An Integrated Computational Model of Visual Search Combining Eccentricity, Bottom-up, and Top-down Cues

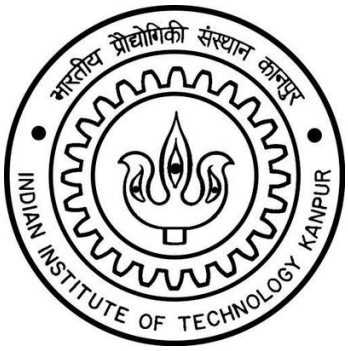
A thesis submitted
in Partial Fulfillment of the Requirements
for the Degree of

B.Tech-M.Tech (Dual Degree)

by

Shashi Kant Gupta

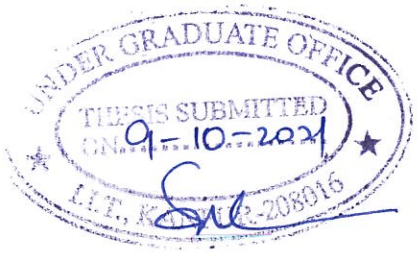
to the

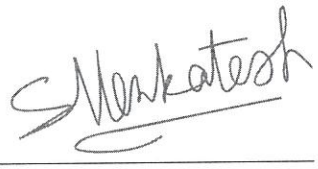


DEPARTMENT OF ELECTRICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY KANPUR
October 2021

CERTIFICATE

It is certified that the work contained in the thesis titled **An Integrated Computational Model of Visual Search Combining Eccentricity, Bottom-up, and Top-down Cues**, by **Shashi Kant Gupta**, has been carried out under my supervision and that this work has not been submitted elsewhere for a degree.





Prof K. S. Venkatesh

Department of Electrical Engineering

IIT Kanpur

October 2021

DECLARATION

This is to certify that the thesis titled **An Integrated Computational Model of Visual Search Combining Eccentricity, Bottom-up,and Top-down Cues** has been authored by me. It presents the research conducted by me under the supervision of Prof. K. S. Venkatesh. To the best of my knowledge, it is an original work, both in terms of research content and narrative, and has not been submitted elsewhere, in part or in full, for a degree. Further, due credit has been attributed to the relevant state-of-the-art and collaborations (if any) with appropriate citations and acknowledgements, in line with established norms and practices.

Signature

Name: Shashi Kant Gupta

Programme: BT-MT (Dual Degree)

Department: Electrical Engineering

Indian Institute of Technology Kanpur

Kanpur 208016

ABSTRACT

Name of student: **Shashi Kant Gupta** Roll no: **16807645**

Degree for which submitted: **B.Tech-M.Tech (Dual Degree)**

Department: **Electrical Engineering**

Thesis title: **An Integrated Computational Model of Visual Search Combining Eccentricity, Bottom-up, and Top-down Cues**

Name of Thesis Supervisor: **Prof K. S. Venkatesh**

Month and year of thesis submission: **October 2021**

Visual search is described as searching for some target object in a given visual scene, having several other non-target objects. Humans are continuously involved in such tasks in their day-to-day life like searching for a specific food item in the mall or searching for their friends at a party. Extensive studies in visual search behaviour have demonstrated the complex interplay of the target object, the search space, and the humans' memory. In parallel, neurophysiological studies have shown how neuronal circuits form complex visual representations. But very little work has been done that links these exciting works from behavioural studies and neuroscience. This thesis introduces an integrated computational model of visual search that incorporates theoretical frameworks from psychology, resembling the architecture from neurophysiology. The proposed model integrates three essential components, an eccentricity-dependent deep convolutional neural network as a visual processor, top-down target modulated activation maps, and bottom-up saliency-based activations. The proposed model can replicate the standard results from several behavioural studies conducted in visual search literature. And at the same time, it is also efficient enough to search for a target object in a complex natural scene. Various autonomous systems can also significantly benefit from the proposed

model, for example, autonomous navigation or visual clinical diagnosis. Most of the previous work on computational modelling of visual search in computer vision involves extensive category-specific training and bear minimal resemblance to biological plausibility. In comparison, the proposed model is self-sufficient and does not require human supervision or extensive task-specific training to search for any new target object. An explainable visual search model that could replicate human visual search behaviour will also bring more trust in such autonomous systems.

To Ma & Papa.

And Artificial & Biological Neurons!

Acknowledgements

I am extremely grateful to Prof. Gabriel Kreiman, Boston Children's Hospital, Harvard Medical School, for his tremendous support and guidance throughout my thesis work. He was always there for thoughtful science discussion and critical examination of my results. I'm also very thankful to Dr Mengmi Zhang. Post. Doc., Kreiman Lab, Boston Children's Hospital, Harvard Medical School for her tremendous support on the computational work and uncountable insightful discussions on this work.

I would also like to thank Prof. Jeremy M. Wolfe and Dr Chia-Chien Wu, Visual Attention Lab, Brigham and Women's Hospital, Harvard Medical School, for providing human psychophysics data from their lab. I'm also very thankful for their comments and suggestions on this work.

I am very thankful to Prof. K. S. Venkatesh, Department of Electrical Engineering, IIT Kanpur, for his support in writing this Thesis and discussions on this work and throughout my Master's program at IIT Kanpur. I would also like to thank him for his constant support and guidance on my career choices during my stay at IIT Kanpur. I'm also very thankful for all the classes which he taught me, his very first lecture on signals and system was something that sparked my interest in signal processing and computer vision.

I would also like to thank the Cognitive Science program at IIT Kanpur, which allowed me to explore and gain deep insights into human cognition and learn various skills needed to conduct research in this domain.

Finally, I would like to thank my family and friends for their emotional support

and encouragement throughout my stay at IIT Kanpur. Their constant encouragement and support helped me thrive in this challenging environment.

Contents

List of Figures

Chapter 1

Introduction

Visual search is described as searching for some target object in a given surrounding having several other non-target objects. Humans are continuously involved in such tasks in their day-to-day life like searching for a specific food item in the mall, searching for their friends at a party, searching for impurities while preparing rice, and many more. Various autonomous systems can also be significantly benefited from a visual search system, such as autonomous navigation or visual clinical diagnosis.

Even though the visual search is fundamental to humans' day-to-day tasks, the brain goes through very complex cognitive processing while performing a visual search task. Extensive studies in visual search behaviour have demonstrated the complex interplay of the target object, the search space, and the humans' memory and attention and how these often affect the search performances across different visual search tasks ([1, 2, 3, 4, 5, 6, 7]). The neurophysiological studies of visual processing in the brain have also demonstrated how different neuronal circuits capture visual representations, deploying attention and the eye moment ([8, 9, 10, 11, 12, 13, 14, 15]). Combining the studies from both worlds, we can see there are several complex steps involved while performing a simple visual search of finding an apple in a fruit basket. At first, humans must have stored some complex visual representation of apples in their visual memory. Furthermore, when they look at the fruit basket, the visual scene in front of their eyes might go through similar visual processing, which was responsible for storing the visual representation of the apple.

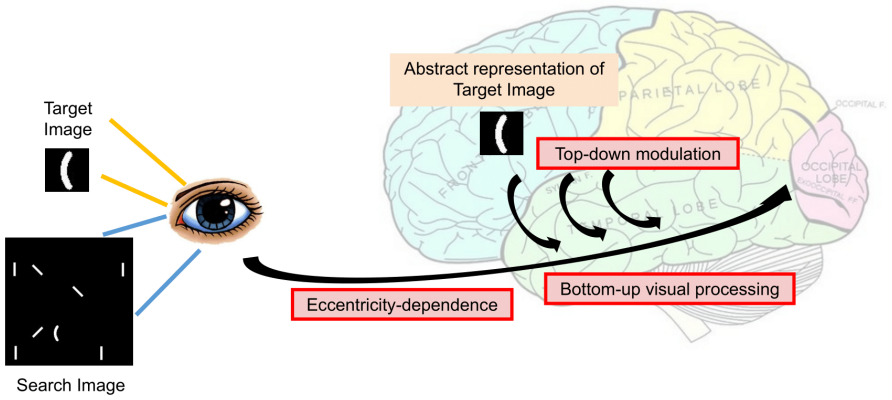


Figure 1.1: Basic cognitive processes involved in visual search

Later the brain might be using these representations to build an attention map to predict the probability of finding the apple at a specific location in the given scene. Based on the predictions, the brain sends some signals to move the eyes to that particular location. However, the search does not necessarily end here. After the eye movement, humans verify if the fixated object is really the target or not, which involves another task, i.e. object recognition. Now, suppose the fixated location is not the target. In that case, the whole process repeats again but with the additional component of "memory" responsible for inhibiting the probability of visiting an already visited location. As we see in the above example, the visual search requires a complex interplay of various cognitive functions of the brain. Thus an extensive amount of neural processing might be undergoing during those tasks (See **Figure 1.1** for illustration).

The computational difficulty of visual search can also be realised from the approaches used in computer vision. One of the most basic and earliest models of attention can be thought of as classic template matching algorithms in computer vision. Even though the template matching algorithm provides a great tool to search for an exact target template in a given image, it struggles with some variations or occlusion in the target objects. The template matching approach can be improved by applying it to specialised features extracted from the original image instead of raw pixels. While several models could explore this possibility, IVSN, a recently in-

troduced model of attention by [16], uses a similar method to present a biologically plausible visual search algorithm. IVSN model has shown that bringing knowledge from neuroscience to build the visual search architecture can significantly improve the performance of such a model. Several other computer vision approaches using the latest trends in deep learning technology have been shown to perform very well in object detection tasks that can be loosely connected to visual attention [17, 18, 19]. But these models are limited by searching for only a fixed number of object classes and requires an extensive amount of training to achieve it finally. On the other hand, humans perform a visual search on a zero-shot basis, i.e., you show them a target object, and even if they have not seen it in the past, they show an efficient performance while searching for that object. In contrast, the object detection model will fail badly in those cases. Moreover, these models usually do not have any resemblance to the neurophysiology of visual search. Note that with tons of computational models published these days in artificial intelligence, one massive shift of focus is to show how trustable these AI systems are. For example, in scenarios like autonomous cars, it becomes a necessary condition to build a trustable system. Building a model that aligns with neurophysiology and cognitive theory of the brain is one way we can build trust in these systems. Because if we can show that the model behaves in the same way a human might behave in a similar situation, it's understandable that the system is as much trustable as a human will be. Thus, one crucial objective that we followed in this thesis was to build a computational model of visual search that could replicate human behaviour. At the same time, it must show comparable performances compared with other computer vision models of attention.

During visual search, humans do not necessarily use target-based attention mechanisms. They sometimes also attend to some other locations based on the general likeability of that location which is purely based on the statistics of the search image. These are termed saliency models in computer vision and are often referred to as bottom-up attention in psychology or neuroscience. Several computer vi-

sion approaches exist, like the itti-Koch model, graph-based visual saliency, and the DeepGaze model [20, 21, 22]. Usually, the architecture of these saliency-based attention models differs significantly from the target-based attention model. To our knowledge, none of the previous work tries at combining both top-down and bottom-up of attention into a single model. But from behavioural studies, we know that both kinds of attention play an essential role during attentional deployment. Furthermore, from neurophysiology, we know that the visual stimuli must be going through similar visual processing during the initial stages for both types of attention deployment. The two methods differ only at the end while assigning attentional priorities in the brain. Thus, another focus of this thesis is to integrate different components of visual attention into a single visual processing stage.

This thesis work introduces an integrated computational model of visual search that incorporates both target-based and saliency-based attention. Both use the same model for initial visual processing. The model is biologically plausible. The proposed model can replicate the standard results from several behavioural studies in visual search literature. And at the same time is also efficient enough to search for a target object in a complex natural scene. The model is self-sufficient and does not require human supervision to search for any new target object. Some part of the model is free from any task-specific training, while some part does incorporate task-specific training, but it does it based on its self-feedback mechanism. In other words, the model knows how to learn and does not need any external human interference.

1.1 Organisation of thesis

Chapter 2 provides the necessary background studies to build the foundational block for the proposed "computational model of visual search". It explains in detail what visual search tasks are and gives examples of some important visual search categories. It then briefs some past work in visual search and describes how visual information is processed in the brain and machines (by machines, SOTA computational model in computer vision). **Chapter 3** describes the proposed modelling

work. **Chapter 4** describes different experiments carried out to compare the model behaviour to that of humans and the model's performance in doing a visual search task in natural image conditions. Along with experiment details, **Chapter 4** also shows the individual results of each of the experiments. A combined result comparison across different tasks is explained in **Chapter 5**. Based on the results described in **Chapters 4** and **5**, **Chapter 6** proposes an additional module to the initial proposed model in **Chapter 3** to better capture human behaviour across all the tasks. Finally, **Chapter 7** provides a conclusion of this thesis work and directions to future work.

Chapter 2

Background

This chapter describes necessary background studies to build the foundational block for building the proposed "computational model of visual search". The chapter starts by describing what visual search is and some of the most popular types of visual search tasks that contributed significantly to the development of visual search literature (**Section 2.1**). Then it introduces some of the influential works in modelling of visual search and attention, which were extensively helpful in building the computational model described in this work (**Section 2.2**). Since visual representation plays a vital role in deploying attention during visual search, the chapter will also briefly cover how visual information is processed in the brain and machines (by machines, we mean SOTA computational model in computer vision) (**Section 2.3**).

2.1 What is visual search?

In the most straightforward words, visual search can be described as searching for some target object in a given surrounding having several other non-target objects. For example, see **Figure 2.1**, the target image is a "curve", and the subject needs to find it among other distractor objects, which are straight lines of varying angles. This is one simple example of visual search in an artificial display. As explained earlier in the introduction, the visual search task is fundamental to day-to-day life. The difficulties can vary a lot depending upon what kind of target and distractors

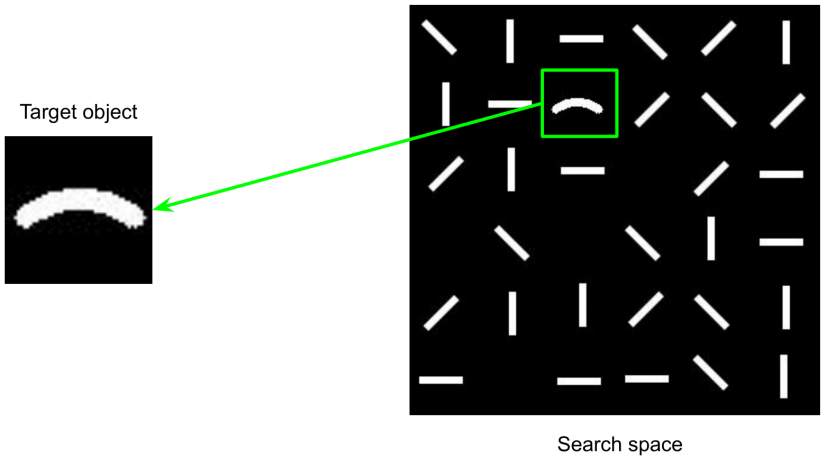


Figure 2.1: An example of visual search task: Searching for a ”curve line” as target item among several straight lines as distractor items.

are present. For example, searching for a red line among green lines could be very easy, but searching waldo in children's book “where's the waldo?” is challenging. Various visual search tasks exist in visual search literature which were used to study the visual search properties. Two very prominent visual search experiments that played a significant role in developing theories in visual search literature are the feature-conjunction search pairs and search asymmetry.

2.1.1 Feature conjunction search pairs

The early theories of the visual search were mainly based on the observation that it is effortless to search for a target object when the target can be distinguished from the distractors on the basis of a single “basic” feature. Here basic features mean colour, shape, motion, depth [23, 24]. These types of search conditions were termed “feature search”. They are also called “parallel search” because in these cases, the subject finds the target very quickly, almost instantly, which seems like all the objects (target and distractors) were attended in parallel and simultaneously. However, when the subjects were given a search task in which they need to combine two or more basic features to distinguish the target from the distractors, the time taken by them increases monotonically with the number of distractor items. Thus,

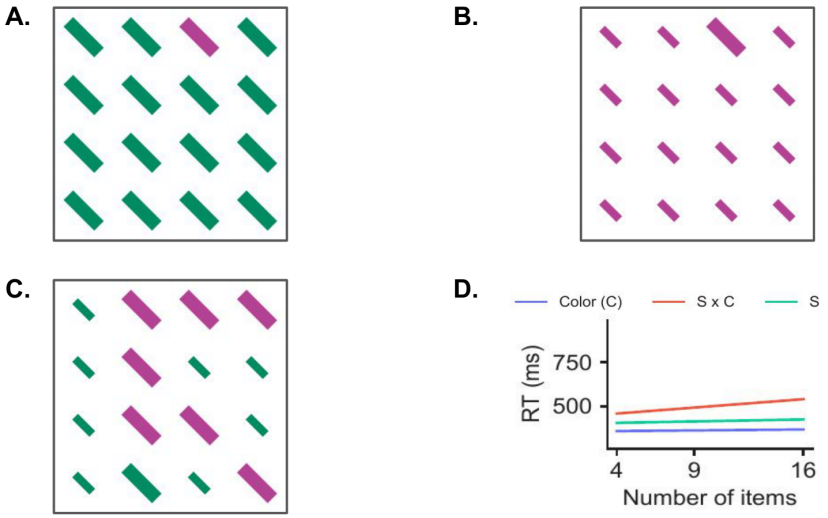


Figure 2.2: Example of feature conjunction search pairs. It's easy to search for singleton feature "a pink bar" in panel **A.** or "a big bar" in panel **B.** While it's difficult to search for a conjunction of two features for example "a big green bar" in panel **C.** Panel **D.** shows the performance of humans on these task where y-axis represents the time taken to find the target object (RT) and x-axis represents the number of distractor items in the display (C - Color; S - Size; S x C = Size and Color conjunction).

these types of search tasks were called "serial search" or "conjunction search". This class of visual search experiments played a significant role in the development of Feature Integration Theory [23] (see **Section 2.2.1** for details).

2.1.2 Asymmetry in visual search

Another important class of visual search experiments are the Visual Search Asymmetries. A search asymmetry is said to occur when searching for object A amidst other objects B is substantially easier than searching for object B amongst multiple objects A. This phenomenon is interesting because, unlike the feature conjunction search here, the distinguishable feature in both search conditions is the same, and despite that, one condition is easier than the other. A simple explanation like searching for basic features or conjunction of those basic features could not explain these asymmetry results. This suggests that the difficulty of visual search not only depends on the difference of features between the target and distractor but also on a specific feature that constitutes making the target object. The guided search

model [25] is an influential psychological model that explains this phenomenon (see **Section 2.2.2** for details).

