absent) images in the present study, a timeout response could in no way be considered a positive aspect of performance. Likewise, since the focus was on accuracy rather than JAFROC, no confidence data was collected after each decision was made. There were 2 practice trials at the start to familiarise participants with the procedure and the whole experiment took approximately 40min.

Results

All data were subjected to a 2 x 2 mixed measures ANOVA with preview (scene, mask) as a within participants factor and expertise (novice, experienced) as between participant factors. For all measures only trials with correct responses were analysed.

Performance

Overall accuracy across all images and conditions averaged just 44% (novices ranging from 2% to 46%; experienced radiographers ranging from 32% to 80%). There was a main effect of preview, $F(1, 39) = 6.64$, $p = .014$, $\eta^2 = .15$, and a main effect of expertise, $F(1, 39)$

= 60.88, $p < .001$, $\eta^2 = .61$, However, there was no interaction, $F(1, 39) = 0.01$, $p = .92$, $\eta^2 < .01$. As expected, experienced radiographers were much better at detecting a range of pathologies (57%) than novice observers (30%). However, as shown in Table 3, rather than finding a scene preview benefit, both groups of observers were significantly worse at detecting pathologies with a scene preview (novice = 28%, experienced = 56%) compared to a mask preview (novice = 32%, experienced = 59%).

RTs across all images and conditions averaged 6977ms and there was no main effect of preview, $F(1, 39) = 2.55$, $p = .12$, $\eta^2 = .06$, and no interaction, $F(1, 39) = 0.06$, $p = .81$, $\eta^2 < .01$. However, there was a main effect of expertise, $F(1, 39) = 9.92$, $p < .01$, $\eta^2 = .20$, with experienced radiographers faster at detecting a range of pathologies (M = 6308ms) than novice observers (M = 7645ms).

<< Insert Table 3 about here >>

Search-related eye movements

The overall search latency averaged 3060ms across all images and conditions. However, unlike the RT findings, there was no main effect of preview, $F(1, 39) = 2.23$, $p = .14$, $\eta^2 = .06$, no main effect of expertise, $F(1, 39) = 0.91$, $p = .35$, $\eta^2 = .02$, and no interaction, $F(1, 39) = 0.02$, $p = .88$, $\eta^2 < .01$. The overall number of fixations before finding pathology averaged 11.46 across conditions and mirrored the search latency non-significant findings. There was no main effect of preview, $F(1, 39) = 1.77$, $p = .19$, $\eta^2 = .04$, no main effect of expertise, $F(1, 39) = 0.43$, $p = .52$, $\eta^2 = .01$, and no interaction, $F(1, 39) = 0.19$, $p = .66$, $\eta^2 < .01$.

First eye movement

Initial saccadic latency averaged 248ms across conditions, however, there was no main effect of preview, $F(1, 39) = 0.30$, $p = .59$, $\eta^2 < .01$, no main effect of expertise, $F(1, 39) = 2.54$, $p = .12$, $\eta^2 = .06$, and no interaction, $F(1, 39) = 0.44$, $p = .51$, $\eta^2 = .01$. Similarly, the initial saccadic amplitude averaged 2.32°, yet there was no main effect of preview, $F(1, 39) = 1.91$, $p = .18$, $\eta^2 = .05$, no interaction, $F(1, 39) = 0.13$, $p = .72$, $\eta^2 < .01$, whereas the main effect of expertise was approaching significance, $F(1, 39) = 3.93$, $p = .054$, $\eta^2 = .09$. Taken together, all eye movement metrics failed to demonstrate a scene preview benefit for novice or experienced observers examining a range of medical images using the FPMW paradigm.

Discussion

Experiment 3 focused on clarifying the weak scene preview benefit observed in Experiment 2 by systematically increasing the range of medical image types and pathologies and thereby maximise the benefit of the scene preview. Consistent with previous research (e.g., Donovan et al., 2005; Donovan & Litchfield, 2013; Manning et al., 2004), Experiment 3 found that experienced radiographers could detect more pathologies and do so faster than novices. However, rather than maximising the scene preview benefit compared to mask preview, Experiment 3 found that both novices and experienced radiographers were worse at detecting pathologies with a scene preview than a mask preview, and all eye movement metrics confirmed there was no search related advantage of the scene preview.

