Topic Paper – Targeted Search in Visual Displays

Dave Schreifels
Grad 6
Michigan Technological University

Abstract

User error resulting from poor UI design has previously been responsible for large-scale disasters that could have otherwise been avoided if the user were more able to quickly find the desired information and controls. Visual saliency has been previously studied and modeled, but contemporary models are often overly complex. The current proposal investigates the possibility of greatly simplifying these models, thereby reducing the amount of necessary predictors and, consequently, enhancing our ability to prevent these disasters in the future by saving money and increasing efficiency.

Introduction

The field of Human-Computer Interaction (henceforth "HCI") is impacted in its entirety by the study of visual search. In every conceivable application of HCI, users must be able to quickly and efficiently locate objects in a visual field, be it an icon or alert in a visual display, a lever or dial on a control module, or something in between (Jacob & Karn, 2003). Given this, research on search as it relates to the visual aspects of UI design within HCI is especially critical to the advancement of the field.

The Level 5 nuclear accident at Three Mile Island demonstrates this point quite well. Three Mile Island was a two-core nuclear power plant in Pennsylvania, and in March of 1979 the second reactor core experienced a partial meltdown caused by a cooling malfunction – the valve used to drain coolant from the reactor tank became stuck open. This in itself would not have been especially problematic; these are the sorts of problems that reactor workers are expected to handle all the time. However, the reactor operators were unable to interpret the information they were given by their own instruments, and in an attempt to fix the problem they took actions that made the problem far worse than it would have initially been. The core was deprived of coolant and melted down (fig. 1), and the power plant is now defunct (United States Nuclear Regulatory Commission, 2014; United States President's Commission on the Accident at Three Mile Island, 1979).

How does an incident like this happen? To get to the root cause of this issue, one must consider the design of the system, in particular where human operators were expected to be involved. There are two major problems with the design of the system that, in and of themselves, caused an otherwise relatively tame anomaly to spiral out of control. The first is that there was no indication of the water level of the coolant, and while this is certainly a problem, it's not as easily solved as one might think. Simply adding a measurement tool would not be adequate; the system already had hundreds of sensors ranging across the entire facility, attempting to account for everything. Indeed, even if the measure of the coolant had been there, there was a bigger issue at play.

When the valve became stuck open, over 300 warning lights, alarms, and notifications went off within the first minute of the incident, with more than 200 more to follow in the next minute. No person can be reasonably expected to account for that many issues, even if they know where everything is and have a suspicion of what the problem might be. The design of the system was to make salient relevant warnings in a timely manner, but this model breaks down at its logical extreme, illustrated by this nuclear accident. There are too many things to pay attention to, too many distractors and things to consider when the seconds matter (Norman, 2013; Steelman, McCarley, & Wickens, 2011). The simple fact of the matter is that making an object salient isn't good enough if umpteen other targets must also be salient simultaneously. This is especially

problematic in the scenario of nuclear plants like Three Mile Island, where simplification of the system is made impossible by the necessity of both the amount and saliency of so many search targets.

