# Active Visual Search: Investigating human strategies and how they compare to computational models

**Tiffany Wu**

A Thesis Submitted to the Faculty of Graduate Studies in Partial Fulfillment of the Requirements for the Degree of Master of Science

Graduate Program in Computer Science
York University
Toronto, Ontario

January, 2024

©Tiffany Wu, 2024

# Abstract

Real world visual search by fully active observers has not been sufficiently investigated. Whilst the visual search paradigm has been widely used, most studies use a 2D, passive observation task, where immobile subjects search through stimuli on a screen. Computational models have similarly been compared to human performance only to the degree of 2D image search. I conduct an active search experiment in a 3D environment, measuring eye and head movements of untethered subjects during search. Results show patterns forming strategies for search, such as repeated search paths within and across subjects. Learning trends were found, but only in target present trials. Foraging models encapsulate subject location-leaving actions, whilst robotics models captured viewpoint selection behaviours. Eye movement models were less applicable to 3D search. The richness of data collected from this experiment opens many avenues of exploration, and the possibility of modelling active visual search in a more human-informed manner.

# Acknowledgements

First and foremost, I would like to thank my supervisor, Dr. John Tsotsos, for his support and guidance throughout this process. I have learned so much about the research process as well as how I think about science through working with you.

I would also like to thank Khatoll Gauss for her hard work in running subjects with me for the experiment. Data collection would have been extremely painful without her help.

I further thank all my lab mates for providing feedback and suggestions for various aspects of this thesis. Your supportive yet useful comments have made my thesis a much better work.

To my supervisory committee, Dr. Michael Brown and Dr. Patrick Cavanagh, thank you both for your time and effort in examining my thesis. I truly value the feedback I have received.

Finally, I thank Hansy, my dog, for being my constant companion throughout writing this thesis. Your constant judgment of my work was invaluable. And of course, Jeremy, for inserting random words in my thesis to make sure I would proofread my own work properly.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1 Overview

Visual search is a ubiquitous task; people search for objects on a daily basis. However, machines have struggled to perform search to the same degree that a human can, despite the many computer vision algorithms for deep learning, reinforcement learning, and robotics implementations that have tried. Utilizing our understanding of human behaviour on search might be useful for generating new algorithms. Although a large body of research exists studying human visual search, most of it has focused on the 2D, passive version of the task. Little research has been conducted to fully understand human search behaviour in an active, 3D environment where subjects are presented with issues such as viewpoint selection and occlusion, yet this is how most visual search tasks are conducted in real life.

## 1.2 Motivation

This thesis aims to address the gap in both human psychophysical studies investigating fully active search, as well as understanding existing computational models and robotics implementations in comparison to human behaviour. By gaining a better understanding of human active visual search behaviour, better computational algorithms can potentially be created to be used on machines such as mobile robots.

Most existing research around human visual search has been conducted on flat computer

screens with very confined search spaces, and thus limited ecological validity. On the other hand, experiments on human active search — which widens the scope to include head and body movements — have focused on simple tasks, without addressing more complex cases involving occlusions and larger movements. Furthermore, most active search studies have been conducted within the confines of a table; no body movements or navigation are needed to find the targets in their tasks. Less is known about subject behaviour when they must navigate around larger obstacles to disocclude potential targets. The experiment conducted for this thesis is the first to address this issue, by gathering eye and head movement data during visual search in a 3D environment when subjects are untethered and free to move around. This data allows us to discover interesting patterns of human active search behaviour that could not otherwise be found in passive search data.

Existing computational models for visual search have not been extensively tested against such human performance. Models are vastly simplified and do not tackle active search, instead focusing on passive search in 2D static images, for which data is readily available. Models implemented on mobile robots, by contrast, tackle the active search problem, but have no comparisons to human behaviour. Having a comparison between robotics models and human behaviour is important, not only to learn more about processes underlying human vision, but to potentially improve model performance. Using the data obtained through this thesis, indirect, high-level comparisons were made to existing computational models. Robotics models, eye movement and saliency models, as well as foraging models were compared to human data. Each group of models were compared to my active search data in different ways. Robotics models were compared in head movements and viewpoint selection, eye movement models were compared to instances where subjects only moved their eyes, and foraging models were compared to the higher-level behaviours of subjects moving between areas.

My thesis work is the first to record eye and head movement data on untethered subjects during a visual search task. The data gathered gives a better understanding of how people conduct search in the real world, and how it differs from passive search on a 2D screen. Comparisons made to models show that existing eye movement models are difficult to apply in the real world, robotics models are more capable of capturing the viewpoint selection

by head movements in a 3D environment, and foraging models show similarities in strategy between real world search and foraging behaviour. Such comparisons reveal which components of the models can be extrapolated to human active search, thus providing a foundation towards developing a more cohesive active search model for humans in the real world.

# Chapter 2

# Related works

There are many fields whose works are connected with this thesis. This chapter captures an overview of the background research conducted in these various fields, emphasizing the gap in the research surrounding active search in the real world. Whilst each field has their approach to address the search problem, there are few attempts to truly understand how people behave in a 3D environment when conducting search.

## 2.1 Human vision

### 2.1.1 Visual search

In psychophysics, a typical visual search task involves asking observers to determine if a particular target object is present among distractors. The observer's response time and accuracy are typically recorded. From trial to trial, the number of items displayed vary so that a search slope can be computed (response time / set size). This search slope gives the average amount of time an observer spends inspecting each item, and it is an indicator of the difficulty of the task (Wolfe, 1998).

The visual search paradigm was greatly popularized by Treisman & Gelade's (1980) Feature Integration Theory (FIT). Visual search formed the basis of FIT and many were intrigued by the possibility of connecting psychophysics to actual brain functions (Nakayama and Martini, 2011). Although the theory itself has since been disproven, the visual search

paradigm remains a powerful tool to gain insight into human visual cognition.

Since then, the visual search paradigm has been very widely used in psychology. It has been utilized to explore many different aspects of vision and the inner workings of visual cognition. The popularity of the paradigm can be seen in Wolfe (1998)'s paper summarizing over a million trials of visual search, spanning around 2500 experiments. Eckstein (2011)'s more recent review of the visual search literature shows widespread use of this paradigm to investigate many tasks, strategies that the brain uses to search, as well as many models that have been proposed for eye movements in visual search.

Eimer (2014) presents four stages of selectivity by attention in visual search tasks. These include preparation, guidance, selection, and identification. Preparation involves goal-directed activations that prime the visual system before the stimulus appears. Guidance occurs once the subject views the search display, and involves feature-based bottom up selection. Selection comes next, specifying the top-down spatial biases to control the next attended location. Finally, identification is the process of identifying if the fixated object is the target. Although he presents these as four serial stages of search, he acknowledges that they may occur in parallel as well.

As visual search is an integral component of everyday tasks in the real world, the importance of investigating and understanding it is evident in the research surrounding the field. Many models of search have emerged since Treisman and Gelade (1980), including Guided Search (Wolfe, 2021; Wolfe and Gray, 2007; Wolfe, 1994; Wolfe et al., 1989), the Dimensional Weighting Model (Liesefeld and Müller, 2020), the Target Acquisition Model (Zelinsky, 2008), and many more.

The more recent Guided Search models (such as Guided Search 6.0) suggest that instead of two distinct types of search, as proposed by Treisman and Gelade (1980), there is a continuum, and the relative performance in each search task depends on several guiding factors including bottom-up saliency, top-down guidance, priming, reward, and scene structure. The Dimensional Weighting Model also focuses on different guiding features, and weights each feature's saliency map to get an overall priority map to guide attention. Context cueing (Chun and Jiang, 1998), the phenomenon of improved efficiency in search when trials have been learned, is an example of a guiding factor, as is scene grammar (Vo et al.,

2019). Zelinsky (2008) focuses on modelling overt eye movements during search in the Target Acquisition Model. Other search models include Bundesen et al. (2011)'s NTVA, Snodgrass and Townsend (1980), and Eckstein et al. (2000)'s signal detection model.

However, the majority of these models encapsulate behaviour in variations of the original paradigm with 2-dimensional displays of artificial items.

Correspondingly, most of the visual search experiments to date have been conducted on a 2D computer screen, using artificial, controlled stimuli made by experimenters. Whilst the ability to control for so many factors is an appeal of the visual search task, it is also a great shortcoming. Not only is the depth of all items limited to one level (unless depth is simulated with various depth cues), the search space is also restricted to a confined area, meaning head or body movements are not needed to perform search.

Some of the contrast between this paradigm and the real world can be resolved using naturalistic images — images from the real world. However, these images are still viewed passively, excluding the active component of selecting viewpoints, something humans need in order to conduct search tasks in everyday life. There are many more degrees of freedom in the real world, a 3D environment, compared to the simplified 2D, confined space search that occurs with this paradigm. These include the possibility of occlusion, depth and size differences, and the need to navigate around obstacles to obtain new viewpoints or retrieve the target item.

Consider the task of searching for your keys. Once a view is processed, if the keys are not fully visible in the field of view, active vision is needed to perform the search. It is used to select the next view that could help you find your keys. If there are objects in view that could occlude your keys, selecting a new viewpoint that will allow you to check if your keys were occluded is needed.

Thus, in order to achieve the goal of gaining a better understanding of human visual search strategies in the real world, we must include active vision.

### 2.1.2 Human active vision

One of the first experiments investigating active search in humans was the classic study by Yarbus (1967), where eye movements were tracked while observing an image given different

questions, as well as during a free-viewing task. He showed that eye movement patterns would differ drastically based on task demands. The significance of this work is that it revealed that eye movements depend on the reason for looking at the image.

Hayhoe and Ballard (2005) provides a review of subsequent eye movement research in natural behaviours. They place emphasis on using active visually guided tasks to anchor meaning behing the fixations, rather than using only fixational data conducted on 2D screens (such as those seen in Yarbus (1967)). Active visually guided experiments can be conducted with the use of portable eye tracking devices, and studies have shown further that fixations are tightly coupled with task demand.

Some experiments have been conducted involving active vision in humans, where the main task includes performing visual search (Ballard et al., 1995; Pelz et al., 2001; Hayhoe et al., 2003). However, not many active vision tasks that focus on search have been conducted. Fully active vision tasks include not just eye movements, but head and body motions as well. The reduced number of experiments in real world search especially stems from the fact that it is difficult to carry out these experiments — active vision has many more degrees of freedom than a typical passive search task. There is an additional dimension of space, and multiple degrees of freedom added with respect to viewpoint selection. Where should the observer be in space to take the view? At what direction and angle should they face? Thus, there are many more environmental variables with active vision that become difficult to measure and control for, allowing for easily confounded results.

There has been limited research conducted that tracks the observers' movements in space around an experimental set-up. In visual search, studies that use tracking are only using eye tracking and do not include head or body motions. Whilst eye tracking does allow for measuring active vision, without tracking head motions, this is akin to measuring the case when people stand completely still and only move their eyes to search. Tracking only eye movements also does not allow for measuring behaviour when objects are occluded, since moving the head or body would be necessary in order to disocclude objects.

Furthermore, typical search paradigms determine the difficulty of a search task based on the computed search slope, whose value depends on the number of items in the scene. In the real world, we encounter the "set size problem": there is no satisfactory definition of a

"set size" or a "number of items" in the real scene. If there was a set of trees in the scene, does each tree count as an item? Does the group of trees count as one item? Or does each branch and each leaf count as an item? There is no accurate way of defining set size in real scenes like that.

Researchers are aware of the gap between the simple classical visual search paradigm and the real world, and have attempted to address this in the past. Many have adopted the approach of using naturalistic images of the real world as a proxy (Peters et al., 2005; Tatler et al., 2005; Wolfe et al., 2011). As Marius't Hart et al. (2009) state, there is an issue with this approach — it isn't known for a fact how informative static 2D images are for inferring how we allocate our gaze in the real world.

### Free-viewing

Marius't Hart et al. (2009) attempt to address this by designing a task to compare gaze allocation between watching a video of free exploration in the real world and looking at a sequence of static frames from the same video. They found that the continuous presentation was indeed better for predicting real world gaze than viewing the static images. This conclusion may also be applicable in visual search, although no studies found have yet addressed this issue.

Foulsham et al. (2011) also conducted a similar study about the differences in gaze allocation between the lab and the real world. By using a mobile eye tracker, they were able to compare gaze allocation of subjects performing a task in the real world, as well as gaze allocation of subjects observing a first-person video of the same event in the lab. They found that subjects had significantly different gaze distributions in the real world compared to video, as well as some significant differences in what subjects were looking at. Thus, there may be important differences that would not allow for results from 2D search to be extrapolated into the real world.

### Active vision in real world tasks

Another aspect of gaze allocation not captured in 2D viewing is gaze for navigation. Matthis et al. (2018) found interesting results regarding gaze allocation and control in sub-

jects walking on different types of terrain. They showed that distinctions can be made in gaze strategies between flat, medium, and rough terrains, in order to adapt their strides and foot placements appropriately. The need to look at where a subject is going is clearly not captured in 2D scenes, as no navigation is necessary. However, depending on the type of terrain, fixations to where a subject is walking may take up a significant proportion of fixations in the environment. This then influences gaze distribution in the real world in comparison to a 2D scene.

Ballard et al. (1995), Pelz et al. (2001), and Hayhoe et al. (2003)'s works combine both eye tracking, head movements, and actions. These experiments investigate eye, head, and hand coordination during a real world task. In Ballard et al. (1995) and Pelz et al. (2001) it was a block copying task, and in Hayhoe et al. (2003), subjects had to make a sandwich. Both of these tasks involved search in the process, as subjects had to search for the correct block to place, or the knife or peanut butter jar when making the sandwich. They discovered many interesting patterns in gaze allocation, as well as the relation of gaze with planning and coordinating movements. These include fixating at an object before picking it up, and fixating at the area it should be placed before dropping it down. Another key insight from these experiments is that subjects acquire the information needed to perform each component of the task just prior to its use, rather than far ahead of time. They also found that the size of head movements is likely related to the constraints of the experiments, and that gaze shifts are almost always coupled with head movements in natural tasks.

Although these works are a step towards understanding active vision in the real world, they have not addressed the importance of body movements to select viewpoints. In all these experiments, subjects are required to stay in the same location, and the search/task display are presented to them on a table, with only head and arm movements needed to complete the task. In the real world, search cannot be conducted sitting down. Objects are frequently occluded given only one location, and people must move around in order to disocclude objects that they may be searching for, particularly in a cluttered environment.

Among the studies considering the visual search task specifically, Howard et al. (2011) conducted a real world search experiment where the search space was a table subjects stood in front of. They used an eye tracker to track subjects‘ gaze when searching for a target in

an array of 38 objects on the table. As they were interested in the formation of distractor representations, they did not put as much focus onto other aspects of the search task. This paradigm limited the search space to one table and there was little to no occlusion, so there was no need for the subject to move at all to find different viewpoints to complete the task.

Another study investigating active search in the real world used LEGO blocks as the target and distractor stimuli instead (Sauter et al., 2020). Similarly, they acknowledge a discrepancy between the classical visual search paradigm and the real world, and they address this by investigating behaviour in the real world and simplifying after. Whilst their paradigm investigated visual search in the real world, it was more focused on integrating haptics and movement by asking observers to grab the specified targets and placing them in a bin. No eye or head movement data was recorded.

Finally, Foulsham et al. (2014) conducted studies on the top-down and bottom-up aspects of active search in the real world involving head and body movement. Subjects were told to retrieve an envelope from a specified mailbox. For some subjects, the mailbox was outlined in bright pink, whilst in others it was the same as the other mailboxes. Again, they used a mobile eye tracker to track subjects' gaze during the trial. They found that the bright pink salient outline of the mailbox did not help speed subjects in finding the correct mailbox, unless it was mentioned in the instructions that the target mailbox was bright pink outlined. This was not a typical search task, as there was only one trial per observer, and there were no trials where the target was absent. However, it showcases some interesting properties of saliency with respect to active real world search.

Thus, not only has it been shown that gaze allocation in a 2D display is markedly different from the real world, there exists a gap in the literature with regard to understanding fully active observation, particularly in the specific task of visual search.

### 2.1.3 Computational models of human search

Directly addressing visual search, Najemnik and Geisler (2005) propose a Bayesian model describing optimal eye movement strategies for search. An ideal Bayesian searcher (IBS) is a model that chooses the next fixation based on maximizing information gain, whilst being able to perfectly integrate information across fixations. They show that humans perform

nearly as well as the ideal observer model when searching for Gabor patches in background noise. They also show that the other dominant model of search at the time, MAP (maximum a posteriori), was less adequate at simulating human search behaviour than the IBS.

Recently, Bujia et al. (2022) extend the work from Najemnik and Geisler, combining it with a saliency map approach to model search in natural images. They propose to feed saliency maps into the IBS as prior information. Using a deep neural network (DNN) to generate saliency maps of the images, they then fed these maps into the IBS to predict scanpaths. Comparing this model to human performance, they found they were able to achieve up to 90% similarity. This shows that the IBS can be extended to naturalistic images and still predict human performance well.

Zhou and Yu (2021) propose a Constrained Continuous Time Entropy Limit Minimization (CTELM) model for search, inspired by Najemnik and Geisler (2009)'s Entropy Limit Minimization (ELM). Using this model, they are able to account not just for the fixation location, but include fixation duration as well, using an evidence accumulation model that depends on the time spent at each fixation. This model chooses the next fixation by maximizing information gain, but rather than having just one choice per fixation as in Najemnik and Geisler (2005), for each fixation, the model must choose whether to continue fixating at the current location to gather more evidence, or to move on to the next fixation location. The constraints imposed were derived from human behaviour, and they were: 1) higher preference for shorter saccades, 2) inaccurate saccade landing position, and 3) limitations to memory. Without these constraints, the model did not perform in a way that was similar to humans. The authors propose that this may be because subjects are optimizing the balance between costs of the given constraints to choosing the optimal eye movement.

Rao et al. (2002) propose an eye movement model for visual search using iconic representations. The key features in their model include a separation of the targeting process from the decision process, meaning that gaze control was separate from target detection. Their model was fitted against human data with high quantitative and qualitative similarities.

### 2.1.4 Foraging

One area of search that directly ties into real world search is foraging. Some view foraging as a task that involves non-visually guided search in the real world (Smith et al., 2008; Gilchrist et al., 2001), whilst others define it as any task involving multiple target objects (Wolfe, 2013; Kristjánsson et al., 2022; Johannesson et al., 2016). Either way, foraging is a more general task than visual search that ties closely to how search is used in the real world.

Smith et al. (2008) conduct an experiment comparing visually guided large-scale search with non-visually guided search. Their experimental setup involved light switches arranged on a 4x4m area, and subjects were required to press a switch on the target light when found. For visually guided search, they found that consistent with search tasks on a 2D screen, feature-present search (red and green light among green) had no linear search slope, whilst feature-absent search (green light among red and green) did. Although they used a real world setup, there was no eye or head movement tracking, and there were no target absent trials.

Likewise, Kristjánsson et al. (2022) also experiment with foraging in 3D. However, instead of using a physical setup, they use VR to increase control of the environment. As a result, their target stimuli were not realistic objects. Instead, they were synthetic-looking balls or cubes of various colours against a plain grey background. They found that observers showed a smaller difference between feature and conjunction foraging compared to 2D foraging tasks, and tended to begin foraging from the bottom of the display and up.

### 2.1.5 Foraging models

There are several models within the field of foraging to explain various behaviours (see Bella-Fernández et al. (2021) for a detailed summary). A popular model for optimal patch-leaving behaviour is the Marginal Value Theorem (MVT), originally used to model animal foraging behaviour (Charnov, 1976). This model states that an optimal predator would leave their current patch to move to the next one once the instantaneous capture rate at that patch reaches the average capture rate of the entire habitat, where the instantaneous capture rate is the net energy intake.

From the MVT, several predictions can be made about optimal foraging behaviour. These include a correlation between inter-patch travel time and patch dwelling time, a correlation between patch dwelling time and patch density, and a decrease in the rate in which foragers find targets over time. These have all been observed in both animal and human behaviour (Bettinger and Grote, 2016; Wolfe, 2013; Turrin et al., 2017). Wolfe (2013) also shows that MVT and human patch-leaving behaviour align for the most basic foraging task. However, in the more general case, for both humans and animals, researchers have found that actual patch dwelling times are longer than what is predicted as optimal by the MVT.

In terms of patch-finding behaviour, Lévy processes are one of the most popular mathematical models. Viswanathan et al. (1999) proposed an optimal model using Lévy walks: if the target is within distance r, the forager approaches that patch for search. If no targets are in sight (within r), then the forager moves in a random direction and distance, sampled from the Lévy flight probability distribution function. They tested their model against bee, albatross, and deer data, and found that these animals behaved consistent to the model, albeit only in areas with sparse target distributions. However, this model does not take into account memory of patch locations or patch qualities, which do play a role in patch finding.

Anderson (1983) has also modelled patch selection behaviour using the Travelling Salesman Problem (TSP). Using this model, he found an alternative explanation for the commonly-seen "trap-lining" behaviour in some animals' foraging behaviour. However, models like Viswanathan et al. (1999) and Anderson (1983) assume that the patch locations are known, although that may not be true in the real world (Bartumeus and Catalan, 2009).

Thus, many mathematical models have been developed to explain various elements of foraging behaviour found in humans and animals in the real world. Since foraging is closely related to visual search, these models may be useful in better understanding search behaviour in the real world.

## 2.2 Robotics/computer vision

### 2.2.1 Active vision from a computational perspective

Active vision is a key component of performing search in the real world.

Bajcsy (1988) defines active sensing as "the problem of intelligent control strategies applied to the data acquisition process which will depend on the current state of data interpretation and the goal or task of the process".

Some problems that are ill-posed and difficult to solve for passive observers become well-posed and easily solved with active vision (Aloimonos et al., 1988). Examples include shape from shading, shape from contour, shape from texture, and structure from motion.

Ballard (1989) summarizes the computational advantages gained from "animate vision" vs. passive vision. Using animate vision, systems must be lightweight and highly adaptable in order to perform in real-time. He states the importance of context in understanding animate vision, as vision is "understood in the context of visual behaviours that the system is engaged in". Any model of active vision, he asserts, must function in real time, being lightweight and highly dependent on adaptive behaviours.

He further mentions the ability of animate vision to utilize coordinate frames external to the agent, in order to simplify computations during the active vision process. Another advantage is the ability to move the camera; he states that getting multiple viewpoints is more effective as well as less costly than conducting visual search with a single image of a scene. Single image search will also fail in cases where occlusion is present, whereas gathering different viewpoints will easily solve this problem.

Chen et al. (2011) also state, from the perspective of robotics systems, that active perception encourages moving sensors (selecting viewpoints) in order to constrain the interpretation of a given environment, and that placement of sensors is key to efficiency and full automation in robotics.