Results from Experiment 2 hinted that scene previews could have unforeseen costs in terms of diagnostic performance, but this was only found for novice observers, not experts. By randomising medical image and pathology type, Experiment 3 replicated the finding that scene previews impaired novice performance, but also that scene previews impaired experienced radiographers that are currently practicing in hospitals. This scene preview impairment goes against our current understanding of how initial scene previews are

supposed to be exploited with experience in order to enhance performance (e.g., Donovan & Litchfield, 2013; Drew et al., 2013a; Krupinski, 2010; Kundel et al., 2007; Wolfe et al., 2011). Whilst we adopted a 100% prevalence rate in Experiment 3 so as to be more consistent with Experiment 1 and to ensure we had adequate power to detect scene preview effects, knowing that there was always pathology could have led to a change in decision thresholds and led to an increase in false positives. However, this changing of decision thresholds purely based on prevalence would still not account for our pattern of results, and specifically why the scene preview impaired detection of targets relative to mask preview.

Examining the accuracy effects in more detail, it is evident from Tables 2 and 3 that there was a clear drop in overall accuracy (approx 15%) between Experiment 2 and Experiment 3. Experienced radiographers frequently demonstrate comparable detection performance to experts in nodule detection tasks (Donovan & Litchfield, 2013; Litchfield et al., 2010; Manning, Barker-Mill, Donovan, & Crawford, 2005; Manning et al., 2006) and fracture detection tasks (Donovan et al., 2005). As such, we do not believe this drop in accuracy was due to a lack of expertise, but instead due to increasing task demands. This is supported by the fact there was also a comparable drop in novice accuracy (from 45% in Experiment 2, to just 30% in Experiment 3). The key question though is how this more demanding task contributed to the impairment in detection for scene preview compared to mask preview.

In all previous research where accuracy has been reported using the FPMW paradigm (e.g., Castelhano & Henderson, 2007; Võ & Henderson, 2010, 2011; Võ & Schneider, 2010) observers have always had to switch between different images and target knowledge across trials and never before has an accuracy impairment been found for scene preview. As such, it is not as if switching creates a generic cognitive-load issue that could give rise to accuracy impairments for scene preview. One potential explanation of how switching between medical

images could have led to the specific scene preview impairment is how observers were able to filter out the inherent distractors in the image under the different preview conditions. In the scene perception tasks, there is always only one search target in the scene and efforts are made to select images carefully so as to minimise target-like distractors. Likewise in medical imaging tasks, there is also often only one search target – a genuine pathology that has been verified by consultant radiologists beforehand. However, what is inevitable with medical images is that normal anatomy can provide many potential target-like distractors that are inherent to the image (Wester et al., 1997). Some abnormalities have poor visual conspicuity (Krupinski, Berger, Dallas, & Roehrig, 2003) and together with the co-presence of target-like distractors this means that even when fixating directly at a target for several seconds it can often be declared as not being an abnormality (Kundel et al., 1978; Manning, Ethell, & Donovan, 2004) or instead that a normal feature is identified as the abnormality (Wester et al., 1997). Indeed, when a suspected abnormality is difficult to disambiguate from normality its spatial location may have to be relied upon to make correct decisions (Carmody, Kundel & Toto, 1984; Donovan & Litchfield, 2013). When a medical image is flashed, the observer processes the gist of the image, and potentially detects pathology, even if the abnormality cannot later be localized (Evans et al., 2013). It may be that flashing the same type of images (chest) in Experiment 2 allowed expert observers (but not novices) to better discriminate between targets and distractors. In contrast, switching between image types in Experiment 3 may have meant that both novice observers and experienced observers were more susceptible to the distractors inherent in the images following the scene preview. Conversely, the mask preview prevented observers from immediately processing the gist as well as potential targets and distractors, and this may have mitigated any further susceptibility to distractors. The implication is that the costs of perceiving this initial glimpse must have outweighed any benefit of coarse image categorization and target detection.

Taken together, these substantial issues in detection accuracy may fundamentally explain why we did not find robust search benefits of scene preview when attempting to apply the FPMW paradigm to this domain-specific task. The unforeseen costs of scene previews and the implications this has are further explored in the general discussion.

General Discussion

The aim of the present study was to establish how the initial scene representation guides search as a function of domain-specific expertise. Using the FPMW paradigm (Castelhano & Henderson, 2007) two experience groups (expert radiologists, psychology students) searched for everyday objects from real-world scenes (Experiment 1), and lung nodules from chest x-ray images (Experiment 2), whereas a second sample of observers (experienced radiographers, psychology students) searched for a variety of pathologies from different medical image types (Experiment 3). Consistent with previous research (e.g., Castelhano & Henderson, 2007; Võ & Henderson, 2010) we found strong scene preview effects for the observers in Experiment 1, as both expert radiologists and psychology students were able to exploit a brief glimpse of the upcoming scene to guide search. However, in this first application of the FPMW to a specific expertise domain, we found only weak effects of scene preview in Experiment 2 using these same participants. This suggests that experts were not substantially better than novices at exploiting the scene preview of medical images. Overall, both groups of observers were able to find abnormalities in medical images faster and with fewer eye movements following a brief glimpse of the scene. However, it was only when we examined the preview effects of each group in isolation that search metrics of experts (but not novices) seemed to improve with scene preview. Moreover, these improvements in search did not translate into benefits in diagnostic performance. Experts identified more nodules than novices in all conditions, but providing a brief glimpse of the

