UNIVERSITY OF CALIFORNIA, SAN DIEGO

A Baysian Framework for Saliency and a Probabilistic Model for Visual Search

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Computer Science
and
Cognitive Science

by

Lingyun Zhang

Committee in charge:

Professor Garrison W. Cottrell, Chair
Professor Serge J. Belongie
Professor Virginia R. de Sa
Professor David J. Kriegman
Professor Terry J. Sejnowski

2007

The dissertation of Lingyun Zhang is approved, and it is acceptable in quality and form for publication on microfilm:

Chair

University of California, San Diego

2007

To my parents.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

It is a great pleasure to thank the many people who made this dissertation possible.

It is not possible to overstate my gratitude to my Ph.D. supervisor, Dr. Garrison Cottrell. He guided me and funded me through the past five years. He provided teaching, advice, company and encouragement throughout my Ph.D. journey. He inspires ideas, encourages independence and promotes originality. From him, I learned not only the scientific skills and the art of research, but also life wisdom. I find myself extremely fortunate and privileged to be one of his graduate students. If time could go back, I would not choose differently.

I wish to thank Professor Javier Mollevan for encouraging me to implement my algorithm on his robots. I had great fun and learned a lot about the care and the effort involved to make a piece of hardware work. Attending his group meetings and discussions greatly opened my eyes beyond my own research.

I am grateful to Professor Terry Sejnowski, Serge Belongie, David Kriegman and Virginia de Sa for many invaluable suggestions and taking the time and energy to serve on my Ph.D. committee.

I feel obliged to many professors from whom I learned valuable mathematical skills. In particular, I want to thank Professor Sanjoy Dasgupta for his teaching in graphical models, Professor Jason Schweinsberg for his teaching in probability theory and Professor Nuno Vasconcelos for his teaching in statistical learning. What I have learned from these classes is reflected in many places throughout this dissertation. More importantly, they showed me the elegance of problem solving and the art of teaching.

I am in debt to Joe McCleery, Matthew Tong, Tim Marks, Nicholas Butko for collaborating with me in many research projects. I have learned a lot from them on not only how to research and how to publish, but also how to work in a team.

I would like to thank Dan Hill and Patrick Gallagher for proof-reading many of my manuscripts and their valuable suggestions. I would also like to thank everyone in GURU for discussions and suggestions, and for their patience of helping me through many practice talks. I also wish to thank everyone in PEN for collaboration and discussion over the last five years. They keep my mind open and encourage me to stay interdisciplinary.

I want to thank Honghao Shan, my best friend here in UCSD. He has helped and supported me in both personal life and research activities for the last six years. He is everything one can expect from a friend and more. I also want to thank my other numerous friends that I can not all name individually. They offer me priceless friendship that I will always cherish.

Lastly, and most importantly, I wish to thank my parents, Wenyu Ling and Chaohui Zhang. They bore me, raised me, supported me, taught me, and loved me. To them I dedicate this dissertation.

Chapter II and Chapter III, in part, is a reprint of the paper in preparation "A Bayesian Framework for Saliency", co-authored with Honghao Shan, Tim K. Marks, Matthew H. Tong and Garrison W. Cottrell. The dissertation author is the primary investigator and author of these two papers.

Chapter IV, in part, is a reprint of the paper in preparation "A Bayesian Framework for Dynamic Scenes", co-authored with Matthew H. Tong, Nicholas J. Butko, Javier R. Movellan and Garrison W. Cottresll. The dissertation author is the primary investigator and author of these two papers.

Chapter V, in part, is a reprint of the paper in preparation "Probabilistic Search: a New Theory on Visual Search", co-authored with Matthew H. Tong and Garrison W. Cottrell. The dissertation author is the primary investigator and author of these two papers.

VITA

Apr. 1983	Born in Harbin, Hei Long Jiang Province, P.R. China
2002	B.S. in Computer Science & B.A. in Management University of Science and Technology of China
2005	M.S in Computer Science University of California, San Diego
2007	Ph.D. in Computer Science & Cognitive Science University of California, San Diego

PUBLICATIONS

Lingyun Zhang, Matthew H. Tong and Garrison W. Cottrell (In preparation) Probabilistic Search: a New Theory on Visual Search.

Lingyun Zhang, Matthew H. Tong, Nicholas J. Butko, Javier R. Movellan and Garrison W. Cottrell (In preparation) A Bayesian Saliency Framework for Dynamic Scenes.

Lingyun Zhang, Honghao Shan, Tim K. Marks , Matthew H. Tong and Garrison W. Cottrell (In preparation) A Bayesian Framework for Saliency.

Honghao Shan, Lingyun Zhang and Garrison W. Cottrell (In preparation) Capturing Visual Structure by Recursive ICA.

Nicholas J. Butko, Lingyun Zhang, Garrison W. Cottrell and Javier R. Movellan (In review) Visual Saliency Model for Robot Cameras.

Lingyun Zhang, Matthew H. Tong and Garrison W. Cottrell (2007) Information Attracts Attention: a Probabilistic Account of the Cross-Race Advantage in Visual Search. In Proceedings of the Twenty-eighth Annual Cognitive Science Society Conference.

Joseph P. McCleery, Lingyun Zhang, Liezhong Ge, Zhe Wang, Eric M. Christiansen, Kang Lee, Garrison W. Cottrell (In review) The roles of visual expertise and visual input in the face inversion effect: Behavioral and neurocomputational evidence. Vision Research.

Honghao Shan, Lingyun Zhang and Garrison W. Cottrell (2006) Recursive ICA. In Advances in Neural Information Processing Systems 18 (NIPS2006), MIT Press, Cambridge, MA.

Lingyun Zhang and Garrison W. Cottrell (2006) Look Ma! No Network!: PCA of Gabor Filters Models the Development of Face Discrimination. In Proceedings of the Twenty-eighth Annual Cognitive Science Society Conference.

Lingyun Zhang and Garrison W. Cottrell (2005). Holistic Processing Develops Because it is Good. In Proceedings of the Twenty-seventh Annual Cognitive Science Society Conference.

Lingyun Zhang and Garrison W. Cottrell (2005). A Computational Model which Learns to Selectively Attend in Category Learning. In Proceedings of the Fourth International Conference of Development and Learning: From Interaction to Cognition.

Lingyun Zhang and Garrison W. Cottrell (2004). Seeing Blobs as Faces or Letters: Modeling Effects on Discrimination. In Proceedings of the Third International Conference of Development and Learning: Developing Social Brains.

Lingyun Zhang and Garrison W. Cottrell (2004). When Holistic Processing is Not Enough: Local Features Save the Day. In Proceedings of the Twenty-sixth Annual Cognitive Science Society Conference.

FIELDS OF STUDY

Major Field: Computer Science and Cognitive Science
Studies in Artificial Intelligence
Professor Garrison Cottrell

ABSTRACT OF THE DISSERTATION

A Baysian Framework for Saliency and a Probabilistic Model for Visual Search

by

Lingyun Zhang
Doctor of Philosophy in Computer Science and Cognitive Science
University of California, San Diego, 2007
Professor Garrison W. Cottrell, Chair

Visual attention reflects the sampling strategy of the visual system. It is of great research interest not only because of its mysterious nature as a biological system, but also because of its potential benefit to computer vision and graphics. Psychologists have investigated visual attention for many decades by psychophysical experiments such as visual search tasks. Sophisticated mathematical models have been built to account for the wide variety of human performance data. With the development of eye movement tracking system, where people fixate when they perform certain tasks can be explicitly recorded and provide straightforward evidence of what people pay attention to. Computational models are emerging fast in recent years that take complex images and videos as input and generate saliency maps which predict what attracts people's attention. In particular, there sees a trend of building principled statistic models that have explicit optimization goals. However, there seems to be a canyon between these two lines of research although both seeks to better understand visual attention. Visual search models are often designed to work with well controlled stimuli with distinct target and distractors, and are not applicable to complex images and videos. On the other hand, saliency algorithms are not supported by theories that can account for the variety of human data in visual search.

In this dissertation, we make an effort of developing a visual attention theory from first principles. Our goal is to have a framework that combines the virtues of both visual attention models and saliency algorithms. We address the following issues to achieve our goal:

(1) We develop a Bayesian framework of saliency by considering what the visual system is trying to optimize when directing attention. Bottom-up saliency emerges naturally as the self-information of visual features. Unlike existing saliency measures, which depend on the statistics of the particular image being viewed, our measure of saliency is derived from natural statistics. Our Bayesian framework also facilitates the incorporation of top-down effects. The measure of overall saliency in visual search, which combines the bottom-up saliency with top-down knowledge of the target's appearance, emerges from our model as the pointwise mutual information between the observed visual features and the presence of a target.

(2) Based on the theory, we implemented bottom-up saliency algorithms for both static images and dynamic scenes. In our model saliency is computed locally, which is consistent with the neuroanatomy of the early visual system and results in an efficient algorithm with few free parameters. They demonstrates good performance at predicting human fixations during free-viewing of images and videos. A real time version of dynamic saliency is implemented on a robotic camera. When the camera is oriented toward salient regions, the chance of seeing people is greatly improved.

(3) Our saliency framework account for feature search, conjunction search and many search asymmetries straightforwardly. We further examine given a saliency map, how attention is directed. We treat this as a multi-bandit decision making problem and propose that attention is directed probabilistically with the strategy of probability matching. We also treat the visual search task as a sequential decision making problem when investigating when subjects terminate a trial. Taken together, we were able to account for many observations of mistakes and response time in visual search tasks.

Together these contributions made efforts toward a unified statistical model of visual attention that not only account for human behavior, but also

allows practical implementation on complex images and videos.

I Introduction

I.A Visual Attention

Visual attention reflects the sampling strategy of the visual system. The surrounding world is of tremendous amount of visual information that the visual system can not fully process. The problem the system thus facing is what to process and what not to, and with how much processing resources. Despite the small amount information the system can handle, sampled by discontinuous saccades, we experience a seamless, continuous world. More importantly, we human, as well as many other animals, survive with this heavily down sampled visual information.

To investigate how visual attention works, is not only of interest to understanding the biological system itself, but also of potential great use for computer vision and graphics. As the biological vision system is currently the only system that "solved" vision, insight of how it manages to heavily down sample the input but successfully recognize objects and configure their spacial positions would be of help to build efficient computational visual systems. Moreover, this down sampling may even enhance performance by filtering out irrelevant information. The understanding of the sampling strategy is thus of interest to computer vision scientists even with unconstrained computational power, which is yet to come. Thus, investigating what the visual system can do and how it achieves it with its resource constraints is of interest to computer vision. On the other hand, understanding the visual system's limitations is of interest to computer graphics. The knowledge about perceptual limitations in visual and audio systems have proven fruitful in image and music compression. Knowing what the visual system care and not care to process can save great effort in computer graphics by rendering images “look” realistic but not necessarily physically correct or even possible.

I.B Previous work

What attracts attention? This question has been of interest to vision researchers for many decades and numerous experiments have been carried out to investigate it. It is well known that a white bar can not hide in the a sea of black bars. No matter how many black distractors you put in there, the white one will “pop out” from the scene and grabs the attention in no time. The reverse works equally fast that a black bar jumps out from a pool of white ones. A vertical bar also pops out from a pool of horizontal bars and vice versa. However, a horizontal white bar does not pop out from a pool of horizontal black bars and vertical white bars. Figure I.1 shows examples of stimuli that illustrate feature search and conjunction search for a horizontal white bar.