It can easily be seen that active vision is an integral component of visual search in the real world. Humans are also much better at search involving active vision compared to current robots. It is therefore important to investigate how humans use active vision to complete search tasks, to see if similar strategies can be applied in robotics.

### 2.2.2 Computer vision algorithms

Bruce and Tsotsos (2009) propose the AIM (Attention based on Information Maximization) model as a visual saliency computational framework. The key idea of this model is that saliency in a visual context is equivalent to a measure of the information present in a local region. They found that this model outperforms classic saliency map models (Itti and Koch, 2000) in predicting eye fixation data based on both the fixation density map and area under ROC. Although it was not designed as a model for visual search, they show how this model is able to explain various behaviours found in human visual search, such as search slopes, the effect of target-distractor and distractor-distractor similarity, and search asymmetries.

More recently, Wloka et al. (2018) propose STAR-FC (Selective Tuning Attentive Reference model — Fixation Control) that combines higher-level (objects) saliency at the central visual field with low-level (features) saliency in the periphery to predict fixation sequences. To obtain central visual field saliency scores, they used SALICON. For the peripheral saliency, they used AIM. They compared the similarity of STAR-FC to human performance against other popular saliency models including AIM, VOCUS2 (Frintrop et al., 2015), and more, in a free-viewing task. They found that this model performed best in terms of distance from human fixations for the first 5 fixations, as well as showing a spatial distribution of fixations most similar in shape to humans. Furthermore, this model is one of few models that predict fixation sequences rather than a heat map of fixations. Such sequences can be utilized to understand the time course of fixations and what they mean in a functional manner, unlike a heat map with no temporal information. Like Bruce and Tsotsos (2009), this model does not directly address the task of visual search, but is able to predict early fixations well.

Frintrop et al. (2015) developed VOCUS2, a model inspired by Itti and Koch (2000)'s saliency model. Using feature channels computed in parallel, a pyramidal structure for computations over multiple scales, and contrast computed using Difference-of Gaussians, they showed that the more traditional, biologically inspired, structure of models is still competitive against other deep learning models. They further state that the building blocks of other modern saliency models is simply the difference of a patch from the local or global

surround, in other words, the contrast. This model is a saliency model rather than a search model.

### 2.2.3 Deep learning search models

Adeli and Zelinsky (2018) propose a deep network, Deep-BCN, to model attention in a visual search task. This model is inspired by Desimone and Duncan (1995)'s Biased Competition Model. The model uses several convolutional layers to model the brain's early visual regions (V1 to V4), with outputs from these layers spatially biased by top-down modulations. They introduce DLPFC nodes (Dorsolateral PFC), and model top down biases from the DLPFC to IT, and from IT to V4 using gradient signals feeding back from these layers.

This model was trained on human fixational data, and is thus able to make eye movements. Comparisons to human behaviour reveal that the model generated similar scanpaths to subjects, as well as a similar number of fixations and distance travelled across the screen. Furthermore, the model was able to predict typically seen effects in search, such as more fixations and higher search slope for target absent trials. However, the search task the model was trained on was a very simple task, consisting of object instances placed onto a blank background. There also seems to be no easy way to extend the model to function on natural image search.

Wei et al. (2016) also propose a neural network, termed Sparse Diverse Regions (SDR) classifier, to model fixation prediction in a visual search task. This model incorporates Inhibition-Of-Return (IOR), a well-studied feature of human search (Klein, 2000), on top of a sparse region ranking SVM (RRSVM). They trained this model to perform image classification without localization data, and report results comparable to State-of-the-art models in predicting fixation density maps during search tasks. However, this type of model only generates a priority map of where people are likely to fixate, without giving a temporally ordered scanpath of fixations. Moreover, the only biological inspiration explicitly included in this model is that of IOR.

A model that is capable of performing scanpath prediction in visual search is Yang et al. (2020)'s Inverse Reinforcement Learning model. They created a dataset COCO-Search18,

with 18 object categories and over 3000 search trials, in order to train their model. Search images were natural scene images, and human subjects found the target within 4-5 fixations most of the time. Although they showed that their model outperformed baseline in this task, it was a very simple search task, it was their own created dataset, and the model's probability of finding the target was still well below human performance. Thus, not only does this show that a deep learning model like this does poorly, the lack of biological inspiration further provided no insight into understanding mechanisms of human search.

Another model predicting fixation sequences is Zhang et al. (2018)'s Invariant Visual Search Network (IVSN). This model predicts fixation sequences with zero-shot learning, achieving near human performance on tasks as difficult as finding Waldo in the classic Where's Waldo search task. The model uses a feed-forward deep network, VGG-16, to process images, and includes top-down modulation to simulate attention by convolving the target representation (obtained from running it through VGG-16) with the search image representation (also obtained from running it through VGG-16). Using this method, they were able to predict fixation sequences with similarity scores to human fixations significantly higher than chance. However, the null models they compared against were very basic, and as such were a poor baseline for comparison.

### 2.2.4 Robotics models

Chen et al. (2011) provide a good overview of research on active vision in robotics, including active visual search. They define active vision as a question of "where to look": a robot's visual sensor must be moved and reconfigured frequently in order to achieve purposeful visual perception. One of the key problems in robotics is viewpoint selection — sensor planning. Indeed, much research has been conducted on the issue of purposive sensing, selecting viewpoints that increase a robot's understanding of the environment. Such sensor planning is evidently relevant in a robot conducting active search. The importance of viewpoints is especially clear in the case of search where parts of the search space are occluded from the camera's current view. Different viewpoints would thus be required in order to find the target. Because a brute force solution is too expensive, sensor planning is needed in order for a robot to complete such a task.

Ye and Tsotsos (1999) defined the sensor planning problem for visual object search. The sensor planning problem involves selecting the correct camera view angles, camera directions, and robot locations, in order to get the target in view for object recognition algorithms to detect the object. Sensor planning is an integral component of any active visual search task.

Shubina and Tsotsos (2010) modelled active real world search for a mobile robot. They propose multiple different cost functions to help the robot select the next action. Ultimately, they found that maximizing the probability of detecting the target whilst minimizing distance to travel yielded the best performance in the robot. However, no comparisons were made to human performance.

Rasouli et al. (2020) discusses attention based active search for mobile robots. This paper proposes to integrate an attention model to help improve robot performance in a visual search task. This involved a bottom-up module using AIM (Bruce and Tsotsos, 2009), and a top-down attention module on the color feature. The color similarities between target and scene were computed using the histogram backprojection technique (Swain and Ballard, 1991). They were able to improve search by up to 42% in structured and 38% in unstructured environments in simulation studies. No comparison was made to human performance again, so it is unknown if the strategy used here corresponds in any way to human search strategies.

Ye et al. (2018) propose recognition-guided action policy learning for active search with mobile robots. Their approach involves a deep neural network for object recognition, as well as deep reinforcement learning for action prediction. In particular, they tested a variety of different reward functions, and found that using a reward proportional to the size of the bounding box containing the target (only if it is larger than previously detected bounding boxes) gave the best performance. Again, they had no comparisons to human performance.

Schmid et al. (2019) also explore a reinforcement learning solution for active visual search with robots, but focus on the different subtasks involved. They define these subtasks as explore, approach, and active termination. They use a deep recurrent Q network (DRQN) to implement a solution that addresses all three subtasks. Exploration involves navigating the agent to a pose such that the target object becomes visible. Approach involves moving the agent closer to the target, starting from a pose where the target is already visible. Active termination involves making a "declaration" action if the approached object is the target,

and not making an action if it is not the target. Using simulation studies, they found that for difficult search scenes (multiple large rooms), exploration was the limiting factor for success. Once again, they made no comparisons to human performance.

Sjöö et al. (2012) used topological spatial relations "on" and "in" to improve robot visual search performance. Using indirect search (first proposed by Garvey (1976), and first implemented by Wixson and Ballard (1994)) with topological spatial relations, they designed an algorithm to search for intermediate objects with "on" or "in" relations to the target, as specified in the task. This allowed for improved performance in indirect search due to the reduced search space, easier detection of intermediate objects due to their larger size, as well as accounting for occlusions from containers for "in" relations to choose better next views. Their experiment showed a success rate of 85% using indirect search, whereas uninformed search had a success rate of only 35%. Thus, it seems that using topological spatial relations can make search easier for a robot.

There are a multitude of studies investigating active search with mobile robots. There are also numerous studies exploring computational models for human visual search. However, there seems to be little overlap in these two areas — the robotics studies do not consider human performance comparisons, and the computational models do not consider active search in 3D environments. A possible contributor to this gap could be the absence of a unified paradigm for measuring human active search, making human performance comparisons in active search difficult.

## 2.3 Summary

In summary, there has been extensive research conducted on human visual search. Nonetheless, there still exists many gaps, especially in the ecological validity of these studies, extending into active vision and the real world. Indeed, there has been a push toward using naturalistic images and real world environments, but there are several avenues that remain unexplored in this area, particularly in understanding search that requires movement in an environment.

Furthermore, recent advances in machine learning and computer vision have contributed

to computational models for search. Although robotics models consider the search problem from an active vision perspective, they have yet to produce results comparable to humans. Eye movement models have shown promising results in predicting fixation sequences on a flat computer screen, but have not been applied to tasks in any 3D, real world environment. Existing deep learning models have little biological inspiration and mostly perform subpar compared to humans. Thus, a clear gap in this area of research is an understanding of human behaviour during fully active observation.

# Chapter 3

# Problem statement

There is a lack of research including fully active observers conducting visual search in the real world, as existing research on active observers have not considered the added dimension of changing viewpoints by moving around obstacles in space. A lack of computational models that predict human behaviour during an active visual search task is also apparent, as existing models have limited generalizability to natural image search tasks, as well as limited biological inspiration. As Ballard (1989) states, animate vision models that are general purpose algorithms must be shunned if they require vast amounts of computational power in order to be practical. Deep learning solutions for search are either trained on vast amounts of data with little biological inspiration, or are limited to the specific stimuli that they have been trained on.

Thus, my thesis aims to address this gap in research using a human active visual search experiment conducted in the real world. My research questions are separated into two sections, the first considering direct results of human subjects performing active search, and second, how those compare to existing computational models of search.

## 3.1 Research questions

### 3.1.1 Investigating human active search

I ran an active search experiment in a controlled 3D space using the PESAO environment (Solbach and Tsotsos, 2020) in order to answer the following research questions:

RQ1. How do eye and head movements during active search in a controlled environment vary by target presence, set size, and visibility?

RQ2. Are there any eye and head movement patterns that can be uncovered corresponding to different search strategies?

These questions will be addressed in Chapter 4. Although results from this experiment do not generalize to all real world search tasks, this is a first step in gaining an understanding of the potential strategies and analyses in the real world that cannot be discovered in 2D.

### 3.1.2 Comparing to computational models

After gathering and analyzing the human data from the active search experiment, I compared these results with computational models, in order to answer the following questions:

RQ1. How do computational models perform in comparison to humans?

RQ2. Are strategies used in computational models for active search comparable to human strategies?

These questions will be addressed in Chapter 7.

# Chapter 4

# Human active search experiment

## 4.1 Methodology

The following section details the methods used to conduct my active visual search experiment. I decided to utilize a physical real world active visual search task for a number of reasons. Firstly, such a task has the highest ecological validity in comparison to viable alternatives, such as VR or first-person-view computer simulations. Even in virtual reality, there is no guarantee that subject eye and head movements in a virtual environment correspond to the same movements in the real world (Zhao, 2011). We also wanted subjects to move around the environment naturally, whereas virtual reality setups would require subjects to move around in space using a controller rather than physically moving. Furthermore, the setup was designed with the possibility for subjects to manipulate objects in the search space, although this variable has been excluded for now. The experiment was approved by the Office of Research Ethics (ORE), with certificate number STU 2023-004.

### 4.1.1 Participants

Subjects were 61 York University students (13 excluded), recruited via printed posters and emails sent through the EECS graduate department. The 48 valid subjects were counterbalanced with a 4x4 between-subjects design. The two factors used for counterbalancing were: 1) table and cage layout, and 2) trial version. Each factor had four levels, adding to

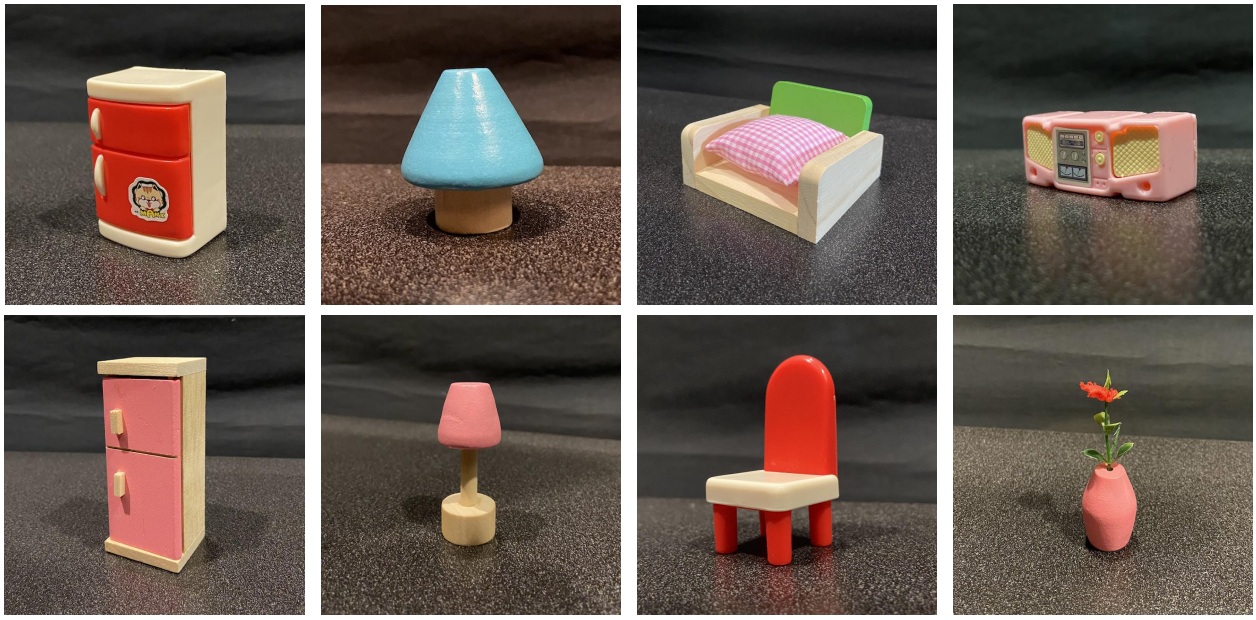


**Figure 4.1:** Example stimuli

16 combinations of the two factors. Thus, three subjects were run on each combination. The number of subjects was specified a priori, in order to ensure enough data was collected in each combination to analyze appropriately. Informed consent was obtained to participate in the experiment, as well as to be recorded with first-person and third-person video cameras.

### 4.1.2 Stimuli

Stimuli were 1:12 size dollhouse furniture of everyday objects, such as chairs, tables, couches, etc (see Figure 4.1 for some examples, all objects in Appendix A). They are made in a variety of colours, materials, and sizes. Stimuli were placed in pre-determined locations in the setup for every trial, and no trial had the same set of objects in the same locations. There were 4 different sets of 12 trials. Figure 4.3 shows an example of a trial with some objects, and a subject inspecting the objects for the target.

### 4.1.3 Setup

The experiment was conducted in the PESAO (Solbach and Tsotsos, 2020) environment, in a 3x4m space with 5 light sources and 6 Optitrack camera trackers. Multiple light sources

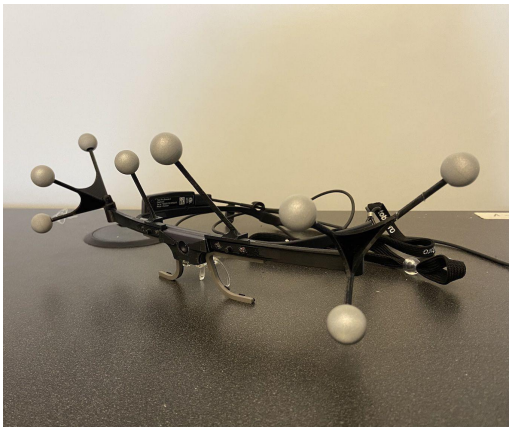


**Figure 4.2:** Tobii Glasses 2 (mobile eye tracker) with OptiTrack reflective markers mounted (for head tracking)

**Figure 4.3:** Example of a subject performing the search task

were used to ensure more uniform lighting, and the Optitrack cameras were used to track the target location as well as the observer's head movements. The subject was asked to wear eye-tracking glasses and a set of passive tracking markers mounted on the glasses (Figure 4.2). Tables and wireframe cages were arranged in pre-determined configurations in the area, such that observers would have to walk and navigate around to obtain different viewpoints during the search task.

We chose wireframe cages so that the Optitrack cameras would still be able to track markers through the shelving, rather than being occluded by solid wall shelves. Black coverings were placed on some faces of the cages, increasing the amount of occlusion, whilst ensuring that enough cameras could still see into each cage to maintain accurate tracking. The tables and wireframe cage locations stayed the same throughout the experiment for each subject, but they varied between 4 different versions across subjects (Figure 4.7). Figure 4.4 shows an example of an empty setup with tables and cages. This forms the basis of the setup, and objects are placed afterwards.

### 4.1.4 Procedure

In order to answer my first two research questions, I designed an active visual search experiment. For this search task, subjects had to determine if a specified target is present or

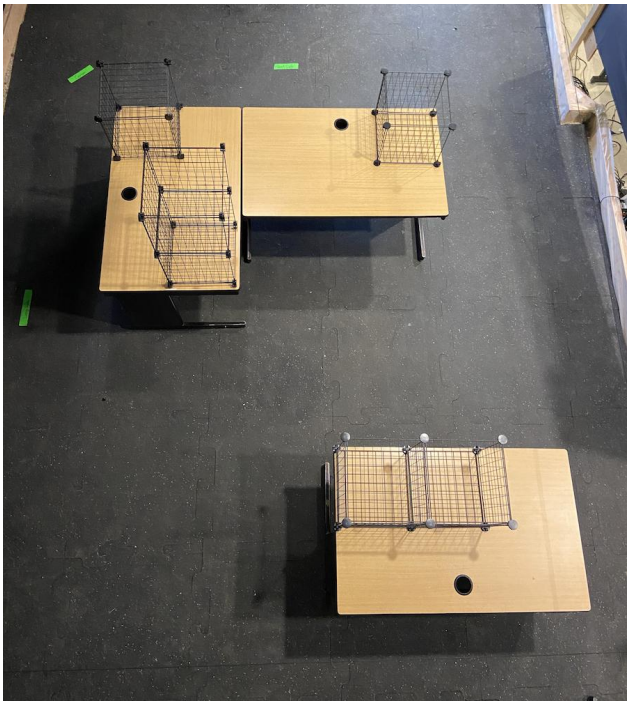


**Figure 4.4:** Empty setup with tables and cages

not. The target was presented as an image on a monitor at the start of each trial for 5 seconds. All targets were presented with their canonical view. Once the image disappeared, they could turn around to begin the search. Subjects could move freely within the experimental setup in order to complete the search (see Figure 4.5). Throughout the experiment, subjects were wearing a set of Tobii Glasses 2 with markers attached to track both their eye and head movements. The eye trackers were calibrated once at the start of the experiment. Once the subject made a decision, they had to verbalize their response. If they believed the target was present, they were told to fixate on the target, whilst saying "I have found the target". If they believed the target was absent, they had to say "I think the target is absent". Subjects began from the same location for every trial (in the far right corner — top right in plots in Figure 4.7), and each subject performed 12 trials in total.

I manipulated three independent variables:

(1) Target presence [present, absent]

(2) Target visibility from starting location [yes, no]

(3) Set size [30, 40, 50, 60]

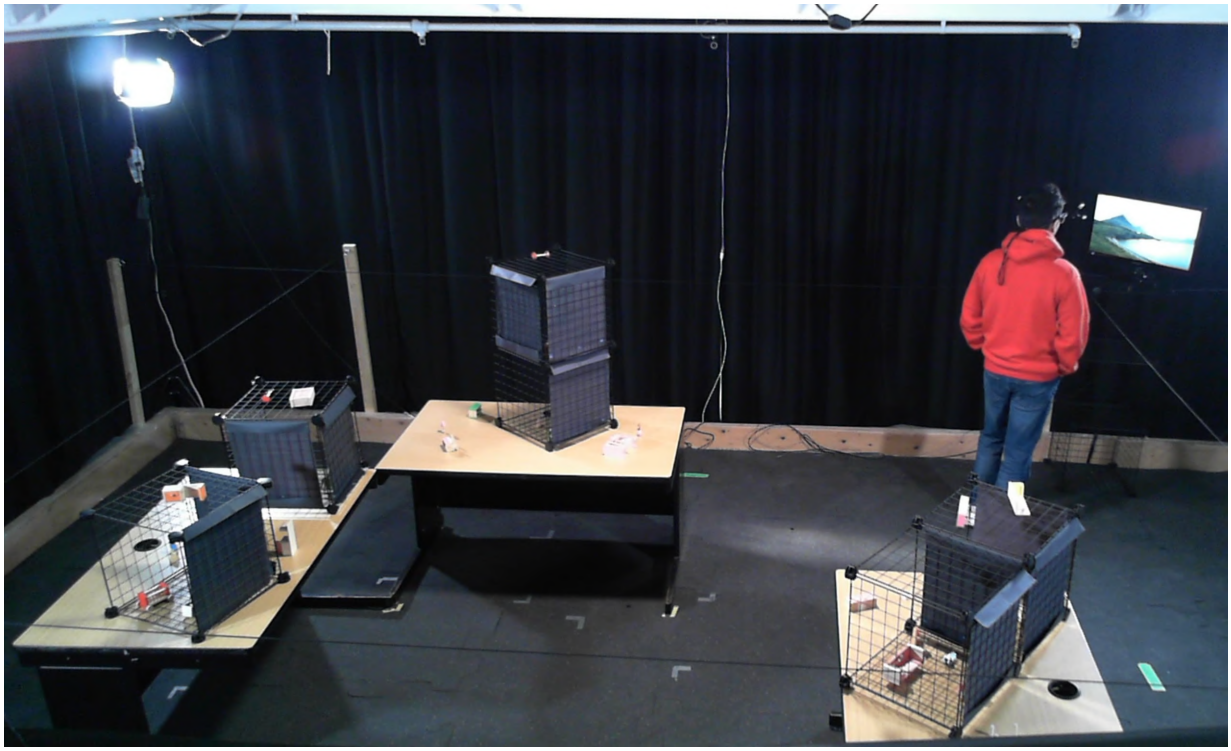


**Figure 4.5:** Experiment environment example — layout 4

The set size variable was included in order to estimate a search slope for the subjects. Of these variables, I fully counterbalanced target presence within the 12 trials for each subject, and counterbalanced the visibility and set sizes between subjects, such that there were always an even distribution of set sizes and target visibilities, but they were paired to different trial orders across the different trial versions.