medical scene did not lead to additional improvements in decision-making, and in fact, further impaired novice performance. Experiment 3 was designed to tease out these weak scene preview effects by requiring novice observers and experienced radiographers to examine a greater range of medical image and target types, thereby maximising the expertise advantage of receiving the scene preview to guide search. Instead, we discovered unexpected findings that corroborate the results of Experiment 2; scene previews of medical images led to impaired accuracy compared to mask preview for both novice and experienced observers and there was still no search benefit for scene preview.

At a descriptive level, the holistic model of medical image perception (Kundel et al., 2007; Nodine & Kundel, 1987) helps account for the well documented expertise differences in search and diagnostic performance (Nodine & Mello Thoms, 2010; Reingold & Sheridan, 2011). The ability to exploit the initial glimpse of the scene is at the core of the holistic model (Kundel et al., 2007) but is also a key component of scene perception research (Castelhano & Henderson, 2007; Torralba et al., 2006; Wolfe et al., 2011). By using the FPMW paradigm to control the contribution of the initial glimpse on subsequent search as a function of expertise, the present study extends previous eye-tracking research that has until now only been able to indirectly investigate these issues, either via tachistoscopic studies (Carmody et al., 1981; Kundel et al., 1975; Evans et al., 2013) or free viewing studies (Donovan & Litchfield, 2013; Kundel et al., 1978; Kundel et al., 2008; Manning et al., 2004). We first discuss why we only observed a weak expertise advantage of processing the scene preview in Experiment 2, and then elaborate on the explanation we put forward in Experiment 3, as to how an initial scene preview of a medical image could impair the detection of targets as found in Experiment 2 and Experiment 3.

First, a major component of how the scene-context guides search is that observers learn the spatial relationships between scenes and the objects within (Castelhano & Henderson,

2007; Castelhano & Heaven, 2010). As a result, the scene-context can guide search towards likely target locations, even when targets were not present in previews. Critically, however, unlike scene perception and the search for everyday objects, the scene-context of the chest x-ray is not particularly predictive as to the location of lung nodules, as these targets can appear anywhere within the lung fields (Båth et al., 2005). Contextual guidance of the scenes (e.g., Torralba et al., 2006) would therefore promote the outer lung fields as highly probable search areas but would not be able to narrow down search guidance much further on a given image, and instead would rely on subsequent search-related eye movements to discount non-target areas.

Second, it would seem there is a better target template in the visual search for real-world scenes than in medical images (Malcolm & Henderson, 2009; Vickery, King, & Jiang, 2009). As mentioned previously, lung nodules can be difficult to identify and the medical images often contains numerous distractors from normal anatomy that closely resemble the features of nodules (Krupinski et al., 2003; Wester et al., 1997). A better target template would allow for greater sensitivity to target signals and attenuation of distractors that do not share target similar features, and can help guide search. Taking both these issues into consideration, one reason why only a weak expertise advantage in search was found for scene preview in Experiment 2 was because the targets were difficult to identify and were not in a predictable location.

Notwithstanding these issues, an alternative explanation for the weak scene preview could have simply been because the same type of image and target type was searched repeatedly throughout Experiment 2. To rule out this alternative explanation, Experiment 3 increased the variety of image types and pathology across trials. However, rather than enhancing the scene preview benefit, there was no scene preview benefit for novice or experienced observers, and instead both groups showed impaired accuracy in this condition

compared to mask preview. In the discussion of Experiment 3 we highlighted that experienced radiographers have comparable performance to radiologists in these specific medical imaging tasks (e.g., Donovan et al., 2005; Donovan & Litchfield, 2013; Manning, et al., 2005, 2006) and so the reason why experienced radiographers were impaired along novices with scene preview is unlikely to be because the radiographers lacked sufficient expertise. Moreover, we put forward that one of the reasons why experts in Experiment 2 were not likewise impaired with scene preview, could have been because repeatedly flashing the same type of images (chest) in Experiment 2 may have allowed expert observers to better discriminate between targets and distractors. Here we elaborate on this explanation by drawing on the mechanisms that may have allowed experts to do this.