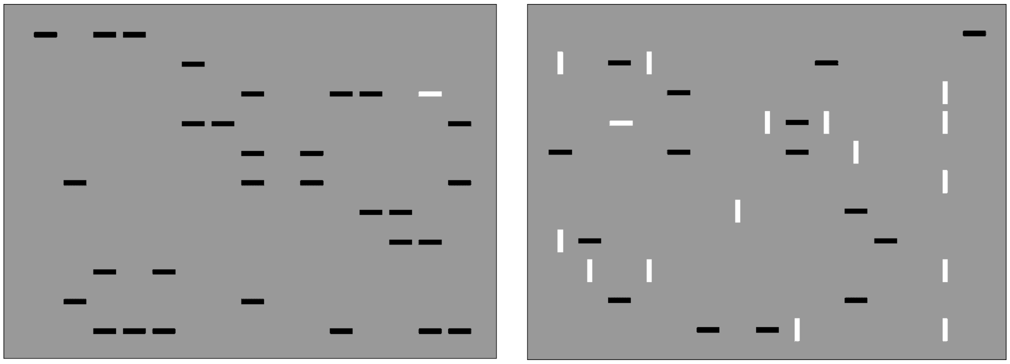

Figure I.1 **Left:** An example of feature search. The white bar pops out and attracts attention automatically. **Right:** An example of conjunction search. The white horizontal bar, although different from all other items, does not attract attention instantly.

Another interesting phenomenon is that sometimes an item A pops out from a pool of item B’s, but an item B does not pop out from a pool of item A’s.

For example, a 10 degree tilted bar takes no time to search from a pool of vertical ones but a vertical in a pool of 10 degree ones requires some effort to notice [75]. Another example with somewhat higher level stimuli is that it takes longer for Caucasians to search for a Caucasian face in a pool of African American faces than the reverse. To make things more complicated, Caucasian basketball fans who are familiar with many African American players do not show this difference [41, 42]. This phenomenon that the searching difficulty changes when the target and the distractors switch the roles has been referred to as search asymmetry [75, 76, 89].

Many computational models have been built to account for various phenomena in the classical paradigm of visual search. We will introduce a few here just to give a flavor of the variety of models in the literature. Treisman's feature integration theory (FIT) [74] and Wolfe's guided search (GS) [86,91,92] are among the most prominent ones and can find their reminiscence in many computational saliency algorithms that will be discussed later.

FIT proposed that processing in one feature space is pre-attentive and parallel while combining features needs attention and is serial. This directly addresses the phenomenon that a feature target pops out but a conjunction target does not. Treisman has also done numerous work in search asymmetry and concluded several categories of search asymmetry including "prototypes do not pop out" and "lack of feature does not pop out" [75, 76].

Guided search model are composed of several components. The input image is first processed in basic feature dimensions in parallel, which resembles FIT. Its output goes through the selective bottleneck of visual attention, which can be mediated by a "guiding representation". Then the selected features/objects proceed to the process of object recognition. GS4 also acknowledges high-level properties such as image statistics and scene analysis, but they are not explicitly modeled. With proper parameters, GS4 accounts for the continuity between parallel search and serial search, and gives the task difficulty a continuous measurement "efficiency". It also accounts for a large body of other observations in visual search.

Bundesen developed a visual attention theory that has a selection component and a classification component [4,5]. The selection follows Luce's selection rule [44] that the probability of an item being attended to is proportional to the product of its sensory evidence and selection bias. The probability of classifying an item as a certain category is proportional to the product of sensory evidence and the category pertinence. This model accounts for many visual attention phenomena as well as the linear response time in visual search tasks.

Zhaoping and colleagues have proposed that pre-attentive computational mechanisms in primary visual cortex create a saliency map [43,96]. In their work, firing rates of output neurons in V1 provide a saliency map, the higher the firing rate, the bigger the salience. Their biologically-based V1 model accounts for a number of phenomena qualitatively, e.g. pop out, the effect of background homogeneity on search difficulty and some of the search asymmetries.

Another family of models are the limited capacity models [36,54,59,85]. The essence of this idea is that all items in the visual field are processed at once by the limited attentional resources. Evidence accumulates at each location for the presence of a target or non-target item. Search terminates when one item crosses the “yes” threshold or all items cross the “no” threshold. The rate of accumulation depends on the amount of attentional resources available to each item. Thus when the amount of resources is fixed, increased set size results in fewer resources per item and slower average speed.

A family of models based on signal detection theory (SDT) can also account for some observations in visual search (see [79] for a review). The assumption here is that the search process monitors the noisy output of a matched filter. The subject's responses and processing time depend on the filtered output of the target vs. the maximum output of the distractors. When the number of distractors increases, there is a larger probability that the maximum of the distractors will come close to that of the target, making the task more difficult. Also, when the distractor is similar to the target, the distribution of outputs from the matched filter overlap and the task again becomes more difficult. Thus SDT models account for both the effect of the number of distractors and the continuum of processing time depending on the discriminability of the target. Indeed, GS4's parallel front end is similar to a SDT model [91].

With the advancement of sophisticated eye movement tracking systems, where visual attention is directed can be explicitly examined by recording where subjects fixate when watching an image or a video. This provides a very different kind of data from classical visual search paradigm. The stimuli are often complex images or videos. Subjects are instructed just to "view" the display or to perform a task such as looking for a target. Their eye fixations are recorded in the meantime. This new kind of data calls for computational models that can take complex stimuli as input. Former visual attention models are mostly developed to account for human behavior in well controlled stimuli such as those shown figure I.1. They are often too specified or complicated to be computational applicable to complex natural images. For example, many models explicitly index each item in the visual search stimuli and examine their orientation, color, etc., which is not possible to apply to complex images without clear boundaries from one item to another. Thus, a relatively new direction of research seen in the recent decade are saliency map algorithms that operates on images and videos. The evaluation, instead of accounting for response time and mistakes in visual search tasks, is to predict people's eye fixations. A saliency model which assign high values to where people look but low values to where people ignore is considered a good model.

Itti and Koch's saliency model [29–31] is one the earliest and the most used for comparison in later work. The model is an implementation of and expansion on the basic ideas first proposed in [38]. The model is inspired by the visual attention literature, such as feature integration theory [74], and ensuring that the model is neurobiologically plausible. The model takes an image as input, which is then decomposed into three channels: intensity, color, and orientation. A center-surround operation, implemented by taking the difference of the filter responses from two scales, yields a set of feature maps. The feature maps for each channel are then normalized and combined across scales and orientations, creating conspicuity maps for each channel. The conspicuous regions of these maps are further enhanced by normalization, and the channels are linearly combined to form the overall saliency map. This process allows locations to vie for conspicuity within each feature dimension, but has separate feature channels contribute to saliency independently. This model has been shown to be successful in predicting human fixations and useful in object detection [30, 31, 55]. However, it has been criticized as being ad hoc, partly because the overarching goal of the system (i.e., what it is designed to optimize) is not specified, and it has many parameters that need to be hand-selected.

Itti and Baldi [28, 32] proposed a Bayesian surprise model for saliency of dynamic scenes. The surprise detectors maintain data model of Poisson distributions at each location over multiple time scales which are updated every time step upon new data. Surprise, which measures how much the current data changes the model, is calculated as KL (Kullbach-Liebler) divergence between the distributions before and after the update.

Several saliency algorithms are based on measuring the complexity of a local region [10, 33, 60, 94]. Yamada and Cottrell [94] measure the variance of 2D Gabor filter responses across different orientations. Kadir and Brady [33] measure the entropy of the local intensity probability distribution. Renninger et al. [60] measure the entropy of local line orientation histograms, and the most salient point at any given time is the one that provides the greatest information gain conditioned on the knowledge obtained during previous fixations. All of these saliency-as-variance/entropy models are based on the idea that the entropy of a feature distribution over a local region measures the richness and diversity of that region (Chauvin et al., 2002), and intuitively a region should be salient if it contains features with many different orientations and intensities. A common critique of these models is that highly textured regions are always salient regardless of their context. For example, human observers find an egg in a nest highly salient, but local-entropy-based algorithms find the nest to be much more salient than the egg [3,18].

Gao and Vasconcelos [18,20] propose an intriguing goal for saliency: classification. That is, a goal of the visual system is to classify each stimulus as belonging to a class of interest or not. This was first used for object detection [18], where a set of features are selected to best discriminate the class of interest (e.g., faces or cars) from all other stimuli, and saliency was defined as the weighted sum of feature responses for the set of features that are salient for that class. This forms a definition that is inherently top-down and goal directed, as saliency is defined for a particular class. In [20], bottom-up saliency is defined using the idea that locations are salient if they differ greatly from their surroundings. They use difference-of-Gaussian (DoG) and Gabor filters, measuring the saliency of a point as the Kullbach-Liebler (KL) divergence between the histogram of filter responses at the point and the histogram of filter responses in the surrounding region. This addresses a problem commonly faced by other models that use linear filter responses as features: highly textured areas always receives high saliency scores. Later, we will discuss a way that our model could address this problem, by using nonlinear features that model complex cells or neurons in higher levels of the visual system.

Oliva and colleagues proposed a probabilistic model for visual search tasks [51,71]. When searching for a target in an image, the probability of interest is the joint probability that the target is present in the current image, together with the target's location (if the target is present), given the observed features. This can be calculated using Bayes rule:

$$p(O=1,L|F,G) = \underbrace{\frac{1}{p(F|G)}}_{\substack{\text{bottom-up saliency}\\ \text{(as defined by Oliva et al.)}}} p(F|O=1,L,G)p(L|O=1,G)p(O=1|G) \quad \text{(I.1)}$$

where $O = 1$ denotes the event that the target is present in the image, L denotes the location of the target when $O = 1$, F denotes the local features at location L, and G denotes the global features of the image. The global features of G represent the scene gist. Experiments show that the gist of a scene can be quickly determined, and the focus of their work largely concerns how this gist affects eye-movements. The first term is independent of the target and is defined as bottom-up saliency; they approximate this conditioned probability distribution using the current image's statistics. The remaining terms respectively address the distribution of features for the target, the likely locations for the target, and the probability of the target's presence, all conditioned on the scene gist. As we will see in Section II.A, our use of Bayes' rule to derive saliency is similar to this approach. However, the probability of interest in the work of Oliva and colleagues is whether a target is present anywhere in the test image, whereas the probability we are concerned with is the probability at each point in the visual field that a target is present. In addition, Oliva and colleagues condition all their probabilities on the values of global features. Conditioning on global features/gist affects the meaning of all terms in equation (I.1), and justifies their use of current image statistics for bottom-up saliency.

Bruce and Tsotsos defined bottom-up saliency based on maximum information sampling [3]. Information, in this model, is computed as Shannon's self-information, $-\log p(F)$, where F is a vector of the visual features observed at a point in the image. The distribution of the features is estimated from a neighborhood of the point, which can be as large as the entire image. When the neighborhood of each point is indeed defined as the entire image of interest, as implemented in [3], the definition of saliency becomes identical to the bottom-up saliency term in the work of Oliva and colleagues in equation (I.1) [51, 71]. It is worth noting, however, that the feature spaces being used are different. Oliva and colleagues used biologically-inspired linear filters of different orientations and scales. These filter responses are known to correlate with each other; for example, a vertical bar in the image will activate a filter tuned to vertical bars but also one tuned to 45 degree tilted bars. The joint probability of the entire feature vector is estimated using multivariate Gaussian distributions [51] and later multivariate generalized Gaussian distributions [71]. Bruce and Tsotsos [3], on the other hand, employed features that were learned from natural images using independent component analysis. These have been shown to resemble the receptive fields of neurons in primary visual cortex (V1), and their responses have the desired property of sparsity. Furthermore, the features learned are approximately independent, so the joint probability of the features is simply the product of the marginal probability of each feature, simplifying the probability estimation without unreasonable independence assumptions.