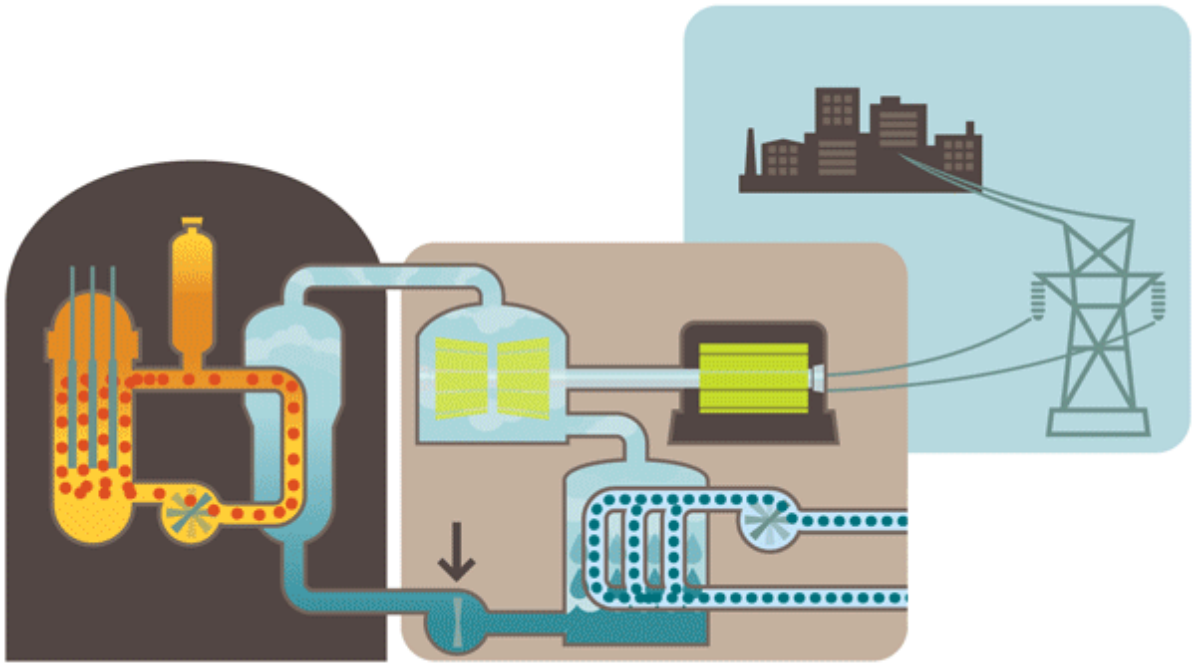

Figure 1 – Faulty Coolant Release Valve

Previous Work in This Area

Research on visual search currently offers an incomplete view of how humans look around a naturalistic display or scene. Traditionally, models of search have revolved primarily around showing a user an image and letting them look around the visual scene, observing by one means or another where they looked and how often (Harel et al, 2015; Seo & Milanfar, 2009; Bylinskii et al, 2015). There has been research since then that has suggested that this approach is incomplete (e.g. Harel, Kosh, & Perona, 2006), but rather than attempt to slowly correct for this incompleteness the field has largely deviated from simple-bottom up approaches in favour of much more complex approaches based on simulated top-down models. These models are largely based around the idea of deep-learning neural networks that can teach the computer what, for example, a dog looks like, so that when it detects a dog or something like a dog in an image it can identify it as salient (Kümmerer, Theis, & Bethge, 2014; Kruthiventi, Ayush, & Babu, 2015; Huan et al, 2015; Pan et al, 2016; Kümmerer, Theis, & Bethge, 2016). Figures 2 & 3 provide examples of saliency predictions from the DeepGaze II neural network.

Figure 2 – Image Submited for Saliency Demonstration

This approach, though reasonably good at predicting saliency, is also quite complex and computationally intensive; by the principle of Occam's Razor, it is at least reasonable to attempt to find a simpler approach, especially if there is reasonable suspicion that one might exist.

This reasonable suspicion comes from an age-old debate between two prominent researchers, Anne Treisman and Jeremy Wolfe, whose nuanced disagreements on the topic are not the subject of this proposal (but which will be discussed later). The commonalities between their research projects over the last few decades has formed the basis for much of the research on search cited in this proposal, and for this reason both models are worth discussing at length. I'll start by discussing the work of Treisman, and then I'll discuss how Wolfe differs.