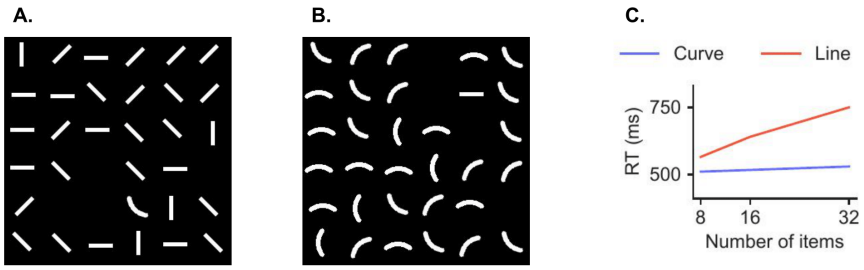


Figure 2.3: Example of visual search asymmetry. It's easy to search for curve among lines (**A**) as compared to the reverse case (**B**). Panel **C.** shows the performance of humans on these task where y-axis represents the time taken to find the target object (RT) and x-axis represents the number of distractor items in the display.

2.2 Models of visual search

This section will shed some light on the two most popular models of visual search in psychology, i.e. Feature Integration Theory and Guided Search Model. Both models were mostly developed based on reaction time (RT) studies from various visual search experiments. RT is defined as the time required by a subject to decide whether the target is present or not present in the shown visual scene (or search image). Mostly psychologist looks at the slope of the RT vs the numbers of distractor objects in the search image as the metric for difficulty while the intercept is considered as bias introduced because of motor response time to perform the experiment, which is not much due to the cognitive process involved while performing a visual search. The higher the slope more difficult the search task is considered. The most common approach in developing these theoretical models is by varying the conditions of the search task and analysing how those changes affect the difficulty of the tasks.

2.2.1 Feature Integration Theory

Feature Integration Theory (FIT) was introduced by Treisman in 1980 [23], and became one of the most influential models of visual attention. According to FIT,

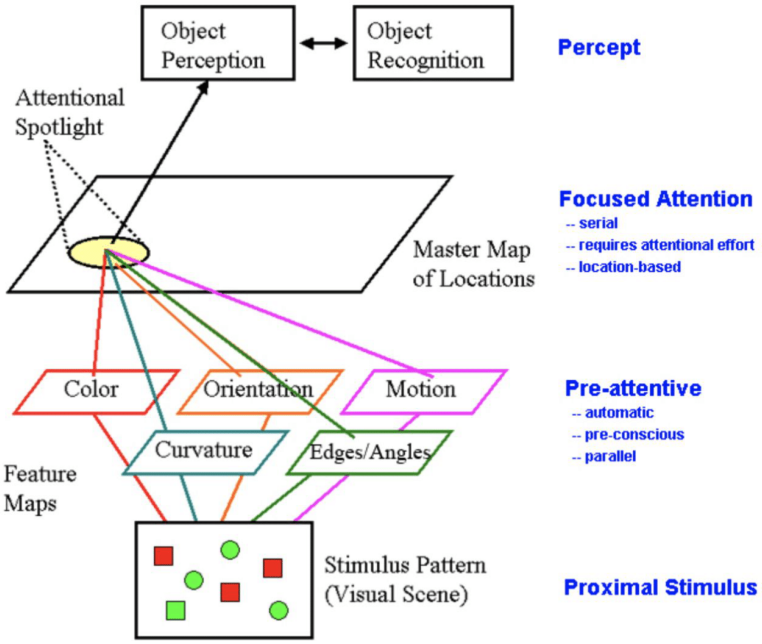


Figure 2.4: Architecture of Feature Integration Theory (Source: https://psyc2016.whatanimalssee.com/visualSearch.html | Date: 30 September, 2021)

the deployment of attention is processed in two stages. In the first stage, i.e. pre-attentive stage, the information about basic features such as colour, shape, motion, depth are collected automatically and parallelly. Parallel means all the features across the whole image are identified simultaneously. The complete objects are identified in the second stage of processing, combining the basic features from the first stage to define the object. This combining of features is done serially. This theory perfectly explains the feature conjunction search, i.e. when the target object can be distinguished only based on the pre-attentive stage, the performance will be dramatically fast, yielding almost a constant zero slope of RT vs Items because, in the pre-attentive stage, all the feature maps are processed simultaneously in parallel. On the other hand, since the conjunction of features will need multiple feature maps from the pre-attentive stage to integrate during the focused attention stage, the performance will not be parallel, and RT will increase with the number of distractors.

2.2.2 Guided Search model

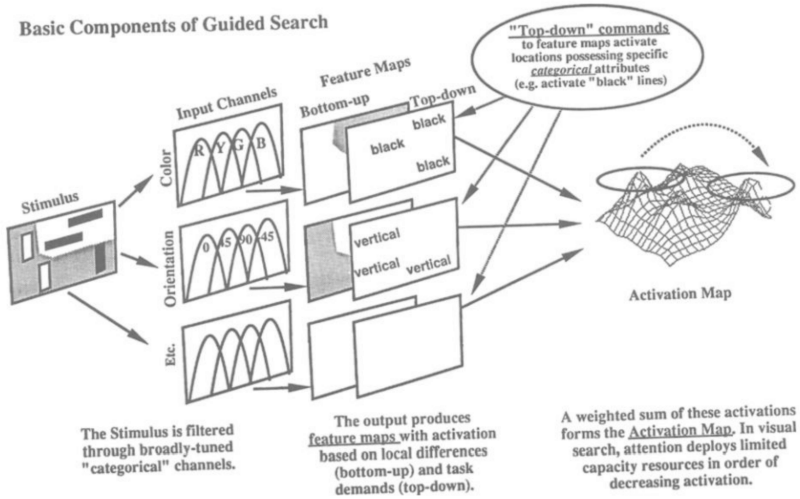


Figure 2.5: Architecture of Guided Search model (Source: [26])

Jeremy Wolfe in 1989 introduced the Guided Search (GS) model, and since the publication of the initial GS model, it has gone through several major revisions [27, 26, 25]. Similar to FIT, the GS model is also involved in two-stage attentional processing. But unlike FIT, GS does not suggest a simple strategy of searching by combining multiple specific features of the target object in the given search image. In GS, the first step is quite similar to FIT, in which input stimulus is filtered by basic "categorical" channels, this step is here termed pre-attentive processing. Generally, these channels are responsible for filtering basic features like colour, orientation, etc. In GS, attention is defined in terms of activations/ priority maps, which defines the priority for attending at a given location. These activation maps/ guidance are based on various factors. The two most important factors that have been introduced since the early versions of the GS model are the bottom-up activation, i.e. stimulus-driven guidance and the top-down activation, i.e. feature-based guidance. In bottom-up activation, the activation maps are generated based on the saliency of the item in the scene. Saliency is a measure of how unusual the object at a given position is, in respect to its surrounding. Clearly, this type of activation does not depend on the

demands of the task, i.e. the information of the target object is not needed, and guidance is purely based on the statistics of the given scene. In top-down activation, guidance is based on features from the target object. In this case, those locations which contain features similar to the features in the target object gets a higher priority. We do not know what kind of features are capable of guiding attention during the top-down guidance, but basic features like colour, orientations, etc are undoubtedly guiding features. An exhaustive list of guiding/ non-guiding features can be found in (Box 1 of [1]). The final activation map is a linear combination of all top-down activations and bottom-up activation and can be represented by an equation. It is possible to answer search asymmetry using the guided search model because in the guided search model, the top-down activation map will not only depend on the feature difference between the target and distractor object but also on what guiding attributes are present in the target object.

2.3 Visual representation in the brain and machines

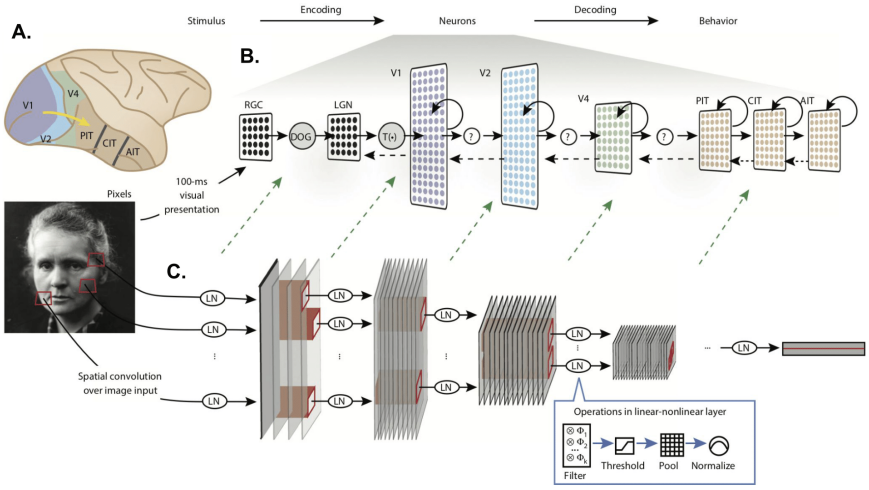


Figure 2.6: Structure of visual pathway in the brain and machine. **A.** shows the anatomical brain regions. **B.** shows the model ventral visual stream. **C.** shows an arbitrary convolutional neural network architectures (Source [28])

2.3.1 The brain, the eye and the ventral stream

In this section, our aim is not to fully explain the visual information processing in the brain but to give a brief overview of the brain's visual circuitry, and furthermore, to make the reader realise how the proposed model of visual search is very much biologically plausible. Note that the biological plausibility of the model is essential because along with predicting the visual search behaviour, we also want our model to align well with neurophysiological studies.

The human brain consists of roughly 100 billion neurons which form the basic building blocks of information processing in the brain. Extensive neuropsychological studies have suggested that different anatomical regions are responsible for different kinds of cognitive functioning (**Figure 2.6**). The occipital lobe of the primary cerebral cortex of the brain is the most dominant region for vision. The human visual system starts with the eyes. The light first enters through the eyes and falls on the retina, and excites the photoreceptor neurons. Then the signals from the photoreceptors reach the retinal ganglion cells (RGCs) via intermediate neuron cells (horizontal, bipolar, and amacrine neurons). Most of the information from the RGCs passes down to the lateral geniculate nucleus (LGN) in the thalamus. LGN then transmits the visual signals to the first visual area in the visual cortex called the "primary visual cortex (or V1). The neurons in V1 mainly respond to low-level features like edges, colours, orientations and directions. The output from V1 separates into two different pathways, popularly known as "what" and "where" pathways. The "where" pathway is called the dorsal stream, which mostly processes object motion and spatial location, whereas the "what" pathway is called the ventral stream is primarily involved in object recognition and forms some complex visual representations of the visual scene in front of the eye. Since the ventral stream is mainly responsible for object recognition and in the processing of a high level of visual representations, we will mostly focus on the properties of the ventral stream. The ventral stream forms sequential and hierarchical stages of visual processing. The information from V1 passes down to V2 and then to V4 and IT along the ventral

stream. Each of these cortical regions integrates information from its previous region to form a more complex and informative representation. As a result, these higher regions respond to more complex stimuli, unlike the V1 cortex, which primarily responds to edges or colours.

Another essential property of the ventral stream that plays a vital role in deploying attentional priorities is the decrease of visual acuity from the fovea to the peripheral region. The human visual system has small receptive field sizes in the foveal part, and these receptive field sizes increase with eccentricity within a given visual area [29]. Due to these differences in visual acuity, humans move their eyes to attend to a specific location in a visual scene. In addition, receptive field sizes also increase from one brain area to the next along the visual hierarchy of the ventral stream [29].

2.3.2 The machines: Artificial Neural Networks

Even though there are vast numbers of computational models for extracting high levels of visual representations and modelling the visual cortex, we will mainly focus on Artificial Neural Networks (ANNs). There are a couple of valid reasons to do so:

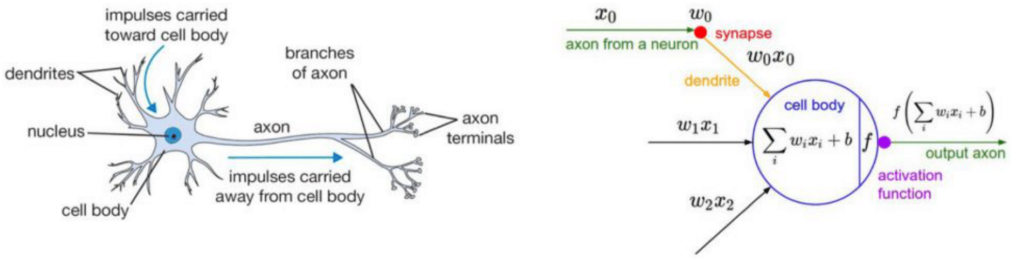


Figure 2.7: Schematics of Artificial Neuron (Source: https://cs231n.github.io/ | Date: 30 September, 2021)

1. The most fundamental building blocks of ANNs is inspired by the biological neurons (**Figure 2.7**) and are often called artificial neurons. The most noticeable difference between artificial neurons and biological neurons is that

artificial neurons communicate in terms of "real values", while on the other side biological neurons communicate using spikes. However, note that most of the information in biological neurons is encoded in the neuronal spiking frequency rather than the exact time of those spikes. Now, if we make an analogy of those "real values" of artificial neurons with spike frequencies in biological neurons, it is fair to consider artificial neurons as a simplified model of biological neurons [30].

2. CNNs, which are the class of ANNs consisting of convolutional layers, gives a fair analogy to the simple and complex cells found in the V1 cortex of the ventral stream. Past studies have shown that, similar to the V1 cortex, the early layers of CNNs also respond mainly to simple features like edges and colours.

3. Like the ventral stream, Deep CNN architecture also goes through sequential and hierarchical processing stages.

4. It has been found that the representations extracted from intermediate layers of Deep CNNs correlate with the neural recordings from the macaque brain. They were also found to correlate with various other forms of non-invasive recordings recorded from the human brain [31, 32, 33]. Some studies in cognitive psychology and deep learning suggest that these Deep CNNs can also capture hidden psychological representations similar to humans [34, 35].

5. Last but not least, state of the art ANN models have broken the record of most of the other forms of computer vision models on several types of perceptual tasks, often attaining better or close to human performance [36, 37, 38, 39].

Chapter 3

Modeling Visual Search

The proposed model is build upon the theories from features integration theories, guided search model, and invariant visual search network [23, 25, 16]. Two most essential components that are responsible in guiding attention in visual search are: the bottom-up stimulus driven guidance and the top-down target feature modulated guidance. For building each of these component we stuck to one unique model of the human ventral visual cortex to extract-eccentricity-dependent visual features from the target and search images. So, this makes our bottom-up and top-down model an integrated model of visual attention unlike any other previous models where only either one of these features were used. For the model of the visual cortex, we used a pre-trained deep convolutional neural network and introduced eccentricity dependent sampling to make it more similar to neurophysilogy.

The model is schematically illustrated in **Figure 3.1**. The model follows the similar stages of visual search that a human might carry out in a search task. The model takes two inputs: a target image (I_t, image of the object to search) and a search image (I_s, image where the target object is embedded amidst distractors). In the initial conditions, the model fixates on the center of the search image. At each fixation n, the model calculate a bottom-up saliency map (S_n) and a top-down attention map (A_n). A weighted linear combination of these two maps results in an overall attention map (O_n). A winner-take-all mechanism selects the maximum of the overall attention map O_n as the location for the $n+1$-th fixation. This process

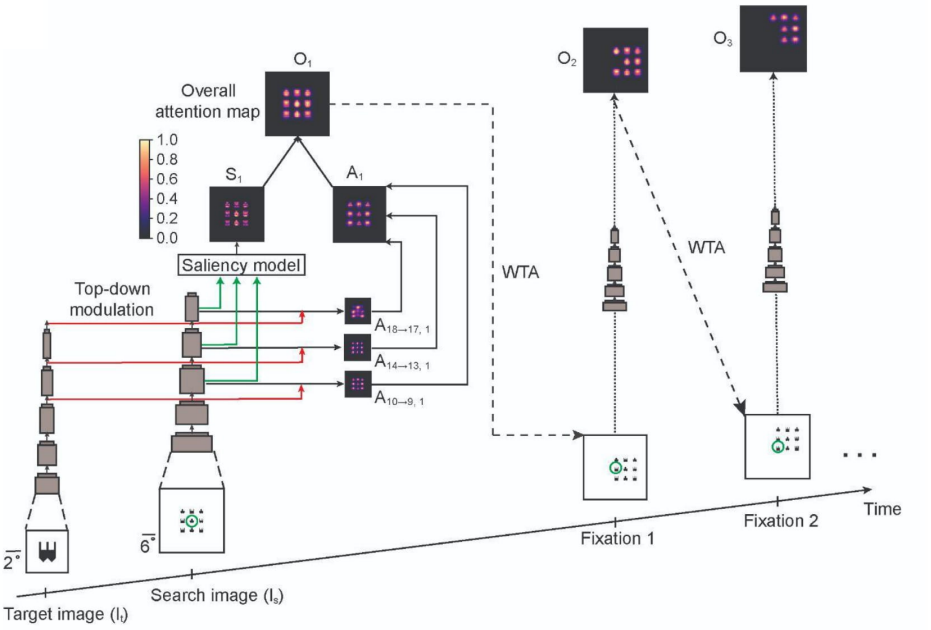


Figure 3.1: Schematic of the computational model of visual search after integrating various components.

iterates until the model finds the target with a total of N fixations. The model has an infinite inhibition of return and therefore does not revisit previous locations. Humans do *not* have perfect memory and do re-visit previously fixated locations, for a more extensive quantification of such return fixations, see [40]. However, for most of the experiments considered here, the total number of fixations is small and therefore the number of return fixations would also be small.