These saliency models are quite successful in accounting human eye fixation data when viewing images and videos. However, the underlying theories are often not able to account for the rich data in visual search. Although the algorithms are capable to predict where people are likely to look at when viewing complex images and videos, they do not predict the comprehensive behavior when viewing simple stimuli used in classical visual search paradigms. Some theoretic principles seem to be missing compare to visual attention models.

Although searching on well controlled simple stimuli and viewing complex scenes are very different tasks, they both reflect how visual attention works. Some saliency algorithms have made some effort to work also with visual search stimuli. However, as some of the tasks demands higher level processing on such as faces, generic algorithms that use linear filters as front end preprocessing are almost impossible to account for the richness of these data. In this dissertation, we make an effort of develop a visual attention theory by considering the goal of the visual system, which is able to account for many visual search behaviors. We then implement efficient algorithms based on the theory to work with complex images and videos. That is, we try to develop saliency algorithms based on a visual attention theory which can account for the visual search data. Although the implementation can not realize all the predictions of the theory itself due to the limitations of computational power and lack of knowledge in how higher level visual processing works with highly nonlinear stimuli, it works as well as state of art saliency algorithms in predicting human eye fixation data while being more efficient and biologically plausible. This is our preliminary effort to combine the virtues of previous work in both visual search models and bottom-up saliency algorithms. We hope to provide some insight into how visual attention works by developing a principled theoretic framework that takes both literature into account.

I.C Dissertation Outline

The remaining chapters are organized as follows.

Chapter II We develop our theoretic framework of saliency. In particular, our framework takes natural statistics into account. Predictions of saliency models using natural image statistics vs. current scene statistics are compared. Our model accounts straightforwardly for feature search, conjunction search, and many search asymmetries.

Chapter III Bottom-up saliency is implemented here for complex static color images. Features and their probability distributions are learned from natural images with linear efficient coding theory. The result is evaluated on human fixation data while free viewing images and is compared to previous bottom-up saliency algorithms. Our implementation performs as well as the state of art algorithm while being more biologically plausible and computationally efficient.

Chapter IV Bottom-up saliency is implemented here on color videos. For computational efficiency, we used linearly separable spatiotemporal filter responses as features, and we designed special temporal filter to allow very fast calculation of filter responses. Our algorithm again performs as well as the start of art algorithm while being much more efficient. We also implemented real time version with some simplification assumptions. Without compromising much in the performance of predicting human fixations, it allows us to implement a saliency oriented robotic camera. The results show great improvement in chances of seeing people.

Chapter V Saliency is not the whole story. Given a pre-attentive saliency map, where to look is a decision to make, as well as when to stop and report target absent the target is not found. In this chapter, we will explore the decision making aspects in visual search tasks. We propose that attention is directed probabilistically according to the saliency map, sharing the characteristic of probably matching that observed in decision making behavior in bandit problems. Furthermore, when to stop searching is treated as a sequential decision making problem. Together we are able to account for many interesting phenomena in visual search tasks qualitatively without fitting any parameters.

Chapter VI We conclude the whole dissertation with a brief summary of contribution and discusses possible directions for future work.

II A Baysian Framework for Saliency: Information Attracts Attention

II.A Saliency is Probability

We propose that one goal of the visual system is to find potential targets that are important for survival, such as food and predators. To achieve this, the visual system must actively estimate the probability of a target at every location given the visual features observed. We propose that this probability is visual saliency.

To formalize this, let z denote a point in the visual field. A point here is loosely defined; in the implementation described in Chapter III a point corresponds to a single image pixel. (In other contexts, a point could refer other things, such as an object [95].) We let the binary variable C_z denote whether or not point z belongs to a target class, let L_z denote the location of point z (i.e., the pixel coordinates of the point z), and let F_z denote the visual features of point z. Saliency is then defined as $p(C_z = 1|F_z = f_l, L_z = l)$, where the value of f_l represents the visual features observed at image location l. This probability can be calculated using

Bayes' rule:

$$\begin{aligned} s_z &= p(C_z = 1 | F_z = f_l, L_z = l) \\ &= \frac{p(F_z = f_l, L_z = l | C_z = 1) p(C_z = 1)}{p(F_z = f_l, L_z = l)}. \end{aligned} \tag{II.1}$$

Assume for simplicity that features and location are independent and conditionally independent given $C_z = 1$:

$$p(F_z = f_l, L_z = l) = p(F_z = f_l) p(L_z = l), \tag{II.2}$$

$$p(F_z = f_l, L_z = l | C_z = 1) = p(F_z = f_l | C_z = 1) p(L_z = l | C_z = 1). \tag{II.3}$$

This entails the assumption that the distribution of a feature does not change with location. For example, (II.2) implies that a random point in the left visual field is just as likely to be green as a random point in the right visual field. Furthermore, (II.3) implies (for instance) that a point on a target in the left visual field is just as likely to be green as a point on a target in the right visual field[1]. With these independence assumptions, (II.1) can be rewritten as:

$$s_z = \frac{p(F_z = f_l | C_z = 1) p(L_z = l | C_z = 1) p(C_z = 1)}{p(F_z = f_l) p(L_z = l)} \tag{II.4}$$

$$= \frac{p(F_z = f | C_z = 1)}{p(F_z = f)} \cdot \frac{p(L_z = l | C_z = 1) \cdot p(C_z = 1)}{p(L_z = l)} \tag{II.5}$$

$$= \underbrace{\frac{1}{p(F_z = f)}}_{\substack{\text{Independent} \\ \text{of target} \\ \text{(bottom-up saliency)}}} \cdot \underbrace{\underbrace{p(F_z = f | C_z = 1)}_{\text{Likelihood}} \cdot \underbrace{p(C_z = 1 | L_z = l)}_{\text{Location prior}}}_{\substack{\text{Dependent on target} \\ \text{(top-down knowledge)}}} \tag{II.6}$$

To compare this probability across locations in an image, it suffices to estimate the log probability (since log is a monotonically increasing function). For this reason, we take the liberty of using the term saliency to refer both to s_z and

[1] The extent to which these two assumptions are true depends on the feature space. For example, illumination may not be invariant to locations: as sunshine normally comes from above, the upper part of the visual field is likely to be brighter. But illumination contrast features, such as the response to a DoG (Difference of Gaussians) filter, would be more invariant to location changes.

to $\log s_z$, which is given by:

$$\log s_z = \underbrace{-\log p(F_z = f)}_{\text{Self-information}} + \underbrace{\log p(F_z = f | C_z = 1)}_{\text{Log likelihood}} + \underbrace{\log p(C_z = 1 | L_z = l)}_{\text{Location prior}}. \quad \text{(II.7)}$$

The first term on the right side of this equation, $-\log p(F_z = f_l)$, depends only on the visual features observed at the point, and is independent of any knowledge we have about the target class. In information theory, $-\log p(F_z = f_l)$ is known as the *self-information* of the random variable F_z when it takes the value f_l. Self-information increases when the probability of a feature decreases—in other words, rarer features are more informative. We have discussed self-information earlier in the context of previous work, but as we will see later, our use of self-information differs from previous approaches.

The second term on the right side of (II.7), $\log p(F_z = f_l | C_z = 1)$, is a log-likelihood term that favors feature values that are consistent with our knowledge of the target. For example, if we know that the target is green, then the log-likelihood term will be much larger for a green point than for a blue point. This definition of the top-down effect when searching for a known target is consistent with the finding that human eye movement patterns during iconic visual search could be accounted for by a maximum likelihood procedure for computing the most likely location of a target [58].

The third term, $\log p(C_z = 1 | L_z = l)$, is independent of visual features and reflects any prior knowledge of where the target is likely to appear. It has been shown that if the observer is given a cue of where the target is likely to appear, the observer attends to that location [57].

After omitting the location prior from (9), the equation for saliency has just two terms, the self-information and the log-likelihood, which can be combined:

$$\begin{aligned}
\log s_z &= \underbrace{-\log p(F_z{=}f_l)}_{\substack{\text{Self-information}\\\text{(bottom-up saliency)}}} + \underbrace{\log p(F_z{=}f_l|C_z{=}1)}_{\substack{\text{Log likelihood}\\\text{(top-down knowledge)}}} && \text{(II.8)}\\
&= \log \frac{p(F_z{=}f_l|C_z{=}1)}{p(F_z{=}f_l)} && \text{(II.9)}\\
&= \underbrace{\log \frac{p(F_z{=}f_l, C_z{=}1)}{p(F_z{=}f_l)p(C_z{=}1)}}_{\substack{\text{Pointwise mutual information}\\\text{(overall saliency)}}} \quad . && \text{(II.10)}
\end{aligned}$$

The resulting expression, which is called the *pointwise mutual information* between the visual feature and the presence of a target, is a single term that expresses overall saliency. Intuitively, it favors feature values that are more likely in the presence of a target than in a target's absence.

When the organism is not actively searching for a particular target (the *free viewing* condition), the organism's attention should be directed to any *potential* targets in the visual field, despite the fact that the features associated with the target class are unknown. In this case, the log-likelihood term in (II.7) is unknown, so we omit this term from our calculation of saliency. (This can also be thought of as assuming that for an unspecified target, the likelihood distribution is uniform over feature values.) In this case, the overall saliency reduces to just the self-information term: $\log s = -\log p(F_z{=}f_l)$. We take this to be our definition of bottom-up saliency. It implies that the rarer a feature is, the more it will attract our attention. This definition of saliency explains many observations in the visual search paradigm, such as the search asymmetry between feature presence versus absence, between prototypes versus non-prototype exemplars, and between other-race versus same-race faces [95]. (See Section II.B for more details.)

Note that all of the probability distributions described here should be learned by the visual system through experience. They should reflect the natural statistics of the environment and the learning history of the organism, rather than just the statistics of the current image.

In summary, calculating the probability of a target at each point in the visual field leads naturally to the estimation of information content. In the free-viewing condition, when there is no specific target, saliency reduces to the self-information of a feature. This implies that when the attention is directed only by bottom-up saliency, moving one's eyes to the most salient points in an image can be regarded as maximizing information sampling. This is consistent with the basic assumption of Bruce and Tsotsos [3]. When a particular target is being searched for, on the other hand, our model implies that the best features to attend to are those that have the most mutual information with the target. This has been shown to be very useful in object detection with objects such as faces and cars [77].

We have been discussing saliency only in the context of two categories: target and non-target. It can be generalized to multiple categories associated with different importance. Assume that there are many categories indexed by i, each associated with reward r_i when correctly identified. Saliency is then defined as the expected reward, or utility, of attending to a point z:

$$\begin{aligned} u_z &= \Sigma_i \; p(C_z = i | F_z = f, L_z = l) \; r_i && \text{(II.11)} \\ &= \Sigma_i \; \frac{p(F_z = f | C_z = i) p(C_z = i | L_z = l)}{p(F_z = f)} \; r_i && \text{(II.12)} \\ &= \frac{1}{p(F_z = f)} \; \Sigma_i \; p(F_z = f | C_z = i) p(C_z = i | L_z = l) \; r_i && \text{(II.13)} \end{aligned}$$

This can again be decomposed to bottom-up saliency $\frac{1}{p(F_z=f)}$ which is independent of categories and top-down component that combines the knowledge of the appearance of each category, their likely locations and associated reward.

II.B Test Image Statistics vs. Natural Scene Statistics

In our framework, the probability terms are not constrained to the current visual scene. The probability distributions are learned by the visual system through experience. They should reflect the natural statistics of the environment and the learning history of the organism.