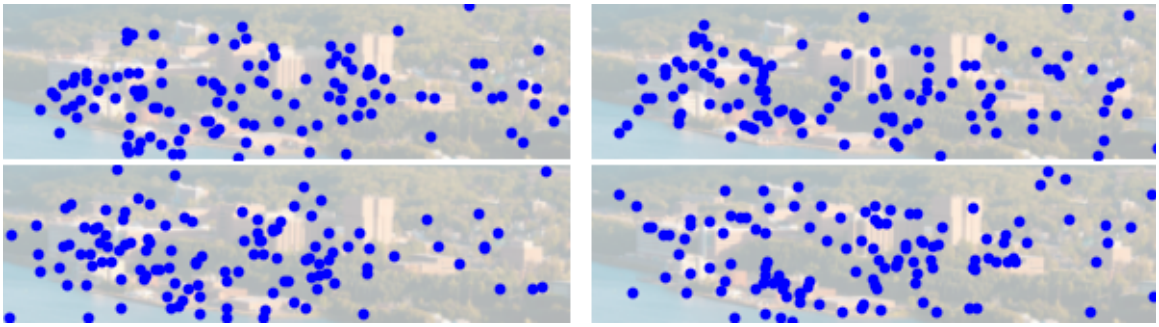

Figure 3 – Four Saliency Maps Generated by DeepGaze II

Feature Integration Theory (henceforth FIT) is a theoretical model of visual attention developed by Anne Treisman (as well as Garry Gelade) that suggests that “features” of objects are perceived automatically and immediately with no need for focused attention, but that these features are only combined into proper objects once focused attention is granted to them. It is further asserted that this combination is later in processing than its detection counterpart, and that the combination happens serially rather than in parallel. The theory has been changed over the years to fit emerging data (as theories often are), but the core principle remains the same (Treisman & Gelade, 1980; Treisman & Gormican, 1988; Treisman, 1996; Treisman, 1998; Treisman, 2006). A representation of this theory can be found in figure 4.

The first stage of this model is perhaps the more interesting portion. Treisman claims that objects are not always perceived as objects, but rather that their features (blue, for example) are, in a sense, free floating in our vision prior to identification. Under Treisman's assertions, if one were to show, say, a previously unseen red coffee mug to a person in the corner of their eye, they wouldn't be able to identify it without focused attention (and, at that angle, they possibly wouldn't be able to identify it at all, but that's beside the point). Instead, one would simply perceive the redness, the approximate size, the location, and the relative brightness of the mug. The primary assertion here is not simply that people won't put the pieces together to identify the object, but rather that they *cannot* do so. These features, it is further claimed, will all be perceived at once, automatically processed in parallel at first detection, and will be in a sense freely unassociated with each other in that part of the vision. It is only when we focus our attention on these features that they become able to perceive them as part of a whole.

The second stage of the model suggests that these features must be combined in serial via focused attention. That is to say that, as opposed to the initial process of detection, combining these features into meaningful objects requires focused attention instead of just an awareness of their existence. Consequently, this would suggest that each object we perceive that has similar features to other objects must be perceived one at a time. Further, these objects must stay together once stuck

together, so to speak, or they'd break apart into a jumble of features again once attention shifted. This unexplained portion of FIT was dubbed the “Binding Problem” and was addressed by Treisman in later papers (e.g. Treisman, 1996)

Figure 4 – Graphic Representation of Feature Integration Theory

Jeremy Wolfe takes a slightly different view on the same theory, dubbing his own theory “Guided Search” and noting that it owes an intellectual debt to Feature Integration Theory. Like Treisman, Wolfe starts with the basic assumption that the visual system simply cannot process all of its input. As well, the assumptions that the visual system discards some input and that it processes information selectively are also held over from FIT. A last assumption, minor but important, is that two messages cannot be processed in different parts of the visual field at the same time. Wolfe's theory, like Treisman's, has also adapted through the decades, providing its own take on visual search as new data as emerged (Wolfe, Cave, & Franzel, 1989; Wolfe, 1994; Wolfe & Gancarz, 1997; Wolfe & Gray, 2007). However, Wolfe's model contrasts with FIT in a couple of important ways, and a representation of the model can be found in figure 5.