After a fixation at a given location, the model needs to verify whether the target is present at that location or not. Since the focus here is on visual search, the model bypasses this recognition step by using an "oracle" recognition system. The oracle checks whether the selected fixation falls within the ground truth target location, defined as the bounding box of the target object. The bounding box is defined as the smallest square encompassing all pixels of the object. In the experiments discussed here, the target is always present and therefore the model will always eventually find the target. To evaluate the effect of the object recognition step, we considered a basic recognition system in which a cosine similarity score is calculated between the features of the target object and the object pointed by the current

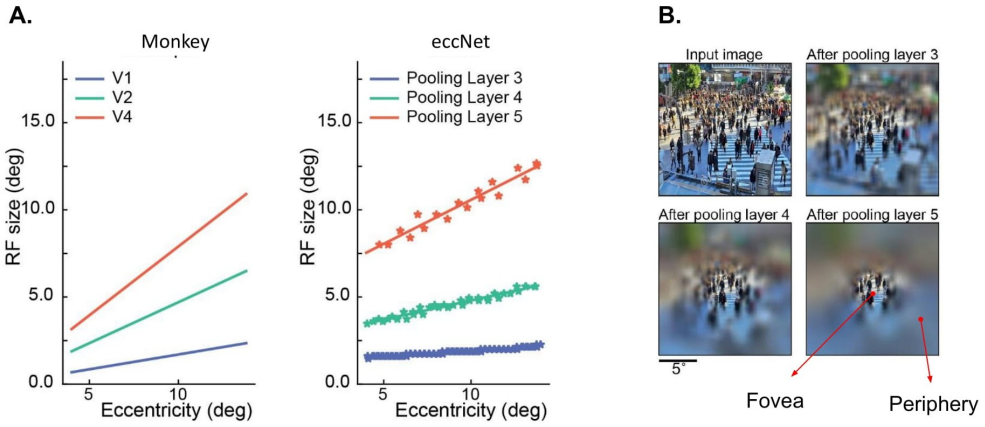


Figure 3.2: Eccentricity-dependent sampling in visual cortex. A. Eccentricity-dependent sampling leads to increasing receptive field sizes as a function of eccentricity for the macaque visual cortex (left, Freemanet al.2011) and the proposed visual cortex model (right). **B.** Example image and illustration of visual acuity as a function of eccentricity at pooling layers 3 through 5 of the proposed model.

fixation. And based on a threshold value, the model decides whether the fixated object is the target or not. It was observed that by manually tuning the threshold value, similar results to the 'oracle' method could be achieved. This suggests that a diffusion-decision model can be used to estimate the right threshold value in an online fashion without any manual tuning. This is out of the scope of this thesis and is proposed as future work in (**Appendice A1**).

3.1 Eccentricity-dependent model of visual cortex

Before building an integrated architecture of visual search, we must have a computational model of the visual cortex which is responsible for transforming the raw-pixels of the images to a higher dimensional feature space. To be able to predict the correct set of human behaviors, this model must have computational similarity with the computation in the brain's visual cortex. And finally, this model should be responsible to act as the backbone for extracting features for different top-down or bottom-up models of visual search. As described in the introduction, we considered focusing on Deep Convolutional Neural Networks (DNNs) because of their high level

similarity with the brain's ventral stream and past studies showing its capabilities in capturing neurophysiological and psychological features. But unlike the current Deep-CNNs model which has a uniform pooling operation across all the space of the image, leading to a uniform size of the receptive field across all eccentricities, for the visual cortical neurons, receptive field sizes increase from one brain area to the next along the visual hierarchy (**Figure 3.1A**, left). This increase is captured by Deep-CNNs through pooling operations. In addition, receptive field sizes also increase with eccentricity *within* a given visual area (summarized in [29]; **Figure 3.1A**, left). So, the proposed model introduces eccentricity-dependent pooling layers that brings similar eccentircity-dependent sampling in these Deep-CNNs as it is found in the brain **Figure 3.1A**, right). Since there isn't any study which accurately maps this dependence in the human brain, we compared this with eccentricity-dependence in a macaque's brain, whose brain architecture is believed to very much resemble that of humans.

We first define notations used in standard average pooling layers in the deep learning literature ([41]). For clarity and simplicity, we described the model components in the units of pixels; and then we provid a scaling factor to convert pixels to degrees of visual angle in the end of this section. In an average-pooling operation at layer l of VGG16 (the layer numbers used here follow the original definitions in the VGG16 architecture excluding the activation layers), unit j in layer $l+1$ takes the average of all input units i in the previous layer l within its local receptive field of size r_{l+1}. Its activation value y is given by:

$$y_{l+1}^{j} = \frac{1}{r_{l+1}} \sum_{i=0}^{r_{l+1}} y_{l}^{i} \tag{3.1}$$

Note that traditionally most VGG16 architecture uses max-pooling instead of average pooling but we considered to use the average-pooling operation instead because of the eccentricity dependence, which brings a significant increase in pooling window size as the eccentricity increases. Note that for very large window sizes, max

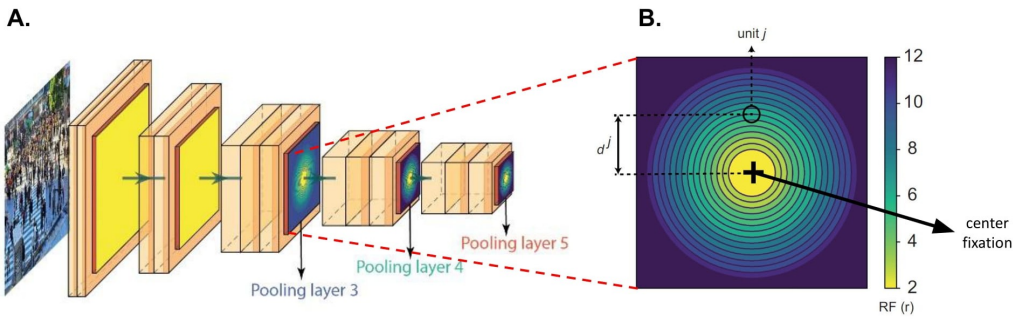


Figure 3.3: Eccentricity-dependent sampling in Deep-CNNs. A. Proposed ecc-NET architecture for visual cortex showing the eccentricity-dependent pooling layers in original VGG16 model. **B.** llustration of eccentricity-dependent pooling layer l, shows the receptive field sizes (rjl,n) for each unit j with distance dj from the centre (the color bar here denotes the size of the pooling window in pixels).

pooling will simply destroy all the information stored in the higher eccentricity region but average pooling layer will retain more information in the average of the region inside the pooling window.

In the eccentricity-dependent operation, the receptive field size $r^j_{l+1,n}$ of input unit j in layer $l+1$ is a linear function of the Euclidean distance $d^j_{l+1,n}$ between input unit j and the current fixation location (fixation number n) on layer $l+1$. The further away the unit j is from the current fixation, the larger the receptive field size (**Figure 3.3B**; in this figure, the fixation location happens to be at the center of the image). Therefore, the resolution is highest in the fixation location and decreases in peripheral regions.

$$r^j_{l+1,n} = \begin{cases} \lfloor \eta_{l+1}\gamma_{l+1}(d^j_{l+1,n}/\eta_{l+1} - \delta) + 2.5px \rfloor, & \text{if } d^j_{l+1,n}/\eta_{l+1} > \delta \\ 2px, & \text{if } d^j_{l+1,n}/\eta_{l+1} < \delta \end{cases} \tag{3.2}$$

The floor function $\lfloor \cdot \rfloor$ rounds down the decimal pool window sizes to the nearest integers. The variable γ_{l+1} is a positive scaling factor for layer $l+1$, defining how fast the receptive field size of unit j expands with respect to its distance from the fixation

at layer $l+1$. Based on the slope of eccentricity versus receptive field size in the macaque visual cortex ([29]), we experimentally set $\gamma_3 = 0.00$, $\gamma_6 = 0.00$, $\gamma_{10} = 0.14$, $\gamma_{14} = 0.32$, and $\gamma_{18} = 0.64$ (see **Figure 3.2A** for the slopes of eccentricity versus receptive field sizes over pooling layers). We define $\delta = 4$ dva as the fovea size. For those units within fovea sizes, we set a constant receptive field size of 2 pixels. Here η_{l+1} is a positive scaling factor which converts the dva of the input image to the pixel units at the layer l. We map the receptive field sizes in units of pixels to units of degrees of visual angle (dva) using 30 pixels/dva. This basically defines the number of pixels in one degree. This value indirectly represents the clarity of vision for our computational model. More the number of pixels in one degree better the clarity of vision. Note that due to computational limitations, we have restrictions on how much clearer vision we can use. Since we have a stride of 2 pixels at each pooling layer, the mapping parameter η from pixel to dva decreases over layers: $\eta_3 = (30/2)$ pixels/dva, $\eta_6 = (30/4)$ pixels/dva, $\eta_{10} = (30/8)$ pixels/dva, $\eta_{14} = (30/16)$ pixels/dva, and $\eta_{18} = (30/32)$ pixels/dva. To achieve better downsampling outcomes, the average-pooling operation also includes the stride ([41]) defining the movement of downsampling location. We empirically set a constant stride to be 2 pixels for all eccentricity-dependent pooling layers.

A visualization for $r^j_{l,n}$ is shown in **Figure 3.3B**; where different colors denote how the receptive field sizes expand within a pooling layer. We illustrate the change in acuity at different pooling layers in **Figure 3.2B**.

These customized eccentricity-dependent pooling layers can be easily integrated into other state-of-the-art object recognition deep neural networks. All the computational steps are differentiable and can be trained end-to-end with other layers. However, since we are interested in testing the generalization of our model from object recognition to visual search, we do not retrain the model and instead use the weights of the original VGG16 architecture. Because there is no training, one might think that the resulting network might show lower performance in object recognition than the original VGG16 results but we argue that the ”main” feature extraction

part of the model remains fairly robust against these changes. Note that generally most Deep-CNNs architectures have two sets of layers, some initial groups of convolutional layers followed by some some linear classification layers. It's general hypothesis that the convolutional layer captures the feature representations and the linear layers use those extracted features to learn the classification model to perform object recognition. So to test whether these changes affects the object recognition performance we freeze the convolutional layers and only trained the top-classifcation layers. We observe that even after bringing in these changes, the model showed performance close to that of the original VGG16 model.

3.2 Bottom-up saliency model

The bottom-up saliency model is based on the information maximization approach (**Figure 3.4**). This method has been previously shown to be effective to find salient regions in an image ([42]). The original implementation used a representation based on independent component analysis. Instead, here we use the feature maps extracted from the computational model of the visual cortex (eccNET). Since the feature maps of eccNET change based on fixation locations, the bottom-up saliency is recomputed at every fixation step. This is different from the existing bottom-up saliency prediction literature where most models take the entire image as input and computation of these saliency maps does not depend on the current fixation. Considering that for the top-down model (see next section), the model already need to compute the features from eccNET, the additional computation of finding the information gain for saliency estimation at each fixation step is negligible. At layer l of eccNET, we extracted feature maps of size $C_l \times H_l \times W_l$, where C_l is the number of channels. and H_l, W_l denote the height and width, respectively. On the c^{th} channel of the feature maps, we define the histogram function $F_{l,c,n}(\cdot)$, which takes the activation values $y^j_{l,c,n}$ as inputs and outputs its corresponding frequency among all individual units j at all $H_l \times W_l$ locations at the n^{th} fixation.

Next, the model calculates the probability distribution for each unit j on the c^{th}

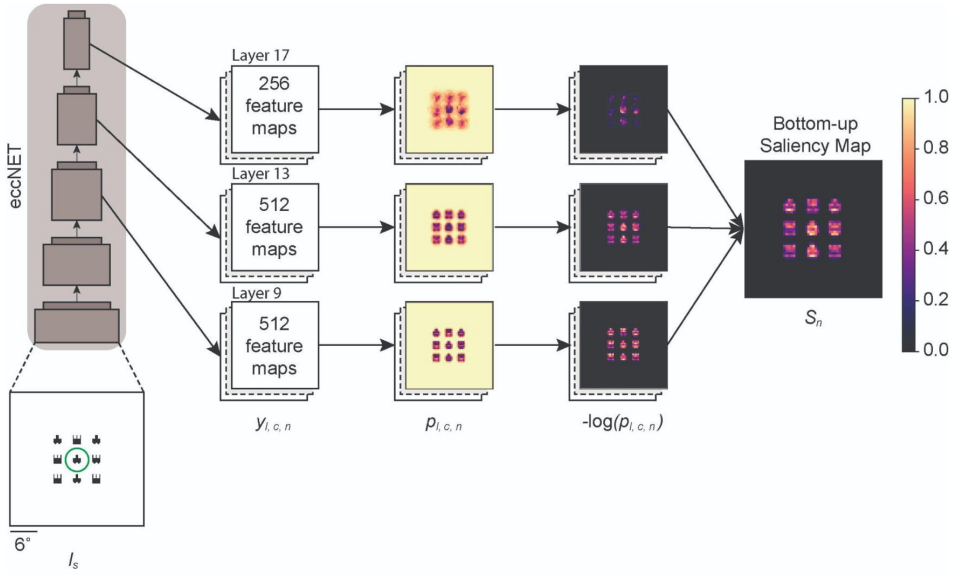


Figure 3.4: Computational model of bottom-up visual saliency maps. At each fixation n, the saliency model extract numbers of feature maps ($y_{l,C,n}$) at layer l with total number of C channels from the visual cortex model (eccNET) and then estimate the probability distribution for individual channel of the feature maps ($p_{l,c,n}$). Then it calculates the self information ($A_{l,c,n}$ =-log($p_{l,c,n}$)), normalizes to [0,1], and adds them to compute the overall salience map (S_n). Heatmaps show example visualization of $p_{l,c,n}$ and $A_{l,c,n}$. See scale bars on the right for activation values on these maps.

feature map at layer l and n^{th} fixation:

$$p^j_{l,c,n} = \frac{F_{l,c,n}(y^j_{l,c,n})}{\sum_{i=0,1,\ldots,W_l \times H_l} F_{l,c,n}(y^i_{l,c,n})} \tag{3.3}$$

where $p^j_{l,c,n}$ denotes how prevalent the activation value $y^j_{l,c,n}$ is over all units j on the c^{th} channel feature map. To capture attention drawn to less frequent visual features on an image, the model uses the normalized negative log probability to compute a saliency map for each channel and then averages the saliency maps over all channels and then over all selected layers $l = 9, 13, 17$ to output the overall saliency map S_n at the nth fixation:

$$S_n = \sum_{l=9,13,17} \sum_{c}^{C_l} \frac{-\log(p^j_{l,c,n})}{p_{max} - p_{min}} \tag{3.4}$$

where:

$$\begin{cases} p_{min} = \min(\{-\log(p^i_{l,c,n}) : i = 1, 2, ..., H_l \times W_l\}) \\ p_{max} = \max(\{-\log(p^i_{l,c,n}) : i = 1, 2, ..., H_l \times W_l\}) \end{cases} \tag{3.5}$$

where the normalization of negative log probability is carried out by taking the difference between the maximum and minimum negative log probability among all the individual units i in the cth channel at layer l. Since not all feature maps at the selected layers are of the same size, we downsampled individual saliency maps in the lower layers $l = 9, 13$ to be of the same size as those at layer $l = 17$.

3.3 Top-down modulated activation maps

The top-down modulation is inspired by the IVSN model ([16], **Figure 3.5**). But we brought in several notable changes to the IVSN model:

1. The base feedforward deep neural network architecture for the visual cortex used in IVSN was VGG16 ([43]), which has uniform receptive field sizes throughout the image. In stark contrast, visual cortex shows strong eccentricity-dependent receptive field sizes. Here we replaced the VGG16 architecture with an eccentricity-dependent model of visual cortex, described above. We refer to the new model as eccNET.

2. Both IVSN and eccNET require computation of an overall attention map in order to decide where to fixate next. Given the uniform sampling in IVSN, the attention map was computed only once and did not change from one fixation to the next, except for inhibition of previously visited locations. Since eccNET outputs different feature maps depending on the fixation location, the top-down modulation attention map also is a function of fixation n.

3. The top-down modulation in IVSN happens only in a single layer (the top layer in the default version of the model). Here the model combines top-down modulated features across multiple layers.

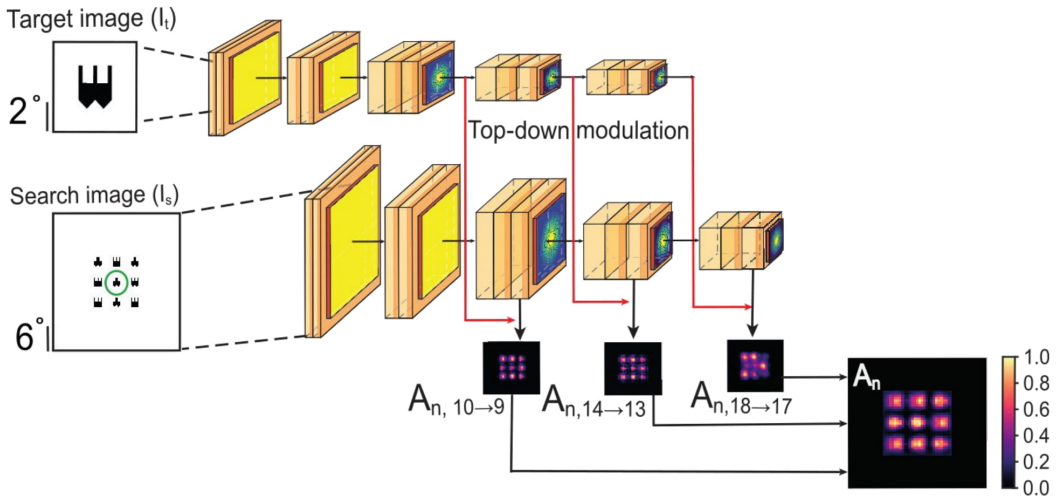


Figure 3.5: Computational model of top-down feature dependent activations. The model takes as input a target image (Im) and a search image (Is) both of which are processed through eccNET. The feature extracted from target image applies a top-down modulation on the features in search image at three different layers producing three different top-down activation maps. These three individual top-down activations are linearly combined to finally produce an overall top-down attention map (An).

Both the target image (I_t) and search image (I_s) are passed through the same model mimicking the extraction of features in the visual cortex. Given the current n^{th} fixation location, the model generates feature maps $\phi^t_{l+1,n}$ at layer $l+1$ for the target image I_t. Correspondingly, the model generates $\phi^s_{l,n}$ in response to the search image I_s. We define the top-down modulation map $A_{l+1\to l,n}$ as:

$$A_{l+1\to l,n} = m(\phi^t_{l+1,n}, \phi^s_{l,n}) \tag{3.6}$$

where $m(\cdot)$ is the target modulation function defined as a 2D convolution with $\phi^t_{l+1,n}$ as convolution kernel operating on the search image feature map $\phi^s_{l,n}$. Note that the layer $l+1$ modulates the activity of layer l. In IVSN, the top-down modulation occurred only in one layer (in the default version, using the top layer). In contrast, here the model takes the weighted linear combination of normalized top-down modulation maps across multiple layers ($l = 9, 13, 17$) to compute the overall top-down modulation map A_n. Since the top-down modulation maps at different layers are of

different sizes, we resize all the top-down modulation maps $A_{10\to 9,n}$ and $A_{14\to 13,n}$ to be of the same size as $A_{18\to 17,n}$.

$$A_n = \sum_{l=9,13,17} w_{l,n} \frac{(A_{l+1\to l,n} - \min A_{l+1\to l,n})}{(\max A_{l+1\to l,n} - \min A_{l+1\to l,n})} \tag{3.7}$$

where $w_{l,n}$ are weight factors that determine how strong the top-down modulation at the lth layer contributes to the overall top-down modulation attention map A_n. These weights $w_{l,n}$ were calculated during the individual search trials using the maximum activation value obtained from each individual top-down modulation map:

$$w_{l,n} = \frac{\max A_{l+1\to l,n}}{\sum_{i=9,13,17} \max A_{i+1\to i,n}} \tag{3.8}$$

In Experiment G (**Figure 4.7**), subjects might employ a unique search strategy where the target features, such as color and orientation, might be weighted equally in top-down modulation. Thus, we empirically set $w_{9,n} = w_{13,n} = w_{17,n} = 1/3$ for this experiment.