I measured subjects' response time, accuracy, eye movement data, and head movement data. I also tracked the location of the tables and cages, as well as target objects.

I had initially intended to include trial difficulty as an independent variable as well. This would include trials with pop-out like search and different levels of conjunction search, to see if similar differences could be found in performance between these types of search as seen in classic 2D search tasks. Using the features size, shape, texture, and color, I tried changing the number of features shared between the target object and distractors to produce different trial difficulties. By this definition, I had trials with the target object sharing 0 to 1 feature with distractors to simulate pop-out difficulty, sharing 1-2 features with distractors for conjunction of 2 features difficulty, and sharing 2-3 features with distractors for conjunction

of more features difficulty. However, it was harder than expected to create trials that truly differed in terms of search difficulty.

In particular, as I was using sets of toy furniture objects, there was a large variety of object shapes, and some objects with different categorized shapes would have similar appearances (for example a stool and a round table, see Figure 4.6a). On the other hand, objects categorized as the same shape could vary wildly in appearance as well (Figure 4.6b). Furthermore, a traditional pop-out search entails more homogeneous distractors than possible given the wide variety but limited number of objects I had. Since similarity between distractors was not controlled for, a large distribution of feature values was possible for them. Thus, even if the target shared no features with any distractors, the heterogeneity of the distractors meant pop-out search would not be possible with this design. A better measure for object similarity, as well as controlling for distractor feature distributions, should be used in the future to properly differentiate easier and harder trials.

Finally, even if I were able to control for and quantify object similarity, the effects of different trial difficulties would likely be minimal in a 3D, real world setting. In a 3D environment, occlusion and pose matter too — if the target object is occluded from your view, you will not find the target there, regardless of the trial's difficulty. If you are presented with the non-canonical pose of the target, you may need to take even more views from different perspectives once a target is in view to confirm the object. Thus, the 3D nature of the task interferes with controlling trial difficulty by feature similarities, and any effects of the target's uniqueness would only come into play once the target object is in the field of view.

**(a)** A stool and a table. These objects' shapes were categorized as different, but they look similar.

**(b)** Two cupboards. These objects were both categorized as cupboards, but they look different.

**Figure 4.6:** Example images of objects of same and different categorized shapes.

### 4.1.5 Choosing object placements/Generating trials

Trials were generated using an algorithm, with the following constraints:

(1) 12 trials total
(2) 6 present trials, 6 absent trials
(3) 3 trials with each set size (30, 40, 50, 60)
(4) the target is in a different table or cage each time
(5) each table or cage has a similar number of items (+/- 2)

First, I randomly varied the order of the set sizes and the target presence, such that each trial ended up with different combinations of these variables. I then generated objects for each trial given the set size, and whether or not the target was present. Targets and distractors were chosen from a pool of 119 objects. A full list of these objects and their attributes (size, shape, colour, texture) can be found in the appendix.

After generating the objects for the first trial, they were randomly assigned a table or cage (chosen from 9 possibilities: table 1, table 2, table 3, cage 1, cage 2, cage 3, cage 4, cage 5, cage 6) such that there were approximately the same number of objects in each.

Subsequent trials were generated such that as many objects were shared between consecutive trials, given the trial order. This was done to minimize the time subjects needed to wait between trials as we changed the stimuli in the setup. If a target was present, it would always be removed in the trial immediately after. Objects new to the next trial are then

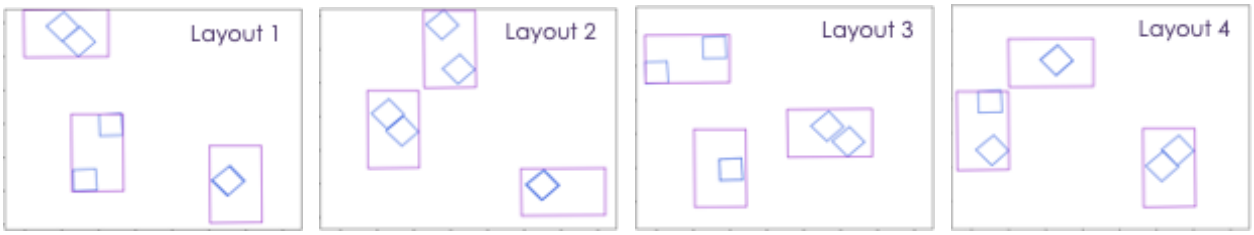



**Figure 4.7:** Experiment table and cage layouts

assigned a location in order to balance out the number of items in each location as much as possible. There were never 2 trials with the same target object.

Four sets of trials were generated, and subjects were run on one of the four trial versions.

### 4.1.6 Choosing layouts

The layouts of the tables and the cages were not chosen randomly. The number of tables and cages was kept constant across the layouts to make for easier analysis and trial generation.

3 tables were chosen to allow for enough space in between the tables, regardless of the many configurations possible, for people to walk through the setup. 6 cage shelves were chosen as that was the number that gave sufficient table space and cage space with the 3 tables. There was always a tower of 2 cages, with the second cage being taller, to increase the vertical height variance of the search space. There was always a unit of 2 cages horizontally connected, to see if this unit would be countes as 1 "location" by the subjects or if they would be discrete. And finally, there were always 2 cages that stayed as individual cages.

Several layouts were attempted during the piloting phase, but the ones used in the experiment were the ones that I decided allowed for the most interesting and large movements during the search task (Figure 4.7). These particular layouts' cages and tables were placed in such a way that subjects needed to move around them in order to look at particular spots that the target could be in.

Black coverings were then added to certain faces of the cages to increase the amount of occlusion from any standing position, in order to further encourage movement through the setup from the subjects. In fact, these coverings ensured that there would be no single

standing location within the experiment setup that would allow the subject to see every location, thus eliminating the optimal standing position to conduct the search across every trial.

There are a vast number of potential configurations of tables, cages, and objects. This particular sampling was chosen in order to make the task tractable. Hopefully, this is a somewhat representative sampling of the configuration space that we want to tackle, and further discussion on this appears below.

# Chapter 5

# Experiment results

This chapter begins with a description of some of the data processing and extraction of behaviours from the eye and head movements, followed by descriptive statistics and a summary of the initial data analysis. Then, I discuss the possibility of comparing my results to 2D search tasks.

I also address research question 1: How do eye and head movements during active search vary by target presence, set size, and visibility?

## 5.1 Eye and head movement metrics

Gaze event data is directly exported from Tobii Pro Lab, including fixations, saccades, their respective durations, as well as their 2D and inferred 3D coordinates from the eye images. The Tobii Glasses 2 used had a sampling rate of 50 Hz, and gaps in the eye movement data up to 75ms were interpolated. A fixation is defined as a sequence of raw gaze points with estimated velocity below 30°/s. Head location and orientation, as well as target location, is directly exported from Motive Optitrack, giving the xyz location as well as orientation of the rigid body. The data is synchronized and merged using PESAO's pipeline (Solbach and Tsotsos (2020)).

For every subject, the data streams from the Tobii Glasses 2 and Motive Optitrack are saved, as is a controller script with trial start and end times. Using their respective timestamps, the three data streams are merged, with missing values interpolated across

time. After the merging, a single large dataset including all eye and head movement data, along with markers for the start and end of each trial, is generated.

Each trial ranges from 4 seconds to 3 minutes long (there was no time limit), and there are a total of 576 trials combined across subjects. This leads to an extremely rich dataset with much to explore. For this thesis, I have chosen to use a select number of metrics that were most relevant to my analyses and discovering strategies.

In each layout of the experiment, there were 3 tables and 6 wireframe cages that the objects were placed in. In the following analyses, I calculated the 3D view vectors of each fixation and logged which of the 9 locations each fixation was pointing at, or if they were not pointing at any.

Aside from the raw fixation, saccade, and subject head location and orientation data, here are the metrics extracted and used for my analysis:

**Crouches** Times when a subject's recorded height is less than 2SD below their average height across the experiment

**Head tilts** Times when a subject's head is tilted at least 30°(pitch) or at least 80°(roll)

**Distance travelled in 3D (m)** Total distance a subject travels across 1 trial

**Revisits** Number of revisits a subject takes to a table or cage in the setup

**Cage location scans** Continuous fixations a subject takes in the same cage location

**Table location scans** Continuous fixations a subject takes in the same table location, including fixations to the cages on the tables

**Exploration (wide scanning)** Sequences of fixations where no more than 3 continuous fixations are directed toward the same location

**K coefficient** Measure of exploration vs. exploitation (Krejtz et al., 2016). Values $< 0$ indicate exploration, values $> 0$ indicate exploitation. K coefficient is calculated as the mean difference between z-scores of each saccade amplitude and their preceding fixation's duration.

## 5.2 Overall performance

Of the 61 subjects ran, 3 were excluded due to low Tobii gaze samples ($< 30\%$) and 10 were excluded due to missing data points transmitted from the Motive Optitrack system. Overall, subjects were accurate during the task (overall accuracy 93.2%, SD=7.5%), though their false alarm rates were higher compared to their miss rates (10% vs. 3.5% respectively). Overall performance across all trials did not indicate any significant learning effect on accuracy. However, when data was split between target present and absent trials, a few significant differences were found. Response time, distance travelled, and number of fixations decreased over time for target present trials, whilst there was no relationship between any of the variables to time in target absent trials (see Figure 5.1). Pearson's correlation coefficients for target present trials were significant for all three metrics (response time: $r = -0.13$, $p = 0.03$, number of fixations: $r = -0.13$, $p = 0.02$, distance travelled: $r = -0.11$, $p = 0.05$), whilst none were significant for target absent trials ($|r| \leq 0.1$ for all).

The difference between target present and absent trials was quite unexpected, as previous search literature has shown learning effects present in both. For example, Neider and Zelinsky (2006) show that scene-constrained objects (jeeps on the ground, blimps in the sky) had improved efficiency in response times and number of fixations for both target present and absent trials. Geyer et al. (2010) also show that contextual cueing improvements were found in both target present and absent trials over different conjunctions of features. In order to remove any additional search cues such as scene grammar/structure (grouping a toilet with a tub and a shower in a "bathroom" area), I explicitly avoided such groupings, and changed the orientation and directions of object placements. This way, there should be no grouping of objects in particular cages or tables that could aid in search efficiency. I also only included 12 trials, a number I thought would be insufficient to cause any learning effects over repetition. It was therefore surprising that any efficiency was seen at all, especially as it visually seems to occur within the first three trials.

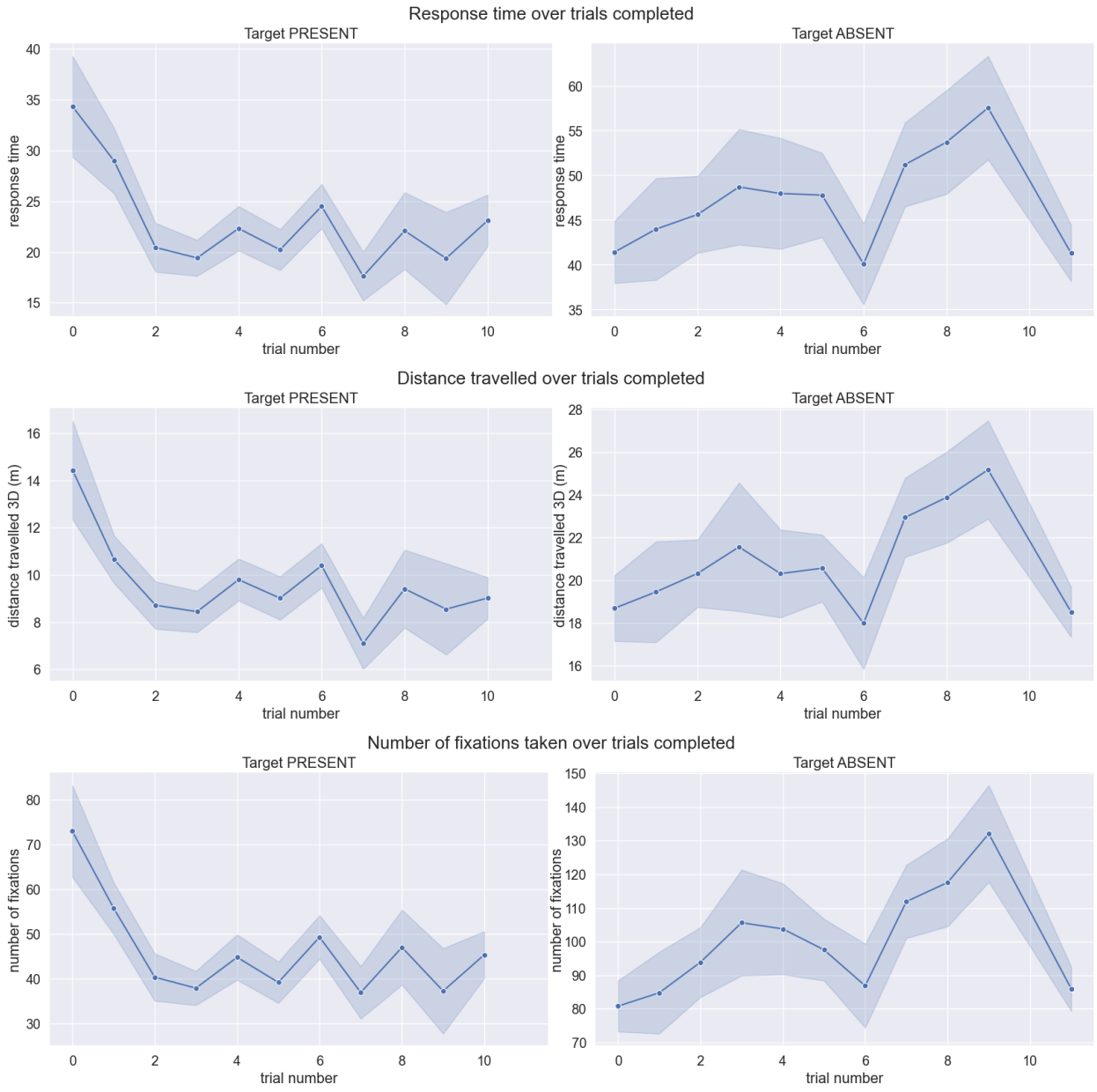



**Figure 5.1:** Response time, number of fixations, and distance travelled plotted over trial number, separated by target presence (Present on left, absent on right). It can clearly be seen that all metrics show a drop in target present trials, indicating a learning effect, whereas no significant pattern is observed in target absent trials. The dip in trial 6 on target absent plots is due to the fact that only one trial version had a target absent trial as trial 6.

To investigate the possible other causes of the gained efficiency for present trials, I analyzed changes in the number of unique locations looked at, the number of revisits, number of cage and table scans, and lengths of cage and table scans over trials. There seemed to be no change over the trials, aside from the number of revisits, which decreased slightly and had an insignificant correlation of -0.11 (p=0.08). The decrease in distance travelled suggests that subjects become more familiar with the setup, thus reducing the amount of walking needed in order to cover the entire search space. Why target absent search did not induce similar learning effects is uncertain, but it is possible that target absent search takes longer to develop efficiency, and thus did not emerge for the 12 trials in my experiment.

After analyzing the accuracies of the trials, I found two particular trials that subjects did poorly on; trial 1 in trial version 3, and trial 1 in version 4 of the task. Figure 5.3 shows the target images shown to subjects for these trials. This was surprising, since many subjects actually had multiple fixations land on the target, yet decided that it was absent. This may have been due to some color issues in the target image presentation, such that subjects believed the target to be a different color than what it really was. Figure 5.2 shows the accuracy of every trial in every trial version. Generally, accuracy for most trials was high, and trial version 4 seems to be the easiest, with all subjects getting all trials correct other than the first one.

In order to gain a better understanding of the overall behaviours of eye movements, I used the K coefficient from Krejtz et al. (2016) to gather a measure for exploration vs.

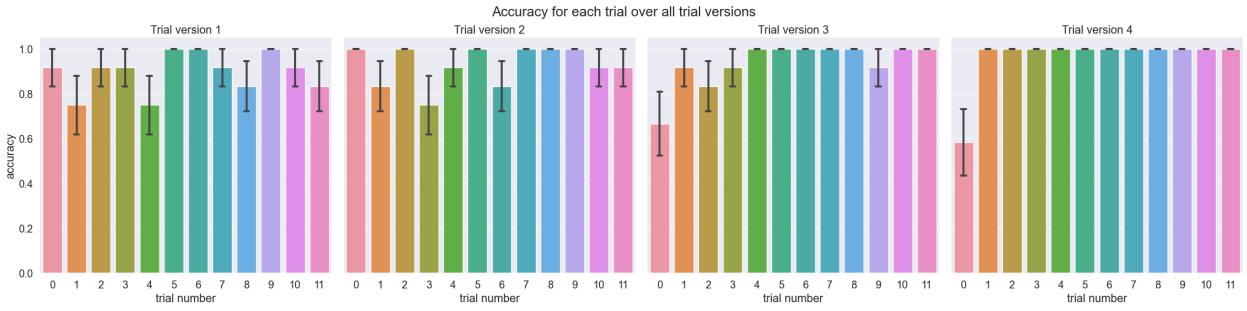



**Figure 5.2:** Accuracy of each trial in every trial version. Accuracy is near or at ceiling in most cases, except the first trials in trial version 3 and 4.

**Figure 5.3:** Target objects for trial 1 in version 3 and version 4. Subjects performed worst on these two.

exploitation in fixations and saccades. The equation to calculate K is as follows:

$$K_i = \frac{d_i - \mu_d}{\sigma_d} - \frac{\alpha_(i+1) - \mu_\alpha}{\sigma_\alpha}, K = \frac{1}{n}\sum_n K_i$$

In this equation, $\mu_d, \mu_\alpha$ are the average fixation durations (d) and saccade amplitudes ($\alpha$), and $\sigma_d, \sigma_\alpha$ are the standard deviations for fixation durations and saccade amplitudes, over all $n$ fixations. The overall K coefficient is thus the average of the $K_i$ coefficients. As mentioned earlier, a positive K coefficient indicates exploitation, whilst a negative K indicates exploration.

There was a slight bias towards exploration in the initial 10% of each trial ($K = -0.04$). This was significantly more negative compared to the rest of the trial ($K = 0.002$, $p < 0.001$), meaning saccade amplitudes were larger and fixation durations were shorter at the start of the trial, so subjects were making more eye movements that travelled further to observe a wider range of the search space.

As for target-facing fixations, I investigated whether their frequencies would change depending on the distance to the target, and the relationship between target distance and time. The target fixation with the largest distance from target was 3.5 meters. Figure 5.4 shows the distribution of the frequency of target facing fixations to distance from target. Looking at the distance over the duration of a trial, I also found a significant negative correlation

between distance and time (see Figure 5.5, $r = -0.29, p = 0$), indicating that subjects moved closer to the target in subsequent target fixations over the duration of the trial.

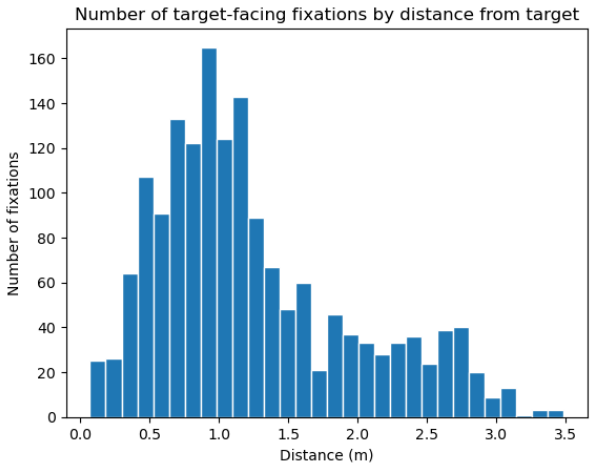



**Figure 5.4:** Number of target-facing fixations by distance to target. Most target facing fixations were between 0 to 1 meters from the target.

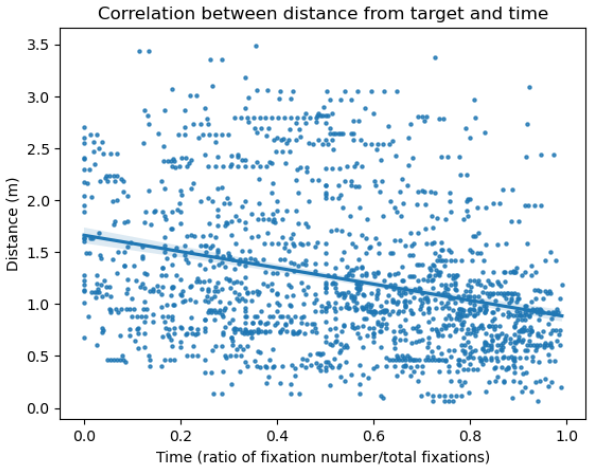



**Figure 5.5:** Correlation between distance from target and time during trial, represented as a ratio of the fixation number over the total number of fixations to normalize the time factor.

I included the 4 different layouts and 4 different trial versions in order to control for the effects that these would have on results. It is interesting to analyze the differences in these after running the experiment. Thus, I conducted a 2-way ANOVA (layout x trial version) on the number of fixations, response time, and distance travelled. There were no significant interactions, but significant main effects of layout and trial versions were found for number of fixations (p=0.01 and 0.04 respectively). No significant pairwise differences were found for either of those main effects. Through this analysis, it seems that any effects of layout and trial version were not so significant that it affected general subject performance. Where these factors come into play is later on, in the strategy analysis (chapter 6).

My first research question deals with whether the three independent variables (target presence, set size, target visibility) affect eye and head movements during the task. I first quantified and broke down the various eye and head movements from the raw data as mentioned above, then conducted two 2-way ANOVAs in order to analyze their effects on these metrics.