The most obvious difference is that Wolfe's model takes a somewhat different view of the preattentive stage of processing. While both Wolfe and Treisman assert that the first stage consists of parallel processing of stimuli, the exact nature of these representations is disputed. In contrast to Treisman's “free floating features” idea, Wolfe proposes something more akin to filtration lenses. It is proposed that rather than free floating, these unattended features are processed coarsely in parallel, but on different channels (e.g. colour, orientation, or size. The idea of these channels (dubbed “Feature Maps” by Wolfe) is that any given channel will account for all “free floating features” on that channel rather than being individual and freely floating (e.g. colour across the whole visual field would be processed together in parallel under this idea as opposed to having, say, a bit of orange on the left and a bit of red on the right being processed as separate).

Guided Search also takes a different stance on the second stage of processing, suggesting not only that attention is not necessarily required for identification of objects but also that the role of parallel processing is to determine which points on the visual field are worthy of attention. That is to say, he goes one step farther than Treisman in saying that attention is drawn by higher levels of activation in the visual field based on these feature maps.

The discussion of these differences should not be taken to mean that Treisman and Wolfe greatly differ on this subject; indeed, their similarities provide the basis not only for this proposal, but also for a wide array of research, both past and present. Rather, these differences must be noted so that it is clear what is more or less settled in the world of visual search and what is not – useful knowledge when basing new research on these studies.

As noted in both theories, targeted search (i.e. searching for a specific target in a field of distractors with similar features) has been done before, but it has not yet gone beyond simple feature conjunctions (e.g. searching for a green O in a field of green Q's and red O's). This lack of investigation of these features in naturalistic imagery constitutes a piece of that puzzle that is easy to overlook but which has great potential for our understanding of visual search. Initial results from Treisman's study not only support FIT; they also suggest that there is an as-yet uninvestigated

difference between targeted search and non-targeted search in naturalistic imagery – imagery such as control panels and user interfaces which are used everyday by millions of people around the world.

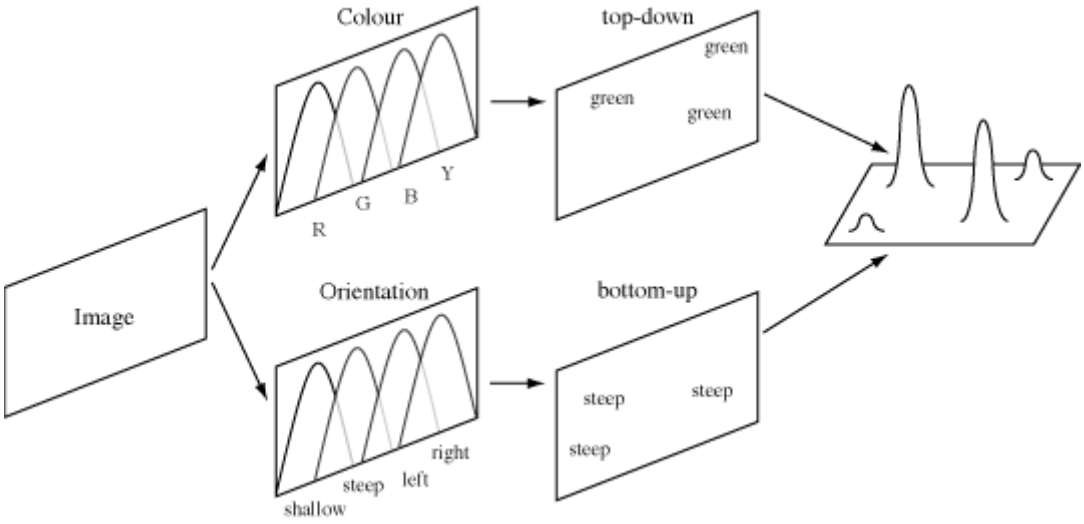

Figure 5 – Graphic Representation of Guided Search Theory

Method

Participants

Participants will be drawn both from the MTU Sona Systems subject pool (consisting of undergraduate students taking the Intro to Psychology course for partial course credit as well as unpaid volunteers from the area.

Materials, Stimuli, & Design

Participants will sit in an adjustable chair with their chin on a chin rest placed 20 inches from the computer screen. The screen will have a resolution of 1920x1080 with a diagonal measurement of 24 inches.