3.4 Integration of top-down and bottom-up activations

Given the overall saliency map S_n and the overall top-down activation map A_n at the nth fixation (see sections above for computation of both maps), we normalize both maps within [0,1] and compute the overall attention map O_n as a weighted linear combination of both maps. $w_{S,n}$ and $w_{A,n}$ denotes the weights applied on the bottom-up saliency map S_n and the top-down modulation map A_n respectively. Previous work suggests that bottom-up saliency plays a more prominent role at the beginning of a trial, before full top-down attention takes place. Also, based on the demands of the task, the relative contribution of bottom-up and top-down can be captured for three different task categories ([44, 45]). Based on these task categories, the model implements three possible schemes:

Scheme 1: This scheme belong to those category of tasks in which search can be benefited more from the bottom-up attention as compared to the top-down attention.

Scheme 2: This scheme belong to those category of tasks in which bottom-up attention and top-down attention both play important roles in finding the target.

Scheme 3: This scheme belong to those category of tasks in which search is penalized more from the bottom-up attention as compared to the top-down attention.

Note that all these three schemes affect the decision bias only at the first and second fixation ($n = 1, 2$) in each individual trial. For the subsequent fixations ($n > 2$), we argue that humans are strongly guided by top-down modulation effect with minimal bottom-up effect; that is, $w_{S,n} = 0$ and $w_{A,n} = 1$ for all $n > 2$ regardless of the nature of visual search experiments. We formulate the computation of overall attention map as follows:

$$O_n = w_{S,n} S_n + w_{A,n} A_n \tag{3.9}$$

where

$$\begin{cases} w_{S,n} = 0, w_{A,n} = 1 & \text{if scheme (1) and } n = 1 \\ w_{S,n} = 0.5, w_{A,n} = 0.5 & \text{if scheme (2) and } n = 1 \\ w_{S,n} = 1, w_{A,n} = 0 & \text{if scheme (3) and } n = 1 \\ w_{S,n} = 0, w_{A,n} = 1 & \text{if scheme (1) and } n = 2 \\ w_{S,n} = 0.37, w_{A,n} = 0.63 & \text{if scheme (2) and } n = 2 \\ w_{S,n} = 0.37, w_{A,n} = 0.63 & \text{if scheme (3) and } n = 2 \\ w_{S,n} = 0, w_{A,n} = 1 & \text{if } n > 2 \end{cases} \tag{3.10}$$

3.5 Fixation to reaction time

The proposed computational model of visual search predicts a series of eye-fixations. The psychophysics experiments 1-9 did not measure eye movements and instead report a *key press reaction time* (RT) whereby subjects presumably found the target.

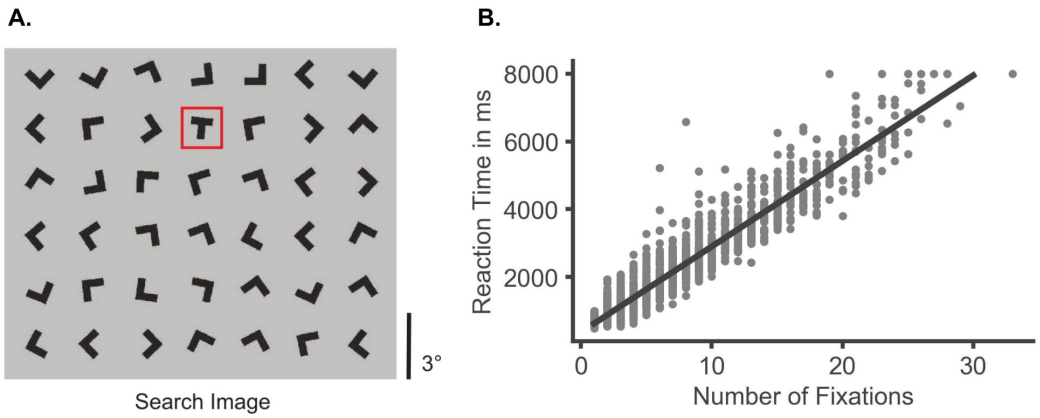


Figure 3.6: Experiment to model reaction time from number of fixations. A. Example from the T vs L visual search task used to evaluate the relationship between reaction times and number of fixations. **B.** Reaction time grows linearly with the number of fixations. Each gray point represents a trial. A line was fit to these data: $R(ms) = \alpha * n + \beta$. A fit using linear least square regression gave $\alpha = 252.359$ ms/fixation and $\beta = 376.271$ ms ($r^2 = 0.90$, $p << 0.001$). This linear fit was used throughout the thesis to convert the number of fixations in the model to reaction time values in milliseconds for comparison with human data.

To relate the model output to RT, we used data from a separate experiment that measured both RT and eye movements (**Figure 3.6A**), see description under "Experiment to convert fixations to key press reaction times". We assume that the key press RT results from a combination of time taken by fixations plus a motor response time. Therefore to calculate key press reaction times in milliseconds from the number of fixations we used the linear fit in Equation 3.11. Here, RT = reaction time in milliseconds, N = number of fixations until the target was found, α = duration of a single saccade + fixation = constant, and β = motor response time = constant.

$$RT = \alpha * N + \beta \tag{3.11}$$

The value for constants α and β were estimated using the linear least-squares regression method on the data obtained from the experiment (**Figure 3.6B**): $\alpha = 252.36$ milliseconds and $\beta = 376.27$ milliseconds. The correlation coefficient was 0.95 ($p < 0.001$). Here we assume that both α and β are *independent of the actual experiment* and use the same constant values for all the figures (see Discussion).

Chapter 4

Visual Search Experiments

This work revisits several important and classical psychophysics experiments in visual search to compare the visual search behaviour predicted by the proposed models against humans'. Most of these experiments were studied in psychophysics to understand why one visual search is more complicated than the other and vice-versa. The task's difficulty in visual search literature is revealed by how long a subject takes to find a target object in a given search display with several distractor objects, often termed reaction time (RT). Usually, RT increases monotonically with the number of distractor items present in the display. Psychologists often compare the slope of the search time vs the number of distractor items as a measure of the difficulty for that specific task. The higher the slope more difficult the task is considered. We consider two well-established properties in the visual search literature for testing our model, the feature conjunction search and the search asymmetries. Within both of the experiment types, we selected several crucial experiments responsible for providing theoretical grounds of feature integration theory and the guided search model. Since most of these experiments have artificial search displays, we considered some additional experiments having natural images to test our model performance in a more natural setting. This chapter will provide the details of each experiment individually and the model performance. The combined results of all the experiments are discussed in the next chapter.

4.1 Visual search asymmetry

Visual search conditions were searching for an object A amidst other objects B is substantially easier than searching for objects B amongst multiple objects A ([46, 47, 48, 49, 50]) is termed as visual search asymmetry. For example, looking for a curved line embedded in a display with multiple straight lines is considerably easier than searching for a straight line in the middle of many curved lines. Search asymmetry is observed in a wide variety of visual search experiments and played a significant role in developing theories in the guided search model. Here we focus on six classical experiments ([46, 47, 48, 49]) to investigate the computational mechanisms that give rise to the emergence of such asymmetries:

4.1.1 Experiment 1: Curvature.

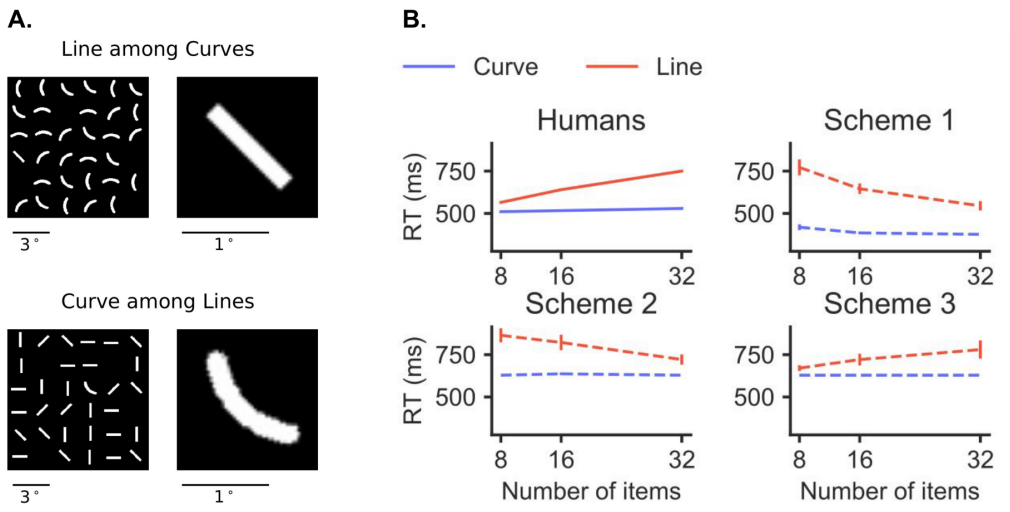


Figure 4.1: Stimuli and RT plots for Experiment 1: Curvature

This experiment is based on [46]. There were two conditions in this experiment: 1. Searching for a straight line among curved lines (**Figure 4.1A**, top) and 2. Searching for a curved line among straight lines (**Figure 4.1A**, bottom). The search image was 11.3 x 11.3 degrees of visual angle (dva). Straight lines were 1.2 dva long and 0.18 dva wide. Curved lines were obtained from an arc of a circle of

1 dva radius, the length of the segment was 1.3 dva, and the width was 0.18 dva. Targets and distractors were randomly placed in a 6 x 6 grid. Inside each of the grid cells, the objects were randomly shifted so that they did not necessarily get placed at the center of the grid cell. This ensures that the inter-object distance does not remain constant. The target and distractors were presented in any of the four orientations: -45, 0, 45, and 90 degrees. Three set sizes were used: 8, 16, and 32. There was a total of 90 experiment trials per condition, equally distributed among each of the set sizes.

Psychophysics study from [46] suggests that it is difficult to search for a straight line among curves as compared to the reverse case (**Figure 4.1B**, Humans). The proposed model showed similar behavior, well all the schemes showed difficulty in searching for straight lines with "Scheme 3" showing the best match to psychophysics data (**Figure 4.1B**). Indicating that it's the top-down modulation that plays an important role in driving asymmetry for search in curve vs line since in "Scheme 3" the relative weight for the bottom-up component is zero. While "Scheme 3" captured the human performance qualitatively, there were quantitative differences. The blue line slope for humans is 0.8, while for the model, it's 0.0, and the slope for the red line for humans is 7.6, but for the model, it's 4.43. It's important to note that the model did not do any specific parameter fitting to achieve the qualitative performance, unlike other computational modelling approaches.

4.1.2 Experiment 2: Lighting direction.

This experiment is based on [47]. There were two conditions in this experiment: 1. Searching for left-right luminance change among right-left luminance changes (**Figure 4.2A**, top). 2. Searching for top-down luminance change among down-top luminance changes (**Figure 4.2A**, bottom). The search image was 6.6 x 6.6 dva. The objects were circles with a radius of 1.04 dva. The luminance changes were brought upon by 16 different levels at an interval of 17 on a dynamic range of [0, 255]. The intensity value for the background was 27. Targets and distractors were

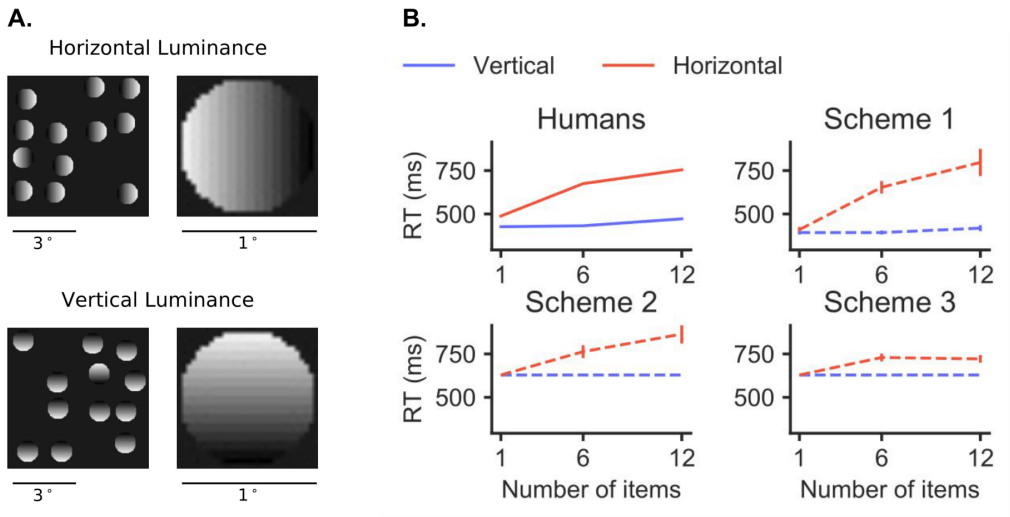


Figure 4.2: Stimuli and RT plots for Experiment 2: Lighting Direction

randomly placed in a 4 x 4 grid. Inside each of the grid cells, the objects were randomly shifted. Three set sizes were used: 1, 6, and 12. There was a total of 90 experiment trials per condition, equally distributed among each of the set sizes.

Psychophysics study from [47] suggests that it is difficult to search for horizontal luminance change as compared to the vertical luminance change condition (**Figure 4.2B**, Humans). All the schemes for the proposed model showed similar behaviour, with the "Scheme 2" showing the best match to psychophysics data (**Figure 4.2B**). Indicating that top-down and bottom-up modulation both play a somewhat equal role in driving asymmetry for search in curve vs line since in "Scheme 2", the relative weight for the bottom-up component is 0.5. For this experiment, the model matched the psychophysics data both qualitatively and quantitatively. The slope for humans is 4.2 and 23.9 for the blue line and red line, respectively. While for "Scheme 2", slopes are 0.0 and 21.3 respectively for the blue and red lines.

4.1.3 Experiments 3: Intersection I

This experiment is based on [48]. There were two different conditions: 1. Searching for a cross among non-crosses (**Figure 4.3A**, top). 2. Searching for a non-cross among crosses (**Figure 4.3A**, bottom). Each of the objects was enclosed in a square

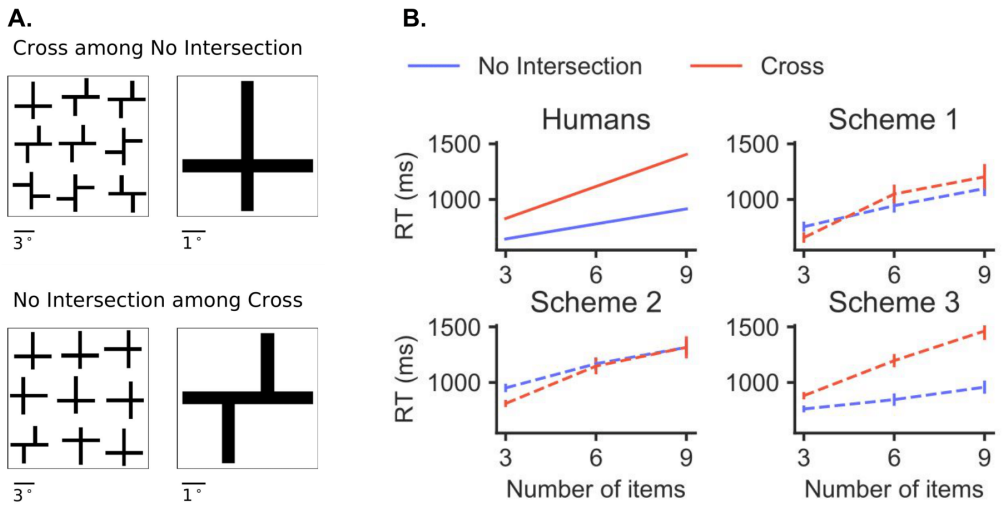


Figure 4.3: Stimuli and RT plots for Experiment 3: Intersection I

of size 5.5 x 5.5 dva. The width of the individual lines used to make the object was 0.55 dva. Non-cross objects were made from the same cross image by shifting one side of the horizontal line along the vertical. The search image spanned 20.5 x 20.5 dva. The objects were randomly placed in a 3 x 3 grid. Inside each of the grid cells, the objects were randomly shifted. The target and distractors were presented in any of the four orientations: 0, 90, 180, and 270 degrees. Three set sizes were used: 3, 6, and 9. There was a total of 108 experiment trials per condition, equally distributed among each of the set sizes.

Psychophysics study from [48] suggests that it is difficult to search for crosses among non-cross as compared to the reverse case (**Figure 4.3B**, Humans). In this experiment, we observed that "Scheme 3" showed the best match to psychophysics data (**Figure 4.3B**). Indicating that it is the top-down modulation that plays an important role in driving asymmetry for cross vs non-cross. Scheme 3 quantitative performance was also close to the humans, with the blue line slope being 45.0 vs 32.7 and the red line slope being 96.0 vs 96.97 for humans vs the model.