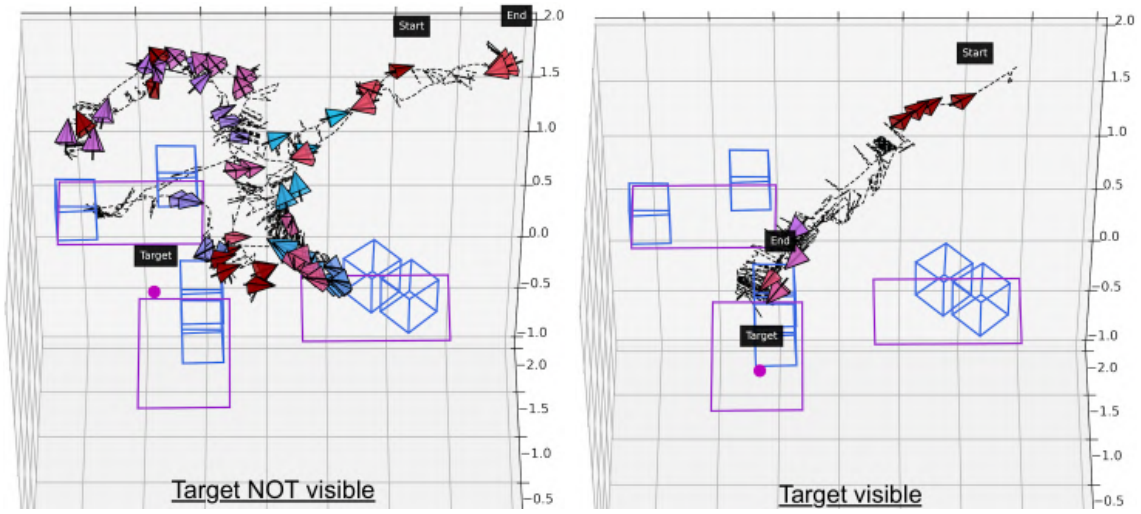
First, isolating target present trials to include the target visibility variable, I conducted a 2-way ANOVA (target visibility x set size) on select metrics (Table 5.1). Target visibility was determined as whether the target was visible or not when standing at the starting location, in the corner of the environment. Overall, there was a significant main effect of set size. Doing Tukey's HSD revealed the main groups that were significantly different were set sizes 30 to 60, and 30 to 50.

Interestingly, there was no significant main effect of visibility from the starting location for any factors. This may have been because objects labeled as "not visible" were only not visible from the starting location. Further analysis shows that subjects move out of the starting location quickly (after an average of 4 fixations), thus making this visibility factor rather trivial, due to the way it is defined.

Taking a further look at the data, I visualized subjects' search-paths in 3D, as shown in Figure 5.6. A search-path is the subject's head trajectory with head directions for each fixation that occurs over the course of a trial. Each fixation is represented by a viewing frustum, and the path of the subject is represented with the dotted line. The first and last fixation frustums are labelled "start" and "end". The temporal order of fixations is represented by color, with blue being the start of the trial, following a gradient to purple, and to a pinkish-red signalling the end. The target, if present, is indicated with a bright magenta dot, labelled "target". Any target-facing fixations are colored dark red.

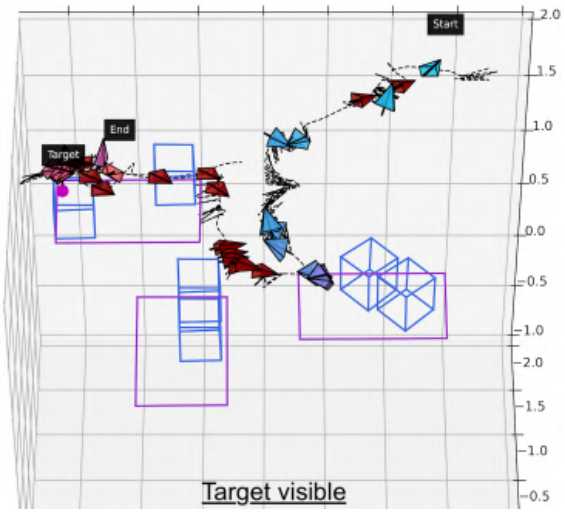
One thing in common that can be noticed across subjects is that the target visibility rarely affects their search-paths at all, until the target is found. The middle image in Figure 5.6 shows a trial where the target was visible from the starting location. However, the subject clearly starts searching in their regular pattern, and misses the target until they reach it in sequence.

Secondly, I conducted a 2-way ANOVA (target presence x set size) on the same select metrics (Table 5.2). Overall, there was a strong significant main effect of target presence. Calculating effect size $\eta^2$ gives large values of around 0.3 for response time, number of fixations, and number of revisits. $\eta^2$ was 0.45 for distance travelled, but only around 0.05 for number of crouches and accuracy. Thus, there is a large effect of target presence on subject performance in these metrics, and smaller but still significant effects of set size.




**(a)** Target not visible from start.

**(b)** Target visible and seen from start.




**(c)** Target visible but not seen from start.

**Figure 5.6:** Not visible vs. visible trials' search-paths. The bottom image shows a visible trial with similar path to not visible, showing that subjects sometimes miss the target and go on their typical search path even when it is visible from the starting point.

| Metrics/variables | Set size | Visibility |
|---|---|---|
| Num fixations | *0.006** | 0.817 |
| Accuracy | *0.0097** | 0.398 |
| Num revisits | *0.000** | 0.837 |

**Table 5.1:** 2-way ANOVA (Set size x Target visibility) on select metrics. No interactions were significant. Cells italicized with an asterisk (*) are significant.

| Metrics/variables | Set size | Target presence | Set size x Target presence |
|---|---|---|---|
| Response time | *0.003** | *0.000** | 0.728 |
| Num fixations | *0.000** | *0.000** | 0.649 |
| Num crouches | 0.203 | *0.000** | 0.731 |
| Distance travelled | 0.054 | *0.000** | 0.898 |
| Accuracy | 0.359 | *0.000** | *0.005** |
| Num revisits | *0.000** | *0.000** | 0.602 |

**Table 5.2:** 2-way ANOVA (Set size x Target presence) on select metrics. Cells italicized with an asterisk (*) are significant.

Thus, to answer my first research question (How do eye and head movements during active search vary by target presence, set size, and visibility?), visibility seems to have little effect on general metrics, whilst target presence has the largest effect, and set size has some effect only for certain pairwise differences. On the surface, the significant effect of target presence is congruent with 2D search, as is the significant effects of set sizes.

Furthermore, to justify this particular sampling of table, cage, and object configurations, we performed ANOVAs showing the weak significant effects of layout (table and cage configurations), as well as trial version (object configurations). Certainly, any discovered behaviours may not extrapolate to edge cases such as environments with only one object, or with no occlusions whatsoever. However, it seems that results from this experiment are able to generalize within configurations that are less extreme.

The next section investigates the similarities and differences between this task and 2D search in detail.

## 5.3 Comparison to 2D visual search

The majority of search literature has been conducted on 2D screens. Naturally, it would be interesting to know how a 3D active search task can compare to this large body of research. However, I have noticed a few issues in being able to directly compare performance and behaviour in these two types of tasks.

Firstly, 3D active search involves many head movements that are unnecessary in a 2D search task. Accordingly, the response time of a 3D search trial is significantly larger, in the range of seconds up to 3 minutes. Computing a search slope of $\frac{\text{response time (ms)}}{\text{set size}}$ to gauge the difficulty of a trial thus has a much different meaning. In 2D search, this computation is done with the assumption that no large movements are made, and the only part of the response time not being used for search is the time it takes for the subject to press a button to respond, which is typically a constant few hundred ms that can be deducted from the response time itself. However, in 3D search, the response time encompasses not only the time it takes to search, but also the time to make eye, head, and body movements, which are much more variable and take longer.

To address this difference, I attempted to compute the search slope using only the aggregated time of all fixations in a trial, thus eliminating the time taken to saccade and make head and body movements. Doing this yielded slopes of 80ms/item for target present trials (SD=302), and 137.4ms/item for target absent trials (SD=242). The large standard deviations in search slope for both target present and absent show that this slope value is not very reliable or meaningful, especially since 17 present and 15 absent slopes were negative. Looking further at the slope ratios of the task, the average slope ratio (absent/present) was 2.76, a reasonable-looking number compared to the typical search task. However, the standard deviation was 7.8, again a very large number, suggesting that this ratio is not a very reliable or stable number, and is thus rather meaningless. Indeed, the lowest ratio was -5, and the highest 37.

Secondly, regarding the fixation location data, there are a significant number of fixations that are not pointing at any potential targets in particular. Classifying each fixation as either "looking at" a location containing objects or not, there are about 21% of fixations

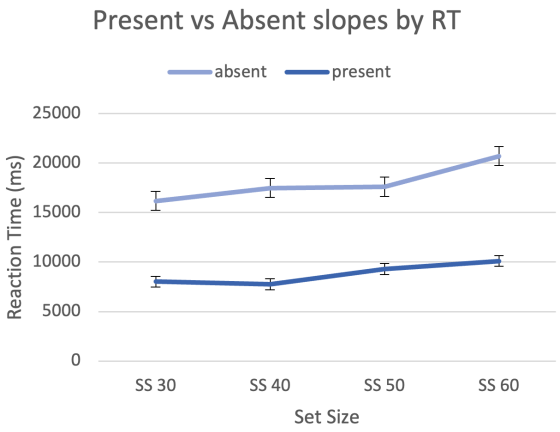



**Figure 5.7:** Response times by set size. There is a clear difference in response times betwen present and absent trials.


**Figure 5.8:** Present vs. absent search slope regression line. The low $r^2$ value indicates little correlation between present and absent slopes.

where subjects do not look anywhere with an object. Although this behaviour exists in typical 2D search tasks, where subjects' fixations do not directly land on a stimulus, they are usually at a location between multiple stimuli and the subject is hypothesized to be looking with a more abstract focus in that general area (Zelinsky, 2012). This was not the case for the fixations I found not pointing at the objects, as subjects were usually looking at the floor or the space between tables instead. This behaviour is likely caused by the fact that navigation in the environment is necessary in order to uncover better viewpoints for the objects. Consequently, it cannot be found in any 2D search task, and thus has no comparison.

Thirdly, looking at the eye movement metrics, differences can be found between 2D and 3D search as well. For 2D search experiments where eye movements are tracked, subjects are typically told to keep their head still so that the eye trackers are able to more accurately track data. With the eye tracking glasses, as well as the possibility of occlusion in my experiment, subjects have much higher freedom of head movement compared to those subjects. As a result, the saccade amplitudes in the active search subjects are significantly lower than in 2D search task subjects, as they can move their heads to change the view rather than making a saccade to a location that is higher degree eccentricity from the central field of view. To convert saccade amplitudes from my experiment into visual angle degrees, I used the averaged

distance from the gaze location of the fixation preceding the saccade and the fixation after to normalize. Average saccade amplitudes were only 0.57 degrees (SD=0.3), compared to a typical search task, which has amplitudes around 5-6 degrees in natural image search (Over et al., 2007). Indeed, Pelz et al. (2001) state that gaze shifts are almost always accompanied by small head movements in natural tasks.

Overall, direct comparisons cannot be made between 2D and 3D search. Although a few metrics may be calculated for both, such as search slope and eye movements, the interpretation of these values differ in a 2D search environment compared to real world 3D search. Slope averages calculated using only fixation durations may seem comparable to some 2D search tasks, but the spread and number of negative slopes say otherwise. Eye movement metrics showed a clear difference in saccade amplitudes between 2D and 3D search. Furthermore, behaviours like navigating through the search space and finding the necessary viewpoints is something that is not captured in 2D search, yet it is an integral component of search in the real world.

# Chapter 6

# Strategy analysis

This section describes the strategy analysis conducted on the eye and head movement data, in order to answer research question 2: Are there any eye and head movement patterns that can be uncovered corresponding to different search strategies?

I split each trial temporally into components first to break down the analysis. Then, once some patterns were identified, I looked for them throughout the entire trial.

It is also important to note that many of the location and looking metrics depend on the layout and trial version, since where the subject is standing and looking depends on where the objects and the tables and cages themselves are placed. Thus, the strategy analysis also breaks down and considers these metrics by layout and trial version.

## 6.1 Data extraction for strategy analysis

In addition to the eye and head movement metrics mentioned in the Results section above, the following location and looking metrics were used to further discover patterns and strategies that subjects are using:

**Look-at** Any fixations that are pointing at any of the tables and cages, where objects are placed

**Target look-at** Fixations where the subject is directly looking at the target object, when the target is present

**Subject head location** Subject's head location in 3D coordinates

**Subject head orientation** Subject's head orientation in roll, pitch, and yaw

**3D fixation point** The projected 3D fixation location given the local 2D coordinates from the Tobii Glasses as well as the subject's head position and orientation

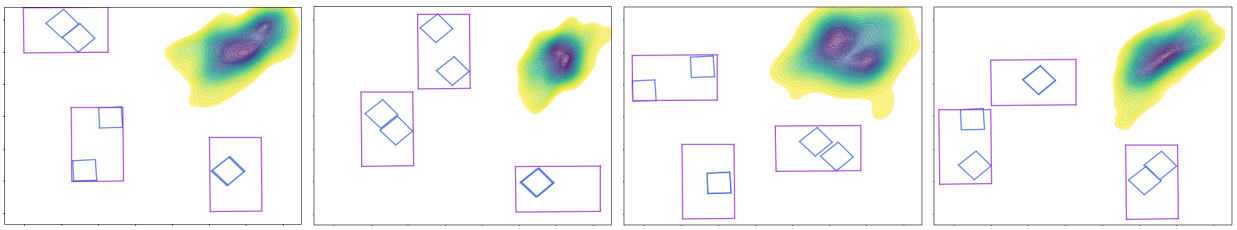
To define where subjects were looking, I divided the environment into cubes, similar to Shubina and Tsotsos (2010)'s formulation of the search region. Each cube was 20cm x 20cm x 20cm. There were a total of 20 (width) x 16 (depth) x 8 (height) cubes to cover the search space. Each table took up around 6 x 4 cubes, at a height of 4 cubes. Each cage took up 1.5 x 1.5 x 1.5 cubes, with all cages except "cage 3" having a height of 5 cubes. "cage 3" would always have a height of 6 cubes, since it is always stacked on top of another cage. Thus, any fixations that landed within these defined cubes were considered to belong to their respective tables and cages. Any further mentions of subjects "looking at" a table or a cage are of fixations landing within cubes that belong to said table or cage.

## 6.2 Initial strategy

The initial strategy describes what subjects do at the start of every trial. In order to separate the start from the remaining trial, I conducted analyses on fixations in the first 5 seconds (or first 25%, whichever is smaller) of each trial, to count as "initial" strategy.

Most subjects are standing close to their starting point (where they are presented with the target image) when their first fixation is recorded (See Figure 6.1 for a distribution of subject head positions). Out of 576 trials, there were 132 trials where the subject's first fixation was pointing at one of the specified locations (table or cage). For the remaining fixations, most were made as the subject was still turning around and the tables or cages were not fully in view yet. By the fifth fixation, two thirds of trials had subjects looking at a table or cage.

To further understand where subjects are beginning their search, we can investigate the first fixation in each trial where the subject is looking at one of the defined locations (table or cage), as well as how consistent this is over the course of the experiment. The most common location subjects looked at first was table 3 (the rightmost table in each layout



**(a)** Layout 1 **(b)** Layout 2 **(c)** Layout 3 **(d)** Layout 4

**Figure 6.1:** Layouts with heatmaps of subject head locations when they take their first fixation. For all layouts, this occurred at or close to where subjects stand when they are presented the target image. Yellow indicates lower frequency, and purple indicates the highest.

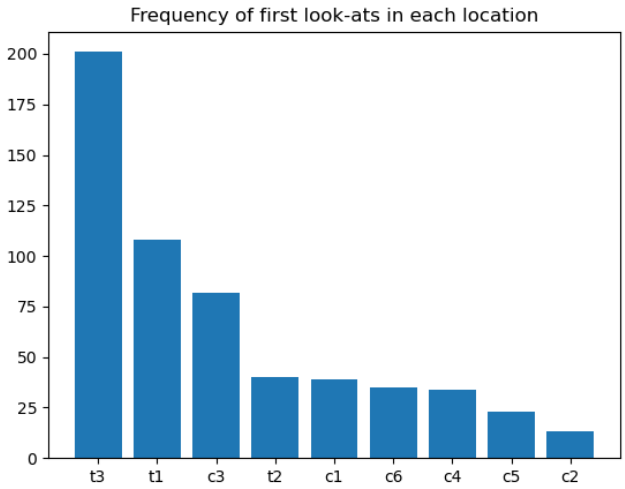



**Figure 6.2:** Frequency of first look-ats in each location. Table 3 (t3) had by far the highest frequency, and corresponds to the first table subjects encounter if they turn clockwise from their waiting position to face the setup.

from bird's eye view, see Figure 4.7), with 35% of all trials starting there. Figure 6.2 shows the frequencies of the other locations. The first location subjects investigate in detail is also most likely to be table 3, with 131 occurrences of first table or cage scans being there, compared to $< 60$ occurrences in all other locations. Looking at the direction subjects turn when the trial starts, we can also see that the majority of trials start with the subject turning clockwise, thus, the first location they encounter is table 3.

It is also interesting to see how consistent subjects are throughout the course of the

experiment. Looking first at some summary statistics, subjects typically have 3.8 (SD=1.3) different starting locations across the 12 trials, though the location with the highest frequency is typically around 5.2 (SD=1.9). Thus, it seems that subjects tend to try a couple different starting locations but stick to one more than the others.

Regarding the amount of exploration done in the initial part of the trial, we can measure the subjects' K coefficient. Average K coefficient was $-0.04$, with $K < 0$ indicating a preference for exploration over exploitation. Although this is a small value, it is consistent with reported values in Krejtz et al. (2016) (their most extreme averages were -0.29 and 0.14), and it is more negative than the remainder of the trial's K coefficient. This is consistent with the fact that the majority of first "scan" actions that a subject takes is the wide scan, with 135 occurrences, compared to 66 cage scans and 24 table scans.

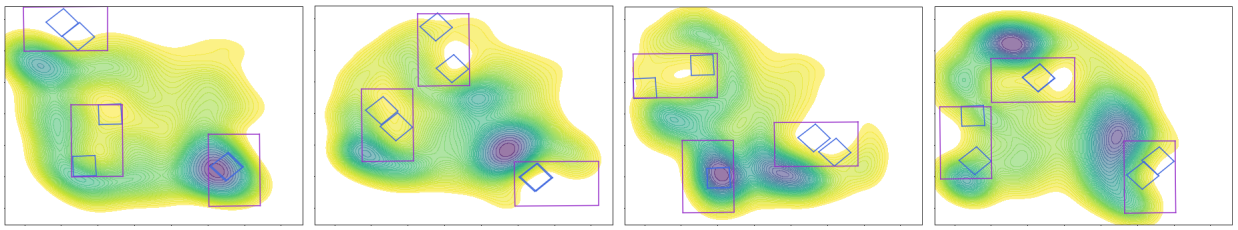
It is interesting to note that although subjects mainly start a trial with a wide scan, defined as scanning across multiple tables and cage locations within a few fixations, they are still likely to miss a target that is visible from the starting location. In many cases, subjects follow the same search-path they normally take, and treat this trial as if the target were not visible from the starting location at all. This may be because the wide scan at the start does not serve the purpose of directly searching for the target, and may be used more to orient the subject to the surroundings as they turn around.

Thus, although I had expected that having a visible target from starting location would change subject behaviour and strategy, that was only the case when subjects happened to look for it and find it when they first turn around. Figure 5.6 shows an example where the visible target caused the subject to deviate from their usual search-path on the right, and no deviation in the middle.

## 6.3 Middle strategy

In order to understand the strategy subjects use for the main body of the task (after initial strategy, until the end of the trial), I first tried to visualize the data to see if anything would emerge.

Indeed, there were some hotspots that I found by plotting subject locations onto each



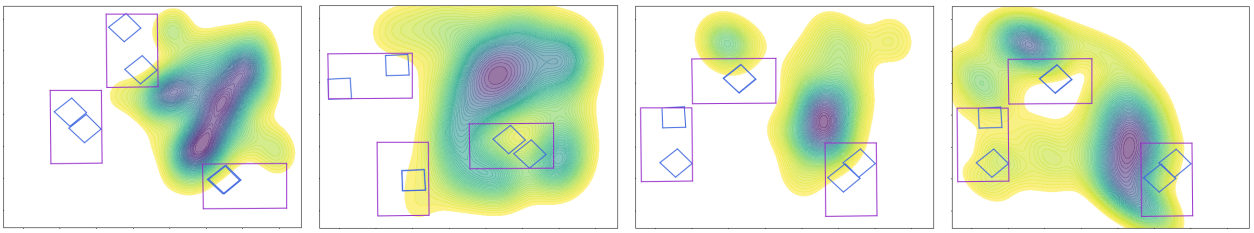
**(a)** Layout 1 **(b)** Layout 2 **(c)** Layout 3 **(d)** Layout 4

**Figure 6.3:** Experiment table and cage layouts with heatmaps of subject head locations overlaid. Visually, there are clear differences between the layouts, especially in which areas are more concentrated. It seems subjects have different "favourite" locations to stand, depending on the layout. Yellow indicates lower frequency, and purple indicates the highest. Hotspots are sometimes over the tables and cages, as subjects will lean over them during their search.

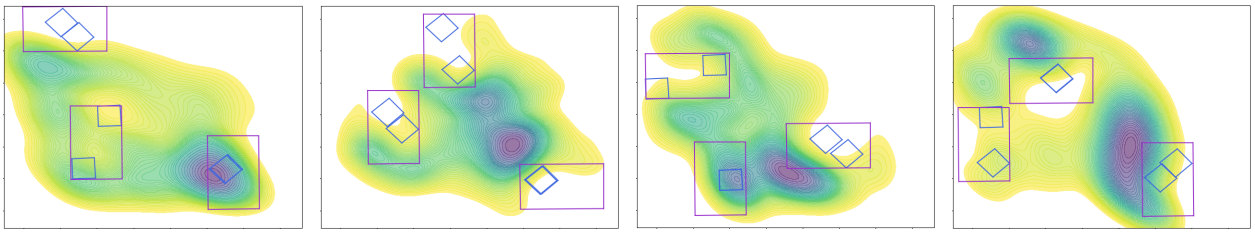
layout. To see if these hotspots were concentrated at a certain time point during the trial, I also plotted the same heatmaps over different slices of time: start-10% of the trial (overlapping with the initial strategy), 10-50%, 50-90%, and finally 90%-end of the trial (See Figure 6.3, 6.4).

To analyze only fixations that occurred in the hotspots, I used DBScan (Ester et al., 1996)to find clusters (epsilon=0.2, min samples = 5% of data). I defined each of the clusters from DBSCAN as a hotspot. For each hotspot, I investigated the distribution of fixations pointing at each location, to see if there were any patterns. Each of the 4 layouts had 3 to 4 hotspots. Generally, each hotspot had the most fixations pointing at one particular table and the cages on that table. Each hotspot location may be the optimal position where subjects stand to get the best views of that particular "island" of table and cages. In layouts 1, 2, and 4, there was no single hotspot where subjects had fixations to every location in the setup. Figure 6.5 shows each hotspot color-coded with their corresponding look-at location frequencies in the bar chart below.