II.B.1 Comparison with previous work

All of the existing bottom-up saliency models described in Section I.B compute saliency by comparing the feature statistics at a point in a test image with either the statistics of a neighborhood of the point or the statistics of the entire test image. When calculating the saliency map of an image (the saliency value at every point in the image), these models only consider the statistics of the current test image. In contrast, our definition of saliency compares the features observed at each point in a test image to the statistics of natural scenes. An organism would learn these natural statistics through a lifetime of experience with the world; in our implementation, we obtained them from a collection of images of natural scenes (see Chapter III). As explained in Section II.A, our definition of saliency was itself derived from a simple intuitive assumption about a goal of the visual system.

Our formula for bottom-up saliency is similar to the one in the work of Oliva and colleagues work [51, 71] and the one in [3] in that they are all based on the notion of self-information. However, the differences between image statistics and natural statistics lead to radically different kinds of self-information. Briefly, the motivation for using self-information with current image statistics is that a foreground object is likely to have features that are distinct from the features of the background. The idea that the saliency of an item is dependent on its deviation from the average statistics of the image can find its roots in the visual search model proposed in [61], which accounted for a number of motion pop out phenomena, and can be seen as a generalization of the center-surround-based saliency found in [38]. Our use of natural statistics for self-information, on the other hand, corresponds to the intuition that since targets are observed less frequently than background during an organism's lifetime, rare features are more likely to indicate targets. The idea that infrequent features attract attention has its origin in findings that novelty attracts the attention of infants [9, 13, 14, 16] and that novel objects are faster to find in visual search tasks (see [89] for a review). This fundamental difference in motivation between our model and existing saliency models leads to very different predictions about what attracts attention.

In the next two sections, we show that by using natural image statistics, our model provides a simple explanation for a number of psychophysical phenomena that are difficult to account for using the statistics of either a local neighborhood in the test image or the entire test image. In addition, since natural image statistics are computed well in advance of the test image presentation, in our model the estimation of saliency is strictly local and efficient.

II.B.2 Feature target is salient but conjunction target is not (with exceptions)

Now we can examine the salience of the target in a traditional feature search and a traditional conjunction search. In a feature search, such as a red dot in a field of green dots will have very strong local contrast in the color dimension, while the green dots do not. High contrast has low probability in natural scenes[2]. Thus the red dot has a high salience while the green dots have low salience, causing the red dot to attract attention instantly; it pops out.

The conjunction search is slightly more tricky. Taking the example of searching for a red horizontal bar is red vertical bars and green horizontal bars. Recall that the salience of the target is $\frac{1}{P(F)} = \frac{1}{P(F_1=f_1,F_2=f_2)}$, where F_1 and F_2 are the two feature dimensions involved in the search (color and orientation in our example). If F_1 and F_2 are independent, $P(F_1 = f_1, F_2 = f_2) = P(F_1 = f_1) \cdot P(F_2 = f_2)$. The conjunction target (the red horizontal bar) is not rare in either color or orientation contrast. Thus it is just as salient as all the distractors and will not pop out. Color and orientation are likely to be independent in natural statistics because

[2] The histograms of the filter response of local (color) contrast features such as difference of gaussians bears a shape of sparse distribution similar to Laplacian distributions.

they seem to be generated by not so related physical processes. Color is related to surface properties of objects while orientation is more related to gravity. Other feature dimensions, however, can be very related and statistically dependent that the decomposition of the joint probability to the product of individual probabilities can no longer a good approximation. For example, McLeod reported that the search for the conjunction target of form and motion can be very efficient [46]. The form and motion can be somewhat related in the nature. For example, almost all animals which move are bilateral symmetric or radius symmetric. The shape of the body is very crucial to how fast the animal can move. For objects that do not move voluntarily, when they move because of gravity or wind etc. they are often aligned in a certain way to the direction of the motion because of the aerodynamics. Thus there are reasons to believe that the statistics from these two feature dimensions are quite correlated. The conjunction target could then potentially have a high salience and pop out from the rest.

It is hard to speculate or measure the statistical dependence on many other feature dimensions. Also there must be other factors affecting when the search is fast or slow in both feature and conjunction searches, not to mention subjects can develop various speed strategies over trials. Our formulation of saliency seems to have the potential of providing part of the underlying driving force. But we are conservative about to what extent it can account for various phenomena in the feature and conjunction search.

However, there is a difference between the predictions on conjunction search when using natural statistics and using current scene statistics. When two features are independent in natural statistics, our model predicts that the conjunction target will not pop out. Saliency models with current scene statistics, on the other hand, predict that the conjunction target will pop out because its joint features are unique in the scene. Take for example the conjunction search stimuli in figure I.1, the target has the features white & horizontal while the distractors have the feature white & vertical or black & horizontal. Thus, the target bears special joint features that is of low probability in the scene, and should stand out from the distractors. Thus, saliency based on current scene statistics can not account for inefficient conjunction search without extra assumptions.

II.B.3 Visual search asymmetry - lower probability item is easier to search

When the probability of a feature is based on the distribution of features in the current test image, as in other models, a straightforward consequence is that if all items in an image are identical except for one, this odd item will have the highest saliency and thus attract attention. For example, if an image consists of a number of vertical bars with one bar that is slightly tilted from the vertical, the tilted bar “pops out” and attracts attention almost instantly [75]; see Figure II.1, left, for an illustration. If, on the other hand, an image consists of a number of slightly-tilted-from-vertical bars with one vertical, the statistics of the current image predicts the same pop out effect for the vertical bar. However, this simply is not the case as humans do not show the same pop-out effect: it requires more time and effort for humans to find a vertical bar within a sea of tilted bars [75]; see Figure II.1, right, for an illustration. This is known in the visual search literature as *search asymmetry*, and this particular type of example corresponds to findings that “prototypes do not pop out” because the vertical is regarded as a prototypical orientation [75, 76, 89].

Unlike saliency measures based on the statistics of the current image or an image neighborhood, saliency based on natural statistics readily predicts this search asymmetry. The vertical orientation is prototypical because it occurs more frequently in natural images than the tilted orientation [78]. Thus, the vertical bar will have smaller salience than the surrounding tilted bars, so it will not attract attention as strongly.

Another visual search asymmetry exhibited by human subjects involves long and short line segments. Saliency measures based on test image statistics

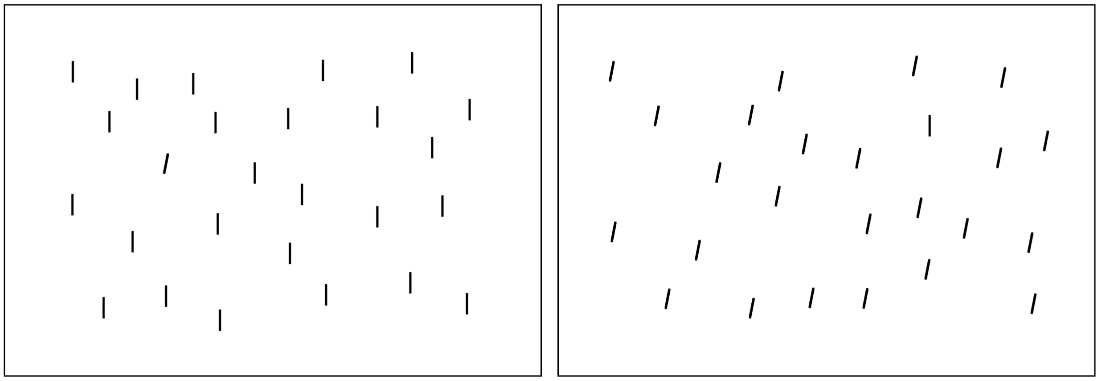

Figure II.1 Illustration of the “prototypes do not pop out” visual search asymmetry [75]. *Left:* A tilted bar in a sea of vertical bars pops out. *Right:* A vertical bar in sea of tilted bars does not pop out.

or local statistics predict that a long bar in a group of short bars (illustrated on the left in Figure II.2) should be as salient as a short bar in a group of long bars (illustrated on the right in Figure II.2). However, it has been shown that humans find a long bar among short bar distractors much more quickly than they find a short bar among long bars [75]. Saliency based on natural statistics readily predicts this search asymmetry, as well. Due to scale invariance, the probability distribution over the lengths of line segments in natural images follows the power law [62]. That is, the probability of occurrence of a line segment of length v is given by $p(V = v) \propto \frac{1}{v}$. Since longer line segments have lower probability in images of natural scenes, our saliency model implies that longer line segments will be more salient.

Visual search asymmetry is also observed for higher level stimuli such as roman letters, Chinese characters, animal silhouettes, and faces. For example, people are faster to find a mirrored letter in normal letters than the reverse [17]. People are also faster at searching for an inverted animal silhouette in a sea of upright silhouettes than the reverse [89], and faster at searching for an inverted face in a group of upright faces than the reverse [50]. These phenomena have

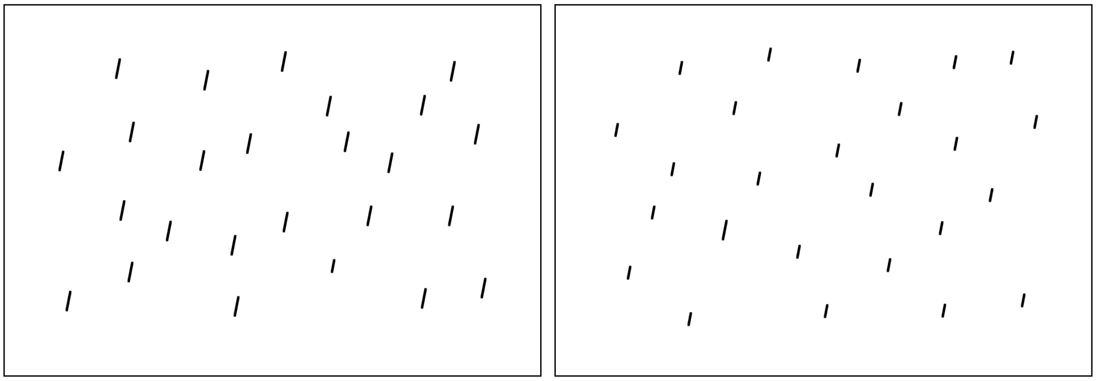

Figure II.2 Illustration of a visual search asymmetry with line segments of two different lengths [75]. *Left:* A long bar is easy to locate in a sea of short bars. *Right:* A short bar in a sea of long bars is harder to find.

been referred to as “the novel target is easier to find.” Here, “novel” means that subjects have less experience with the stimulus, indicating a lower probability of encounter during development. This corresponds well with our definition of bottom-up saliency, as novel items are more salient by definition.

If the saliency of an item depends upon how often it has been encountered by an organism, then search asymmetry should vary among people with different experience with the items involved. This seems to indeed be the case. Modified/inverted Chinese characters in a sea of real Chinese characters are faster to find than the reverse situation for Chinese readers, but not for non-Chinese readers [66, 84]. Levin found an “other-race advantage” as American Caucasians are faster to search for an African-American face among Caucasian faces than to search for a Caucasian face among African-American faces [41]. This is consistent with what our model would predict for American Caucasian subjects that have more experience with Caucasian faces than with African-American faces. However, Levin also found that Caucasian basketball fans who are familiar with many African-American basketball players do not show this other-race search advantage [42]. These seem to provide direct evidence that experience plays an important role in saliency, and that the statistics of the current image alone cannot possibly be the whole story.