Apart from the global-local processing already discussed, medical image perception must be reliant on additional processes such as the way the visual system adapts to images and consistent patterns of stimulation (Webster, 2011). For example, Webster and colleagues (Kompaniez-Dunigan, Abbey, Boone, & Webster, 2015) recently demonstrated that when pathology was easy to discriminate, encouraging adaptation of visual processes by repeated exposure did not increase detection performance. However, when pathology was more difficult to distinguish from the background, adaptation via repeated exposure to images did enhance performance. In some respects, this weighting of signals is similar to what in medical image perception is known as the application of a pre-whitening filter (Eckstein, Pham, Abbey, & Zhang, 2006), which enables the observer to discount the normal anatomic background noise (De Vries, Hooge, Wertheim, & Verstraten, 2013).

Tying all these aspects together, experts have better visual sensitivity to abnormalities than novices (cf. Sowden et al., 2000) and generally outperform them (Donovan & Litchfield, 2013; Nodine & Mello Thoms, 2010; Reingold & Sheridan, 2011), but also, by repeating the same image and target type in Experiment 2, experts may have been able to offset the impact

of processing the distractors visible in the preview, and therefore, maintained the same detection performance for both scene preview and mask preview (cf., Kompaniez-Dunigan et al. 2015). In contrast, by manipulating the image type in Experiment 3 we may have disrupted the observer's ability to become sensitive to pathology and attenuate distractors within the experiment. Just like not appropriately applying a pre-whitening filter (Eckstein et al., 2006), these processes could have a significant impact on decision-making performance in medical imaging, yet are not so crucial in target detection for real-world scenes. For example, the detection of targets in real-world scenes can make use of other channels of information, such as colour, which are irrelevant to grey-scale medical images, but can nonetheless affect how the gist is exploited in the first place (Nijboer, Kanai, de Haan, & van der Smagt, 2008).

One FPMW study that is worth pointing out here in relation to the costs of processing the initial glimpse is Võ and Schneider (2010). Although they did not find the same accuracy impairments we observed in the present study, they made some very relevant discussion points as to the relationship between local and global processing, which overlap in many ways with our own points. Võ and Schneider (2010) examined the individual differences in the ability to process scene previews by comparing those participants that could later describe the differences between the previews conditions (the 'conscious-report' group), with those that could not (the 'no report group'). Again, showing that the FPMW paradigm can lead to surprising results, the conscious-report group did not gain a search benefit from viewing an identical preview containing the background scene context and the inherent objects within, but instead this group could exploit a preview that only consisted of the background scene context. In contrast, the no report group benefited mostly from identical previews. Võ and Schneider (2010) argued that this meant the additional objects in the identical preview may have undergone enhanced processing by the conscious-group, which resulted in competing

scene priors from the local and global pathway. In other words, there was such a thing as too much information in the preview. In this context, our mask preview could be considered a convenient way of blocking out the interference of competing global and local signals and allowed the observers to just focus search and decision-making to what can be seen through the moving window. This suggests that aside from the distinct visual processes used to optimise medical imaging performance, individual differences of how previews are encoded and exploited needs further research, as our current findings also force us to question the real benefits of scene previews, particularly when targets are not so easy to detect.

Overall, we have found with the FPMW paradigm that the scene preview can disrupt observer performance in unexpected ways. Although the processing that takes place based on the initial glimpse of the scene is thought to be integral to expert medical image perception performance (Kundel et al., 2007), research has already highlighted that there are limits as to the types of decisions that can be made solely based on this initial glimpse (Carmody et al., 1981; Evans et al. 2013). Our present study goes further than this by demonstrating that when the initial glimpse is controlled, but search is still allowed, decision-making can sometimes be better when no prior glimpse is available. Given that the ability to exploit the initial glimpse of the scene is at the core of the holistic model (Kundel et al., 2007) and a key component of scene perception research (Castelhano & Henderson, 2007; Torralba et al., 2006; Wolfe et al., 2011) our findings add to the existing FPMW research (e.g., Castelhano & Henderson, 2007; Hillstrom, et al., 2012; Võ & Henderson, 2011; Võ & Schneider, 2010), but highlight that further studies are needed to systematically investigate the target-distractor relationship within scenes.

In addition, this study represents just one domain of visual expertise. As we have found, we cannot easily generalize findings from scene perception research using FPMW to our domain of medical image perception. Likewise, our findings may not straightforwardly apply

to other visual expertise domains. Instead, how an initial scene is processed and exploited for search as function of expertise may depend on the specific parameters of the search task, and as such, we encourage research using the FPMW paradigm in other domains.

Using a medical image perception task, we found weak expertise benefits in search from scene previews, but such search benefits were later overshadowed by scene preview impairments in detection. Medical image perception is a difficult task, but experienced observers find ways of dealing with the consistent patterns of stimulation to help reach better decisions. Clearly, there is a real impetus to both further understand and improve performance on such tasks (Beam et al., 2006; Field, 2014; Krupinski, 2010), and in doing so, this also provides new perspectives on scene perception and expertise. What is refreshing is that with the FPMW paradigm (Castelhano & Henderson, 2007), we may now have the better tools in which to test out potential explanations, and identify ways that enhance or impair performance.