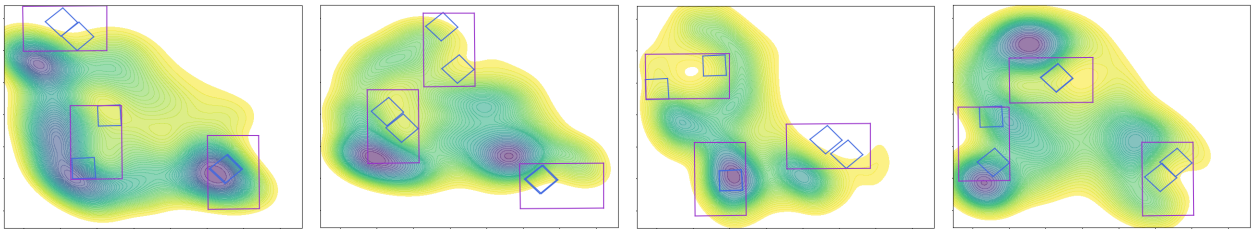
Interestingly, layout 3 (third column in Figure 6.5) had one particular hotspot where there were subject fixations to every location. There were over 100 fixations to every location from this hotspot. This hotspot corresponds to the blue-colored hotspot in the figure, and it was also the most prominent hotspot from 50% to end of trial. Thus, it seems there may have been some kind of "optimal" position that subjects found where they could get a good



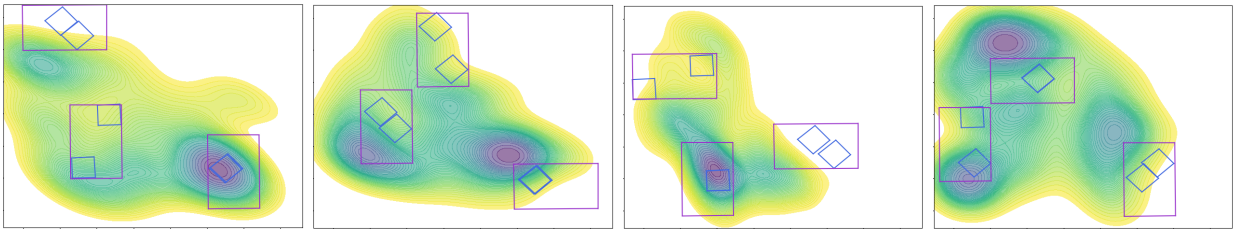
**(a)** Initial 10% of trial. Corresponds to where subjects' head locations likely are for the Initial Strategy section (section 6.2).



**(b)** 10-50% of trial



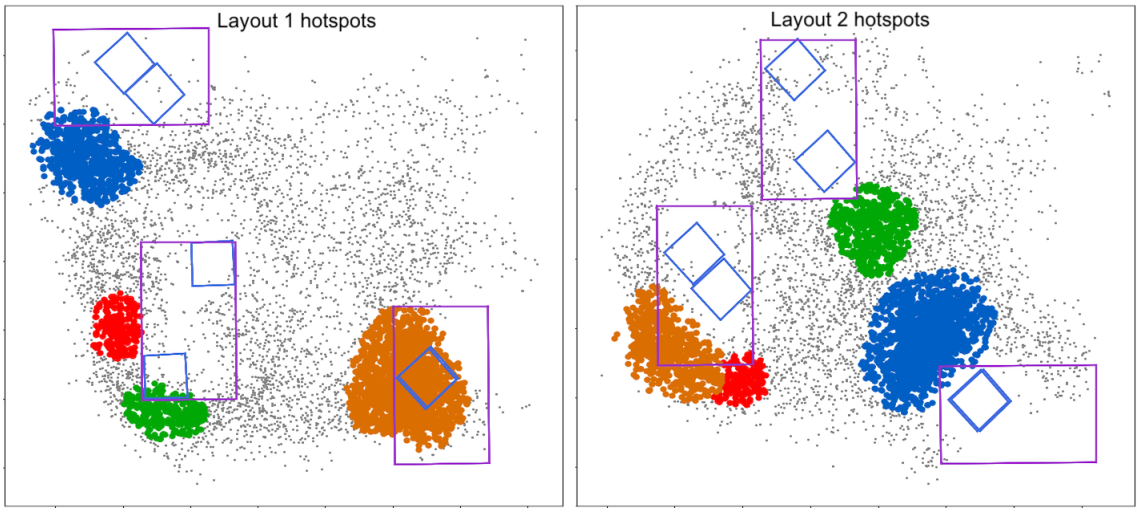
**(c)** 50-90% of trial



Layout 1 Layout 2 Layout 3 Layout 4

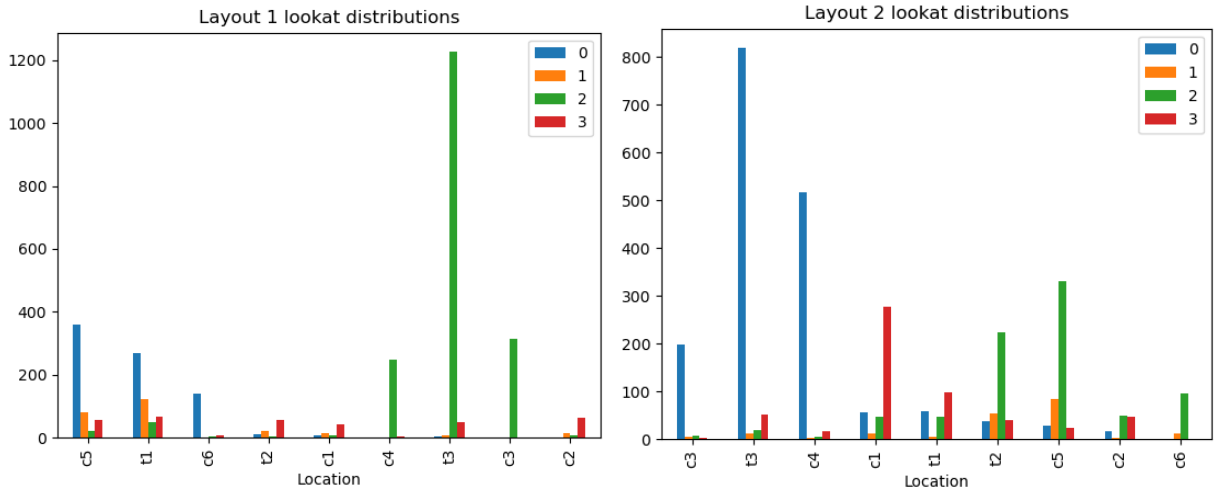
**(d)** Last 10% of trial

**Figure 6.4:** Heatmaps divided over the course of the trial. Different hotspots can be seen across each of the time intervals. Subjects seem to have different preferred head locations at different points of the trial. This could be seen as a coarse piecewise representation of the subjects' "typical" search-paths. Yellow indicates lower frequency, and purple indicates the highest.




**(a)** Layout 1 hotspots

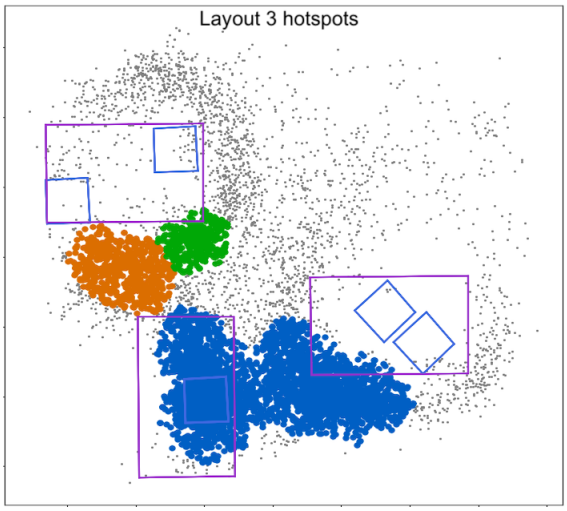
**(b)** Layout 2 hotspots



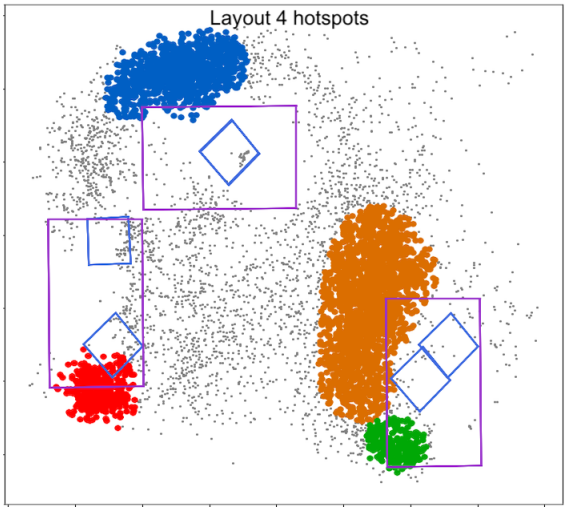

**(c)** Layout 1 corresponding lookat distribution

**(d)** Layout 2 corresponding lookat distribution

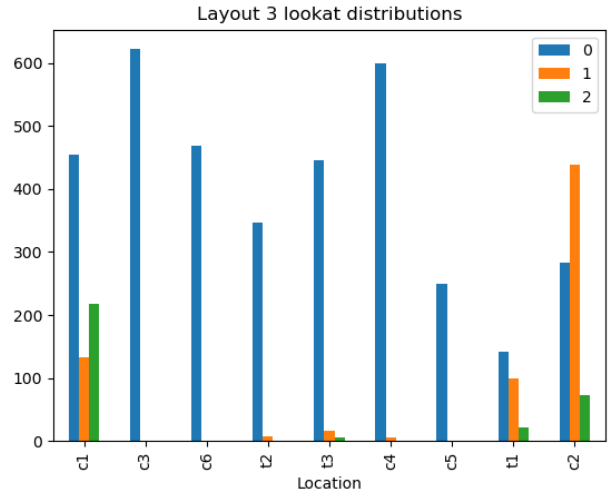
**Figure 6.5:** Hotspots (clusters of subject head locations) and their corresponding distribution of look-at locations for each layout. Greyed out dots are subject locations that did not belong in any hotspots. Large spikes in the bar plots suggest that a particular hotspot was used specifically to inspect objects placed in that location. This plot shows the same head location information as 6.3, except that the particular look-ats are clustered with corresponding look-at information.



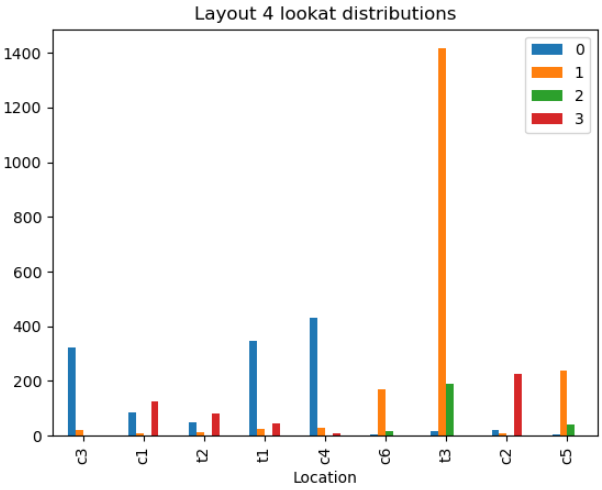

**(e)** Layout 3 hotspots




**(f)** Layout 4 hotspots




**(g)** Layout 3 corresponding lookat distribution




**(h)** Layout 4 corresponding lookat distribution

**Figure 6.5:** Hotspots (clusters of subject head locations) and their corresponding distribution of look-at locations for each layout. Greyed out dots are subject locations that did not belong in any hotspots. Large spikes in the bar plots suggest that a particular hotspot was used specifically to inspect objects placed in that location.

enough view of all locations. The existence of this hotspot, however, did not reduce the subjects' distance travelled. In fact, there was no significant difference in distance travelled between any of the layouts, so there seems to be no improvement in efficiency from this.

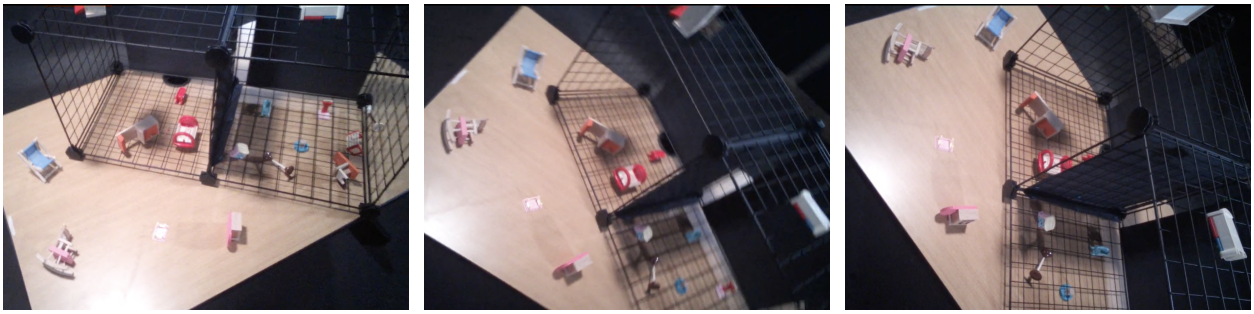
Each hotspot (aside from the one "optimal" hotspot) had a higher concentration of look-ats at particular areas of the setup. For example, in layout 2, the blue hotspot was where subjects looked at table 3 and cage 3 and cage 4 the most, the red hotspot for cage 1 and table 1, and the green hotspot for table 2, cage 5, and cage 6. Thus, the hotspots are able to capture the locations which subjects preferred to use to look at particular areas, typically the ones closest to that area. The location of the hotspot is also informative — it gives an indication of which angle allows subjects to see most of those particular locations.

Comparing target present to target absent trials, it seems like there is little difference. There are similar hotspots for both, so they seem to be the go-to locations for subjects to stand at regardless of whether the target is present or not.

In contrast, when the first 6 trials of each subject were compared to the last 6 trials, different hotspots were found. The later hotspots are the ones that are more prominent in Figure 6.5. This suggests that over time, subjects may have "learned" the best locations to stand in each of the layouts to search more efficiently.

Another component to consider in subjects' strategies is the effect of object placements. Object facing direction and orientation were manipulated throughout the trials, such that many trials had targets placed in positions and orientations different from what was presented at the start. Thus, I wanted to see whether the orientation differences would induce any particular head and body motions, such as head tilts or crouching. These two metrics were defined earlier in section 5.1.

Indeed, in each trial, 12.6 crouches were recorded on average, along with 28.5 rolls and 32.3 pitches. Target absent trials typically induced twice as many crouches, rolls, and pitches, compared to target present trials, likely due to the doubled amount of time subjects spend on absent trials as well. Figure 6.6 shows an example of a subject tilting their head to look at the orange stove, which was the target object.



**(a)** Head tilt example

**(b)** Target object

**Figure 6.6:** Example of a head tilt in frames, target object shown in second row. The subject tilted their head to match their view of the object to its canonical orientation, as presented at the start of the trial.

## 6.4 Terminating strategy

The terminating strategy refers to what subjects do near the end of the trial, whether it is signalling that they have found the target, or that the target is absent.

### 6.4.1 Target present trials

To investigate target present strategies, I use the target-facing fixations, defined as fixations whose central view (within 30 degrees eccentricity) intersects the target object's location, accounting for occlusion from the placement of black paper barriers on the cage setups. This allows me to extract several metrics, such as the **number of fixations** pointing at the target before the subject's response is declared, as well as the **distance of the subject from the target** when each fixation occurs.

The subject fixating once on the target does not immediately signal the end of the trial.

In fact, subjects tend to take around 8.5 target-facing fixations (SD=8) before declaring that the target is present. The average number of revisits to the target location is 1.8 (SD=1), suggesting that subjects do not always find the target the first time the target is within their central field of view.

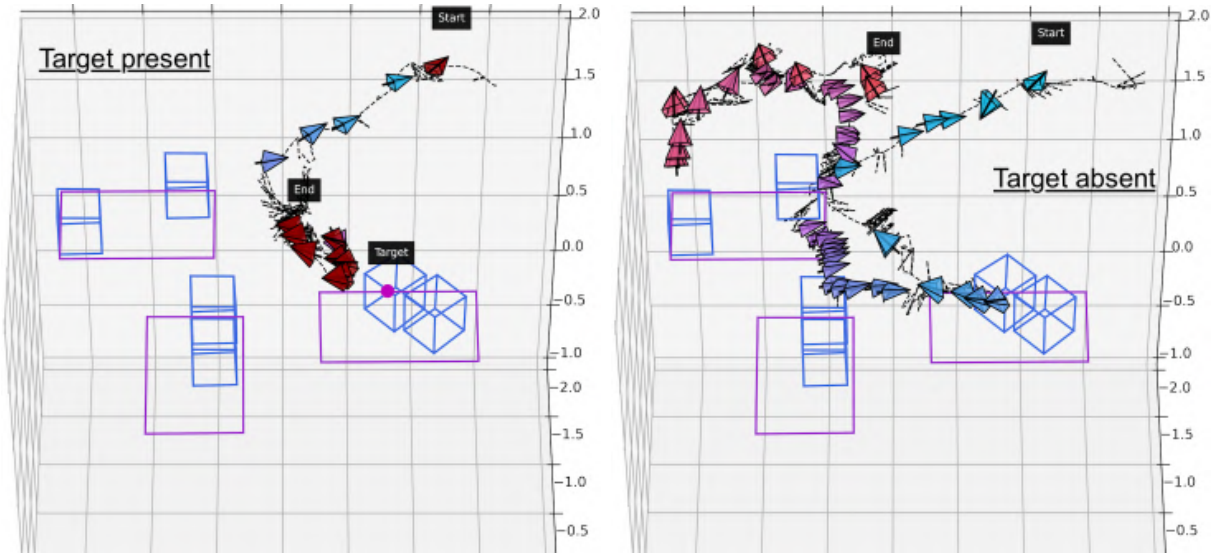
Subjects do not always find the target when they are near it. Sometimes, they may spot the object as a potential target from afar and need to approach it to further inspect whether it is the target. There is a significant negative correlation between distance to target and time ($r = -0.29$, $p < 0.001$), meaning the subject does get closer to the target over time.

An interesting thing to note is the difference between trials where the subject was correct and found the target, vs. trials where the subject missed the target. There were 29 trials in total that were miss trials. Table 6.1 shows the number of target-facing fixations on average for both groups, as well as the number of revisits to the target location. The miss trials clearly had a higher average on both, so subjects truly saw the target but believed that it was not present. This may be due to a largely different orientation of the object, such that the subject decided this object was in fact not the target, or perhaps lighting issues of the target image presented (as discussed earlier). Another possibility is the occlusion of the target object by other objects, such that a very specific angle of viewpoints would be needed in order to find the target, which these subjects did not find.

| Group | Target-facing fixations | Target location revisits |
|---|---|---|
| Correct | 8.1 | 1.8 |
| Incorrect (miss) | 11.7 | 2.8 |

**Table 6.1:** Number of target-facing fixations and number of target location revisits for target present trials, separated by correct and incorrect (miss) trials

Once subjects notice the target object, they spend on average 3.9 fixations looking at the target consecutively before responding.




**(a)** Target present trial **(b)** Target absent trial

**Figure 6.7:** A typical target present. target absent trial search-path. Most target absent trials have a longer search-path with more overlapping frustums, indicating more revisits.

### 6.4.2 Target absent trials

Target absent trials have on average over double the number of fixations than target present trials (103 vs. 42 average). In fact, the minimum number of fixations taken for a target absent trial is 14, showing that the subject must take multiple views before deciding the target is absent. Figure 6.7 shows an example of a target present trial's search-path compared to a target absent trial from the same subject.

Subjects visit 7.7 of the 9 locations on average (SD = 1.5). In fact, 40% of all absent trials had subjects visiting all 9 locations. The average time taken is also around 48 seconds, with the shortest trial being 7.7 seconds. This indicates that subjects truly need to walk around and observe every location before deciding that a target is absent, and cannot decide from a few simple glances.

Other than visiting all the locations compared to stopping once a target is found, there were also more revisiting behaviours found in target absent trials.

There are 18.5 location revisits on average. In contrast, target present trials see only an average of 6 revisits, showing that once they have spotted the target, subjects are much less likely to continue looking at other locations. Doing a more detailed analysis, over all

layouts, the most revisited locations are table 1 and table 3, with an average of 3.1 revisits for both. Subjects revisit 6 distinct locations on average (SD = 2.3), showing that they almost revisit the entire setup twice before declaring the target is absent. The number of revisits to each location, as well as the number of distinct locations revisited, depends on the layout. Table 6.2 shows the most revisited locations for each layout. Layout 1 seems to have the most concentrated revisits to only 3 locations, whilst layout 2's revisits seem to be quite well distributed. Layout 4 seems to just have fewer revisits in general.

| Layout | Top 1 revisited (num) | Top 2 revisited (num) | Top 3 revisited (num) |
|---|---|---|---|
| 1 | Table 1 (4.5) | Cage 5 (3.9) | Table 3 (3.2) |
| 2 | Table 3 (3.8) | Table 2 (3.3) | Cage 1 (3.1) |
| 3 | Cage 1 (4.5) | Cage 2 (2.9) | Table 1 (2.8) |
| 4 | Table 3 (3.2) | Table 1 (2.5) | Cage 1(2.1) |

**Table 6.2:** Most revisited locations in each layout

The k coefficient is also on average negative from 50% to the end of the trial, signalling a slightly higher tendency for exploration. This coefficient is also significantly different from the average k coefficient of the second half of present trials (present k = 0.011, absent k = -0.0038, p=0.026), suggesting a difference in gaze strategies in the later half, increasing in amounts of exploration. This may be because subjects start to look in more sporadic locations, instead of going through each location's objects in order after looking through them once.

## 6.5 Overall strategy

First, looking at the time each subject spends standing in each location, we can see that they typically take 4 fixations before moving, and look at ≈1.6 unique discrete locations.

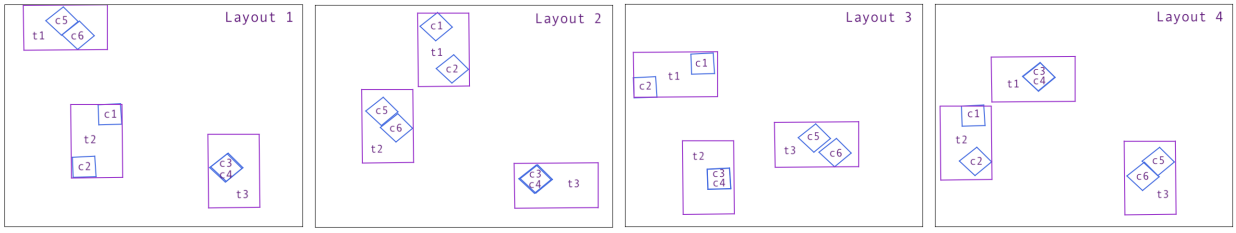
Looking at the identified cage scans, scan lengths were typically 4.6 fixations long, and subjects moved 28.66cm on average in each scan.