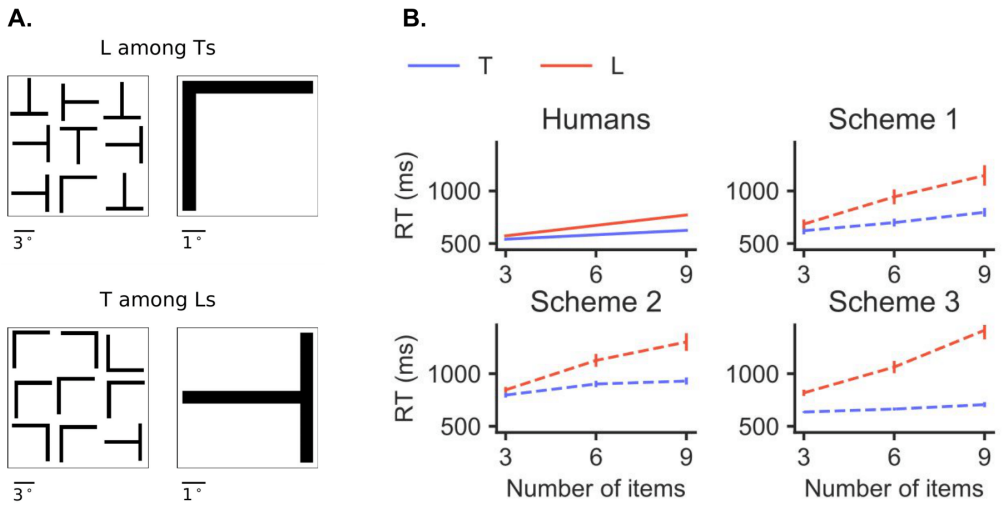


Figure 4.4: Stimuli and RT plots for Experiment 4: Intersection II

4.1.4 Experiments 4: Intersection II

This experiment is based on [48]. There were two different conditions: 1. Searching for an L among Ts (**Figure 4.4A**, top). 2. Searching for a T among Ls (**Figure 4.4A**, bottom). Each of the objects was enclosed in a square of size 5.5 x 5.5 dva. The width of the individual lines used to make the object was 0.55 dva. Non-cross objects were made from the same cross image by shifting one side of the horizontal line along the vertical. The search image spanned 20.5 x 20.5 dva. The objects were randomly placed in a 3 x 3 grid. Inside each of the grid cells, the objects were randomly shifted. The target and distractors were presented in any of the four orientations: 0, 90, 180, and 270 degrees. Three set sizes were used: 3, 6, and 9. There was a total of 108 experiment trials per condition, equally distributed among each of the set sizes.

Psychophysics study from [48] suggests that it is difficult to search for L among Ts as compared to the reverse case (**Figure 4.4B**, Humans). In this experiment also the model was able to capture the asymmetry property using all the proposed schemes, but the quantitative performance was not good. In this experiment, we observed that "Scheme 2" showed the best match to psychophysics data (**Figure**

4.3B), indicating that top-down and bottom-up modulations both play a somewhat equal role in driving asymmetry for search in curve vs line.

4.1.5 Experiments 5: Orientation I

This experiment is based on [49]. There were four different conditions: 1. Searching

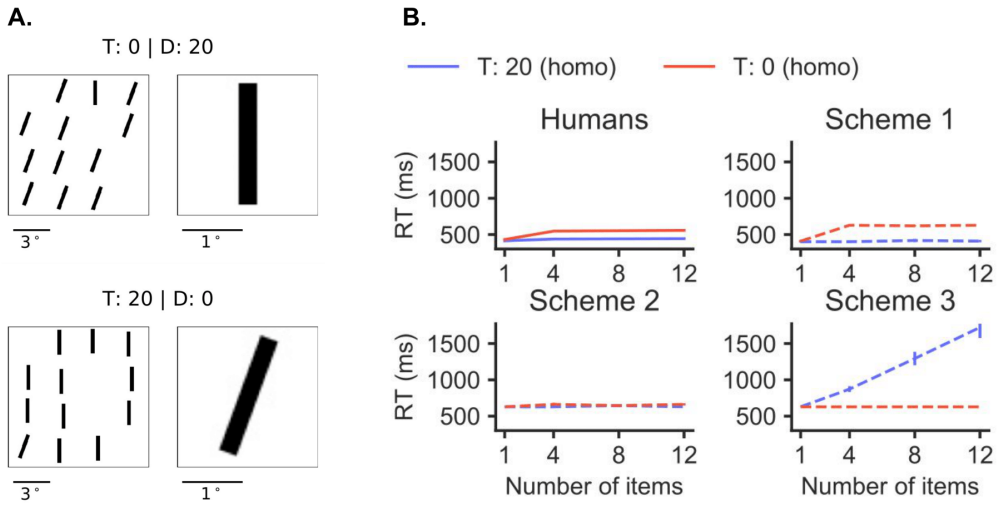


Figure 4.5: Stimuli and RT plots for Experiment 5: Orientation I

for a vertical straight line among 20-degrees-tilted lines (**Figure 4.5A**, top). 2. Searching for a 20-degree-tilted line among vertical straight lines (**Figure 4.5A**, bottom). Each of the objects was enclosed in a square of size 2.3 x 2.3 dva. The lines were of length 2 dva and width 0.3 dva. The search image spanned 11.3 x 11.3 dva. Targets and distractors were randomly placed in a 4 x 4 grid. Inside each of the grid cells, the objects were randomly shifted.

Psychophysics study from [49] suggests that it is difficult to search for vertical lines among tilted lines as compared to the reverse case (**Figure 4.5B**, Humans). In this experiment, we observed that "Scheme 1" showed the best match to psychophysics data (**Figure 4.3B**) while others were failing to replicate the human asymmetry pattern, indicating that it's the bottom-up modulation that plays a vital role in driving asymmetry in such case.

4.1.6 Experiments 6: Orientation II

This experiment is based on [49]. There were two different conditions: 1. Searching

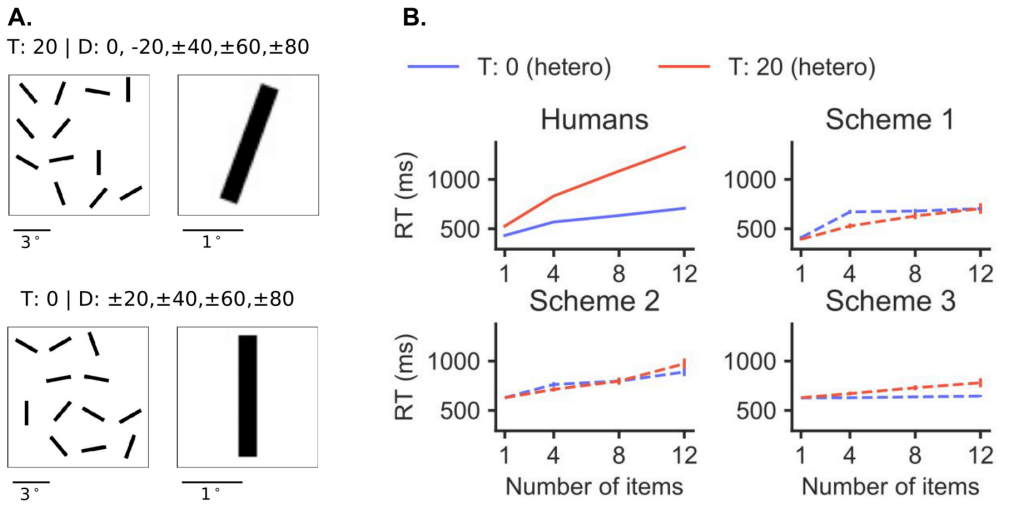


Figure 4.6: Stimuli and RT plots for Experiment 6: Orientation II

for a 20-degree tilted line among tilted lines of angles -80, -60, -40, -20, 0, 40, 60, 80 (**Figure 4.6A**, top). 2. Searching for a vertical straight line among tilted lines of angles -80, -60, -40, -20, 20, 40, 60, 80 (**Figure 4.6A**, bottom). Each of the objects was enclosed in a square of size 2.3 x 2.3 dva. The lines were of length 2 dva and width 0.3 dva. The search image spanned 11.3 x 11.3 dva. Targets and distractors were randomly placed in a 4 x 4 grid. Inside each of the grid cells, the objects were randomly shifted. Distractors were selected such that the proportions of individual distractor angles were equal. Four set sizes were used: 1, 4, 8, and 12. There was a total of 120 experiment trials per condition, equally distributed among each of the set sizes.

Unlike the previous experiments, in this experiment, the psychophysics study from [49] suggests that it is difficult to search for tilted lines line among verticals as compared to the reverse case (**Figure 4.5B**, Humans). The difference here is that the distractors are not homogenous and thus creating more trouble in search. In this experiment, none of the schemes performed similarly to humans. However,

the asymmetry was loosely captured by ”Scheme 3”. The observation makes sense if we compare this case with the previous experiment where the distractors were homogenous. Since in Experiment 5, the distractor is homogenous, there's more room for the target becoming more salient than the distractor and thus, 'scheme 1' shows a better match in that situation. But in this experiment, the saliency was destroyed because of heterogeneous distractors and therefore, ”Scheme 3” proving to be a better match.

4.2 Feature conjunction search

Another prominent example of drastic changes in the difficulty of visual search is the effect of feature conjunctions. Early studies of visual search noted that it is relatively easy to find objects in cases where a single feature distinguishes them from the distractors ([23]). For example, in a display consisting of diagonal blue lines, subjects can rapidly spot a diagonal green line. The target seems to "pop out," and the search times are almost independent of the number of distractors (here, the blue lines). In contrast, it is more challenging to detect a 45° green line in a display consisting of 45° blue lines and minus 45° green lines. In this case, distinguishing the target requires the conjunction of two features, orientation and colour; the search times are longer and increase substantially with the number of distractors. There are multiple examples of such conjunction effects. Here we focus on three classical experiments ([24, 51]) to investigate computational mechanisms underlying the decreased efficiency of a feature conjunction search.

4.2.1 Experiment 7: Conjunction search.

The stimuli in this experiment are based on the conjunction search experiments in [24]. We considered three features to test our model: color, orientation, and size. These features gave us a total of 6 different visual search conditions, three with singleton feature search and three with the conjunction of two features. And thus this makes three different experiment pairs for feature vs conjunction search expeirments, (1) Size and Orientation; (2) Color and Orientation; and (3) Size and Color (4.7). The tested colors were pink (RGB value: 180, 67, 149) and green (RGB value: 0, 140, 99).

The background color was white (RGB value: 255, 255, 255). The orientations were 45° and 135°. The sizes were 0.5 x 1.6 dva and 0.3 x 0.9 dva. The search image was 8.5 x 8.5 dva. Three set sizes were used: 4, 9, and 16. Objects were randomly placed in a 2 x 2 grid (4 objects), 3 x 3 grid (9 objects), or 4 x 4 grid (16 objects).

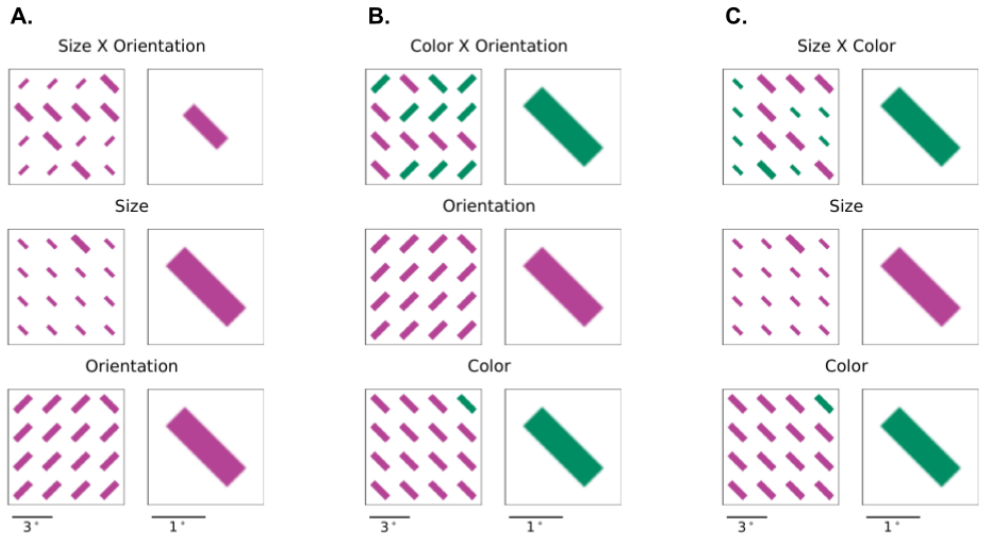


Figure 4.7: Stimuli for Experiment 7: Conjunction search

Objects were placed at the center of the grids. There was a total of 120 experiment trials per condition, equally distributed among each of the set sizes.

In these experiments, "Scheme 2" best captured the relative performance of humans in the three conditions (4.8, A, B, C), i.e. conjunction searches being difficult than the feature-based searches. Also, the absolute reaction time is quite comparable in orientation vs size conjunction (4.8, A) and "orientation vs color" conjunction (4.8, B). But the model performed poorly on matching the absolute reaction time in "size vs colour conjunction" (4.8, C). If we take a close look at "scheme 3" (having no bottom-up) for all three conditions, we can see the model shows some poor performance even on singleton features "size" and "colour". That suggests the

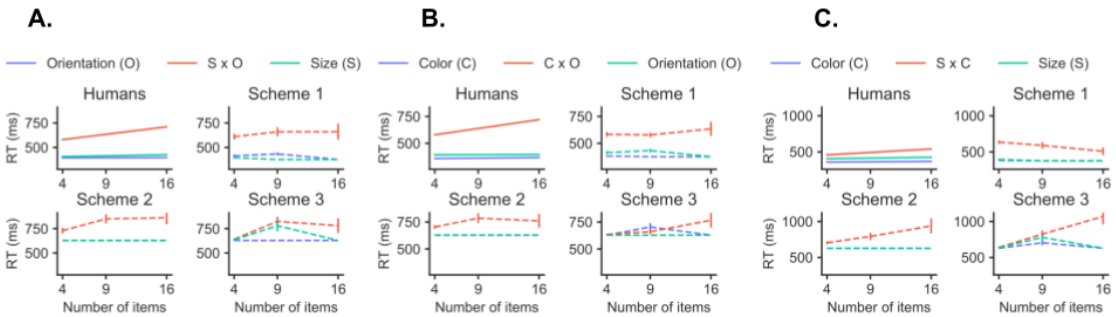


Figure 4.8: RT plots for Experiment 7: Conjunction search

target based feature itself is creating some trouble in identifying these features. Note that these features are fundamental, and Deep CNN models were generally trained to overcome some of these feature differences to improve classification performance during occlusion or variations in the sought image. And probably that is the reason why we got this observation. Since the top-down feature itself creates trouble in finding size and color features, it's not possible to improve the performance during conjunction search. For singleton features, the bottom-up model highlights those features because of their salience.

4.2.2 Experiment 8: Shape.

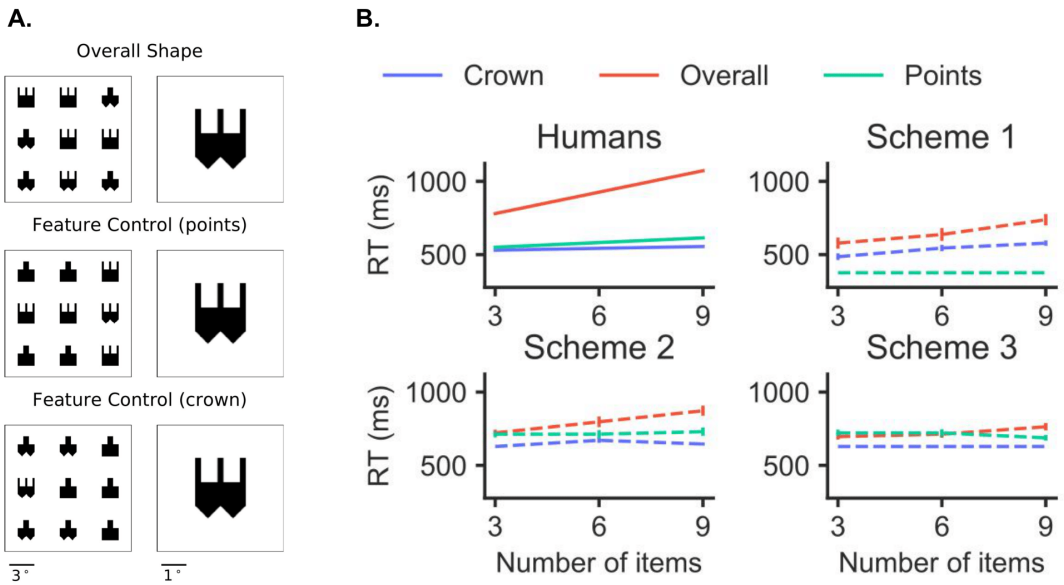


Figure 4.9: Stimuli and RT plots for Experiment 8: Shape

This experiment was based on [51]. Two types of shape features were used: crown (three vertical lines pointing up), and points (two triangles pointing down) (**Figure 4.9A**). The target can be distinguished from the distractors based on one single shape feature in the control conditions. In contrast, in the conjunction search condition, both shape features must be used to find the target. Three set sizes were used: 3, 6, and 9. The search image was divided into 3 x 3 grid cells and covered 15 x 15 dva. The target and distractors were of size 2.5 x 2 dva and were placed

randomly at the centre of the grid cells. There was a total of 90 experiment trials per condition, equally distributed among each of the set sizes.

Similar to the previous experiment, in this also, ”Scheme 2” best captured the performance pattern (**Figure 4.9B**). In this case, also absolute slopes have differences but are comparable. The slope for humans is 48.7, 10.7, and 4.2, respectively, for ”overall”, ”points”, and ”crown” conditions. While for the models with ”Scheme 2”, the corresponding slopes are 25.2, 2.8, and 2.8.

4.2.3 Experiment 9: Preattentive.

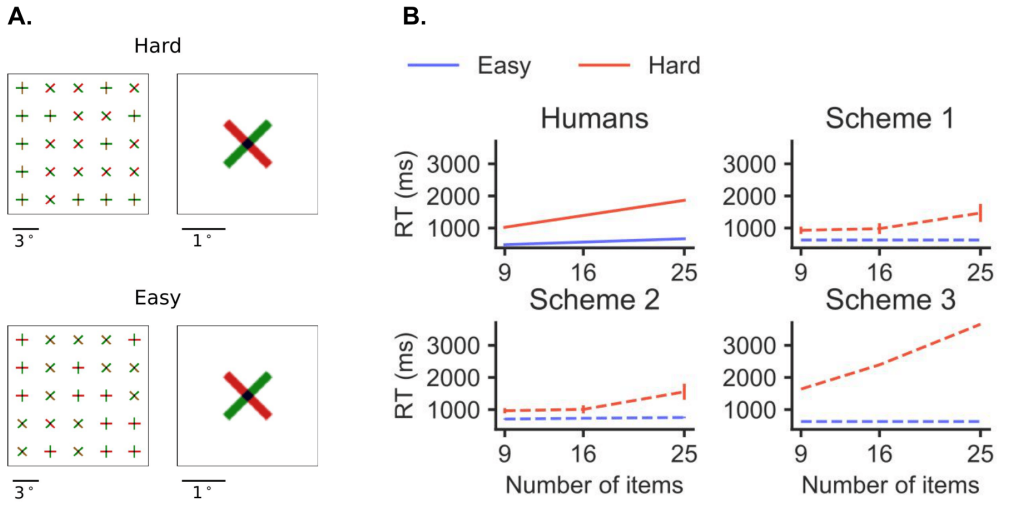


Figure 4.10: Stimuli and RT plots for Experiment 9: Preattentive

This experiment was based on [51]. In the “easy” condition, preattentive objects for the target and distractors were different (**Figure 4.10A**, top) while in the “hard” condition, preattentive objects for the target and distractors were the same (**Figure 4.10A**, bottom). Three set sizes were used: 9, 16, and 25. The search image covered 16 x 16 dva. Objects were enclosed in a square of size 1.4 x 1.4 dva. Objects were randomly placed in a 3 x 3 (9 objects), 4 x 4 (16 objects), or 5 x 5 (25 objects) grid. Objects were placed at the center of the grids. There was a total of 90 experiment trials per condition, equally distributed among each of the set sizes.