II.C Discussion

In this chapter, we hypothesized that a goal of the visual system is to find potential targets such as prey as predators. We inferred what should be calculated to achieve this goal and proposed that, without knowledge of target's location, bottom-up saliency is self-information of visual features and overall saliency is pointwise mutual information between the visual features and the target when a target is being searched. The bottom-up saliency is the in this session. It shares the same formula $\frac{1}{p(F=f)}$ as that in [3, 51, 71]. However, the probability distribution of the features is drawn from natural statistics in our model, but is constrained to the image of question in previous works. We showed our framework straightforwardly account for feature and conjunction search, as well as many search asymmetries, while saliency based on current scene saliency can not predict the same phenomena without extra assumptions. This provides some evidence that current image statistics alone is not sufficient and natural statistics and developmental experience also plays an important role.

II.D Acknowledgment

Chapter II, in part, is a reprint of the paper in preparation "A Bayesian Framework for Saliency", co-authored with Honghao Shan, Tim K. Marks, Matthew H. Tong and Garrison W. Cottrell. The dissertation author is the primary investigator and author of these two papers.

III Static Image Saliency

In this chapter, we will implement a bottom-up saliency algorithm which takes color static images as input and calculates their saliency maps (the saliency at every pixel in an image). Given a probabilistic formula for saliency, such as the one we derived in Section II.A, there are two key factors that affect the final results of a saliency model when operating on an image. One is the feature space, and the other is the probability distribution over the features.

In most existing saliency algorithms, the features are calculated as responses of biologically plausible linear filters, such as DoG (difference of Gaussians) filters and Gabor filters [18, 20, 30, 31, 51, 71]. In [3], the features are calculated as responses filters learned from natural images using ICA (independent component analysis). In this paper, we conduct experiments with both kinds of features.

We describe our algorithm for estimating the bottom-up saliency that we derived in Section II.A, $-\log p(F_z = f_l)$. Here, a point z corresponds to a pixel in the image (and l represents the location of that pixel in the image). For the remainder of the paper, we will drop the subscripts z and l for notational simplicity. In this algorithm, F is a random vector of filter responses, $F = [F_1, F_2, \ldots]$, where the random variable F_i represents the response of the ith filter at a pixel, and $f = [f_1, f_2, \ldots]$ are the values of these filter responses at this location.

III.A Experiment 1: DoG filters

Many existing models use a collection of DoG (difference of Gaussians) and/or Gabor filter responses as the first step of processing the input images [18,20,30,31,51,71]. These filter responses are popular due to their resemblance to the receptive fields of neurons in the early stages of the visual system, namely the lateral geniculate nucleus of the thalamus (LGN) and primary visual cortex (V1).

Let r, g and b denote the red, green, and blue components of an input image pixel. The intensity (I), red/green (RG), and blue/yellow (BY) channels are calculated as:

$$I = r + g + b, \qquad RG = r - g, \qquad BY = b - \frac{r+g}{2} - \frac{\min(r,g)}{2}. \tag{III.1}$$

We apply difference of Gaussians (DoG) filters to each of these channels and use the filter responses as features. The DoG filters are generated by[1]

$$g = \frac{1}{\sigma^2}\exp\left(-\frac{x^2+y^2}{\sigma^2}\right) - \frac{1}{(1.6\sigma)^2}\exp\left(-\frac{x^2+y^2}{(1.6\sigma)^2}\right). \tag{III.2}$$

To each of the 3 channels (I, RG, and BY), we apply 4 DoG filters (shown in Figure III.1), using equation (III.2) with 4 scales ($\sigma = 4, 8, 16$ or 32 pixels), resulting in 12 filters in total. This led to a total of 12 feature response maps, 4 from each channel. By computing these feature response maps on a set of 138 images of natural scenes (photographed by the first author), we obtained an estimate of the probability distribution over the observed values of each of the 12 features. To this estimated distribution for each feature F_i, we used an algorithm proposed by Song [67] to fit a zero-mean generalized Gaussian distribution, also known as an exponential power distribution:

$$p(f;\sigma,\theta) = \frac{\theta}{2\sigma\Gamma(\frac{1}{\theta})}\exp\left(-\left|\frac{f}{\sigma}\right|^{\theta}\right). \tag{III.3}$$

In this equation, Γ is the gamma function, θ is the shape parameter, σ is the scale parameter, and f is the filter response. This resulted in one shape parameter, θ_i,

[1] Equation (III.2) is adopted from the function filter_DOG_2D, from *Image Video toolbox for Matlab* by Piotr Dollar. The toolbox can be found at http://vision.ucsd.edu/∼pdollar/toolbox/doc/.

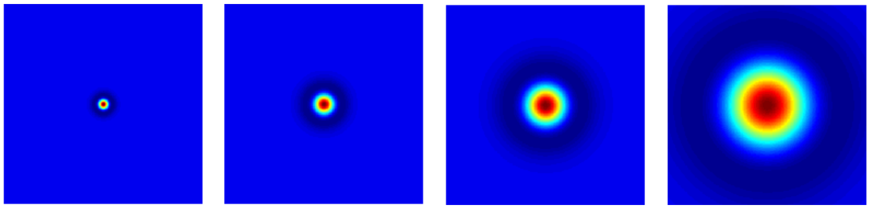

Figure III.1 The four scales of difference of Gaussians (DoG) filters that are applied to each channel.

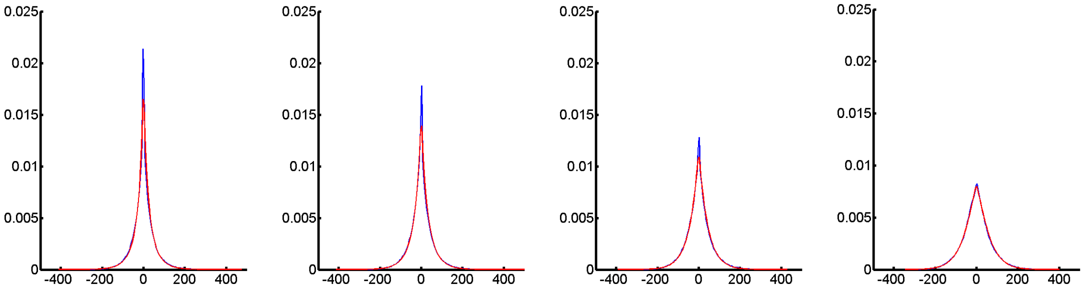

Figure III.2 The graphs show the distribution of filter responses for the 4 DoG filters on the intensity channel collected from the set of natural images (blue line), and the fitted generalized Gaussian distributions (red line).

and one scale parameter, σ_i, for each of the 12 filters: $i = 1, 2, ..., 12$. Figure IV.2 shows the distributions of the 4 DoG filter responses on the intensity (I) channel across the training set of natural images, and the fitted generalized Gaussian distributions. As the figure shows, the generalized Gaussians provide an excellent fit to the data.

Taking the logarithm of (III.3), we obtain the log probability over the possible values of each feature:

$$\log p(F_i = f_i) = \log \theta_i - \log 2 - \log \sigma_i - \log \Gamma\left(\frac{1}{\theta_i}\right) - \left|\frac{f_i}{\sigma_i}\right|^{\theta_i} = -\left|\frac{f_i}{\sigma_i}\right|^{\theta_i} + const. \quad \text{(III.4)}$$

To simplify the computations, we assume that the 12 filter responses are independent. Hence the total bottom-up saliency of the point takes the form:

$$-\log p(F = f) = \sum_{i=1}^{12} \left| \frac{f_i}{\sigma_i} \right|^{\theta_i} + const. \tag{III.5}$$

III.B Experiment 2: Linear ICA Filters

In our final formula for bottom-up saliency (IV.8), we assumed independence between the filter responses. However, this assumption does not always hold. For example, a bright spot in an image will generate a positive filter response for multiple scales of DoG filters. In this case the filter responses, far from being independent, are highly correlated. It is not clear how this correlation affects the saliency results when a weighted sum of filter responses is used to compute saliency (as in [30,31]) or when independence is assumed in estimating probability (as in our case). Torralba et al. [71] used a multivariate generalized Gaussian distribution to fit the joint probability of the filter responses. However, although the response of a single filter has been shown to be well fitted by a univariate generalized Gaussian distribution, it is less clear that the joint probability follows a multivariate generalized Gaussian distribution. Also, much more data is necessary for a good fit of a high-dimensional probability distribution than for one-dimensional distributions. It has been shown that estimating the moments of a generalized Gaussian distribution has its limitations even for the one-dimensional case [67], and it is much less likely to work well for the high-dimensional case.

To obtain the linear features used in their saliency algorithm, Bruce and Tsotsos [3] applied independent component analysis (ICA) to a training set of natural images. This has been shown to yield features that qualitatively resemble those found in the visual cortex [2,52]. Although the linear features learned in this way are not entirely independent, they have been shown to be independent up to third-order statistics [83]. Such a feature space will provide a much better match for the independence assumptions we made in (IV.8). Thus, in this method we follow [3] and derive complete ICA features to use in our measure of saliency. It

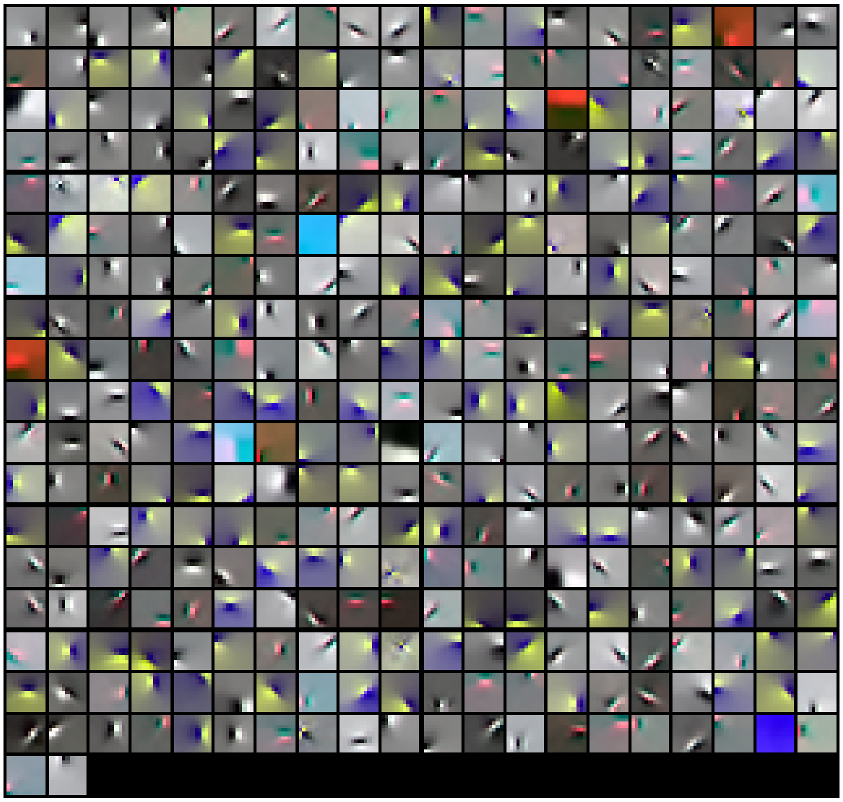

Figure III.3 The 362 linear features learned by applying a complete independent component analysis (ICA) algorithm to 11 × 11 patches of color natural images from the Kyoto dataset.

is worth noting that although Bruce and Tsotsos [3] use a set of natural images to train the feature set, they determine the distribution over these features solely from a single test image when calculating saliency.

We applied the FastICA algorithm [26] to 11-pixel × 11-pixel color natural image patches drawn from the Kyoto image dataset [82]. This resulted in $11 \cdot 11 \cdot 3 - 1 = 362$ features[2]. Figure III.3 shows the linear ICA features obtained from the training image patches.