The overall search-path within each subject is quite predictable. Subjects tend to pick a path to traverse and stick with it for the entire experiment. There are also a few most

"popular" paths that subjects take for each layout.

In order to investigate subjects' overall search-paths, I used the Apriori algorithm (Agrawal et al., 1994) to find the most common sets of locations that the subjects looked at in each trial. I used a minimum support of 0.1, meaning itemsets had to occur with probability over a threshold 10% to be included. I divided the trials by layout, as search-paths grouped across layouts may not have as much meaning. After doing that, I found that layouts 2-4 had more similar search-paths compared to layout 1. There were more itemsets of length 5 and more, and there were itemsets up to length 7, but in layout 1 trials there were itemsets only up to length 5.

Generally, the most common itemsets are quite similar to each other. For example, the two most common itemsets of length 6 in layout 2 were: {c1, t1, c4, t2, t3, c5} and {c1, c4, c3, t2, t3, c5}. As shown in Figure 6.8, c1 is a cage on table 1, and c3 and c4 are stacked vertically. By grouping each of the locations to their respective tables, these two itemsets become {t1:(c1, t1), t2: (c4, t2), t3: (t3, c5) } and { t1: (c1), t2: (c4, c3, t2), t3: (t3, c5) }. Therefore, these two search-path itemsets, at the table-level precision, are quite similar. There are itemsets of length 5 that are found in up to 51 trials within layout 2.




**(a)** Layout 1 **(b)** Layout 2 **(c)** Layout 3 **(d)** Layout 4

**Figure 6.8:** Layouts with location abbreviations indicated. c3 was always above c4 in the vertical stack of cages.

After investigating the locations of common itemsets on the layouts, we can also see certain common path directions that subjects take in each layout. In layout 4, it is clear that most subjects start at table 3, then go to table 2, and finally table 1. Because of the layout, this means that subjects are moving typically in a clockwise direction in one sweep to investigate the locations. Looking at subject head locations, we can see that subjects do end up behind and in the corner between table 1 and 2 at the back, although that doesn't

seem to be part of the "main" search-path, and is a path that subjects may only take if they need to take a closer look at parts of the cages and tables that can only be seen from those viewpoints.

Layout 3 had the highest similarity in paths between subjects, with the highest number of common itemsets length 5 and above. Subjects typically start their search at table 3, then table 2, and finally table 1. Again, it seems subjects are searching in a clockwise order, congruent with the direction they turn around to start each trial.

Revisiting the 3D search-path plots, visually inspecting them also confirms the existence of similarities within and between subjects. Most notably, many subjects had only 1 or 2 main pathways they traversed for the entire experiment, and those pathways were similar across subjects. The top row of Figure 6.9 shows a sequence of trials from the same subject that have similar search-paths. Subject 24 in Figure 6.9 also showed a similar search-path to those found in subject 23, with the signature half circle around the left side of the tables and retracing back to the other side. Finally, subject 25 in Figure 6.9 shows a contrasting search-path that was found for the same layout.

## 6.6 Summary

The research question this section addresses is question 2, are there any eye and head movement patterns that can be uncovered corresponding to different search strategies? Indeed, there seem to be many such strategies that were discovered in the data.

To summarize all the strategies from the several parts of the trial, we can say that subjects typically start by turning clockwise and looking at table 3. They start the trial by exploring, meaning that they take longer saccades and shorter fixations, although they typically explore one location at a time. The first scan subjects take is usually a wide scan.

Then, they typically take 4 fixations and look at 1.6 unique locations before moving at the start of each trial, and visit on average 5.4 locations during present trials, and 7.8 during absent trials. There are 3 to 4 hotspots that they tend to take more time at, and they have fixations to most locations from those hotspots. Thus, it seems possible that subjects find certain spots in an environment that may reduce travel time and distance, and continuously

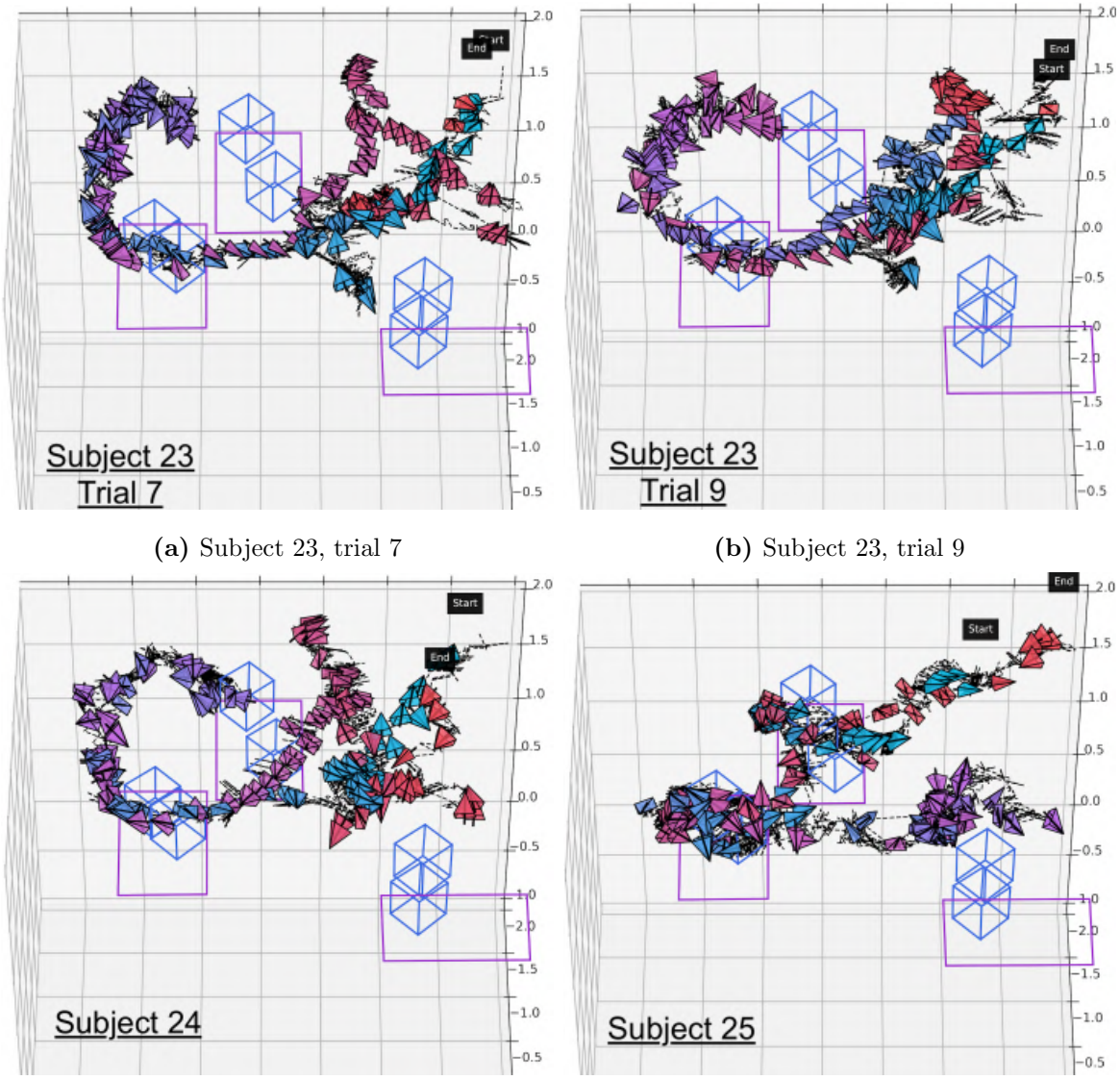
return to those spots to conduct search.

The search-paths taken mostly follow a similar sequence across subjects, and depend on the layout of the setup. For some trials where the target is visible from the starting location, if the subject spots the potential target at the start, their path will deviate from the usual sequence and head straight towards the target. However, most subjects do not notice the visible target, and will instead search on their usual path until they find it. It is interesting that different subjects find common search-paths in each layout, suggesting that the paths may be more determined by the environment than individual differences.

Once a subject thinks they have spotted the target, they approach to inspect and verify. If they decide that it is not the target, they leave and continue the search, following the common path. If they are confident it is the target, they signal the end of the trial. This occurs after 8 target-facing fixations on average. After around 18.5 location revisits without finding the target, the subject declares the target to be absent.

From these strategy analyses, we can glean some information about visual search behaviour in general. When beginning a search, we normally take a few fixations to orient ourselves in the search environment before moving. Our search-paths also seem to be quite dependent on the structure of the environment. Once we locate a potential target, we typically take multiple target-facing fixations, some from different angles, before confirming that the object is the target. Finally, absent target searches elicit significantly more revisits, fixations, and movement than present search.

In the next section, I aim to see if any existing computational models are able to encompass the behaviour described by these strategies. If not the overarching behaviour, are the models able to describe certain sections or parts of these strategies?




**(a)** Subject 23, trial 7

**(b)** Subject 23, trial 9

**(c)** Subject 24. Same pattern as above

**(d)** Subject 25. Different pattern from above

**Figure 6.9:** 3D search-paths of 3 different subjects on the same layout. The top row shows a subject with similar search-paths across trials, bottom left image shows a similar search-path between this subject and the subject above, and the bottom right shows a contrasting search-path. This pattern occurs across many subjects and layouts. It can be seen that similar search-paths can be found between and within subjects for the same layout, but there are multiple of these "typical" search-paths to be found.

# Chapter 7

# Comparison to computational models

In this section, I aim to address my next two research questions:

RQ1. How do computational models perform in comparison to humans?

RQ2. Are strategies used in computational models for active search comparable to human strategies?

I compared the human strategy data obtained to many computational models, from different fields of research. Models used include MVT (Charnov, 1976) from optimal foraging literature, Shubina and Tsotsos (2010) and Rasouli et al. (2020)'s models from robotics, Ideal Bayesian Observer (Najemnik and Geisler, 2005) in psychophysics literature, using the natural images extension made by Bujia et al. (2022), and AIM by Bruce and Tsotsos (2009) in saliency.

## 7.1 High-level differences

Robotics models have the most similar world environment model as the active search task performed by subjects in my experiment. These models will thus be easier to compare to the subjects' overall movement and strategies, although any direct comparison of performance is out of the scope of this thesis, as that would involve either a detailed simulation of my experiment, or implementing a robot to perform the search in my physical environment.

MVT is a broader model from foraging that assumes each patch is visited in turn, and that one cannot stand in the same location to investigate multiple patches. However, it is an interesting model to compare, as each of my possible target locations could be counted as a patch for search.

The Ideal Bayesian Observer is a model for active search on a 2D screen. Thus, it assumes no head movements, only eye movements, and that the search space is restricted to the screen only. Seeing as much of the strategies discovered in my active search experiment involve head and body movements along with eye movements, it is more difficult to find examples to compare this model with the results from my experiment. However, I am able to use Bujia et al. (2022)'s natural images version of the model on frames of the trials where the subject does not move their head or body, but make eye movements. This allows comparison to a very small part of the dataset.

I also wanted to include AIM as a saliency model, to see if subjects' fixations would tend to land on more salient regions. Although the scanpaths would not be comparable, I can see the proportion of fixations that land on regions that are identified by the algorithm as salient, to get a better understanding of how much humans might be guided by saliency in search.

## 7.2 Robotics models

In order to compare robotics models to my experiment data, I compared the discovered strategies and behaviours to cost functions that the models used. Direct comparisons of model performance to human performance is, as mentioned, out of the scope of this thesis, as that would involve either simulating my experiment environment in detail, or implementing a robot to perform search in my setup.

### 7.2.1 Shubina and Tsotsos (2010) model

4 cost functions were tested in Shubina and Tsotsos (2010) for the robot to use during search. These include:

(1) Choose action with the largest detection probability

(2) Explore current position, then choose next position by maximizing detection probability

(3) Explore current position, then choose next position by maximizing detection probability while minimizing distance

(4) Explore current position, then combine utility functions from (2) and (3) for choosing next position

To understand which cost function human strategies are most similar to, I investigated the distribution of fixations and look-ats where the subject is standing in one location without moving much, the distance between each of these locations, as well as the length of the gap between fixations that are pointing at one of the specified locations.

One common pattern found in subjects is the tendency to take a few fixations every step along a search-path. This is somewhat similar to cost function (3) in action, as this leads to behaviour like scanning viable angles in one location before moving to the next closest location. In Shubina and Tsotsos (2010), the robot seemed to take 3 to 4 snapshots before moving to the next location. In my experiment data, subjects take on average 8.8 fixations in each standing location, looking at on average 2.7 unique discrete locations, before moving.

One interesting human behaviour that is not emulated in Shubina and Tsotsos (2010)'s model is the tendency to track one location as the subject is moving. For example, while the subject is moving, they may be getting continuous viewpoints on one particular cage or table by walking around it. There were 503 examples of 3 fixations or more where subjects would move their head or body, but keep the same 3D gaze point. There were many more (2099 to be exact) examples of this with a length of 2 fixations. Figure 7.1 shows one such example, where the subject fixated at the same gaze location but moved around it in order to disocclude some objects that were hidden behind the black side of the cage. This behaviour does not correspond to any of the cost functions in Shubina and Tsotsos (2010) as it was not a capability of the system, and it may not even be a desired behaviour in a robot system.

Thus, we can see that some strategies taken by the subjects in my experiment can be reflected by some of these cost functions, in particular cost (3). There is little to no evidence for behaviour supporting cost function (1) or (2) as subjects tend to continually take views

**Figure 7.1:** Example of subject keeping the same gaze location but moving around in order to disocclude objects. In this case it was objects hidden by the black coverings in the lower right corner of the cage.

even as they are moving in the environment, unlike a robot that needs to stop in order to take a snapshot. The foveated nature of human vision means that subjects must take many more snapshots, each with a different fixation location, in order to construct the environment that a robot can do in one snapshot, and this inherently changes the behaviour between the two systems. Furthermore, Shubina and Tsotsos (2010)'s robot does not have good image stabilization, making taking images and processing them while travelling much more difficult. The robot is also only given the target as a 2D image, and all images gathered of the scene are 2D. As such, 3D information is not available to Shubina and Tsotsos (2010)'s robot in the same way that it is for a human observer. Due to these differences in the image acquisition system between humans and robots, it is only reasonable that the strategies used are different as well.

One common strategy not accounted for in the cost functions is the tendency for subjects to gather viewpoints of one location consecutively, before moving on to the next location. This may be a strategy developed as it is easier to keep track of the objects in one location at a time, as well as for remembering which locations and objects have been visited.

Another difference between robots and humans is memory — robot models do not have to account for a decaying memory, yet some behaviours that humans do could be used to optimize and minimize working memory usage. Ballard et al. (1995), for example, found that subjects' fixation patterns were consistent with them acquiring information in the scene that was needed only just prior to being used.

Thus, although there are some similarities in broad strategies from humans comparable to the cost functions, the more fine-grained details, such as the order and search-path of the subjects, or the number of fixations to individual objects, are not captured in this robotics model.

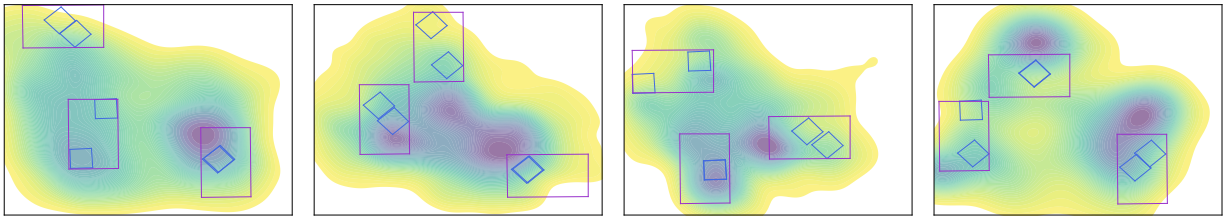
### 7.2.2 Rasouli et al. (2020) model

In Rasouli et al. (2020)'s model, they defined bottom up attention as the AIM algorithm trained on images of office environments to recognize salient regions. Top-down attention was defined using the color feature, implemented by backprojection. They used 4 different models with various combinations of bottom-up and top-down attention to perform search. These include:

(1) Bottom up only
(2) Top down only
(3) Bottom up + top down
(4) Bottom up + top down on the image filtered by bottom up

Due to the setup of the experiment, the only relevant locations for search were the tables and the cages. Other locations in the space were colored in black exclusively, thus no other locations would have similar color features to the target. This meant that the top-down color module contains some of the bottom up attention to relevant regions, since only relevant regions could have similar colors as the target. Therefore, there is no data in my experiment that could correspond with model (2) using purely top down attention. Here, I explore the similarities between the other 3 models and my experiment's data.

Color was used as the primary guiding feature for top-down attention. According to Wolfe and Horowitz (2017), color is also a guiding attribute in search for humans. Thus, to investigate whether human subjects use colour to guide their search in a 3D environment as well, I looked at the correlation between the number of objects sharing the same colour with the target in each location, and the number of look-ats at that location. Using Spearman's rank correlation coefficient, across all trial versions, this correlation was insignificant (r=0.03). However, when considering the separate trial versions, there was a significant



**(a)** Layout 1 **(b)** Layout 2 **(c)** Layout 3 **(d)** Layout 4

**Figure 7.2:** Distribution of subject fixation locations for each layout. Each layout has a unique distribution.

correlation (r=0.21, p=0.025) for trial version 3.

I also computed relationships between other measured features of my objects (shape, size, texture) and the number of look-ats, and none had significant correlations. This shows that out of the mentioned object features, color seems to guide subjects' fixation selection slightly more. Congruent with Rasouli et al. (2020), the effects of the top-down attention module depended on characteristics of the target. Targets in trial version 3 seemed to elicit more look-ats to locations with objects of similar color, although the same could not be said for targets in the other trial versions.

To compare the effects of bottom up saliency in guiding attention, I looked at the distribution of fixations in my subjects' data. Out of the average 74.7 fixations, 49 were pointing at a table or cage, and the remaining were not. According to the AIM algorithm used, salient regions are regions that are unexpected given the local surroundings. These regions are equal to regions marked as discrete locations, which are the only areas to contain objects in my setup. Accordingly, it seems that more importance is placed in these regions, as a higher percentage of fixations land here than in the non-salient areas of the setup, suggesting that bottom up saliency as defined is used to guide fixation selection in human subjects. This can be more clearly shown in Figure 7.2, where it is clear that more fixations are made to locations on the tables and cages than elsewhere.

## 7.3 Foraging

This section discusses one optimal foraging theory, the Marginal Value Theorem, for patch-leaving strategies.

Although my experiment contains only one target among the many distractors in one setup, I can still attempt to formulate it as a foraging problem. Take each table and cage in my experiment as a patch. There are a total of 9 patches. The resources in each location are then the number of objects there. Scanning one object equates to "picking" it in foraging.

### 7.3.1 Patch-leaving strategy

As stated earlier, Marginal Value Theorem (MVT) by Charnov (1976) is a popular model in optimal foraging for patch-leaving strategy. The model provides a prediction for the optimal time a forager should leave the current patch to explore the next, depending on the energy gain from foraging the current patch, the cost of said foraging, the energy cost to travel between patches, and the time needed to travel to the next patch. It states that the optimal time to leave the patch is when the foraging rate at the current patch drops to the average rate of all patches. According to Wolfe (2013), human subjects tend to match predictions given by MVT for the basic foraging task, with uniform distribution of targets and depletable resources. However, in most general cases, subjects tend to stay in a patch longer than predicted by MVT (Bella-Fernández et al., 2021).

I simulate the foraging environment in my experiment by using each table and cage as a patch, and the number of objects in each patch as the patch's density of targets. I compute the rate of energy gain from each patch as the proportion of objects fixated on in a table or cage (we refer to this as patch from now on). This rate is the **instantaneous rate of return**. The **overall rate of return** is given by the number of objects not observed minus the current patch's objects, divided by the total number of objects. The MVT uses instantaneous and overall rates of return to predict an optimal leaving strategy.

Let $N$ be the total number of objects in a trial.

Let $n_{is}$ be the total number of objects at patch $i$ in visit number $s$ (where $i$ is the $s^{th}$ patch being visited)

Let $o_{it}$ be the number of objects observed at patch $i$ at time $t$.

The instantaneous rate of return at time $t$ in visit number $s$ is then:

$$\frac{n_{is} - o_{it}}{n_{is}} \tag{7.1}$$

The overall rate of return at patch $i$ in visit number $s$ is:

$$\Sigma_{j \neq i} \frac{n_{js}}{N} \tag{7.2}$$

An optimal observer according to MVT would leave the current area and forage in the next once the instantaneous rate drops below the overall rate. In this context, the optimal observer should leave the current patch once the probability of the target being outside the current patch is higher than the probability of the target being in the patch. Indeed, this leaving criteria is the same as the criteria for "when to move next" in Shubina and Tsotsos (2010)'s model. The difference is in how the probability of the target's true location is being computed.

After manually annotating 12 subjects' object observations with the tables or cages visited, I compared their actual patch-leaving behaviour compared to what is predicted by the above formulation and MVT. Even with only 12 subjects, the difference between MVT-predicted performance and subject performance is already highly significant. Overall, subjects would stay in patches significantly longer than predicted by MVT (2.9 fixations longer on average). Performing a 2-way ANOVA revealed that there was a significant main effect of trial version and layout, as well as a significant interaction between them (see Table 7.1). Performing a pairwise Tukey's HSD showed significant differences between most pairs of trial versions, but only layout 3 and layout 2 being significantly different out of all layout pairs.

| Metrics/variables | Trial version | Layout | Trial version x Layout |
|---|---|---|---|
| Difference in rate (optimal - observed) | *0.000** | *0.001** | *0.000** |

**Table 7.1:** 2-way ANOVA (Trial version x Layout) on difference in patch leaving rate (optimal - observed). Generally, subjects had a larger leaving rate, corresponding to spending more fixations than optimal before leaving each location. Cells italicized with an asterisk (*) are significant.