All the schemes happened to replicate the human pattern with ”scheme 1” and

"scheme 2" having a better match in terms of absolute reaction time (4.10, B). Scheme 3 also captures the relative performance trend that is the difficult search being more difficult and easy search being easier. But the exact slope for difficult case (**Figure 4.10A, top**) for the model with "scheme 3" is relatively very high.

4.3 Visual search in natural images

These experiments were based on [16]. In these experiments, we directly use the target and search images provided by the authors. This class has three different experimental conditions: Experiment 10: Object arrays; Experiment 11: Natural design; and Experiment 12: Finding Waldo. These sets of experiments were chosen based on increasing difficulty conditions and benchmark the model on different types of visual search conditions relevant to the day-to-day life of humans. Similar to [16], we also followed the following four metrics to evaluate the proposed model on these tasks:

1. **Cumulative performance:** This is defined as the fraction of tasks in which the target was found within the given fixation number. For example, say we have a total of 300 tasks in object arrays, and the target was found only for 100 different cases within two fixations, then cumulative performance at fixation number 2 will be 0.34. Now consider the model finds the target for 100 more cases at 3rd fixation, then cumulative performance at fixation number 3 will become 0.67.

2. **Scanpath similarity score:** This determines the similarity between the scanpath predicted by the model with the scanpaths observed on humans. The metric score was computed using the ScanMatch Toolbox of Matlab.

3. **Saccade size distribution:** During visual search, humans are found to have one particular pattern of saccade sizes. They tend to make higher numbers of smaller saccades as compared to larger ones. The saccade distribution was plotted for both the model and humans side-by-side and were manually compared to test whether and if the model also shows similar behaviour.

4. **Fixation to target distance:** Another interesting observation in human visual search is that the last six fixations generally tend to come closer and closer to the target object, i.e. the last fixation (L-0) being the closest to the

target, and then L-1, L-2 and so on. For a computational model to predict similar behaviour and performance would be a fruitful and exciting direction in studying visual search. So here, we compared the distribution of the euclidean distance between the target and the last six fixations.

The proposed model was compared against the IVSN model ([16]), which shows best performances in these set of the task as compared to other models of visual search. Thus, we took it as the baseline model and compared the performance of the proposed model against all the four different comparison metrics.

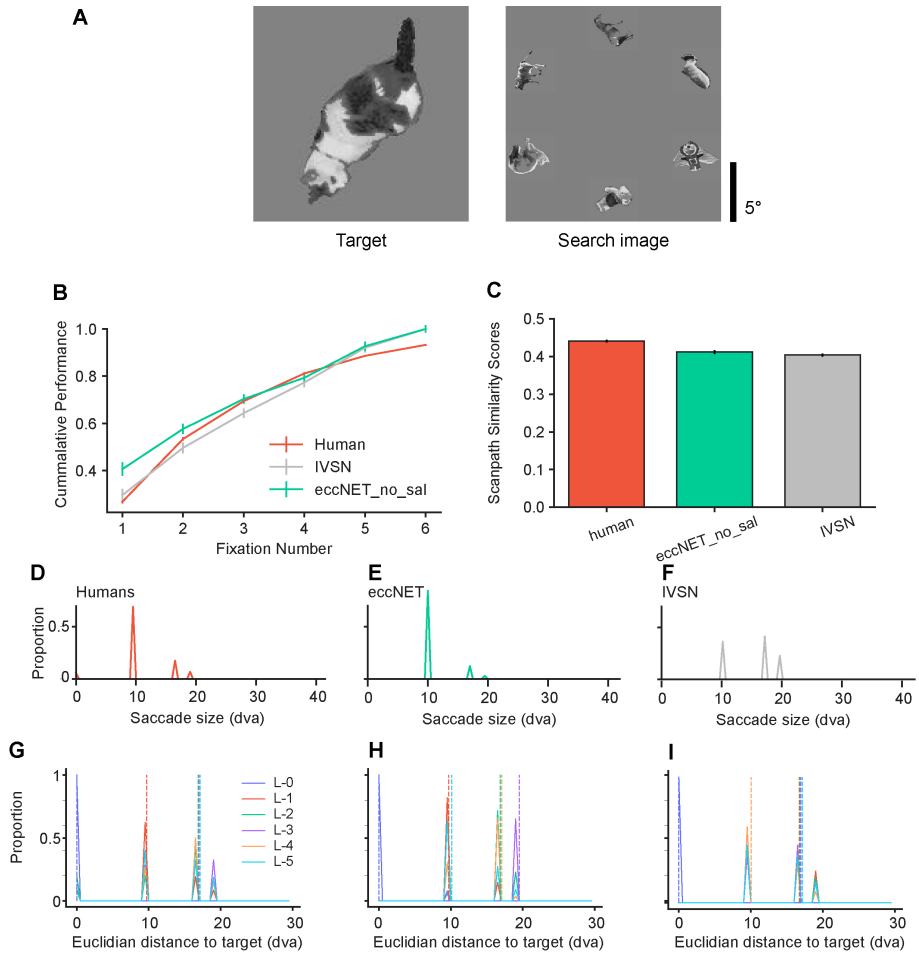


Figure 4.11: The eccNET model matches previous visual search experiments with object arrays ([16]). **A.** Example target and search images **B.** Cumulative search performance as a function of fixation number for humans (red), eccNET (green) and IVSN (gray). IVSN is the model proposed in [16]. **C.** Scanpath similarity scores between humans (red), between humans and ECCnet (green), and between humans and IVSN (gray). The scanpath similarity score measures the similarity between two two eye movement sequences ([52, 16]). **D-F.** Distribution of saccade sizes. **G-I**. Distribution of Euclidean distance from target location to either of the last six fixation locations.

4.3.1 Experiment 10: Object arrays

In the object arrays condition, a target image of a natural object was searched within an array of six objects distributed along a circle (**Figure 4.11**). Sample stimuli and results are shown in **Figure 4.11**. The cumulative performance of the proposed model follows the pattern of humans. Also, the model has a higher area under the curve as compared to IVSN, showing better performance in finding a natural object in arrays (**Figure 4.11B**). Note that despite the introduced eccentricity, the model showed better performance than IVSN, indicating that eccentricity could provide some advantage in devising search strategies in such conditions. This is just an observation, and this thesis does not intend to make any stronger claim on this but recommends further work in this direction. The model also shows comparable scan paths similarity score to humans and the IVSN model (**Figure 4.11C**). The major advance came in the comparison of saccade distribution where the model showed more similar behaviour to humans as compared to the IVSN model, suggesting we have a better model to explain human behaviour in visual search task (**Figure 4.11D-F**). The model captured the human pattern in determining fixation to target distance until the last (L-0) and second last fixation (L-1) which, on the other hand, IVSN failed to do.

4.3.2 Experiment 11: Natural design

In the natural image condition (Experiment 11), a target object was searched in a natural scene (**Figure 4.12**). Sample stimuli and results are shown in **Figure 4.12**. The cumulative performance of the proposed model follows the pattern of humans. The model is having a slightly lower area under the curve as compared to IVSN, showing poor performance in finding a natural object in natural scenes (**Figure 4.12B**). Note that the difference is very small and can be attributed to the degradation in the quality of images due to the introduced eccentricity. The model also shows a comparable scan paths similarity score to humans and the IVSN

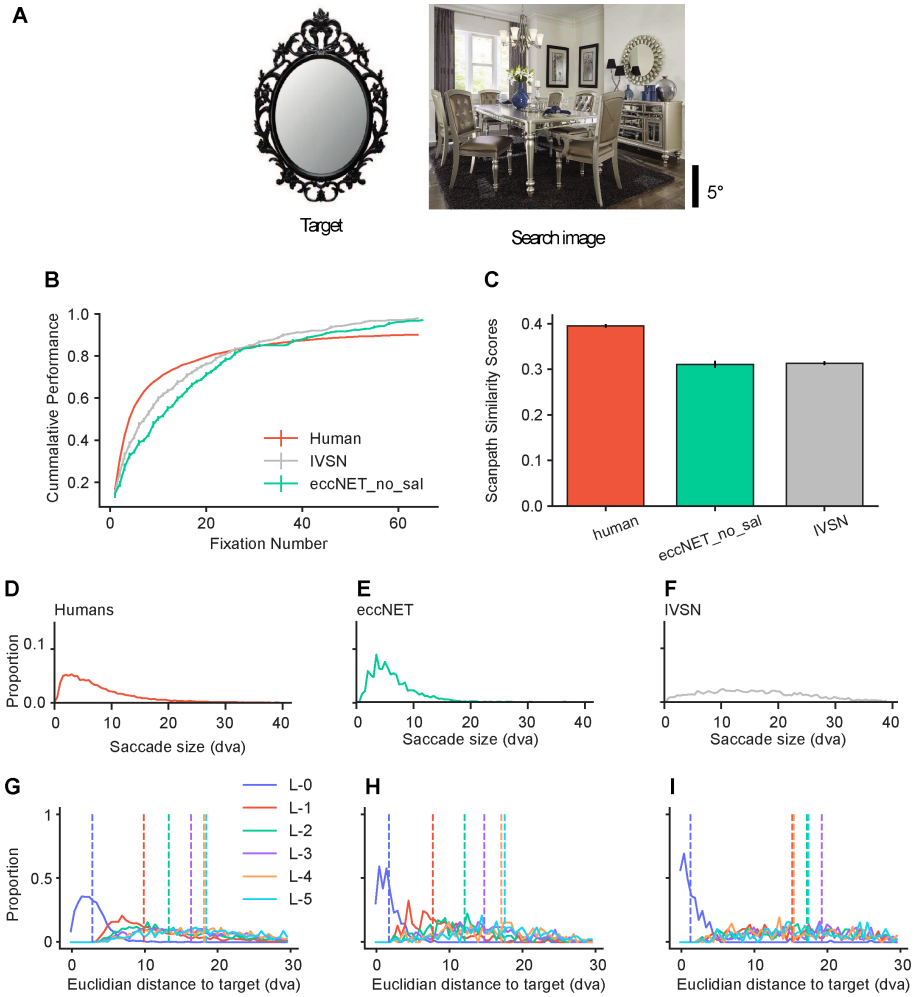


Figure 4.12: The eccNET model matches previous visual search experiments with natural images ([16]). The format and conventions in this figure are the same as in **Figure S2**.

model (**Figure 4.12C**). Similar to results on objects arrays, the major differences came in the comparison of saccade distribution where the model showed more similar behaviour to humans as compared to the IVSN model, suggesting that this was a better model for explaining human behaviour in visual search tasks (**Figure 4.12D-F**). Unlike the object arrays case, the model captured the human pattern in determining fixation to target distance very accurately for all the last six fixations, while the IVSN model fares badly in comparison.

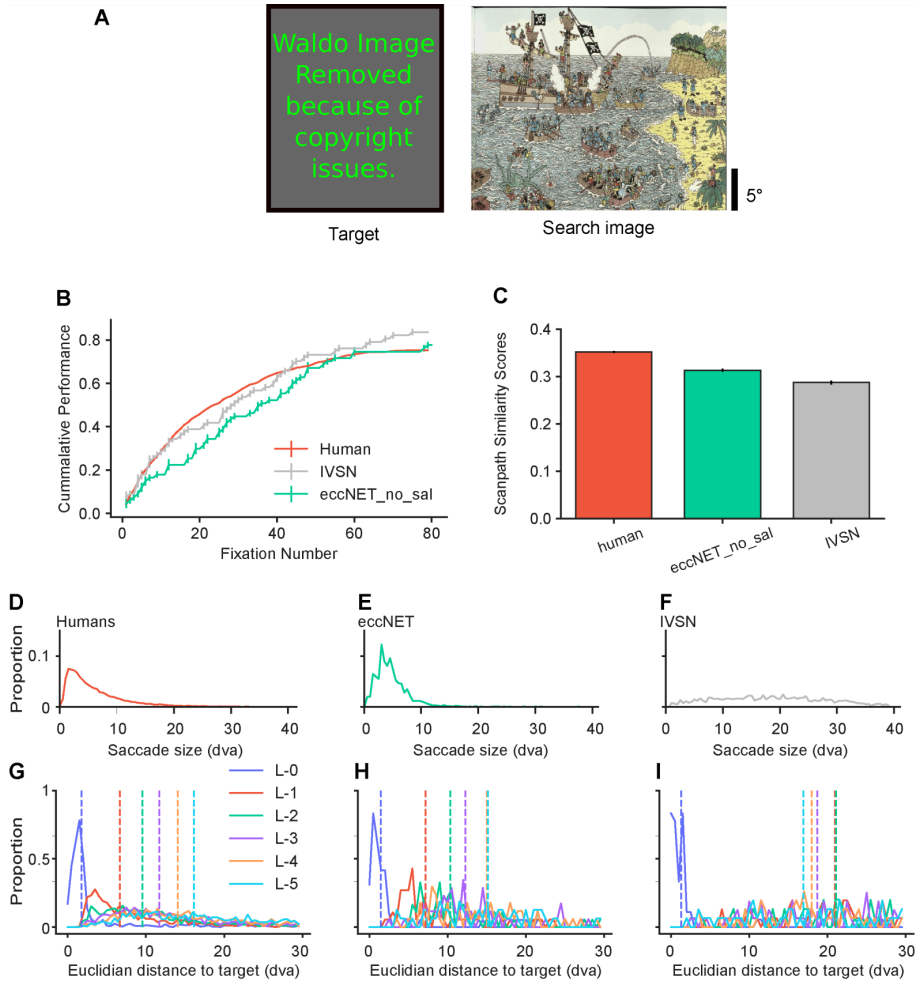


Figure 4.13: The eccNET model matches previous visual search experiments with Waldo images ([16]) The format and conventions in this figure are the same as in **Figure S2**. The target image of 'waldo' is removed because of copyright issues. A similar-looking image can be retrieved at this link (Date: 30 September, 2021).

4.3.3 Experiment 12: Finding waldo

In the Waldo condition, the search task followed the famous search game of finding Waldo in a complex image with multiple similar objects (**Figure 4.13**). Sample stimuli and results are shown in **Figure 4.13**. The cumulative performance of the proposed model follows the pattern of humans. The model is having a slightly lower area under the curve as compared to IVSN, indicatng poor performance in finding waldo (**Figure 4.13B**), though the performance is yet quite comparable to IVSN on the whole. The model also shows significant improvement in scanpaths similarity score to the IVSN model (**Figure 4.12C**) and lies midway between the similarity

score of the IVSN model and Inter-Human. Similar to results in the previous two cases, the significant differences came in the comparison of saccade distribution where the model showed more similar behaviour to humans as compared to the IVSN model (**Figure 4.12D-F**). In determining fixation to target distance in the finding Waldo case, the model accurately matches the human pattern upto the last five fixations but fails for the last sixth fixation, while in comparison, the IVSN model fails badly.

Chapter 5

Combined Results and Discussions

This chapter compares the results obtained by the proposed model against some baseline and ablated versions of the model. Different metrics were used to compare the models. The comparison are done on three different metrics/ conditions: 1. Ability to reproduce the asymmetry property (using a self-introduced Asymmetry Index to measure the level of asymmetry across all the asymmetry search experiments) 2. Quantitative comparison of the search cost slope of reaction time vs the item count (by computing the correlation values between the slopes predicted by model and observed in human psychophysic studies). 3. Comparison based on the visual search performance on Natural Images. Before moving forward, let us briefly introduce the baseline and ablated models used for comparison.

The following four models were used as baselines for comparisons:

1. **Chance prediction:** No attention maps were predicted and the successive fixations were predicted by uniform random sampling. Sampling was done such that it incorporates IOR and do not sample any fixations point inside the inhibited regions.

2. **pixelMatching:** Attention map is computed using raw image pixels of the target and search image. It follows the template matching approach where the target image was moved over the whole search image with a stride of 1×1 but without normalization.

3. **GBVS:** We computed the bottom-up saliency map as proposed in ([53]) and used this map for fixation prediction.

4. **IVSN:** A single top-down activation map based on the features extracted using the top layer of VGG16 is computed for fixation prediction, as described in the IVSN paper ([16]).

Following model ablations were done for comparisons:

1. **eccNET_noecc:** Eccentiricity component of the model was removed to study the effect of eccentricity.

2. **eccNET_0_0_1:** Instead of using multiple layers only the top-most layer was used as visual feature similar to the IVSN model except for the inclusion of the additional component of eccentricity.

3. **eccNET_no_sal:** The bottom-up saliency component of the model was removed.

4. **eccNET_no_topdown:** The top-down target modulation component of the model was removed.

5. **eccNET_Rot90:** Instead of training on original sets of image in ImageNET the Deep-CNN model was trained on rotated ImageNET data.

6. **eccNET_MNIST:** Instead of training on original sets of image in ImageNET the Deep-CNN model was trained on rotated MNIST digit classification data.

5.1 Visual search asymmetry

As we saw in the previous chapter, the proposed model qualitatively captures most of the asymmetry properties of visual search. Irrespective of the nature of task-specific training, the model replicates similar behaviour which humans show. On the other hand, the baseline models failed to capture those asymmetry properties.