Like the DoG features from Section III.A, the ICA feature responses to

[2] The training image patches are considered as $11 \cdot 11 \cdot 3 = 363$-dimensional vectors, z-scored to have zero mean and unit standard deviation, then processed by principal component analysis (where one dimension is lost).

natural images can be fitted very well using generalized Gaussian distributions, and we obtain the shape and scale parameters for each ICA filter by fitting its response to the ICA training images. The formula for saliency is the same as in Method 1 (equation IV.8), except that the sum is now over 362 ICA features (rather than 12 DoG features).

Some examples of bottom-up saliency maps computed using the algorithms from Methods 1 and 2 are shown in Figure III.4. Each row displays an original test image, the same image with human fixations overlaid as red crosses, and the saliency maps on the image computed in Method 1 and Method 2. This figure is included for the purpose of qualitative comparison; the next section provides a detailed quantitative evaluation.

III.C Results

III.C.1 Evaluation method and the center bias

ROC area

Several recent publications [3,20,22,35] use the ROC area metric proposed by Tatler et al. [69] to evaluate eye fixation prediction. Using this method, the saliency map is treated as a binary classifier on every pixel in the image; pixels with larger saliency values than threshold are classified as fixated while the rest are classified as non-fixated. Human fixations are used as ground truth. By varying the threshold, an ROC curve can be drawn and the area under the curve indicates how well the saliency map predicts actual human eye fixations. This measurement has the desired characteristic of transformation invariance, in that only the rank matters.

Assessing performance in this manner runs into problems because most human fixation data sets collected with head mounted eye tracking system have a strong center bias. This bias is partly due to factors related to the set up of the experiment, such as subjects being centered with respect to the center of the

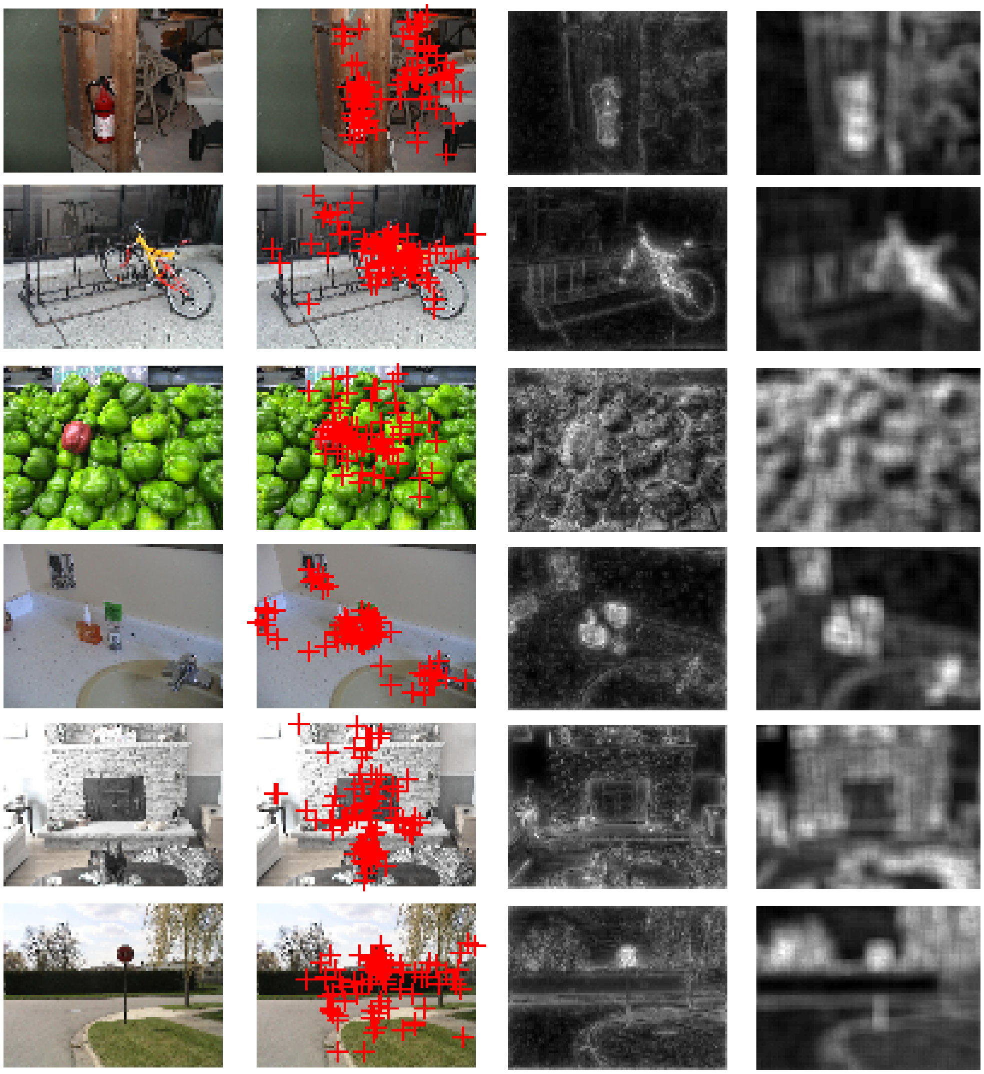

Figure III.4 Each row contains, from left to right: An original test image; the same image with human fixations (from [3]) shown as red crosses; the saliency map produced by our algorithm with DoG filters and with ICA features.

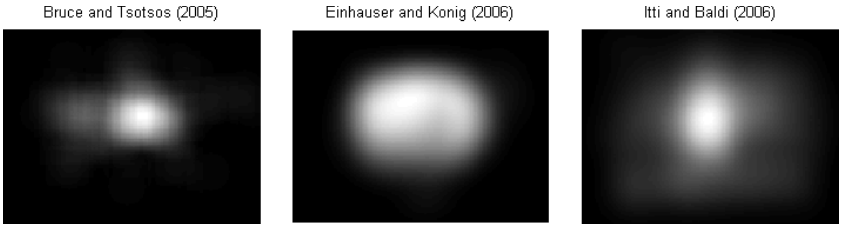

Figure III.5 Plots of all human eye fixation locations in three data sets. *Left:* Subjects viewing color images [3]; *Middle:* Subjects viewing grey images [81]; *Right:* Subjects viewing color videos [28].

screen and framing effects caused by the monitor, but also reflects the fact that human photographers tend to center objects of interest [56,69]. Figure III.5 shows the strong center bias of eye fixations from free-viewing color static images (data from [3]), gray static images (data from [81]) and videos (data from [28]). In fact, simply using a Gaussian blob centered in the middle of the image as the saliency map produces excellent results. For example, on the data set collected in [3], a Gaussian blob fitted to the human eye fixations for that set has an ROC area of 0.80, exceeding the reported results of 0.75 [3] and 0.77 [19] on this data set.

KL divergence

Itti and colleagues make use of the Kullback-Leibler (KL) divergence between the histogram of saliency sampled at eye fixations and that sampled at random locations as the evaluation metric for their dynamic saliency [28,32]. If an algorithm is performing significantly better than chance, the saliency computed at human-fixated locations should be higher than that computed at random locations, leading to a high KL divergence between the two histograms. The KL divergence between two distributions, similar to the ROC measurement, has the desired property of transformation invariance. In [28,32], the random locations are drawn from a uniform spatial distribution over each image frame. Like the ROC performance measurement, the KL divergence awards excellent performance to a

Gaussian blob due to the center bias of the human fixations. The Gaussian blob discussed earlier (trained on the [3] data) yields a KL divergence of 0.44 on the data set of Itti and Baldi [28], exceeding their reported result of 0.24. Thus, both the ROC and KL measurements are strongly sensitive to the effects of the center bias.

Edge effects

These findings imply that models which make use of a location prior (discussed in Section II.A) would better model human behavior. Since all of these models [3, 20, 29, 31] calculate saliency at each pixel without regard to the pixel's location, it would appear that both the ROC measurement and the KL divergence provide a fair comparison between models—no model takes advantage of this additional information.

However, both measures are corrupted by an edge effect due to variations in the handling of invalid filter responses at the borders of images. When an image filter lies partially off the edge of an image, the filter response is not well defined and various methods are used to deal with this problem. Figure III.6 shows the average of all the saliency maps using each of the algorithms of [3, 20, 30] on the data set of Bruce and Tsotsos [3]. It is clear from Figure III.6 that all three algorithms have borders with decreased saliency, but to varying degrees. These border effects introduce an implicit center bias on the saliency maps; these "cool borders" result in the bulk of salience being located at the center of the image. Because different models are affected by these edge effects to varying degrees, it is difficult to determine using the aforementioned measures whether the difference in performance between models is due to the models themselves, or merely due to edge effects[3].

Figure III.7 illustrates the impact that varying amounts of edge effects can

[3] When comparing different feature sets within same model, edge effects can also make it difficult to assess which features are best to use; larger filters result in a smaller valid image after convolution, which can artificially boost performance.

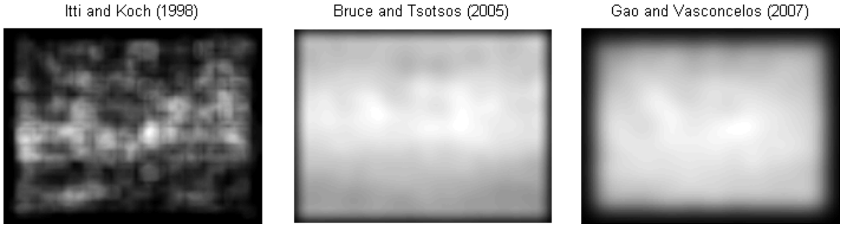

Figure III.6 The average saliency maps of three recent algorithms on the stimuli (120 color images) used in collecting human fixation data by Bruce and Tsotsos [3]. Left: Itti et al. [31]; Middle: Bruce and Tsotsos [3]; Right: Gao and Vasconcelos [20].

have on the ROC area evaluation score by examining the performance of dummy saliency maps that are all 1's except for a border of 0's. The map with a four-pixel border yields an ROC area of 0.62, while the map with an eight-pixel border has an area of 0.73. All borders are small relative to the 120 by 160 pixel saliency map and for these measurements, we assume that the border points are never fixated by humans, which corresponds well with actual human fixation data. A dummy saliency map of all 1's with no border has a baseline ROC area of 0.5.

The KL measurement, too, is quite sensitive to how the filter responses are dealt with at the edges of images. Since the human eye fixations are rarely near the edges of the test images, the edge effects primarily change the distribution of saliency of the random samples. For the dummy saliency maps used in Figure III.7, the baseline map (of all 1's) gives a KL divergence of 0, the four-pixel-border map gives a KL divergence of 0.12, and the eight-pixel-border map gives a KL divergence of 0.25.

While this dummy example presents a somewhat extreme case, we have found that in comparing algorithms on real data sets (using the ROC area, the KL divergence, and other measures), the differences between algorithms are dwarfed by differences due to how borders are handled.

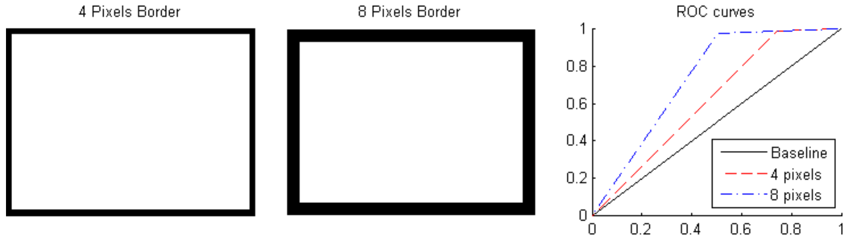

Figure III.7 Illustration of edge effects on performance. *Left:* A saliency map a four-pixel-wide border. *Center:* A saliency map of an eight-pixel-wide border. *Right:* The ROC curves of these two dummy saliency maps, as well as for a baseline saliency map.