Comparing average difference between optimal and observed rate for each trial version (see Table 7.2), we can see that in trial version 1, subjects tended to have close to optimal

leaving rates. However, in all other trial versions, the difference between optimal and observed is significantly below 0, meaning that observed leaving rates were longer than optimal. This behaviour where subjects tend to stay in patches longer than predicted by MVT has been found by others in both humans and animals, and they have suggested it may be due to factors such as differences in qualities between patches (Wolfe, 2013). This is a possible explanation in my data too, since the number of features each object in a location shares with the target are not uniformly distributed between tables and cages. Thus, some tables or cages may have more target-like objects than others.

| Metric | Trial ver. 1 | Trial ver. 2 | Trial ver. 3 | Trial ver. 4 |
|---|---|---|---|---|
| Mean | -0.086 | -1.96 | -5.25 | -4.72 |
| SD | 3.68 | 5.31 | 7.48 | 5.57 |
| Significance (from 0) | 0.75 | *0.000** | *0.000** | *0.000** |

**Table 7.2:** Difference between optimal and observed leaving rate (optimal - observed), mean and standard deviation for each trial version. Trial version 1 was the only version where leaving rates were similar between optimal and observed. The remaining trial versions showed a longer leaving rate for observed values, meaning subjects spent more fixations than optimal before leaving each location. Significance was calculated as a z test comparing to a null hypothesis of mean 0. Cells italicized with an asterisk (*) are significant.

Aside from the exact time to leave a patch, the MVT makes a few other predictions about optimal foraging behaviour. Firstly, it predicts that the growth of information (or currency) found diminishes over time, as resources become depleted. Secondly, it predicts that there exists a correlation between the inter-patch travel time and time spent in a patch, as well as a correlation between patch density and time spent in a patch.

By correlating travel time between different locations and time spent in those locations, I found a significant overall correlation of $r = 0.8$ (see Figure 7.3). Splitting between layouts, there was a significant correlation in each layout with slightly smaller magnitude (layout 1: $r = 0.48$, layout 2: $r = 0.48$, layout 3: $r = 0.64$, layout 4: $r = 0.62$).

Similarly, there is a significant correlation of distance between locations and the time spent in the current location ($r = 0.57$, Figure 7.4). Since each layout has different distances

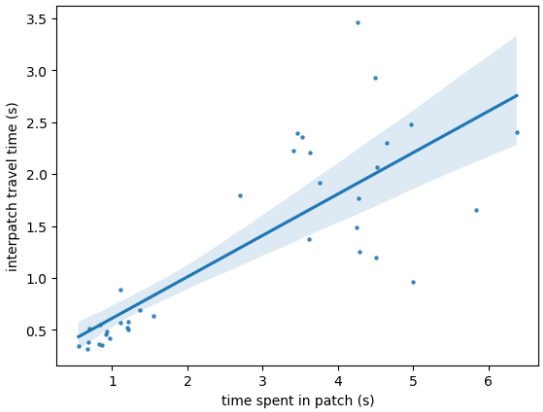



**Figure 7.3:** Correlation between time spent in one location and time spent travelling between locations ($r = 0.8$)

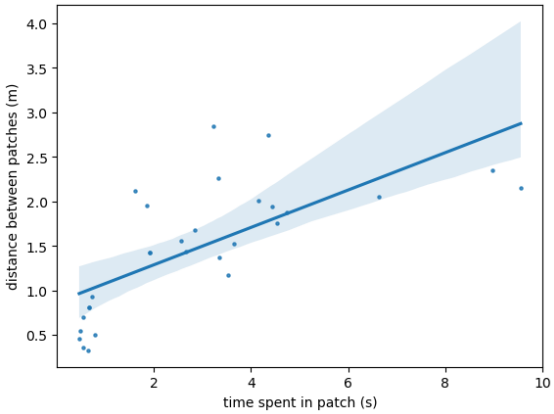



**Figure 7.4:** Correlation between time spent in one location and distance to next location in layout 2 ($r = 0.57$)

and walking paths for the corresponding locations, the within-layout correlation was mostly higher than the overall (layout 1: $r = 0.5$, layout 2: $r = 0.7$, layout 3: $r = 0.64$, layout 4: $r = 0.67$).

Thus, my experiment data is consistent with the MVT prediction that the further the next patch is (whether measured by distance or time taken to travel), the longer the subject spends in the current patch.

In contrast, I found practically no correlation between each location's object density to the time spent there. Overall, the correlation coefficient was $r = 0.06$. This may be because there was little variation in the density of objects between locations, as I tried to equalize them between locations. The minimum number of objects in one location was 2, and the maximum was 7. Therefore, I may have simply too little variation to find any relationship between density and time spent, if it did exist.

To conclude, my experiment data is generally in line with predictions from the MVT, in that these real world observers behave similar to existing literature on human and animal foraging. Most of the same correlations are found, and the tendency to leave a patch slightly after the optimal rate is also found. Although MVT is able to model when subjects leave each location, it is unable to model behaviour within each location, such as which viewpoints subjects are taking. Thus, it is predicting the high-level behaviour only.

## 7.4 Eye movement models

Models for eye movements on an image are difficult to compare with results from my experiment in terms of strategies, since there are a limited number of strategies that are captured only through eye movements. Since most people would rather move their head to change gaze than moving their eyes only, there is limited data to work with. Of the 576 trials recorded, only 70 trials had examples of this (thresholding consecutive fixational head movements to under 20cm and rotation less than 10 degrees for roll, pitch, and yaw), and run lengths were on average 2.7 (SD=1), meaning most of these examples were only 2 fixations long. There were 90 examples in total, but only 37 had a sequence of length 3 or more. After extracting the frames of the corresponding fixations, some were blurry, fixating on the floor, or had the subject's limbs in frame. These were further excluded from being used.

Thus, I took 30 usable examples of length 3 and greater and ran them through the Ideal Bayesian Searcher and AIM, to test if the predicted fixations from IBS corresponded with the actual next fixations that subjects performed, and whether subjects fixated on regions AIM identified as salient.

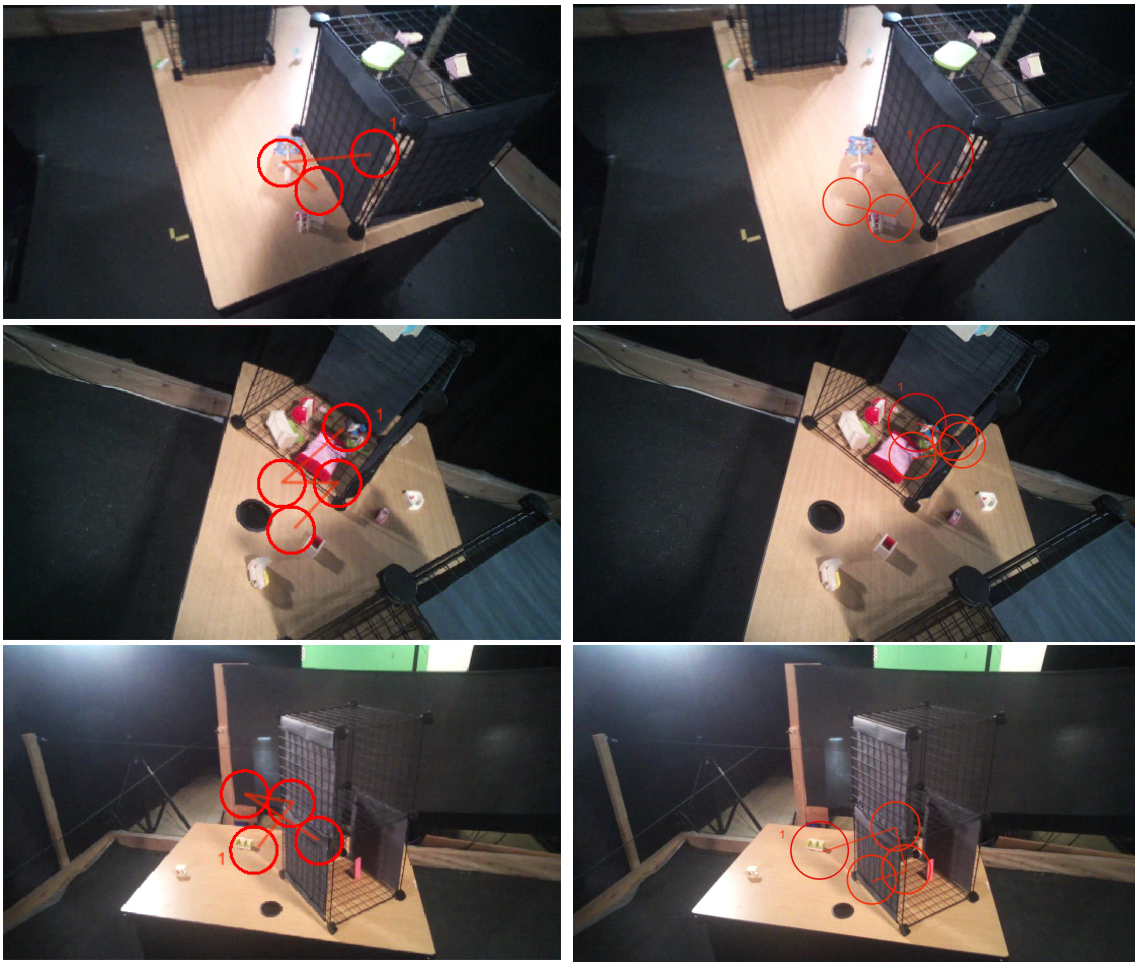
### 7.4.1 Ideal Bayesian Observer

I used the Bujia et al. (2022)'s implementation of Najemnik and Geisler (2005)'s Ideal Bayesian Observer on natural images to predict the ideal fixation sequence given image frames. I fed in the examples where subjects had consecutive fixations without much head movements. Of the examples extracted, only 3 examples had the target in view, the rest were either target absent trials, or the target was present but was not in the current frame.

As the Ideal Observer model assumes that a target is present, I placed a dummy target location in the corner of the image in order to run the model on trials where the target was not visible in the frame. I used the subject's first fixation in each sequence as the initial fixation location. The number of saccades the model made was capped using the number of saccades the subject had for each example. Results from the generated scanpaths were then compared to my human data.

I tested 30 examples that had 3 to 5 fixations in sequence (see Figure 7.5 for a few

examples). Running Bujia et al. (2022)'s algorithm, I used the "geisler" dynamic model for their implementation of Najemnik and Geisler (2005)'s model, with a center bias, and a max number of saccades equal to the length of the fixation sequence in the corresponding human data. The input data image was the frame from the first fixation of the sequence in the human data. I also passed in the same target image as the one provided to subjects at the start of the trial.



**(a)** Ideal Searcher fixations **(b)** Subject fixations

**Figure 7.5:** Fixation sequences generated by Ideal Searcher Model (left) vs. real subject sequences (right). Although the Ideal Searcher provides generally valid fixation locations, the order that it predicts these fixations in is different from a human observer.

The model then generated a sequence of fixations, which I plotted onto my input image. To compare against human data, I plotted the human fixations on each frame as well. Figure

| Subject/example | Fixation 2 | Fixation 3 | Fixation 4 |
|---|---|---|---|
| 15 | 384 | 482 | n/a |
| 19 | 235 | 156 | n/a |
| 25 | 360 | 221 | 301 |
| 32 | 154 | 406 | 152 |
| 42 | 272 | n/a | n/a |

**Table 7.3:** Euclidean distances (px) between predicted and real subject fixation locations in sequence. Some examples have fewer than four fixations, so there is no distance to report. These cells are filled with n/a.

7.5 shows three examples of the IBS generated fixations against the human fixations. I then computed the Euclidean distance in pixels of the predicted algorithm from the model and the fixation centre of the human data. Table 7.3 summarizes some results from the examples. In general, examples were around 243 pixels away from the true human fixation. The smallest difference was 41 pixels, and the largest was up to 620 pixels away.

With the limited samples obtained in the dataset, it can be seen that there are significant differences in predicted fixation locations by IBS compared to the real subjects' fixations. Although the predicted fixations are still somewhat in close proximity to the real fixations (Euclidean distance $< 500$px, with a 1920x1080 image), they are not particularly predictive. Scanpaths were not replicated, simply the general vicinity of the fixation location. Most of the examples were like those in Figure 7.5, where the generated fixations were somewhat in a similar area of the image, but the order in which those fixations were generated was not the same as the subject's fixation sequence. Thus, not only are there very few examples where subjects are standing still, moving their eyes without moving their head, the predictions given by the model in these few examples are not very accurate.

Investigating the target in view trials, we can see that the algorithm performed slightly better than humans. For 2 of the 3 trials, the algorithm found the target within 3 fixations, whereas the subject only found the target in 1 of the 3 trials. There was a smaller average distance between predicted and true fixation locations (147 pixels), but the pattern of fixation sequences were still dissimilar to the human sequences. However, there were only 3 examples

that I could use in this case, thus it is likely not representative of the true comparison between the model's predictions and the subjects' fixations.

New information about the scene is provided with every change in head direction, something that the IBS model does not account for. Even in static natural images, like those used in Bujia et al. (2022), the benefit and natural tendency to move the head to survey a scene is not captured. Perhaps this is why the model performs poorly at predicting fixations in my experiment, as it optimizes finding fixations with the most information in a static image rather than a dynamic moving scene.

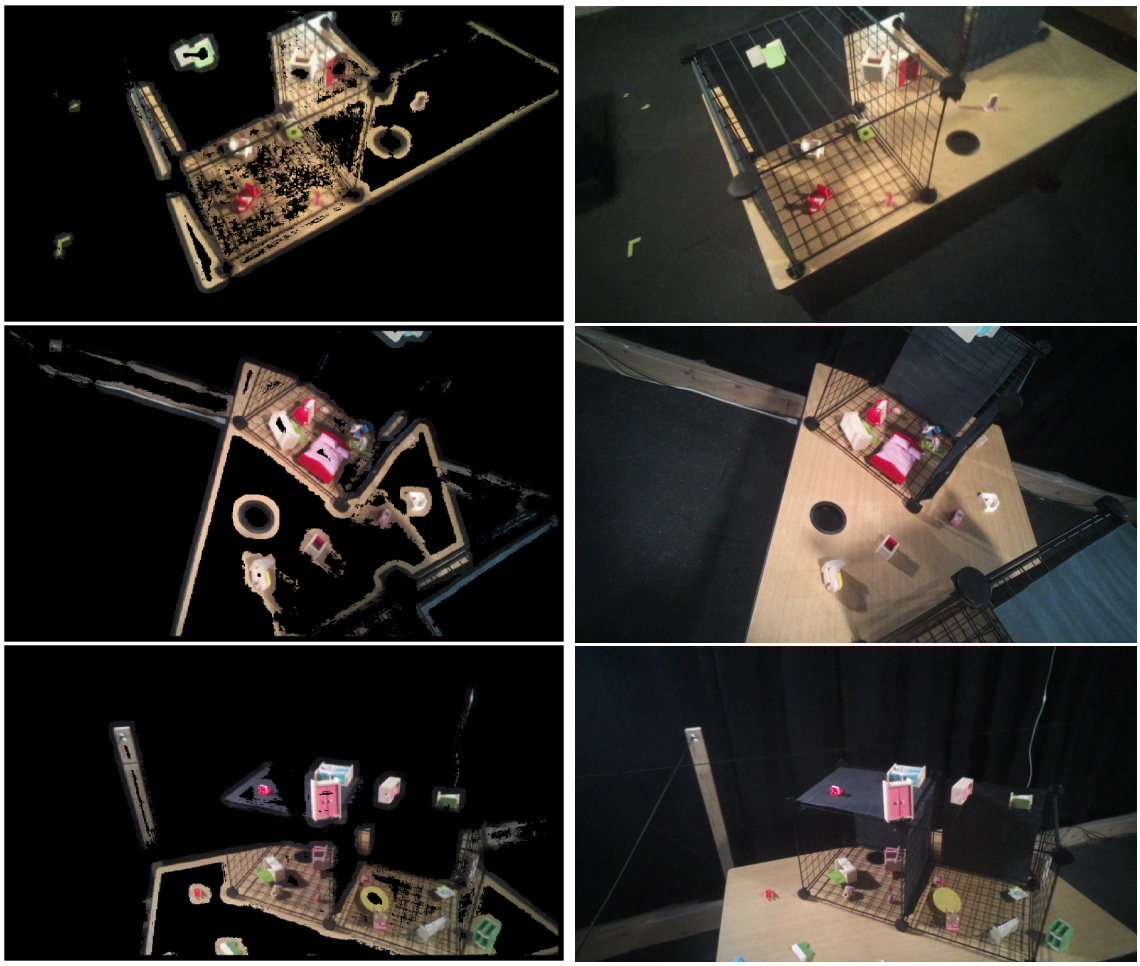
### 7.4.2 AIM

The AIM algorithm by Bruce and Tsotsos (2009) does not provide a sequence of fixations; it generates a saliency map. However, I wanted to see if the salient regions predicted by the algorithm would correlate well with subjects' fixation locations. Thus, I ran the same examples used above through the AIM algorithm, to obtain saliency maps for each example.

Three examples of the generated saliency maps are shown in Figure 7.6, with the original image shown side by side. In each case, the table and cages, in particular the objects on them, are labelled as parts of the salient region, while the floor and the walls are not salient. Empty regions of the tables are also labelled as not salient. Indeed, it seems to highlight any area with features as salient, regardless of whether or not they may be relevant to the search task. Thus, AIM's recall is high, but the precision is low.

Looking at fixation distributions, the majority of fixations (76%) are spent looking at a salient region, either a table or a cage in the setup. The remaining are fixations that do not look at any table or cage, and could be attributed to the subjects needing to look at where they are going in order to navigate the space. Thus, it seems that this saliency map is able to pinpoint the locations that subjects spend the majority of their time looking. However, it also highlighs some areas that subjects do not look at, such as the black circle holes on each table.

AIM is unlikely to miss any regions that may be important to search, although it provides a rather generous amount of false alarms. Consequently, it may be a good starting point as a foundation for robotics models to use, but would benefit with being more selective.



**(a)** AIM saliency maps **(b)** Original images

**Figure 7.6:** 3 examples of saliency maps, original image shown on right. AIM tends to label the object stimuli as salient, although it also seems to label empty areas of the cages as salient.

## 7.5 Summary

I had two main research questions when approaching the comparisons to computational models. There was not always a direct way to address the first question, but for each model, I attempted to find a means to compare performance. I was not able to compare performance between robotics models and humans, as doing so would have required a reconstruction of my setup in a simulation, or implementing a real world robot with an algorithm to search in my environment. Thus, I was only able to compare behaviours between robotics models and humans, through analyzing each model's cost functions. In foraging, I was able to easily

transform my experiment data into a foraging problem, and feed it into the MVT. Through that, I was able to compare the difference in performance of an optimal patch-leaving strategy compared to humans. Similarly to other foraging tasks, subjects seem to spend more time than is optimal at each patch before leaving. For the IBS, again, I was able to find cases where subjects were moving only their eyes and compare performance of the model against human fixations. Although there is no better or worse fixation sequence, the comparison of fixation locations as well as the sequence show that IBS does not predict subjects' fixations in the few examples that I found.

In order to answer my second research question, "are strategies used in computational models for active search comparable to human strategies?" I found a few different ways to compare elements of my experiment data with the 3 main types of models. In robotics models, I compared the movement and fixation selection strategies. In foraging, I compared patch/area leaving strategies, and in eye movement models, I compared fixation sequence strategies. In summary, my experiment data is most closely in line with predictions made in foraging, particularly MVT, as well as models found in the robotics field. The few valid examples of still subjects moving their eyes only were not in line with eye movement predictions given by the Ideal Bayesian Observer, although it has been shown to work well to predict eye movements in a typical 2D search task.

# Chapter 8

# Conclusion

## 8.1 Implications

There are several implications of the work presented in my thesis. First and foremost, this is the first experiment to record eye and head movement data in a visual search task done in a real world environment, where the subjects are untethered and free to move around during the task. Importantly, this meant that I was able to investigate the importance of selecting viewpoints during a visual search task. The results from the human experiment allows us to gain a better understanding of how people are conducting active search in the real world, highlighting some key differences between results from real world search and those found in 2D search tasks.

Due to differences in the ability to determine search slope, the need for viewpoint selection, as well as the difficulty and impracticality of defining trial difficulty using object features, popular search models from the 2D literature cannot be compared or used for data collected from this experiment. Moreover, these differences exist for many search tasks in the real world when compared to a 2D search. Thus, models from 2D search are likely unsuitable for describing visual search behaviour in a 3D, real world environment.

Strategies uncovered from the human data further show the importance of considering patterns of behaviour using metrics like eye and head movements together, rather than simple psychophysics metrics like response time, accuracy, and search slope.

Comparisons to computational models allow us to show differences as well as similarities

in behaviour between models and humans. Robotics models are able to capture some general behaviours exhibited by subjects, such as taking multiple views in one location before moving. Foraging models predict subject behaviour for leaving the various tables and cages in the setup. Eye movement models were limited in data for comparison, and the few examples that were comparable showed poor predictability for subject eye movements.

Overall, each of the models were comparable to certain components of the human subject data, but none were able to encompass the search behaviour as a whole. Search may seem like a simple task for humans, but it involves many different decisions and actions that need to come together in order to work.

## 8.2 Limitations and Future Directions

Although the gathered dataset of experimental data is incredibly rich, it is still lacking in power. Due to my interest in testing the many variables including target presence, visibility, set size, layout, and trials, I was only able to gather the data of 3 subjects for each layout and trial combination, and only about 1 to 2 trials per subject on the combinations of the other variables. This makes it difficult to make any statistical conclusions, especially on the variables manipulated within subjects. Compared to 2D search tasks with hundreds of trials, I only ran 12 trials. This was due in part to the long setup time for each trial, as well as the high variability in trial completion time for each subject. As well, subjects would noticeably begin to fidget with the tracking glasses after about 30 minutes (typical duration of our task with 12 trials was 25-35 minutes, not including setup and calibration), especially subjects that were not accustomed to wearing glasses. In order to maintain an acceptable level of comfort for subjects, we chose not to extend the experiment duration.

The presentation of images of the target as a prompt was also another limitation. Subjects were only able to see a canonical view of the target, as well as no reference for size. The colors in the images could also be distorted from lighting when the image was taken. This could cause the target to look like a different color than seen in the setup.

There were many more variables that could have been manipulated to gather more interesting data, such as including non-uniform or non-random distributions of objects across the

locations, varying the density of objects across the setup, having the subject start the trial in different locations, using scene grammar to structure the layouts, and so on. Limitations on time frame would not allow us to do these experiments, although the results may be interesting. The novelty of this experiment design thus allows for many previously unanswered questions to be answered.

On the flip side, certain manipulated variables, such as target visibility from starting location and some set sizes, could have been eliminated to simplify the experiment. Perhaps three levels of set sizes with a larger difference ((30, 45, 60) instead of (30, 40, 50, 60)) could have been used.

Conducting an active search experiment in a virtual environment could also have been simpler. With a virtual environment, certain variables become much easier to control, and switching between trials becomes trivially easy. 3D projections of gaze coordinates also become much simpler. Perhaps a virtual version of the experiment could be conducted as well, although implications from such an experiment may not extend as strongly to the real world as the work in this thesis. Furthermore, Mon-Williams and Wann (1998) found that long durations of viewing would lead to deficits in stereoscopic depth, thus limiting the practical duration of experiments conducted in virtual reality.