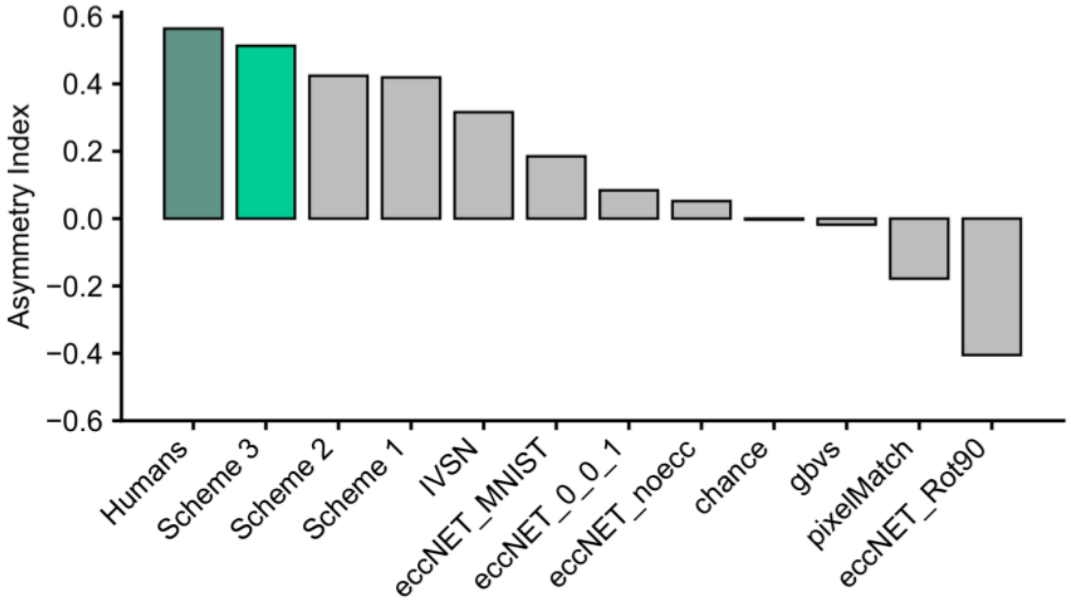


Figure 5.1: Asymmetry Indexes for eccNET, ablated and other models. Average Asymmetry Index for humans (dark green), eccNET (Scheme 3, light green), and alternative models (light gray)

To better compare all the models on asymmetry properties, an asymmetry index is introduced to measure the asymmetry in a given task. The **Asymmetry Index** is defined as $(H - E)/(H + E)$ for each experiment. Here H and E are the computed slope of the RT versus item count plots for the hard (H, larger search slopes) and easy (E, lower search slopes) conditions within each experiment. If a model follows the human asymmetry patterns for a given experiment, it will have a positive Asymmetry Index. The Asymmetry Index takes a value of 0 if there is no asymmetry, and a negative value indicates that the model shows the opposite behavior to humans. The individual asymmetry indices were computed for each experiment, and then the average is taken across all the experiments to compare different models. The results are shown in **Figure 5.1**. It can be clearly seen that the proposed model performed significantly better than the other baseline models (0.513, Scheme 3, **Figure 5.1**). The model was also ablated to test what were the essential components contributing to search asymmetry. This revealed that both of the major changes that were brought in the IVSN model to build the top-down component of the proposed model in this thesis played a significant role in arriving at this result. Without the eccen-

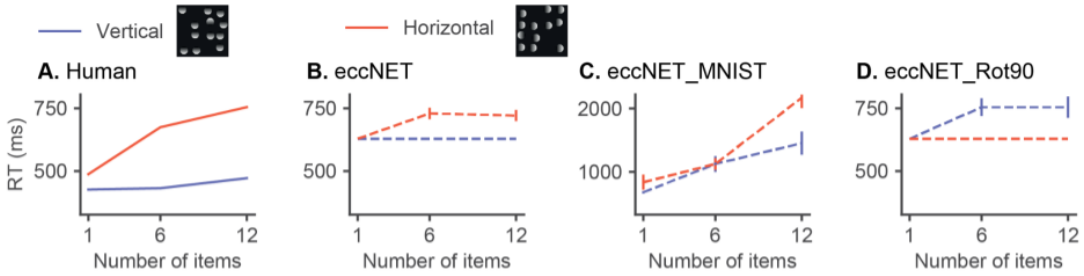


Figure 5.2: Training data alters the performance of the visual search or biases the polarity of search asymmetry. Reaction times as a function of the number of items for Experiment 2.

tricity dependence, the model scored only 0.052 (**Figure 5.1, *eccNET_noecc***), and without the multiple top-down modulations, the model scored 0.084 (**Figure 5.1, *eccNET_0_0_1***). But this still does not necessarily mean that only the visual architecture contributes towards search asymmetry. The thesis suggests that the statistics of the visual data used to train the object recognition model are also responsible for the search asymmetry. This was shown by performing the same sets of experiments but after altering the training data; in one case, we rotated the training dataset by 90 degrees to the model (**Figure 5.1, *eccNET_Rot90***). In the second case, we considered using low feature images of MNIST to train the model (**Figure 5.1, *eccNET_MNIST***). It was observed that changing the training data also alters the performance of the visual search asymmetry, and in some cases, it even alters the polarity **Figure 5.1** and **5.2**.

5.2 Quantitative comparison of search cost slope

To better compare the model's performance with other baseline models, we computed the correlation scores between the slope of the reaction time vs item count between the model and the human for all the visual search conditions. The best scheme out of the three for each experiment were chosen manually and was named *eccNET_optimal*. The results are shown in the **Figure 5.3**. Additionally, the result for *eccNET_DM* model, which is independent of different integration scheme

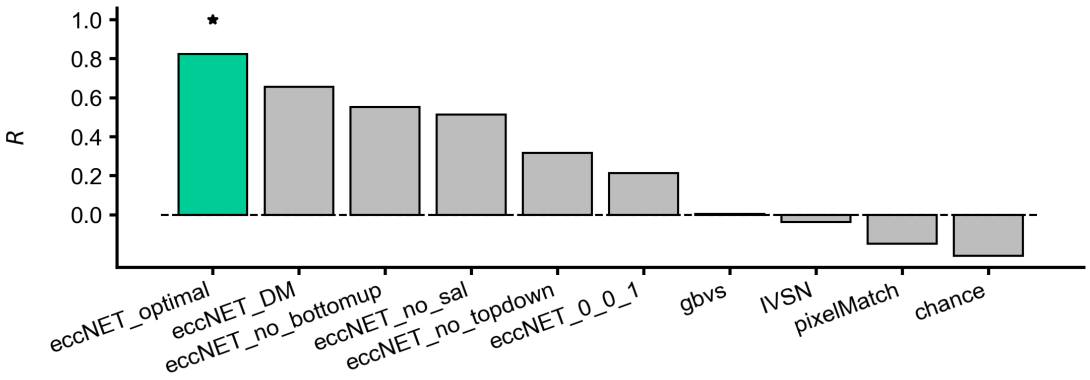


Figure 5.3: Correlation score for search cost slopes between baseline models and humans over all experiments Here we show the Pearson's correlation score obtained on comparing search cost slopes in the reaction time versus number of objects plots of the model with that of humans. The gray bars are score for baseline models compared against our final proposed model in green. $*$ denotes $p < 0.01$ for zero correlation.

(see next **Chapter 6**) and does not require any human interference is also shown in the same bar plot. Again the proposed model performed best, showing the highest correlation score across all the tasks as compared to other baseline models. Similar to search asymmetry, here also it was found that the ablated model scores were significantly lower than the final proposed model, thus indicating that each of the new components that was brought in the proposed model compared to IVSN has significantly helped improve the correlation scores.

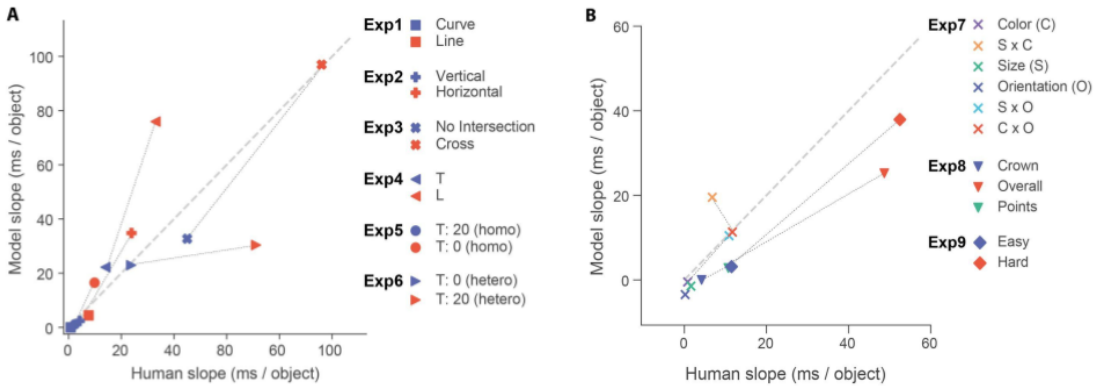


Figure 5.4: The search cost slopes for the model match humans'. For each of the experiments and experimental conditions, the search cost slope in the reaction time versus number of objects plots were computed. This figure shows the slopes for the model (y-axis) versus humans (x-axis) for the search asymmetry experiments **(A)** and the conjunction search experiments **(B)**. The gray dashed line indicates the identity line. The dotted line join the different conditions within the same experiment.

Considering the low number of data points, looking only at the correlation values may not be a good idea. So, we also looked at the exact scatter plots of slope of reaction time vs the item count for each of the experiment conditions. It was found that most of the point are close to the $y = x$ identity line, indicating that the predicted slopes are close to the human slopes. The slope scatter plot is shown in **Figure 5.4**.

5.3 Visual search on natural images

The results from experiments no.s 11-13 show that despite bringing multiple changes in the top-down component of the model compared to IVSN, the model is decent enough to perform equally well on the three sets of visual search tasks considered in the IVSN paper. Note that these tasks were smartly chosen to benchmark different search models on varying difficulty of the search tasks in natural image settings. Along with showing comparable performance, the proposed model in this work showed a much better prediction on saccade size distribution where the IVSN model fails to do so. These results further suggest that along with showing good performance in a natural setting, the model predicts human behaviour much better than any previous models of visual search. The results are shown in **Figure 4.11-4.13** in **Chapter 4**.

Chapter 6

Predicting Task-Dependent Saliency Bias

The observation on various experiments using different modelling schemes suggests that possibly humans are involved in some online learning process during which they predict the relative weight for the bottom-up and top-down maps. Further, these weights do not seem to be the same for all the sought experiments and could probably vary on different stimuli, demands of the task, etc. No single scheme appears to explain all the search experiments. We tried to introduce an additional component to the proposed model, which brings similar specs into it. Before moving forward, we would like first to state all the essential points that we must consider while building such a mechanism:

1. The mechanism should not try to predict relative weights ($W1$) by using the ground truth RT data from humans because while performing the same experiment, humans did not have those. Note that we can always do a parameter fitting to get the best match for the relative weight based on the experimental data, and say that this is the weight that explains the results, but the purpose of this thesis is not only to prove that but to also explain how humans determine that relative weights.

2. The same method should be followed for all the different sets of experiments.

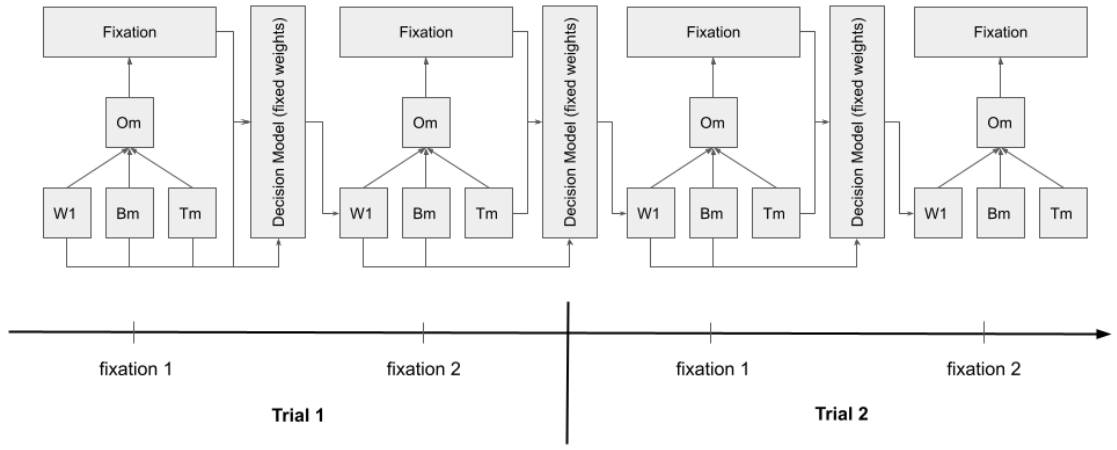


Figure 6.1: The sequential process illustrating the module for determining the saliency bias. At each fixation the module sends four signals to the decision model *fixation point, current saliency bias ($W1$), bottom-up saliency map, and top-down target-modulated map.*. The decision module updates it's internal bias parameters and use it to predict the saliency bias ($W1$) for next fixation.

Any new parameters introduced in the model for this additional mechanism should be the same across all the experiments and should not vary depending on the search task.

Considering all the above points, we introduced a sequential decision mechanism that uses the information at the current fixation to predict the relative weight for the next fixation. To be precise, the model uses the current relative weight W1, top-down map, bottom-up map, and feedback from the object recognition model to know whether it found the target or not. And then, based on these four pieces of information, it updates its belief about the W1. The process is shown in **Figure 6.1**. Whenever the model is given a new experiment, it starts the search by putting $W1 = 0.5$, i.e. giving an equal contribution to both the maps. And then, after each fixation, it updates its belief about W1 and continues this until the end of all the experiment trials. The exact process by which the model updates the belief about '$W1$' is shown in (**Figure 6.2**). The process can be described in following steps:

1. The model assumes that 'W1' follows some probabilistic distribution with parameters alpha and beta.

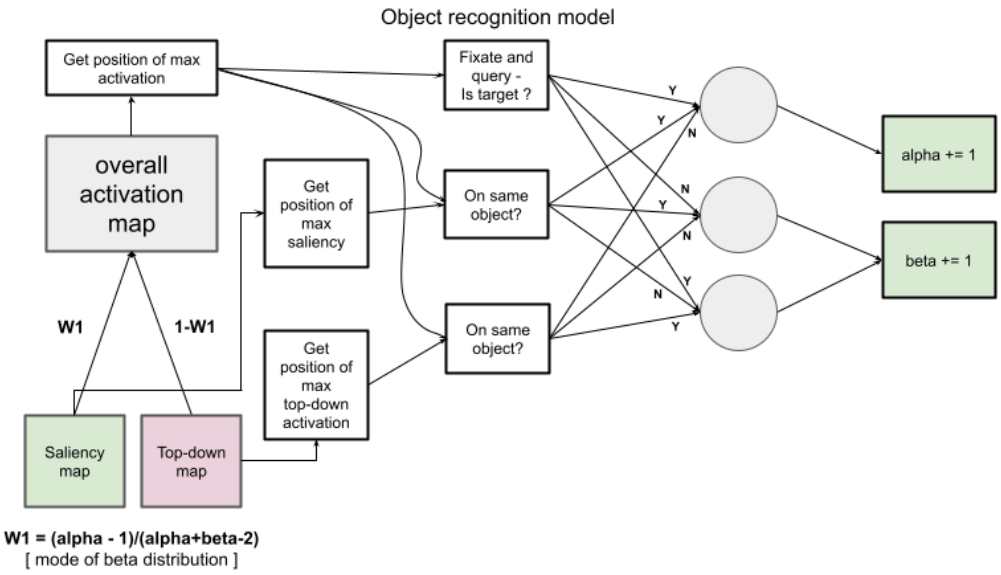


Figure 6.2: The decision model of the module predicting saliency bias The figure shows the flowchart of how the internal parameters are updated at each fixation and how it is used to predict the saliency bias ($W1$)

2. The model starts the search with the prior belief of $alpha = 1$ and $beta = 1$ for each experiment condition.

3. At each fixation, during the trial, the model updates the value of alpha and beta based on what it has seen, i.e. the current fixation, saliency map and top-down attention map.

4. The updated value of alpha and beta is used prior to the next trial in the same experiment condition.

5. The update is a continuous process, and updating of alpha and beta is done during all the search trials in that specific condition.

6. The model will start again from the beginning assigning $W1 = 0.5$, i.e. giving equal weights to each map for any new experiment condition.