Acknowledgements

This work was supported by the Edge Hill University Research Investment Fund and a Phillips Medical Systems grant. We would also like to sincerely thank the action editor, James Enns, as well as Jeremy Wolfe and the two anonymous reviewers for their constructive comments.

References

Båth, M., Håkansson, M., Börjesson, S., Kheddache, S., Grahn, A., Ruschin, M., ... & Månsson, L. G. (2005). Nodule detection in digital chest radiography: introduction to the RADIUS chest trial. *Radiation Protection Dosimetry, 114*, 85-91.

Beam, C. A., Krupinski, E. A., Kundel, H. L., Sickles, E. A., & Wagner, R. F. (2006). The place of medical image perception in 21st-century health care. *Journal of the American College of Radiology, 3*, 409-412.

Bertram, R., Helle, L., Kaakinen, J. K., & Svedström, E. (2013). The effect of expertise on eye movement behaviour in medical image perception. *PloS one, 8*, e66169.

Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology, 14*, 143–177.

Carmody, D. P., Kundel, H. L., & Toto, L. C. (1984). Comparison scans while reading chest images: Taught but not practiced. *Investigative Radiology, 119*, 462–466.

Carmody, D. P., Nodine, C. F., & Kundel, H. L. (1980). Global and segmented search for lung nodules of different edge gradients. *Investigative Radiology, 15*, 224–233.

Carmody, D. P., Nodine, C. F., & Kundel, H. L. (1981). Finding lung nodules with and without comparative visual search. *Perception & Psychophysics, 29*, 594–598.

Castelhano, M. S., & Heaven, C. (2010). The relative contribution of scene context and target features to visual search in scenes. *Attention, Perception, & Psychophysics, 72*, 1283–1297.

Castelhano, M. S., & Heaven, C. (2011). Scene context influences without scene gist: Eye movements guided by spatial associations in visual search. *Psychonomic Bulletin & Review, 18*, 890-896.

Castelhano, M. S., & Henderson, J. M. (2007). Initial scene representations facilitate eye movement guidance in visual search. *Journal of Experimental Psychology: Human Perception and Performance, 33,* 753-763.

Chakraborty, D. P., & Berbaum, K. S. (2004). Observer studies involving detection and localization: Modeling, analysis and validation. *Medical Physics, 31,* 2313–2330.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New Jersey: Lawrence Erlbaum.

De Vries, J. P., Hooge, I. T., Wertheim, A. H., & Verstraten, F. A. (2013). Background, an important factor in visual search. *Vision research, 86,* 128-138.

Donovan, T., & Litchfield, D. (2013). Looking for cancer: Expertise related differences in searching and decision making. *Applied Cognitive Psychology, 27,* 43-49.

Donovan, T., Manning, D. J., Phillips, P. W., Higham, S., & Crawford, T. (2005). The effect of feedback on performance in a fracture detection task. *SPIE, 5749,* 79-85.

Drew, T., Evans, K. K., Võ, M. L. H., Jacobson, F. L., & Wolfe, J. M. (2013a). What can you see in a single glance and how might this guide visual search in medical images? *Radiographics, 33*, 263–274.

Drew, T., Võ, M. L. H., Olwal, A., Jacobson, F., Seltzer, S. E., & Wolfe, J. M. (2013b). Scanners and drillers: Characterizing expert visual search through volumetric images. *Journal of Vision, 13*, 3.

Eckstein, M. P., Pham, B. T., Abbey, C. K., & Zhang, Y. (2006). The efficiency of reading around learned backgrounds. *SPIE, 6146,* 170-178

Evans, K. K., Georgian-Smith, D., Tambouret, R., Birdwell, R. L., & Wolfe, J. M. (2013). The gist of the abnormal: Above-chance medical decision making in the blink of an eye. *Psychonomic Bulletin & Review, 20,* 1170-1175.

Field, J. K. (2014). Perspective: The screening imperative. *Nature, 513*(7517), S7-S7.

Gegenfurtner, A., Lehtinen, E., & Säljö, R. (2011). Expertise differences in the comprehension of visualizations: a meta-analysis of eye-tracking research in professional domains. *Educational Psychology Review, 23*, 523-552.

Green, D. M., & Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*. New York: Wiley.

Greene, M. R., & Oliva, A. (2009). The briefest of glances. The time course of natural scene understanding. *Psychological Science, 40*, 464-472.

Hillstrom, A. P., Schloley, H., Liversedge, S. P., Benson, V. (2012). The effect of the first glimpse at a scene on eye movements during search. *Psychomic Bulletin & Review, 19*, 204-210.