40 images will be selected from the MIT300 (Bylinskii et al, 2015) and 40 more will be created for the purposes of this study. Of the created images, 10 will be of websites; 10 will be of program interfaces; and 20 will be real-world control modules and interfaces. For each of the 80 search images, 4 target images will be created; 2 of these will be present within the image, and 2 will be fabricated to mimic existing parts of the base images. Finally, for each group of 2 images, one will be a search target that is easy to determine the presence of at a glance, and the other will require a more careful search through the image.

We will employ these divisions in a 2x2x2 design, dividing search images by the presence or absence of the target image in the search image, the difficulty (easy or hard) of finding the target image (or determining its absence), and whether the target image is presented before or after the search image. Each participant will see each base image exactly once, and therefore each will see exactly one quarter of the target images, counterbalanced orthogonally across the three aforementioned conditions. Order of images in the before/after conditions will be randomized respectively, but for the sake of consistency these tasks will be kept separate.

Procedure

Participants will sit at a computer using a chin rest and adjustable-height chair to keep a consistent viewing distance and angle. They will be told to adjust their chairs such that their chins rest comfortably on the chin rest, and then will be told they will be participating in what is effectively a simplified Where's Waldo game. There are two parts – a targeted search task, and a non-targeted search task. They will first perform the targeted search task, followed by the non-targeted search task (described below). Each of these tasks will involve searching a visual field and identifying the presence or absence of target images within a larger base image. Prior to each task, participants will first carry out three tutorial trials.

Target Image First

For each trial in this task, participants will first be shown the target image that they will be searching for, and they will be told to study it for as long as desired. When ready, they will click the mouse, whereupon the larger base image will appear. Upon either finding the target image within the stimulus or determining its absence, they will again click the mouse and search time will be recorded. They will then either click where, approximately, they found the target image in the base image, or they will click a button labeled "ABSENT."

Base Image First

For each trial in the second task, participants will first be shown the base image, and will be told to study it for as long as they desire. When ready, they will click the mouse, whereupon search time will recorded and the target image will be displayed along with two buttons labeled "PRESENT" and "ABSENT". Participants will then indicate using the buttons whether or not they saw the probe in the stimulus image.

Expected Results

The primary expectation from the proposed research is that we'll find evidence of a difference in how people search around naturalistic imagery and UI imagery based on whether or not they have a target to search for. It is further expected that the different types of stimuli (naturalistic imagery from the MIT300, interfaces, websites, and real-world control systems) will yield different search times due to the differing nature and purpose of the subject matter.

Implications for Future Research

What would results like this mean? It depends on how and how meaningfully these differences present themselves. For the purposes of speculation, it is assumed all differences will be significant. With this assumption in mind, the most mimportant result would be that there is in fact a difference in how people look around a naturalistic scene when they're looking for something versus simply being told to "take in the scene," so to speak. If this is true, it would imply that our understanding of visual search can be enhanced by much simpler methods than contemporary approaches using neural networks. Instead of having, for example, 20 predictors that fit one data set extremely well, we could have, say, 10 predictorss that predict that data set almost as well as the 20-predictor model but which are much more broadly applicable to other data sets, too. This sort of simplification – where one can apply the same model to very similar work – is almost always preferred, and is certainly preferred in an area such as visual search which has a lot of similar but not redundant applications.

Furthermore, if there is a meaningful difference between the different types of stimuli regardless of whether the target was shown before or after the base image, this has implications for

necessary research that might have otherwise been left uninvestigated by a complex, neural-network-based paradigm.

The primary question in all of these cases would be: why? Why would, for example, a commercial website be faster or slower to search than the control module for a nuclear device? And further, should it be? And is there any way we can change that? For each pairing, these questions arise if there's a meaningful difference in response time between the two. Surely the first step would be to validate the results with a larger scale study of the same style focusing on just those two types of interfaces; beyond that, the type of difference (e.g. interfaces versus real-world control modules) must determine how it's further investigated. If for example interfaces are more quickly and easily searched through than physical control modules, should we then replace the controllers at nuclear power plants with these interfaces? Perhaps only the most important controls should be interface-based, and the secondary and tertiary controls should not, so as to avoid another incident like the nuclear accident at Three Mile Island. These are just a small sampling of foreseeable scenarios, but in the light of positive results, the possibilities become endless.