7. At any point of time, W1 will be the mode of the beta distribution. Therefore, when $alpha = beta \implies W1 = 0.5$. Otherwise: $W1 = (alpha - 1)/(alpha + beta - 2)$

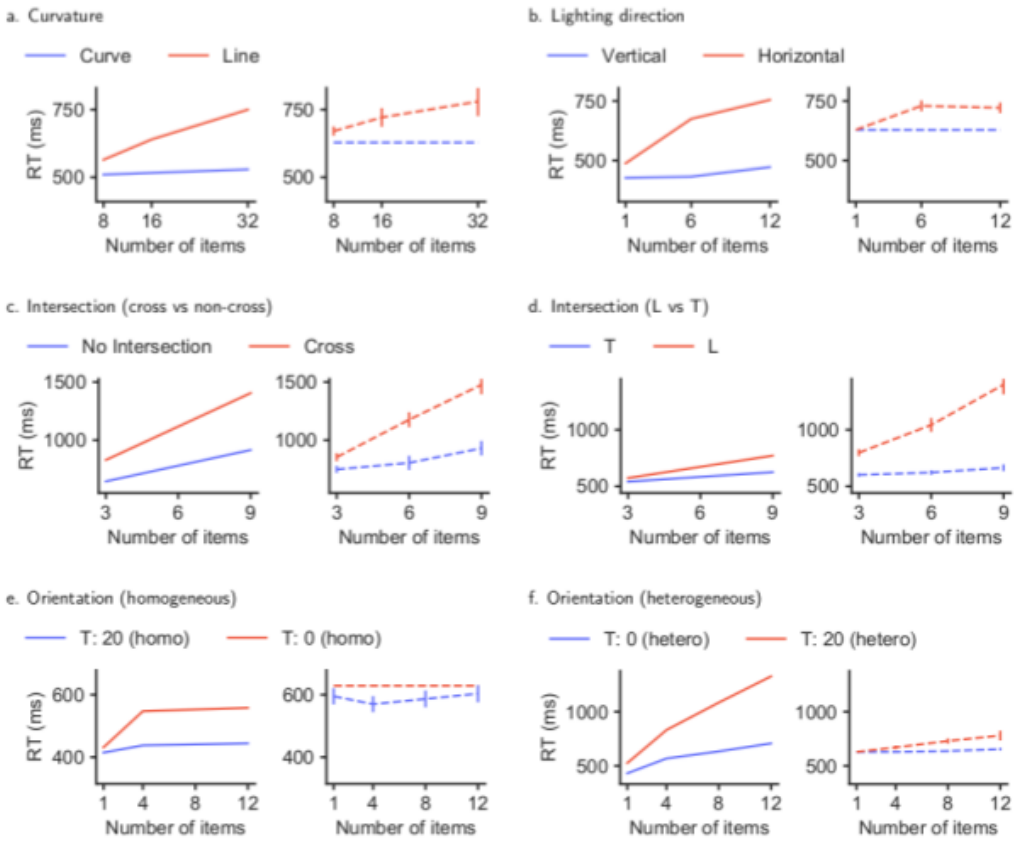


Figure 6.3: Reaction Time predicted by the model for asymmetry search experiments after incorporating the module for predicting saliency bias

6.1 Results and discussions

After including the module predicting the Task-Dependent Saliency Bias in the visual search model, all the experiments were repeated from the asymmetry and feature-conjunction classes. The model accurately predicted the relative pattern of difficulty for each of the experiments giving a completely self-dependent model capable of predicting the trends of human performance across all the sets of the experiment. Though the absolute comparison of slope gave slightly poor performance if compared against a manually selected salience scheme (0.656 correlation score compared to the 0.824 correlation score of manually selected salience schemes), there is clearly significant room for improving the model in the matter of predicting the relative contribution of top-down and bottom-up component. But considering

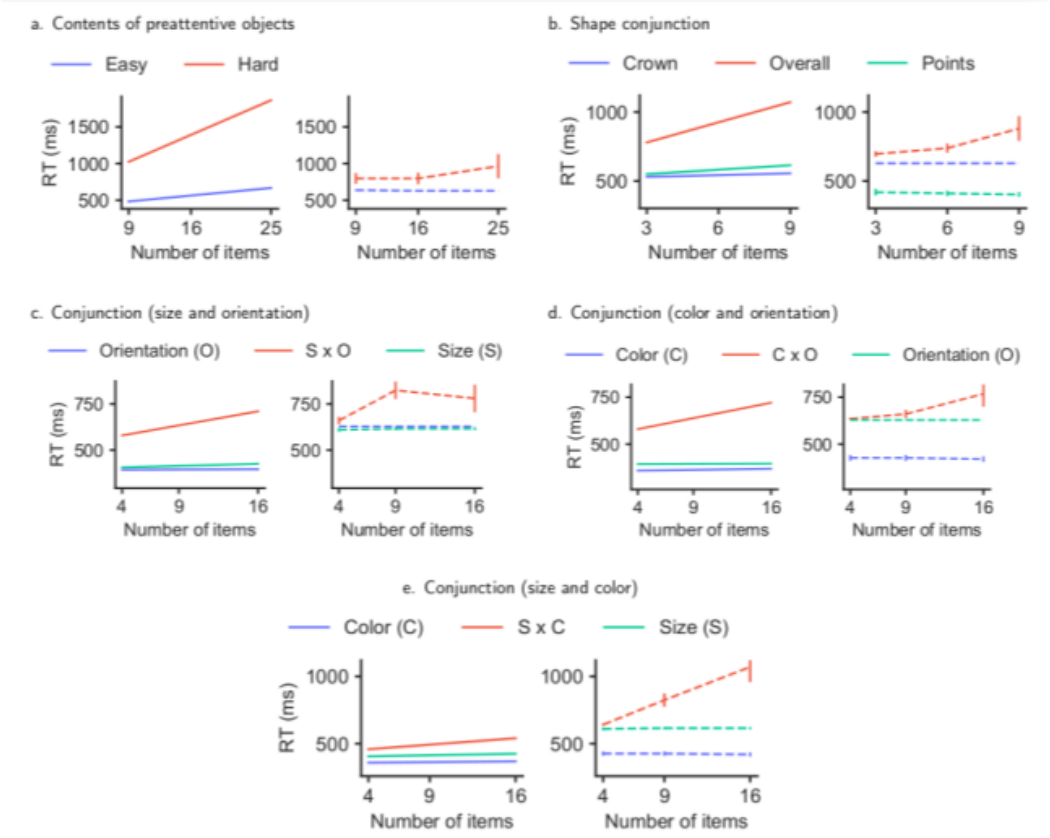


Figure 6.4: Reaction Time predicted by the model for feature-conjunction search experiments after incorporating the module for predicting saliency bias

the simplification and the fact that the model was not given any external information regarding the experiments, the results are excellent. This reaction time search cost slope predicted by the improved model showed a correlation of 0.656 against the human slope, where none of the other ablated models or baseline models reached this score. The reaction time plots are shown in **Figure 6.3** and **6.4**

Chapter 7

Conclusion

This thesis introduces an integrated computational model of visual search that incorporates theoretical frameworks from psychology, resembling the architecture from neurophysiology. The proposed model integrates three essential components, an eccentricity-dependent deep convolutional neural network as a visual processor, top-down target modulated activation maps, and bottom-up saliency-based activations. The thesis argues that to build a universal model of visual search, it's essential to evaluate the model on its ability to replicate human behaviour in classical visual search tasks and show comparable efficiency while performing a search task in natural scenes. Thus, the proposed model was evaluated on multiple visual search tasks against humans in terms of search performance and capturing human behaviour. It was found that the proposed model qualitatively predicted human behaviour in most of the considered experiments and showed efficient performance on search tasks in natural images. It's important to note that the model had no exposure to any of the images considered in these experiments and was only exposed to images in the ImageNET dataset, which was used to pre train the Deep-CNN model.

Moreover, the model is self-sufficient and does not require human supervision or task-specific training to search for any new target object. Some part of the model is free from any task-specific learning, while some part does incorporate task-specific training, but it does it based on its own self-feedback mechanism (see **Chapter 3.2**). In other words, the model knows how to learn and does not really need any human

interference. To our knowledge, none of any other models in literature has considered the depth of psychological and neurophysiological ideas in building such a model for visual search. Moreover, no other work shows an in-depth study on replicating human behaviour across an exhaustive list of visual search tasks considered in this thesis. This thesis also compared the proposed model against other models of visual search. We found that none of them could replicate human behaviour across a decent number of tasks.

Although the model showed excellent qualitative performance, the model falls behind in quantitatively matching the exact reaction time of humans across all the search tasks. Yet, the model brings various simplifications and assumptions. And despite that, the success of the model in providing a good qualitative fit encourages further investigation and opens up a path for exciting future work. It lays down a strong evaluation methodology for any new visual search model. It exhibits the ability to give decent efficiency for visual search in natural scenes: the model suggests exciting directions to study its application in computer vision problems such as in visual surveillance system, autonomous systems, or in visual design, where predicting the visual direction where a human might look is important.

References

[1] JeremyM Wolfe and ToddS Horowitz. "Five factors that guide attention in visual search". In: *Nature Human Behaviour* 1.3 (2017), pp. 1–8.

[2] Jacob L Orquin and Simone Mueller Loose. "Attention and choice: A review on eye movements in decision making". In: *Acta psychologica* 144.1 (2013), pp. 190–206.

[3] Alexander C Schütz, Doris I Braun, and Karl R Gegenfurtner. "Eye movements and perception: A selective review". In: *Journal of vision* 11.5 (2011), pp. 9–9.

[4] Sabine Kastner Ungerleider and Leslie G. "Mechanisms of visual attention in the human cortex". In: *Annual review of neuroscience* 23.1 (2000), pp. 315–341.

[5] James W Bisley. "The neural basis of visual attention". In: *The Journal of physiology* 589.1 (2011), pp. 49–57.

[6] Marie-Hélène Grosbras, Angela R Laird, and Tomás Paus. "Cortical regions involved in eye movements, shifts of attention, and gaze perception". In: *Human brain mapping* 25.1 (2005), pp. 140–154.

[7] Sofia Paneri and Georgia G Gregoriou. "Top-down control of visual attention by the prefrontal cortex. functional specialization and long-range interactions". In: *Frontiers in neuroscience* 11 (2017), p. 545.

[8] Maximilian Riesenhuber and Tomaso Poggio. "Hierarchical models of object recognition in cortex". In: *Nature neuroscience* 2.11 (1999), pp. 1019–1025.

[9] Thomas Serre et al. "A quantitative theory of immediate visual recognition". In: *Progress in brain research* 165 (2007), pp. 33–56.

[10] Gabriel Kreiman and Thomas Serre. "Beyond the feedforward sweep: feedback computations in the visual cortex". In: *Annals of the New York Academy of Sciences* 1464.1 (2020), p. 222.

[11] Robert Desimone and John Duncan. "Neural mechanisms of selective visual attention". In: *Annual review of neuroscience* 18.1 (1995), pp. 193–222.

[12] T Serre, A Oliva, and T Poggio. "Feedforward theories of visual cortex account for human performance in rapid categorization". In: *PNAS* 104.15 (2007), pp. 6424–6429.

[13] John H Reynolds and Leonardo Chelazzi. "Attentional modulation of visual processing". In: *Annu. Rev. Neurosci.* 27 (2004), pp. 611–647.

[14] Maurizio Corbetta. “Frontoparietal cortical networks for directing attention and the eye to visual locations: identical, independent, or overlapping neural systems?” In: *Proceedings of the National Academy of Sciences* 95.3 (1998), pp. 831–838.

[15] Narcisse P Bichot et al. “A source for feature-based attention in the prefrontal cortex”. In: *Neuron* 88.4 (2015), pp. 832–844.

[16] Mengmi Zhang et al. “Finding any Waldo with zero-shot invariant and efficient visual search”. In: *Nature communications* 9.1 (2018), pp. 1–15.

[17] Joseph Redmon and Ali Farhadi. “YOLOv3: An Incremental Improvement”. In: *arXiv* (2018).

[18] Shaoqing Ren et al. “Faster R-CNN: Towards real-time object detection with region proposal networks”. In: *Advances in neural information processing systems* 28 (2015), pp. 91–99.

[19] Nicolas Carion et al. “End-to-end object detection with transformers”. In: *European Conference on Computer Vision*. Springer. 2020, pp. 213–229.

[20] Laurent Itti, Christof Koch, and Ernst Niebur. “A model of saliency-based visual attention for rapid scene analysis”. In: *IEEE Transactions on pattern analysis and machine intelligence* 20.11 (1998), pp. 1254–1259.

[21] Laurent Itti and Christof Koch. “A saliency-based search mechanism for overt and covert shifts of visual attention”. In: *Vision research* 40.10-12 (2000), pp. 1489–1506.

[22] Matthias Kümmerer, Lucas Theis, and Matthias Bethge. “Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet”. In: *arXiv preprint arXiv:1411.1045* (2014).

[23] Anne M Treisman and Garry Gelade. “A feature-integration theory of attention”. In: *Cognitive psychology* 12.1 (1980), pp. 97–136.

[24] Anne Treisman and Sharon Sato. “Conjunction search revisited.” In: *Journal of experimental psychology: human perception and performance* 16.3 (1990), p. 459.

[25] Jeremy M Wolfe and W Gray. “Guided search 4.0”. In: *Integrated models of cognitive systems* (2007), pp. 99–119.

[26] Jeremy M Wolfe. “Guided search 2.0 a revised model of visual search”. In: *Psychonomic bulletin & review* 1.2 (1994), pp. 202–238.

[27] Jeremy M Wolfe, Kyle R Cave, and Susan L Franzel. “Guided search: an alternative to the feature integration model for visual search.” In: *Journal of Experimental Psychology: Human perception and performance* 15.3 (1989), p. 419.

[28] Daniel LK Yamins and James J DiCarlo. “Using goal-driven deep learning models to understand sensory cortex”. In: *Nature neuroscience* 19.3 (2016), pp. 356–365.

[29] Jeremy Freeman and Eero P Simoncelli. “Metamers of the ventral stream”. In: *Nature neuroscience* 14.9 (2011), pp. 1195–1201.

[30] Shashi Kant Gupta. "A More Biologically Plausible Local Learning Rule for ANNs". In: *arXiv preprint arXiv:2011.12012* (2020).

[31] Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. "Deep supervised, but not unsupervised, models may explain IT cortical representation". In: *PLoS computational biology* 10.11 (2014), e1003915.

[32] Daniel LK Yamins et al. "Performance-optimized hierarchical models predict neural responses in higher visual cortex". In: *Proceedings of the National Academy of Sciences* 111.23 (2014), pp. 8619–8624.

[33] Martin Schrimpf et al. "Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like?" In: *bioRxiv* (2020). DOI: 10.1101/407007. URL: https://www.biorxiv.org/content/early/2020/01/02/407007.

[34] Brenden Lake et al. "Deep neural networks predict category typicality ratings for images". English (US). In: *Proceedings of the 37th Annual Conference of the Cognitive Science Society*. Ed. by R Dale et al. Cognitive Science Society, 2015.

[35] Joshua C. Peterson, Joshua T. Abbott, and Thomas L. Griffiths. "Adapting Deep Network Features to Capture Psychological Representations: An Abridged Report". In: *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence* (Aug. 2017). DOI: 10.24963/ijcai.2017/697. URL: http://dx.doi.org/10.24963/ijcai.2017/697.

[36] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems*. 2012, pp. 1097–1105.

[37] S. Ren et al. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.6 (2017), pp. 1137–1149.

[38] K. He et al. "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification". In: *2015 IEEE International Conference on Computer Vision (ICCV)*. 2015, pp. 1026–1034.

[39] Ian J. Goodfellow et al. "Generative Adversarial Nets". In: *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*. NIPS'14. Montreal, Canada: MIT Press, 2014, pp. 2672–2680.

[40] Mengmi Zhang et al. "Look Twice: A Computational Model of Return Fixations across Tasks and Species". In: *arXiv preprint arXiv:2101.01611* (2021).

[41] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

[42] Neil Bruce and John Tsotsos. "Saliency based on information maximization". In: *Advances in neural information processing systems* 18 (2005), pp. 155–162.

[43] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556* (2014).

[44] Steven Yantis and Howard E Egeth. "On the distinction between visual salience and stimulus-driven attentional capture." In: *Journal of experimental psychology: Human perception and performance* 25.3 (1999), p. 661.

[45] Wolfgang EinhÃ, Ueli Rutishauser, and Christof Koch. “Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli”. In: *Journal of vision* 8.2 (2008), pp. 2–2.

[46] Jeremy M Wolfe, Alice Yee, and Stacia R Friedman-Hill. “Curvature is a basic feature for visual search tasks”. In: *Perception* 21.4 (1992), pp. 465–480.

[47] Dorothy A Kleffner and Vilayanur S Ramachandran. “On the perception of shape from shading”. In: *Perception & Psychophysics* 52.1 (1992), pp. 18–36.

[48] Jeremy M Wolfe and Jennifer S DiMase. “Do intersections serve as basic features in visual search?” In: *Perception* 32.6 (2003), pp. 645–656.

[49] Jeremy M Wolfe et al. “The role of categorization in visual search for orientation.” In: *Journal of Experimental Psychology: Human Perception and Performance* 18.1 (1992), p. 34.

[50] Jeremy M Wolfe. “Asymmetries in visual search: An introduction”. In: *Perception & psychophysics* 63.3 (2001), pp. 381–389.

[51] Jeremy M Wolfe and Sara C Bennett. “Preattentive object files: Shapeless bundles of basic features”. In: *Vision research* 37.1 (1997), pp. 25–43.

[52] Ali Borji and Laurent Itti. “State-of-the-art in visual attention modeling”. In: *IEEE transactions on pattern analysis and machine intelligence* 35.1 (2012), pp. 185–207.

[53] Jonathan Harel, Christof Koch, and Pietro Perona. “Graph-based visual saliency”. In: *Advances in neural information processing systems* 19 (2006), pp. 545–552.

[54] Roger Ratcliff and Gail McKoon. “The diffusion decision model: theory and data for two-choice decision tasks”. In: *Neural computation* 20.4 (2008), pp. 873–922.

Appendices

A1 Object recognition module

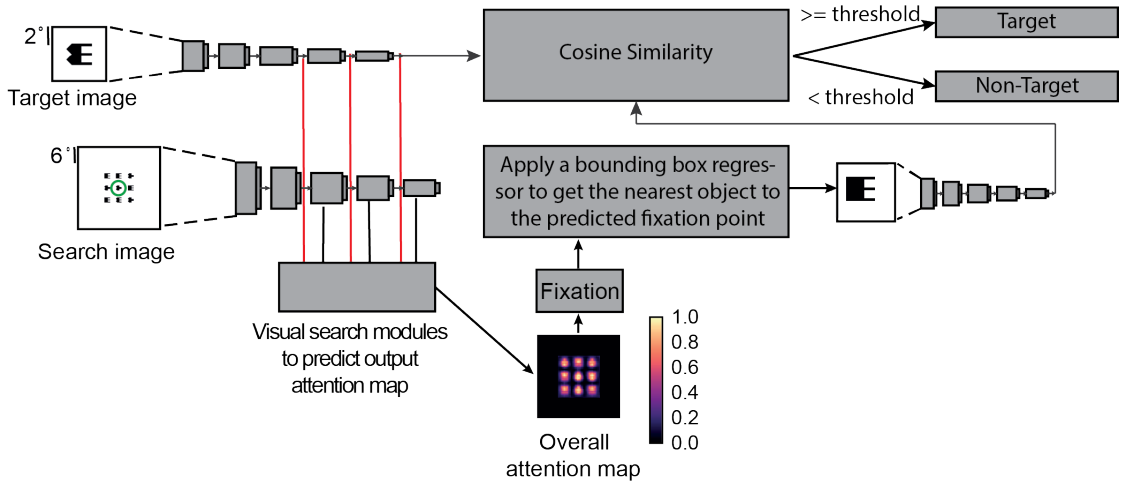


Figure A1: Schematics for implementing the object recognition network in the proposed visual search model

In the context of visual search, the main task of the object recognition model is to decide whether the fixated object is the target object or not which is a simple two choice decision task. Thus, it can be implemented by finding a similarity between the features of the fixated object and the target object and then based on a threshold value, the model will decide whether the fixated object is target or not. The schematic for the architecture is shown in **Figure A1**. Note that, to properly implement this idea, the model will also need a bounding box regression model to find the closest object near the fixated point. Once the bounding box of the fixated object is found, the model can pass down the cropped image of the fixated object down the eccNET. Then a cosine similarity score will be calculated between the tar-

get image's feature and the fixated object. Finally based on the threshold value the model will decide whether the fixated object is target or not. The main difficulty comes on deciding the threshold value, which will vary for different task because not all of them have similar looking objects. To test whether this system works similar to the 'oracle' one we manually tuned this threshold parameter to replicate the results similar to oracle model. It was observed that by choosing a right set of threshold parameter the same results can be obtained. Thus, a diffusion-based decision model [54] could be used to retrieve the threshold parameter in an online fashion. In which the model starts with a same initial belief for the threshold value for each task and subsequently update it's belief after each fixation depending on the feedback of whether it arrived at the target or not. But incorporating all these in the current model is a complex task and is out scope of this thesis.