Hollingworth, A. (2009). Two forms of scene memory guide visual search: Memory for scene context and memory for the binding of target object to scene location. *Visual Cognition, 17*, 237-291.

Jaarsma, T., Jarodzka, H., Nap, M., Merrienboer, J. J., & Boshuizen, H. (2014). Expertise under the microscope: processing histopathological slides. *Medical Education, 48*, 292-300.

Kompaniez-Dunigan, E., Abbey, C. K., Boone, J. M., & Webster, M. A. (2015). Adaptation and visual search in mammographic images. *Attention, Perception, & Psychophysics, 77*, 1081-1087.

Krupinski, E. A. (2010). Current perspectives in medical image perception. *Attention, Perception, & Psychophysics, 72*, 1205-1217.

Krupinski, E. A., Berbaum, K. S., Caldwell, R. T., Schartz, K. M., & Kim, J. (2010). Long radiology workdays reduce detection and accommodation accuracy. *Journal of the American College of Radiology, 7*, 698-704.

Krupinski, E. A., Berger, W. G., Dallas, W. J., & Roehrig, H. (2003). Searching for nodules: what features attract attention and influence detection? *Academic Radiology, 10*, 861-868.

Krupinski, E. A., Tillack, A. A., Richter, L., Henderson, J. T., Bhattacharyya, A. K., Scott, K. M., ... & Weinstein, R. S. (2006). Eye-movement study and human performance using telepathology virtual slides. Implications for medical education and differences with experience. *Human Pathology, 37*, 1543-1556.

Kundel, H. L., & La Follette, P. S. (1972). Visual search patterns and experience with radiological images. *Radiology, 103*, 523–528.

Kundel, H. L., & Nodine, C. F. (1975). Interpreting chest radiographs without visual search. *Radiology, 116*, 527–532.

Kundel, H.L., Nodine, C.F., & Carmody, D.P. (1978). Visual scanning, pattern recognition, and decision making in pulmonary nodule detection. *Investigative Radiology, 13*, 175–181.

Kundel, H. L., Nodine, C. F., Conant, E. F., & Weinstein, S. P. (2007). Holistic component of image perception in mammogram interpretation: gaze-tracking study 1. *Radiology, 242*, 396-402.

Kundel, H. L., Nodine, C. F., Krupinski, E. A., Mello-Thoms, C. (2008). Using gaze-tracking data and mixture distribution analysis to support a holistic model. *Academic Radiology, 15*, 881–886.

Litchfield, D., Ball, L. J., Donovan, T., Manning, D. J., & Crawford, T. (2010). Viewing another person's eye movements improves identification of pulmonary nodules in chest x-ray inspection. *Journal of Experimental Psychology: Applied, 16*, 251-262.

Malcolm, G. L., & Henderson, J. M. (2009). The effects of target template specificity on visual search in real-world scenes: Evidence from eye movements. *Journal of Vision, 9*, 1-13.

Manning, D., Barker-Mill, S. C., Donovan, T., & Crawford, T. (2005). Time-dependent observer errors in pulmonary nodule detection. *British Journal of Radiology*, 78, 1-5.

Manning, D. J., Ethell, S. C., & Donovan, T. (2004). Detection or decision errors? Missed lung cancer from the posteroanterior chest radiograph. *British Journal of Radiology*, 77, 231-235.

Manning, D. J., Ethell, S. C., Donovan, T., & Crawford, T. J. (2006). How do radiologists do it? The influence of experience and training on searching for chest nodules *Radiography, 12*, 134-142.

Mugglestone, M. D., Gale, A. G., Cowley, H. C., & Wilson, A. R. M. (1995). Diagnostic performance on briefly presented mammographic images. *SPIE, 2436*, 106–115.

Nakashima, R., Kobayashi, K., Maeda, E., Yoshikawa, T., & Yokosawa, K. (2013). Visual search of experts in medical image reading: the effect of training, target prevalence, and expert knowledge. *Frontiers in Psychology, 4*, 1-8.

Nijboer, T. C., Kanai, R., de Haan, E. H., & van der Smagt, M. J. (2008). Recognising the forest, but not the trees: An effect of colour on scene perception and recognition. *Consciousness and Cognition, 17*, 741-752.

Nodine, C. F., & Krupinski, E. A. (1998). Perceptual skill, radiology expertise, and visual test performance with NINA and WALDO. *Academic radiology, 5*, 603-612.

Nodine, C. F., & Kundel, H. L. (1987). The cognitive side of visual search in radiology. In J. K., O'Regan, A. Levy-Schoen (Eds). *Eye movements: From physiology to cognition* (pp. 573–582). Amsterdam: Elsevier.