References

Bylinskii, Z., Judd, T., Borji, A., Itti, L., Durand, F., Oliva, A., & Torralba, A. (2015). MIT saliency benchmark.

Harel, J., Koch, C., & Perona, P. (2006). Graph-based visual saliency. In *NIPS 1*(2), *5*.

Henderson, J. M., Brockmole, J. R., Castelhano, M. S., & Mack, M. (2007). Visual saliency does not account for eye movements during visual search in real-world scenes. *Eye movements: A window on mind and brain*, 537-562.

Huang, X., Shen, C., Boix, X., & Zhao, Q. (2015). Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. *Proceedings of the IEEE International Conference on Computer Vision*, 262-270.

Jacob, R. J., & Karn, K. S. (2003). Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. *Mind*, *2*(3), 4.

Kruthiventi, S. S., Ayush, K., & Babu, R. V. (2015). Deepfix: A fully convolutional neural network for predicting human eye fixations. *arXiv preprint arXiv:1510.02927*.

Kümmerer, M., Theis, L., & Bethge, M. (2014). DeepGaze I: Boosting saliency prediction with feature maps trained on imagenet. *arXiv preprint arXiv:1411.1045*.

Kümmerer, M., Wallis, T. S., & Bethge, M. (2016). DeepGaze II: Reading fixations from deep features trained on object recognition. *arXiv preprint arXiv:1610.01563*.

Norman, D. A. (2013). *The design of everyday things: Revised and expanded edition*. Basic books.

Pan, J., Sayrol, E., Giro-i-Nieto, X., McGuinness, K., & O'Connor, N. E. (2016). Shallow and deep convolutional networks for saliency prediction. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 598-606.

Seo, H. J., & Milanfar, P. (2009). Static and space-time visual saliency detection by self-resemblance. *Journal of vision*, *9*(12), 15-15.

Steelman, K. S., McCarley, J. S., & Wickens, C. D. (2011). Modeling the control of attention in visual workspaces. *Human factors*, *53*(2), 142-153.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive psychology*, 12(1), 97-136.

Treisman, A., & Gormican, S. (1988). Feature analysis in early vision: Evidence from search

asymmetries. *Psychological review*, *95*(1), 15.

Treisman, A. (1996). The binding problem. *Current opinion in neurobiology*, *6*(2), 171-178.

Treisman, A. (1998). Feature binding, attention and object perception. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *353*(1373), 1295-1306.

Treisman, A. (2006). How the deployment of attention determines what we see. *Visual cognition, 14*(4-8), 411-443.

United States President's Commission on the Accident at Three Mile Island. (1979). *The need for change, the legacy of TMI: report of the President's Commission on the Accident at Three Mile Island*. The Commission.

United States Nuclear Regulatory Commission. (2014). Backgrounder on the Three Mile Island Accident. *http://www.nrc.gov/reading-rm/doccollections/fact-sheets/3mile-isle.html*, *10*.

Wolfe, J.M., Cave, K. R., Franzel, S.L. (1989). Guided Search: An Alternative to the Feature Integration Model for Visual Search. *Journal of Experimental Psychology: Human Perception and Performance,* 15(3), 419-433.

Wolfe, J. M. (1994). Guided search 2.0 a revised model of visual search. *Psychonomic bulletin & review*, *1*(2), 202-238.

Wolfe, J. M., & Gancarz, G. (1997). Guided Search 3.0. *Basic and clinical applications of vision science*, 189-192. Springer Netherlands.

Wolfe, J. M., & Gray, W. (2007). Guided search 4.0. *Integrated models of cognitive systems*, 99-119.