Eliminating border effects

Parkhurst and Niebur [56] and Tatler et al. [69] have pointed out that random locations should be drawn from the distribution of actual human eye fixations. In this paper, we measure the KL divergence between two histograms: the histogram of saliency at the fixated pixels of a test image, and the histogram of saliency at the same pixel locations but of a randomly chosen image from the test set (effectively shuffling the saliency maps). This method of comparing models has several desired properties. First, it avoids the aforementioned problem that a static saliency map (such as a centered Gaussian blob) can receive a high score even though it is completely independent of the input image. By shuffling the saliency maps, any static saliency map will give a KL divergence of zero—for a static saliency map, shuffling has no effect, and the salience values at the human fixated pixels are identical to those from the same pixel locations at a random image. Secondly, shuffling saliency maps also diminishes the effect of variations in how borders are handled since few eye fixations are located near the edges.

The potential problem with the shuffling method is that because photos taken by humans are often centered on interesting objects, the center is often genuinely more salient than the periphery. As a result, shuffling saliency maps can bias the random samples to be at more salient locations, which leads to an underestimate of a model's performance [8]. However, this does not affect the validity of this evaluation measurement for comparing different models, and its properties make for a fair comparison that is free from border effects.

III.C.2 Performance

We evaluate our free-viewing saliency algorithm on human fixation data from [3]. Data were collected from 20 subjects free-viewing 120 color images for 4 seconds each. As described in Section III.A and Section III.B, saliency maps are calculated for each image using DoG filters (Method 1) and linear ICA features (Method 2). We also obtained saliency maps for the same set of images using the algorithms of Itti et al. [31, obtained from Bruce and Tsotsos][4], Bruce and Tsotsos [3, implemented by the original authors[5]], and Gao and Vasconcelos [20, implemented by the original authors]. The performance of these algorithms evaluated using the KL measure described in Section III.C.1 is summarized in Table III.1. For each algorithm, the shuffling of the saliency maps is repeated 100 times. Each time KL divergence is calculated between the histograms of unshuffled saliency and shuffled saliency on human fixations. The mean and the standard errors are reported in the table.

The results show that our algorithm with DoG filters significantly outperforms Itti and Koch's algorithms ($p < 10^{-57}$), and Gao and Vasconcelos' algorithm ($p < 10^{-14}$), where significance was measured with a two-tailed t-test over different random shuffles. Between Method 1 (DoG features) and Method 2 (ICA features), the ICA features work significantly better ($p < 10^{-32}$). There are further advantages to using ICA features: efficient coding has been proposed as one of the fundamental goals of the visual system [1] and linear ICA has shown to gen-

[4] The saliency maps that produce the score for Itti et al. in Table III.1 come from Bruce and Tsotsos [3] and were calculated using the online Matlab saliency toolbox (http://www.saliencytoolbox.net/index.html) using the parameters that correspond to [31]. Using the default parameters of this online toolbox generates inferior binary-like saliency maps that give a KL score of 0.1095 (0.00140).

[5] The results reported in the paper used ICA features of size 7 by 7. The results reported here, obtained from Bruce and Tsotsos, used features of size 11 by 11, which the authors say achieve better performance.

Table III.1 Performance in predicting human eye fixations when viewing color images. Comparison of our algorithm (Method 1 with DoG filters and Method 2 with linear ICA features) with previous algorithms. Higher values therefore denote better performance.

Model	**KL (std. error)**
Itti et al. [31]	0.1130 (0.00115)
Bruce and Tsotsos [3]	0.2029 (0.00173)
Gao and Vasconcelos [20]	0.1535 (0.00155)
Method 1 (DoG filters)	0.1723 (0.0122)
Method 2 (linear ICA filters)	0.2097 (0.00157)

erate receptive fields akin to those found in V1 [2,52]. In addition, generating the feature set using natural image statistics means that both the feature set and the distribution over features can be calculated simultaneously. However, it is worth noting that the online computations for Method 1 take significantly less time since only 12 DoG features are used, compared to 362 ICA features in Method 2. There is thus a trade off between efficiency and performance in our two methods.

Our algorithm with linear ICA features performs significantly better than Bruce and Tsotsos' algorithm ($p = 0.0035$) on this data set, though the KL divergence scores are numerically quite close. This similarity in performance is not surprising, for two reasons. First, both algorithms construct their feature sets using ICA, the feature sets are qualitatively similar. Secondly, although our saliency algorithm uses the statistics learned from a training set of natural images and Bruce and Tsotsos [3] calculates these statistics using only the current test image, the response distribution for a low-level feature on a single image of a complex natural scene will generally be close to overall natural scene statistics. In addition, our algorithm is more efficient than that of Bruce and Tsotsos [3]. In our algorithm, the probability distributions of features are pre-computed offline from the training set, while in their algorithm the probability distributions have to be estimated for every image.

Table III.2 Some computational components of the algorithms. Notably, our algorithm requires only offline probability distribution estimation and no global computation over the image in calculating saliency.

Model	**Statistics calculated using**	**Global operations**	**Statistics calculated on image**
Itti et al. (1998)	N/A	Sub-map normalization	N/A
Bruce and Tsotsos (2006)	Current image	Probability estimation	Once for each image
Gao and Vasconcelos (2007)	Local region	None	Twice for each pixel
Ours	Training set of natural images	None	None

Table III.2 summarizes some computational components for several algorithms. Computing the statistics offline using a data set of natural images allows our algorithm to compute saliency quickly compared with algorithms that require calculations of statistics on the current image. In addition, our algorithm requires strictly local operation, which is easier for biological systems to compute.

III.D Discussion

We developed a simple bottom-up saliency algorithm which is a single equation expressed in (IV.8). We applied this algorithm to two different set of features and compared their performance to several previous bottom-up saliency algorithms. The performance shows that we works as well as state of art algorithms with some efficiency advantages. In our experiments, we found that linear ICA features works better than hand picked DoG features. As efficient coding has been proposed as one of the fundamental goals of the visual system [1] and linear ICA has shown to generate V1 cell like receptive fields. Using ICA features seems to be a more principled component for our framework. Besides the independence of the features, the probability distribution of these features also come free when the features are learned from natural images.

The visual search asymmetry phenomena described in Section II.B.3 also seem to suggest that the statistics of observed visual features are estimated by the visual system at many different levels, including basic features such as color and local orientation as well as higher-level features. The question of exactly what feature set is employed by the visual system is beyond the scope of this paper. In the current implementation of our algorithm, we only consider linear filter responses as features for computational efficiency. This use of linear features (DoG or linear ICA features) causes highly-textured areas to have high saliency, a characteristic shared with complexity-based algorithms [10, 33, 60, 94]. In humans, however, it is often not the texture itself but the change of texture that attracts attention. Saliency algorithms that use local region statistics, such as [20], address this problem explicitly.

Our model could resolve this problem implicitly by using a nonlinear feature space. Whereas linear ICA features learned from natural images respond to discontinuities in illumination or color, higher-order nonlinear ICA features are found to respond to discontinuity in textures [34, 53, 64]. Figure III.8 shows an image, the response of a linear DoG filter to that image, and the response of a nonlinear feature inspired by the higher-order features learned in [64]. Perceptually, the white hole in the image attracts attention [3]. Whereas the linear feature has zero response to this hole, the higher-order feature responds strongly in this region. We will explore the use of such features in future work.

In conclusion we developed a simple algorithm that can be expressed as a single equation (IV.8). We applied this algorithm using two different set of features, difference of Gaussians (DoG) and ICA-derived features, and compared the performance to several existing bottom-up saliency algorithms. Not only does our

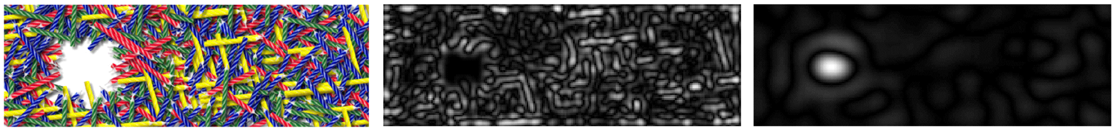

Figure III.8 Left: the input image (adapted from (Bruce2005); Middle: the response of a DoG filter; Right: the response of a non-linear feature, constructed by a DoG filter, whose output is nonlinearly transfered before another DoG is applied.

algorithm performs as well as or better than the state-of-the-art algorithms, but it is also more computationally efficient. In its use of self-information to measure bottom-up saliency, our algorithm is similar to those in [3,51,71], but stems from a different set of intuitions and is calculated using different statistics. In our model, the probability distribution over features is learned from natural statistics (which corresponds to an organism's visual experience over time), whereas these previous saliency models compute the distribution over features from each individual test image. We showed that several search asymmetries which may pose difficulties for models based on test image statistics can be accounted for when feature probabilities are obtained from natural statistics.

In future work, we intend to incorporate the higher-level features. In addition, our definition of overall saliency includes a top-down term that captures the targets features. Although this goes beyond the present dissertation in scope, we plan on examine top-down influences on saliency in future work; preliminary work with faces shows promise.

III.E Acknowledgment

Chapter III, in part, is a reprint of the paper in preparation "A Bayesian Framework for Saliency", co-authored with Honghao Shan, Tim K. Marks, Matthew H. Tong and Garrison W. Cottrell. The dissertation author is the primary investigator and author of these two papers.

IV Dynamic Scene Saliency

IV.A Implementation of Bottom-up Saliency on Dynamic Scenes

In this section, we implement an algorithm that estimates the bottom-up saliency in videos. Although ICA features were shown to perform better than DoG features in predicting human fixations when viewing static images. They are computationally expensive to learn from training videos and to calculate on test videos. For practical reasons, we use separable linear filters instead. We decompose a video to three channels as in section III.A and apply a bank of spatiotemporal filters. The probability distribution of the spatiotemporal features are learned from natural videos. Then for any video, we calculate its features and estimate the bottom-up saliency of each point using $-\log p(F = f)$, as given by the first term of equation (II.7).

The spatial temporal filters we used are separable linear filters. The feature response function has the form $F = V * g * h$, where V is a channel of the video, g is the component that applies only along the spatial dimension and h is the component that applies only along the temporal dimension. The filter responses are used as features.

To keep the algorithm as efficient as possibile, Difference of Gaussians (DoG) filters are again used as the spatial filters, generated by equation III.2. We applied DoG filters to all three channels (I, RG, and BY) with 5 scales ($\sigma = 2, 4, 8, 16$ or 32 pixels), resulting in 15 spatial filters in total. This is a small subset

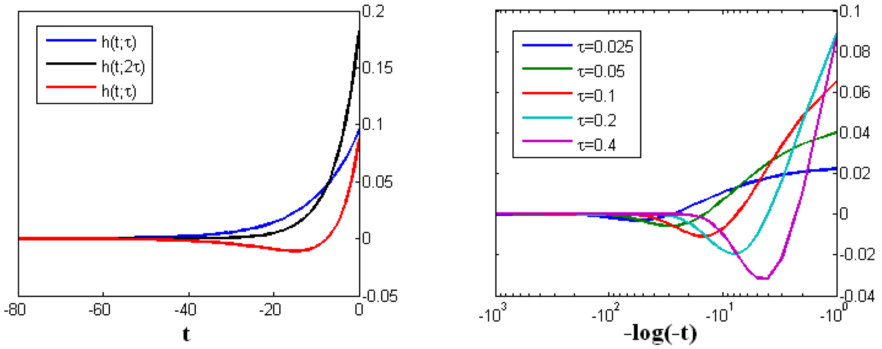

Figure IV.1 On the left is the temporal filter when $\tau = 0.1$. Plotted are $\hat{h}(t;\tau)$ (blue line), $\hat{h}(t;2\tau)$ (black line) and $h(t;\tau)$ (red line). The right plot shows the temporal filters for the five time scales used (values of τ of 0.025, 0.05, 0.1, 0.2, and 0.4).

of the spatial features used in [28, 31].