Nodine, C. F., & Mello-Thoms, C. (2010). The role of expertise in radiologic image interpretation. In E. Samei, E. Krupinski (Eds.), *The Handbook of Medical Image Perception and Techniques* (pp. 139–156), New York: Cambridge University Press.

Oestmann, J. W., Greene, R., Bourgouin, P. M., Linetsky, L., & Llewellyn, H. J. (1993). Chest "gestalt" and detectability of lung lesions. *European Journal of Radiology*, *16*, 154-157.

Pietrzyk, M. W., McEntee, M. F., Evanoff, M. E., Brennan, P. C., & Mello-Thoms, C. R. (2014). Direction of an initial saccade depends on radiological expertise. *SPIE*, *9037*, 90371A-90371A.

Phillips, P., Boone, D., Mallett, S., Taylor, S. A., Altman, D. G., Manning, D., ... & Halligan, S. (2013). Method for Tracking Eye Gaze during Interpretation of Endoluminal 3D CT Colonography: Technical Description and Proposed Metrics for Analysis. *Radiology*, *267*, 924-931.

Reingold, E. M., & Sheridan, H. (2011). Eye movements and visual expertise in chess and medicine. In S. P. Liversedge, I. D. Gilchrist, S. Everling (Eds). *The Oxford Handbook of Eye Movements* (pp. 767–786). Oxford: Oxford University Press.

Russell, B. C., Torralba, A., Murphy, K. P., & Freeman, W. T. (2008). LabelMe: a database and web-based tool for image annotation. *International Journal of Computer Vision*, *77*, 157-173.

Sowden, P. T., Davies, I. R., & Roling, P. (2000). Perceptual learning of the detection of features in X-ray images: a functional role for improvements in adults' visual sensitivity? *Journal of Experimental Psychology: Human Perception and Performance*, *26*, 379–390.

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review, 113*, 766-786.

Vickery, T. J., King, L. W., & Jiang, Y. (2005). Setting up the target template in visual search. *Journal of Vision, 5*, 81-92.

Võ, M. L. H., & Henderson, J. M. (2010). The time course of initial scene processing for eye movement guidance in natural scene search. *Journal of Vision, 10*, 1-13.

Võ, M. L. H., & Henderson, J. M. (2011). Object-scene inconsistencies do not capture gaze: evidence from the flash-preview moving window paradigm. *Attention, Perception & Psychophysics, 73*, 1742–1753

Võ, M. L. H., & Schneider, J. M. (2010). A glimpse is not a glimpse: Differential processing of flashed scene previews leads to differential target search benefits. *Visual Cognition, 18*, 171-200.

Võ, M. L. H., & Wolfe, J. M. (2012). When does repeated search in scenes involve memory? Looking at versus looking for objects in scenes. *Journal of Experimental Psychology: Human Perception and Performance, 38*, 23-41.

Webster, M. A. (2011). Adaptation and visual coding. *Journal of Vision, 11*,1–23.

Wester, C., Judy, P. F., Polger, M., Swensson, R. G., Feldman, U., & Seltzer, S. E. (1997). Influence of visual distractors on detectability of liver nodules on contrast-enhanced spiral computed tomography scans. *Academic Radiology, 4*(5), 335-342.

Wolfe, J. M., Võ, M. L.-H., Evans, K. K., & Greene, M. R. (2011). Visual search in scenes involves selective and non-selective pathways. *Trends in Cognitive Sciences, 15*, 77-84.

Figure 1. Trial sequence depending on whether participants were searching for everyday objects (Experiment 1) lung nodules (Experiment 2, as shown), or a range of pathologies (Experiment 3).