I would also like to ultimately form a computational model for search from the obtained data. Then, this model could be implemented as an algorithm for a robot, to conduct search in the exact same environment, such that direct comparisons can be made between the human results and robot results. Developing and utilizing computational models with a better understanding of human active search also allow users to comprehend mobile robot actions better, making them more suited for real world applications such as search-and-rescue, advertising, and exploration.

## 8.3 Conclusion

In this thesis, I set out to address a few key questions: How do subjects behave in a real world visual search task? How well do existing models encompass these behaviours? Through my experiment, I gathered a large volume of data regarding human eye and head

movements during real world search. I was then able to extract several strategies from the data, including common scanpaths, search-paths, hotspots, and lower-level behaviours such as location tracking and revisits. Overall behaviour was also discovered, such as an increase in efficiency over only target present trials, and a similarity to 2D search in the doubling of response times and number of fixations for target absent trials. By comparing these behaviours to computational models, I found that the MVT from foraging literature did well in predicting patch-leaving behaviours, AIM saliency was overly generous in highlighting salient regions for each snapshot, IBS performed somewhat poorly in predicting eye movements, and robotics models shared some similarities in cost function, but behave differently on a fundamental basis.

Results from this experiment bring us one step closer to understanding how visual search is conducted in the real world, and which elements of the traditional 2D search task are truly translated into a 3D environment. The state of our interpretation of real world search can also somewhat be reflected in the existing computational models of search. Although 3D human search has not been explicitly modelled, there are several relevant models that can be compared. By examining the differences in human behaviour to predicted behaviour, we are able to get a better grasp of components from the 2D search, foraging, or robotics fields, that are consistent with human behaviour in the real world, in order to begin shaping a computational model for human real world search.

# Bibliography


Adeli, H. and Zelinsky, G. (2018). Deep-BCN: Deep networks meet biased competition to create a brain-inspired model of attention control. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition workshops*, pages 1932–1942.

Agrawal, R., Srikant, R., et al. (1994). Fast algorithms for mining association rules. In *Proceedings of the 20th International Conference on Very Large Data Bases, VLDB*, volume 1215, pages 487–499. Santiago, Chile.

Aloimonos, J., Weiss, I., and Bandyopadhyay, A. (1988). Active vision. *International Journal of Computer Vision*, 1(4):333–356.

Anderson, D. J. (1983). Optimal foraging and the traveling salesman. *Theoretical Population Biology*, 24(2):145–159.

Bajcsy, R. (1988). Active perception. *Proceedings of the IEEE*, 76(8):966–1005.

Ballard, D. H. (1989). Animat vision. In *Computer Vision: A Reference Guide*, pages 52–57. Springer.

Ballard, D. H., Hayhoe, M. M., and Pelz, J. B. (1995). Memory representations in natural tasks. *Journal of Cognitive Neuroscience*, 7(1):66–80.

Bartumeus, F. and Catalan, J. (2009). Optimal search behavior and classic foraging theory. *Journal of Physics A: Mathematical and Theoretical*, 42(43):434002.

Bella-Fernández, M., Suero Suñé, M., and Gil-Gómez de Liaño, B. (2021). Foraging behavior in visual search: A review of theoretical and mathematical models in humans and animals. *Psychological Research*, pages 1–19.

Bettinger, R. L. and Grote, M. N. (2016). Marginal Value Theorem, patch choice, and human foraging response in varying environments. *Journal of Anthropological Archaeology*, 42:79–87.

Bruce, N. D. and Tsotsos, J. K. (2009). Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision*, 9(3):5–5.

Bujia, G., Sclar, M., Vita, S., Solovey, G., and Kamienkowski, J. E. (2022). Modeling human visual search in natural scenes: A combined Bayesian searcher and saliency map approach. *Frontiers in Systems Neuroscience*, 16.

Bundesen, C., Habekost, T., and Kyllingsbæk, S. (2011). A neural theory of visual attention and short-term memory (NTVA). *Neuropsychologia*, 49(6):1446–1457.

Charnov, E. L. (1976). Optimal foraging, the Marginal Value Theorem. *Theoretical population biology*, 9(2):129–136.

Chen, S., Li, Y., and Kwok, N. M. (2011). Active vision in robotic systems: A survey of recent developments. *The International Journal of Robotics Research*, 30(11):1343–1377.

Chun, M. M. and Jiang, Y. (1998). Contextual cueing: Implicit learning and memory of visual context guides spatial attention. *Cognitive Psychology*, 36(1):28–71.

Desimone, R. and Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, 18(1):193–222.

Eckstein, M. P. (2011). Visual search: A retrospective. *Journal of Vision*, 11(5):14–14.

Eckstein, M. P., Thomas, J. P., Palmer, J., and Shimozaki, S. S. (2000). A signal detection model predicts the effects of set size on visual search accuracy for feature, conjunction, triple conjunction, and disjunction displays. *Perception & Psychophysics*, 62:425–451.

Eimer, M. (2014). The neural basis of attentional control in visual search. *Trends in Cognitive Sciences*, 18(10):526–535.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, volume 96, pages 226–231.

Foulsham, T., Chapman, C., Nasiopoulos, E., and Kingstone, A. (2014). Top-down and bottom-up aspects of active search in a real-world environment. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, 68(1):8.

Foulsham, T., Walker, E., and Kingstone, A. (2011). The where, what and when of gaze allocation in the lab and the natural environment. *Vision Research*, 51(17):1920–1931.

Frintrop, S., Werner, T., and Martin Garcia, G. (2015). Traditional saliency reloaded: A good old model in new shape. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 82–90.

Garvey, T. D. (1976). Perceptual strategies for purposive vision.

Geyer, T., Shi, Z., and Müller, H. J. (2010). Contextual cueing in multiconjunction visual search is dependent on color-and configuration-based intertrial contingencies. *Journal of Experimental Psychology: Human Perception and Performance*, 36(3):515.

Gilchrist, I. D., North, A., and Hood, B. (2001). Is visual search really like foraging? *Perception*, 30(12):1459–1464.

Hayhoe, M. and Ballard, D. (2005). Eye movements in natural behavior. *Trends in Cognitive Sciences*, 9(4):188–194.

Hayhoe, M. M., Shrivastava, A., Mruczek, R., and Pelz, J. B. (2003). Visual memory and motor planning in a natural task. *Journal of Vision*, 3(1):6–6.

Howard, C. J., Pharaon, R. G., Körner, C., Smith, A. D., and Gilchrist, I. D. (2011). Visual search in the real world: Evidence for the formation of distractor representations. *Perception*, 40(10):1143–1153.

Itti, L. and Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40(10-12):1489–1506.

Johannesson, O. I., Thornton, I. M., Smith, I. J., Chetverikov, A., and Kristjánsson, A. (2016). Visual foraging with fingers and eye gaze. *i-Perception*, 7(2):2041669516637279.

Klein, R. M. (2000). Inhibition of return. *Trends in Cognitive Sciences*, 4(4):138–147.

Krejtz, K., Duchowski, A., Krejtz, I., Szarkowska, A., and Kopacz, A. (2016). Discerning ambient/focal attention with coefficient K. *ACM Transactions on Applied Perception (TAP)*, 13(3):1–20.

Kristjánsson, T., Draschkow, D., Pálsson, Á., Haraldsson, D., Jónsson, P. Ö., and Kristjánsson, Á. (2022). Moving foraging into three dimensions: Feature-versus conjunction-based foraging in virtual reality. *Quarterly Journal of Experimental Psychology*, 75(2):313–327.

Liesefeld, H. R. and Müller, H. J. (2020). A theoretical attempt to revive the serial/parallel-search dichotomy. *Attention, Perception, & Psychophysics*, 82:228–245.

Marius't Hart, B., Vockeroth, J., Schumann, F., Bartl, K., Schneider, E., König, P., and Einhäuser, W. (2009). Gaze allocation in natural stimuli: Comparing free exploration to head-fixed viewing conditions. *Visual Cognition*, 17(6-7):1132–1158.

Matthis, J. S., Yates, J. L., and Hayhoe, M. M. (2018). Gaze and the control of foot placement when walking in natural terrain. *Current Biology*, 28(8):1224–1233.

Mon-Williams, M. and Wann, J. P. (1998). Binocular virtual reality displays: When problems do and don't occur. *Human Factors*, 40(1):42–49.

Najemnik, J. and Geisler, W. S. (2005). Optimal eye movement strategies in visual search. *Nature*, 434(7031):387–391.

Najemnik, J. and Geisler, W. S. (2009). Simple summation rule for optimal fixation selection in visual search. *Vision Research*, 49(10):1286–1294.

Nakayama, K. and Martini, P. (2011). Situating visual search. *Vision Research*, 51(13):1526–1537.

Neider, M. B. and Zelinsky, G. J. (2006). Scene context guides eye movements during visual search. *Vision Research*, 46(5):614–621.

Over, E., Hooge, I., Vlaskamp, B., and Erkelens, C. (2007). Coarse-to-fine eye movement strategy in visual search. *Vision Research*, 47(17):2272–2280.

Pelz, J., Hayhoe, M., and Loeber, R. (2001). The coordination of eye, head, and hand movements in a natural task. *Experimental Brain Research*, 139:266–277.

Peters, R. J., Iyer, A., Itti, L., and Koch, C. (2005). Components of bottom-up gaze allocation in natural images. *Vision Research*, 45(18):2397–2416.

Rao, R. P., Zelinsky, G. J., Hayhoe, M. M., and Ballard, D. H. (2002). Eye movements in iconic visual search. *Vision Research*, 42(11):1447–1463.

Rasouli, A., Lanillos, P., Cheng, G., and Tsotsos, J. K. (2020). Attention-based active visual search for mobile robots. *Autonomous Robots*, 44(2):131–146.

Sauter, M., Stefani, M., and Mack, W. (2020). Towards interactive search: Investigating visual search in a novel real-world paradigm. *Brain Sciences*, 10(12):927.

Schmid, J. F., Lauri, M., and Frintrop, S. (2019). Explore, approach, and terminate: Evaluating subtasks in active visual object search based on deep reinforcement learning. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5008–5013. IEEE.

Shubina, K. and Tsotsos, J. K. (2010). Visual search for an object in a 3D environment using a mobile robot. *Computer Vision and Image Understanding*, 114(5):535–547.

Sjöö, K., Aydemir, A., and Jensfelt, P. (2012). Topological spatial relations for active visual search. *Robotics and Autonomous Systems*, 60(9):1093–1107.

Smith, A. D., Hood, B. M., and Gilchrist, I. D. (2008). Visual search and foraging compared in a large-scale search task. *Cognitive Processing*, 9(2):121–126.

Snodgrass, J. G. and Townsend, J. T. (1980). Comparing parallel and serial models: Theory and implementation. *Journal of Experimental Psychology: Human Perception and Performance*, 6(2):330.

Solbach, M. D. and Tsotsos, J. K. (2020). PESAO: Psychophysical Experimental Setup for Active Observers. *arXiv preprint arXiv:2009.09933*.

Swain, M. J. and Ballard, D. H. (1991). Color indexing. *International Journal of Computer Vision*, 7(1):11–32.

Tatler, B. W., Baddeley, R. J., and Gilchrist, I. D. (2005). Visual correlates of fixation selection: Effects of scale and time. *Vision Research*, 45(5):643–659.

Treisman, A. M. and Gelade, G. (1980). A Feature-Integration theory of attention. *Cognitive Psychology*, 12(1):97–136.

Turrin, C., Fagan, N. A., Dal Monte, O., and Chang, S. W. (2017). Social resource foraging is guided by the principles of the Marginal Value Theorem. *Scientific Reports*, 7(1):11274.

Viswanathan, G. M., Buldyrev, S. V., Havlin, S., Da Luz, M., Raposo, E., and Stanley, H. E. (1999). Optimizing the success of random searches. *Nature*, 401(6756):911–914.

Vo, M. L.-H., Boettcher, S. E., and Draschkow, D. (2019). Reading scenes: How scene grammar guides attention and aids perception in real-world environments. *Current Opinion in Psychology*, 29:205–210.

Wei, Z., Adeli, H., Nguyen, M. H., Zelinsky, G., and Samaras, D. (2016). Learned region sparsity and diversity also predicts visual attention. *Advances in Neural Information Processing Systems*, 29.

Wixson, L. E. and Ballard, D. H. (1994). Using intermediate objects to improve the efficiency of visual search. *International Journal of Computer Vision*, 12(2-3):209–230.

Wloka, C., Kotseruba, I., and Tsotsos, J. K. (2018). Saccade sequence prediction: Beyond static saliency maps. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wolfe, J. M. (1994). Guided Search 2.0 A revised model of visual search. *Psychonomic Bulletin & Review*, 1(2):202–238.

Wolfe, J. M. (1998). What can 1 million trials tell us about visual search? *Psychological Science*, 9(1):33–39.

Wolfe, J. M. (2013). When is it time to move to the next raspberry bush? Foraging rules in human visual search. *Journal of Vision*, 13(3):10–10.

Wolfe, J. M. (2021). Guided Search 6.0: An updated model of visual search. *Psychonomic Bulletin & Review*, 28(4):1060–1092.

Wolfe, J. M., Alvarez, G. A., Rosenholtz, R., Kuzmova, Y. I., and Sherman, A. M. (2011). Visual search for arbitrary objects in real scenes. *Attention, Perception, & Psychophysics*, 73(6):1650–1671.

Wolfe, J. M., Cave, K. R., and 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.

Wolfe, J. M. and Gray, W. (2007). Guided Search 4.0. *Integrated Models of Cognitive Systems*, pages 99–119.

Wolfe, J. M. and Horowitz, T. S. (2017). Five factors that guide attention in visual search. *Nature Human Behaviour*, 1(3):1–8.

Yang, Z., Huang, L., Chen, Y., Wei, Z., Ahn, S., Zelinsky, G., Samaras, D., and Hoai, M. (2020). Predicting goal-directed human attention using inverse reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 193–202.

Yarbus, A. L. (1967). *Eye Movements and Vision*. Springer.

Ye, X., Lin, Z., Li, H., Zheng, S., and Yang, Y. (2018). Active object perceiver: Recognition-guided policy learning for object searching on mobile robots. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6857–6863. IEEE.

Ye, Y. and Tsotsos, J. K. (1999). Sensor planning for 3D object search. *Computer Vision and Image Understanding*, 73(2):145–168.

Zelinsky, G. J. (2008). A theory of eye movements during target acquisition. *Psychological Review*, 115(4):787.

Zelinsky, G. J. (2012). TAM: Explaining off-object fixations and central fixation tendencies as effects of population averaging during search. *Visual Cognition*, 20(4-5):515–545.

Zhang, M., Feng, J., Ma, K. T., Lim, J. H., Zhao, Q., and Kreiman, G. (2018). Finding any Waldo with zero-shot invariant and efficient visual search. *Nature Communications*, 9(1):3730.

Zhao, Q. (2011). 10 scientific problems in virtual reality. *Communications of the ACM*, 54(2):116–118.

Zhou, Y. and Yu, Y. (2021). Human visual search follows a suboptimal Bayesian strategy revealed by a spatiotemporal computational model and experiment. *Communications Biology*, 4(1):1–16.

## 8.4 Appendix A

**Figure 8.1:** Part 1 of all objects used in experiment. Sorted in alphabetical order of object name.

**Figure 8.2:** Part 2 of all objects used in experiment. Sorted in alphabetical order of object name.

## 8.5 Appendix B

| Object | colour | size | shape | texture |
|---|---|---|---|---|
| babychair1 | red, white | small | babychair | plastic |
| babychair2 | aqua, brown | medium | babychair | wood |
| babychair3 | pink, brown | large | babychair | wood |
| basket | pink | small | basket | plastic |
| bed1 | pink, yellow | small | bed | plastic |
| bed2 | brown, aqua | medium | bed | wood |
| bed3 | pink, brown | large | bed | wood |
| bed4 | pink, brown | large | bed | wood |
| bed5 | red, pink, white brown | xlarge | bed | wood |
| chair1 | aqua, yellow | small | chair | plastic |
| chair10 | green, brown | xlarge | chair | wood |
| chair11 | green, brown | xlarge | chair | wood |
| chair12 | green, brown | xlarge | chair | wood |
| chair13 | green, brown | xlarge | chair | wood |
| chair2 | red, white | small | chair | plastic |
| chair3 | red, white | small | chair | plastic |
| chair4 | red, white | small | chair | plastic |
| chair5 | red, white | small | chair | plastic |
| chair6 | orange, brown | medium | chair | wood |
| chair7 | orange, brown | medium | chair | wood |
| chair8 | pink, brown | large | chair | wood |
| chair9 | blue, brown | xlarge | chair | wood |
| clothesline1 | white, red | small | clothesline | plastic |
| computer1 | pink, yellow | small | computer | plastic |
| cot1 | blue, pink | small | cot | plastic |
| cot2 | aqua, brown | medium | cot | wood |
| cot3 | pink, brown | large | cot | wood |

| | | | | |
|---|---|---|---|---|
| cot4 | red, brown | xlarge | cot | wood |
| couch1 | yellow, pink | small | couch | plastic |
| couch10 | green, brown, white, pink | xlarge | couch | wood |
| couch2 | pink, yellow | small | couch | plastic |
| couch3 | yellow, pink | small | couch | plastic |
| couch4 | white, red | small | couch | plastic |
| couch5 | green, brown, white, red | medium | couch | wood |
| couch6 | green, brown, white, red | medium | couch | wood |
| couch7 | green, brown, white, red | medium | couch | wood |
| couch8 | green, brown, white, pink | xlarge | couch | wood |
| couch9 | green, brown, white, pink | xlarge | couch | wood |
| cupboard1 | pink, aqua, yellow | small | cupboard | plastic |
| cupboard2 | yellow, pink, aqua | small | cupboard | plastic |
| cupboard3 | red, white, blue | small | cupboard | plastic |
| cupboard4 | pink, brown | large | cupboard | wood |
| cupboard5 | pink, brown | large | cupboard | wood |
| cupboard6 | pink, brown | large | cupboard | wood |
| cupboard7 | green, brown | xlarge | cupboard | wood |
| cupboard8 | green, brown | xlarge | cupboard | wood |
| dresser1 | yellow, pink, aqua | small | dresser | plastic |
| dresser2 | red, blue, brown | xlarge | dresser | wood |
| fridge1 | red, white | small | fridge | plastic |
| fridge2 | orange, brown | medium | fridge | wood |
| fridge3 | pink, brown | large | fridge | wood |
| fridge4 | green, brown | xlarge | fridge | wood |
| hanger1 | red, yellow | small | hanger | plastic |
| hanger2 | red, yellow | small | hanger | plastic |
| hanger3 | brown, white | small | hanger | plastic |
| iron1 | pink | small | iron | plastic |
| ironboard1 | yellow, blue, pink | small | ironboard | plastic |

| | | | | |
|---|---|---|---|---|
| lamp1 | yellow, pink | small | lamp | plastic |
| lamp2 | blue, brown | medium | lamp | wood |
| lamp3 | pink, brown | large | lamp | wood |
| lamp4 | pink, brown | large | lamp | wood |
| lamp5 | yellow, brown, blue, green | xlarge | lamp | wood |
| lamp6 | red, brown | xlarge | lamp | wood |
| lamppost1 | red | small | lamppost | plastic |
| lamppost2 | brown, white | small | lamppost | plastic |
| lamppost3 | brown, white | small | lamppost | plastic |
| mat1 | yellow | medium | mat | wood |
| microwave | green, grey, brown | xlarge | microwave | wood |
| plant | green, brown | xlarge | plant | wood |
| radio1 | pink, yellow | small | radio | plastic |
| rockinghorse1 | pink, brown | large | rockinghorse | wood |
| scale1 | pink, red, yellow | small | scale | plastic |
| shower1 | white, red | small | shower | plastic |
| shower2 | brown, yellow | medium | shower | wood |
| shower3 | pink, brown | large | shower | wood |
| shower4 | brown, yellow, red, white | xlarge | shower | wood |
| sink1 | white, brown, red, blue | small | sink | plastic |
| sink2 | brown, yellow | medium | sink | wood |
| sink3 | pink, brown | large | sink | wood |
| sink4 | yellow, white | xlarge | sink | wood |
| stand1 | yellow | small | stand | plastic |
| stool1 | blue, brown | xlarge | chair | wood |
| stool2 | blue, brown | xlarge | chair | wood |
| stool3 | blue, brown | xlarge | chair | wood |
| stool4 | blue, brown | xlarge | chair | wood |
| storage1 | yellow, pink | small | storage | plastic |
| stove1 | orange, brown | medium | stove | wood |

| | | | | |
|---|---|---|---|---|
| stove2 | pink, brown | large | stove | wood |
| stove3 | green, brown | xlarge | stove | wood |
| stroller1 | blue, purple, white | medium | stroller | wood |
| swing1 | blue, pink, brown | xlarge | swing | wood |
| table1 | pink | small | table | plastic |
| table10 | green, brown | xlarge | table | wood |
| table2 | white, red | small | table | plastic |
| table3 | brown | medium | table | wood |
| table4 | green, brown | medium | table | wood |
| table5 | brown, green | medium | table | wood |
| table6 | brown | large | table | wood |
| table7 | brown | large | table | wood |
| table8 | yellow, brown | xlarge | table | wood |
| table9 | brown, red | xlarge | table | wood |
| toilet1 | pink, red, white | small | toilet | plastic |
| toilet2 | yellow, brown | medium | toilet | wood |
| toilet3 | pink, brown | large | toilet | wood |
| toilet4 | yellow | xlarge | toilet | wood |
| tub1 | white, red | small | rub | plastic |
| tub2 | brown, yellow, red | medium | tub | wood |
| tub3 | pink, brown | large | tub | wood |
| tub4 | brown, blue, red | xlarge | tub | wood |
| tv1 | pink, yellow | small | tv | plastic |
| tv2 | brown, green | medium | tv | wood |
| tv3 | blue, brown, yellow | xlarge | tv | wood |
| umbrella | blue, brown, pink | xlarge | umbrella | wood |
| vase1 | pink, red, green | medium | vase | wood |
| vase2 | pink, green, red | medium | vase | wood |
| wardrobe1 | yellow, pink | small | wardrobe | plastic |
| wardrobe2 | pink, brown | large | wardrobe | wood |

| | | | | |
|---|---|---|---|---|
| wardrobe3 | red, blue brown | xlarge | wardrobe | wood |
| washer1 | pink, brown | large | washer | wood |

**Table 8.1:** Table of all objects and their attributes, sorted in alphabetical order of object name.