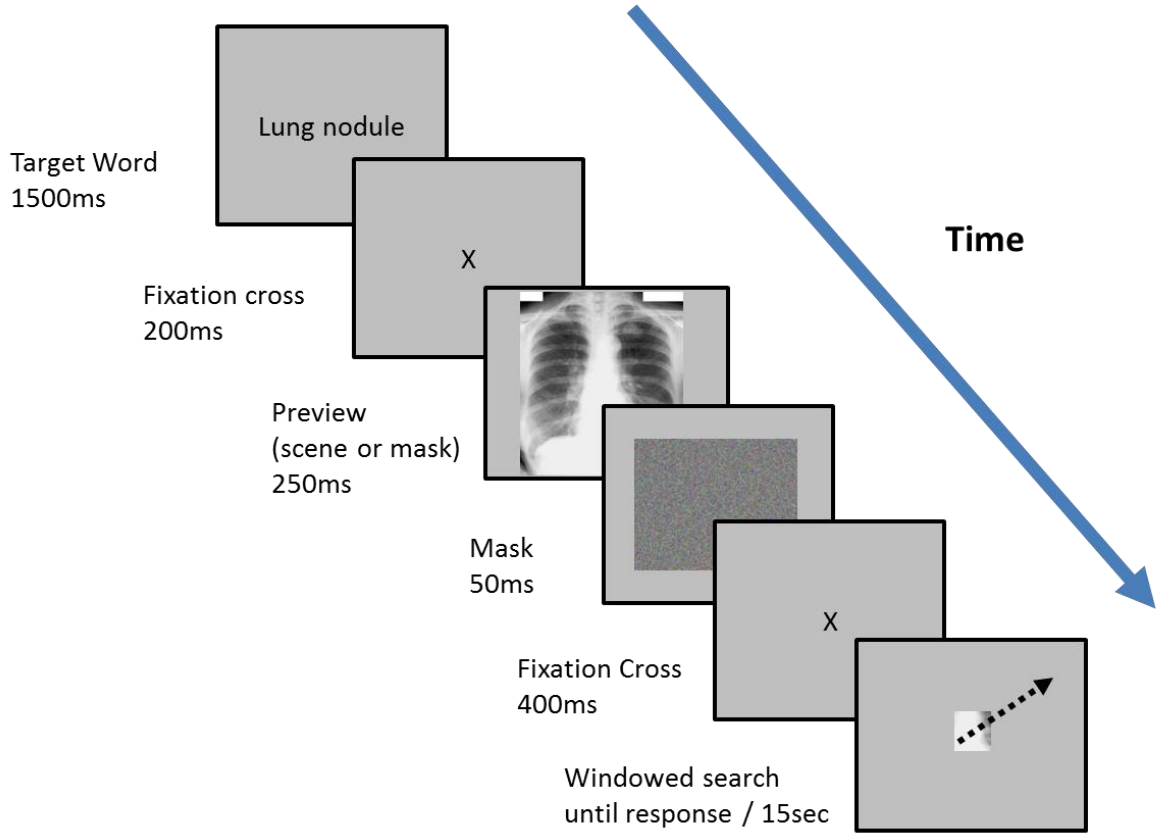

Table 1. *Observer performance and eye movement measures when searching for everyday objects from real-world scenes in Experiment 1.*

Variable	Novice Observers		Expert Radiologists	
	Scene Preview	Mask Preview	Scene Preview	Mask Preview
Accuracy (%)				
M	84.29	91.73	87.14	88.21
SE	2.19	1.29	1.29	2.35
RT (ms)				
M	3107	4737	3185	4552
SE	218	326	204	181
Search Latency				
M	2197	3816	2275	3557
SE	230	265	189	157
Number of Fixations				
M	8.07	13.58	8.74	13.25
SE	0.56	0.60	0.80	0.61
Initial Saccadic Latency				
M	195	282	194	272
SE	12	23	7	12
Initial Saccadic Amplitude				
M	2.57	1.97	2.67	1.88
SE	0.14	0.09	0.16	0.11

Table 2. *Observer performance and eye movement measures when searching for lung nodules from medical images in Experiment 2.*

Variable	Novice Observers		Expert Radiologists	
	Scene Preview	Mask Preview	Scene Preview	Mask Preview
Accuracy (%)				
M	43.21	49.64	68.57	70.57
SE	3.64	3.10	2.90	3.23
JAFROC				
M	.51	.55	.64	.64
SE	.02	.02	.01	.02
RT (ms)				
M	7355	7797	6759	7256
SE	393	335	331	431
Search Latency				
M	4530	4791	3956	4722
SE	243	279	253	269
Number of Fixations				
M	15.04	16.25	13.64	16.19
SE	0.99	1.10	0.87	0.83
Initial Saccadic Latency				
M	217	234	223	259
SE	14	14	20	23
Initial Saccadic Amplitude				
M	2.38	2.23	2.58	2.38
SE	0.21	0.15	0.11	0.19

Table 3. *Observer performance and eye movement measures when searching for a range of pathologies (lung nodules, brain tumors, bone fractures) from a variety of medical images (chest x-rays, brain images, skeletal x-rays) in Experiment 3.*

Variable	Novice Observers		Experienced Radiographers	
	Scene Preview	Mask Preview	Scene Preview	Mask Preview
Accuracy (%)				
M	28.18	31.73	55.68	58.95
SE	2.43	1.91	2.94	3.34
RT (ms)				
M	7872	7418	6474	6141
SE	301	352	345	388
Search Latency				
M	3025	3311	2779	3126
SE	218	222	210	214
Number of Fixations				
M	11.37	12.02	10.58	11.86
SE	0.73	0.71	0.74	0.69
Initial Saccadic Latency				
M	251	230	237	236
SE	13	18	8	15
Initial Saccadic Amplitude				
M	2.62	2.39	2.20	2.07
SE	0.19	0.13	0.16	0.15