We design a special temporal filter efficient calculation. The temporal filter h takes the form:

$$h(t;\tau) = \hat{h}(t;2\tau) - \hat{h}(t;\tau) \tag{IV.1}$$

where $\hat{h}(t;\tau) = \frac{\tau}{1+\tau} \cdot (1+\tau)^t$ where $t \in (-\infty, 0]$ is the frame number relative to the current frame (0 is the current frame, -1 is last frame, etc.) and τ is a temporal scale parameter that determines the shape of the temporal filter. We used 5 temporal scales in our implementation $\tau = 0.025, 0.05, 0.1, 0.2, 0.4$. Figure IV.1 shows how $h(t;\tau)$ is formed and how it varies with τ.

We will refer to $h(t;\tau)$ as a DoE (Difference of Exponentials) filter due to $\hat{h}(t;\tau)$'s similarity with the exponential distribution. We choose DoE as the temporal filter for the following reasons:

- $\lim_{t\to-\infty} h(t;\tau) = 0$. Therefore frames in the distant past do not contribute to the current saliency.

- $\Sigma_{-\infty}^{0} h(t;\tau)dt = 0$. If a part of the scene does not change for a extended period of time, it ceases to be salient.

- $h(t;\tau)$ is largest near $t = 0$ and falls off rapidly. This says that DoE has a strong response to onset and offset of objects.
- It bears some resemblance to the temporal profile of LGN cells [7].
- Using DoE as temporal filters enables the the efficient online estimation of the current saliency map (shown below). Only the spatial filter responses at the current frame and spatial temporal responses at the last frame is necessary for calculation of the current saliency map, removing the need for memory of earlier frames or filter responses.

With the exception of the last property, these properties are all shared with the DoG.

Because all filters are linear:

$$\begin{aligned} F(\tau) &= V * g * h(\tau) && \text{(IV.2)} \\ &= V * g * (\hat{h}(2\tau) - \hat{h}(\tau)) && \text{(IV.3)} \\ &= V * g * \hat{h}(2\tau) - V * g * \hat{h}(\tau) && \text{(IV.4)} \\ &= \hat{F}(2\tau) - \hat{F}(\tau) && \text{(IV.5)} \end{aligned}$$

where $\hat{F}(\tau) = V * g * \hat{h}(\tau)$.

Let $\hat{F}_k(\tau)$ denote frame k of $\hat{F}(\tau)$. Let $R_k = V_k * g$ denote the reponse of the video frame k to spatial filter g. Note the difference: $\hat{F}_k(\tau)$ is to apply a spatial temporal filter to the video first and then take a frame from the response; R_k is to take a frame of the original video and then apply a spatial only filter.

Then we have

$$\hat{F}_{k+1}(\tau) = \frac{\hat{F}_k(\tau)}{1+\tau} + \frac{\tau}{1+\tau} \cdot R_{k+1} \qquad \text{(IV.6)}$$

Hence to estimate $\hat{F}_{k+1}(\tau)$, we shrink the $\hat{F}_k(\tau)$ by a factor of $1 + \tau$ and add R_{k+1} scaled by a factor of $\frac{\tau}{1+\tau}$. Then the final response can be easily calculated by $F_{k+1}(\tau) = \hat{F}_{k+1}(2\tau) - \hat{F}_{k+1}(\tau)$. This leads to a fast online calculation of the feature response and consequently efficient saliency estimation.

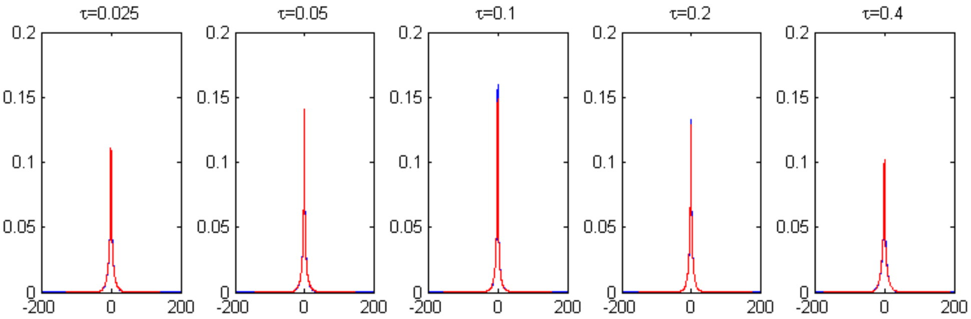

Figure IV.2 The distribution of filter responses for the middle scale DoG filter with all the temporal scales on the intensity channel collected from the set of natural videos (blue line), and the fitted generalized Gaussian distributions (red line).

IV.A.1 Learning the distribution

As described above, there are 15 features on the spatial dimension: 5 from each channel. On the temporal dimension there are 5 scales and they are combined with each spatial feature. Thus there are in total 75 feature responses.

By computing these feature responses on natural videos (about 2 hours of animal/plant documentary videos), we obtained an estimate of the probability distribution over the observed values of each of 75 features. These distributions are again modeled by generalized Gaussian distributions given by equation III.3.

This resulted in one shape parameter, $\theta_{i,j}$, and one scale parameter, $\sigma_{i,j}$, for each of the 75 filters: $i = 1, 2, ..., 15$ is the index for spatial filters, and $j = 1, 2, ..., 5$ is the index for temporal scales. By visual inspection the generalized Gaussians again provide an excellent fit to the data (figure IV.2). [1]

Taking the logarithm of equation (III.3), we obtain the log probability

[1] We also computed these feature responses on videos mainly of indoor scenes and street scenes (40 minutes of soap TV show). The results are very similar to those from the natural videos (with slightly increased variance in color channels, probably due to the colorfulness of soap TV shows). Thus we are confident that the distribution of these low level features are not affected much by the high level theme of the videos and that we have obtained good estimation of the probability distribution for the features in the natural statistics.

over the possible values of each feature:

$$\log p(F_z^{i,j} = f^{i,j}) = \log \theta_{i,j} - \log 2 - \log \Gamma\Big(\frac{1}{\theta_{i,j}}\Big) - \left|\frac{f^{i,j}}{\sigma_{i,j}}\right|^{\theta_{i,j}} = - \left|\frac{f^{i,j}}{\sigma_{i,j}}\right|^{\theta_{i,j}} + const. \tag{IV.7}$$

These feature responses are not independent. But we proceed as if they are for simplicity:

$$-\log p(F_z = f) = \sum_{j=1}^{5} \sum_{i=1}^{15} \left|\frac{f^{i,j}}{\sigma_{i,j}}\right|^{\theta_{i,j}} + const. \tag{IV.8}$$

IV.B Results

We evaluate our saliency algorithm on the human fixation data from [27]. Eye movements were recorded from 8 subjects viewing 50 videos from indoor and outdoor scenes, television broadcasts, and artificial environments totaling over 25 minutes of video at 640 × 480 (at 60.27 Hz, a viewing distance of 80 cm, and with a field of view of 28 deg ×21 deg). Data was collected using a ISCAN RK-464 tracking the right eye. Two hundred eye movement traces were used (four subjects for each video clip). See [27] for more details.

In [28], Itti and Baldi report results of their saliency measure (Bayesian surprise) on this data set. As discussed in section III.C.1, their saliency map was sampled at the target location of a saccade at the time the saccade was initiated. By histogramming the number of fixations for each value of salience, a distribution of saliency was formed for human fixations. This could be compared with the distribution of fixations over saliency for random saccades chosen uniformly over the image by looking at the KL divergence between the two distributions. Both KL divergence and ROC area have the desired property of transformation invariance, which is also shared by ratio above median: how much of the saliency score of human fixations is above the median score of the entire saliency maps. Here we provide the measurement in all these three evaluation metrics.

Again, human eye fixation data has a strong center bias (figure III.5) and how borders are processed has a big effect on the results because modifying the

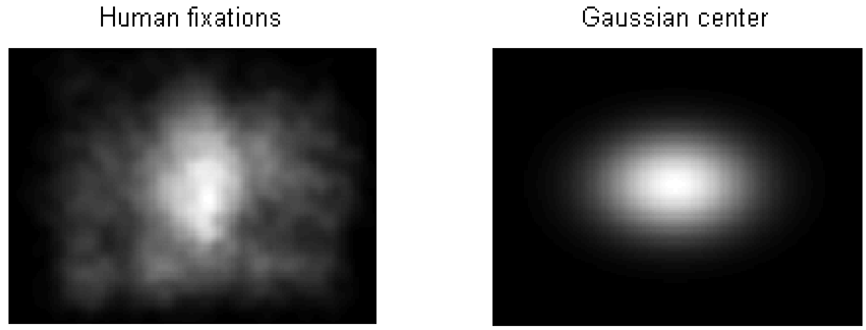

Figure IV.3 A demonstration of a Gaussian fit to [3] fits the overall trend of fixations of [27]. This suggests a strong center bias, and indicates why removing borders have such a large effect.

Table IV.1 Summary of initial results. Using a static Gaussian blob saliency map outperforms other methods.

Method	**KL**	**ROC area**	**% above median**
Baseline	0	0.5	50%
Bayesian Surprise [28]	0.1332 [2]	0.6472	70.91%
Dynamic Saliency	0.1001	0.6262	70.91%
Dynamic Saliency (w/border)	0.1815	0.6596	75.37%
Centered Gaussian	0.4415	0.7641	86.89%

border had large effects on the random-saccading distribution of salience, but little effect on the distribution of salience for human saccades. The simple Gaussian blob saliency map (figure IV.3) discussed in section III.C.1 drastically outperformed our results and the the surprise model [28]. These results are summarized in Table IV.1.

We again altered the way in which the baseline random distribution for the KL divergence is measured. The fundamental problem is that sampling uniformly is not at all indicative of how human saccades tend to be distributed. Instead of sampling uniformly, we used the same pattern of fixations but shuffled the frames of the saliency maps over the whole sequence of movies, destroying all correlation between human fixations and the salience measure at the time of

Table IV.2 KL scores, ROC curve area and percentage above median when the frames of the saliency map are shuffled. This has the benefit of using a random distribution based on the distributions of human saccades and not assuming a uniform distribution.

Method	**KL**	**ROC area**	**% above median**
Baseline	0	0.5	50%
Bayesian Surprise [28]	0.0344	0.5808	61.66%
Dynamic Saliency	0.0409	0.5818	62.39%

sampling. However, as pointed out in [8] this will serve as an *underestimate* of performance since the center of the screen for pictures and video genuinely tends to be the most salient part of the scene when a human is behind the camera.

Nevertheless, our method continues to do better than chance, and as well as Itti and Baldi's surprise model [28], as shown in Table IV.2. The histogram of the saliency score on human fixations are shifted toward larger numbers than that on the shuffled fixations. Figure IV.4 shows the saliency maps on some frames of different videos.

IV.C Real Time Implementation

Saliency algorithms are of potential interest to social robotics. A robot that orients its eyes in a manner similar to humans is likely to give an impression of intelligent behavior and facilitate interaction with humans. Furthermore, such models may be used as interest point operators to orient the robot towards regions of the visual scene that are likely to be relevant.

As part of the RUBI project [47, 48] for the past three years, Movellan's laboratory has been conducting field studies with social robots immersed at the Early Childhood Education Center at UCSD. The goal of these studies is to explore the possibilities of social robots to assist teachers in early childhood education (figure IV.5). One critical aspect of these robots is to be able to find and orient

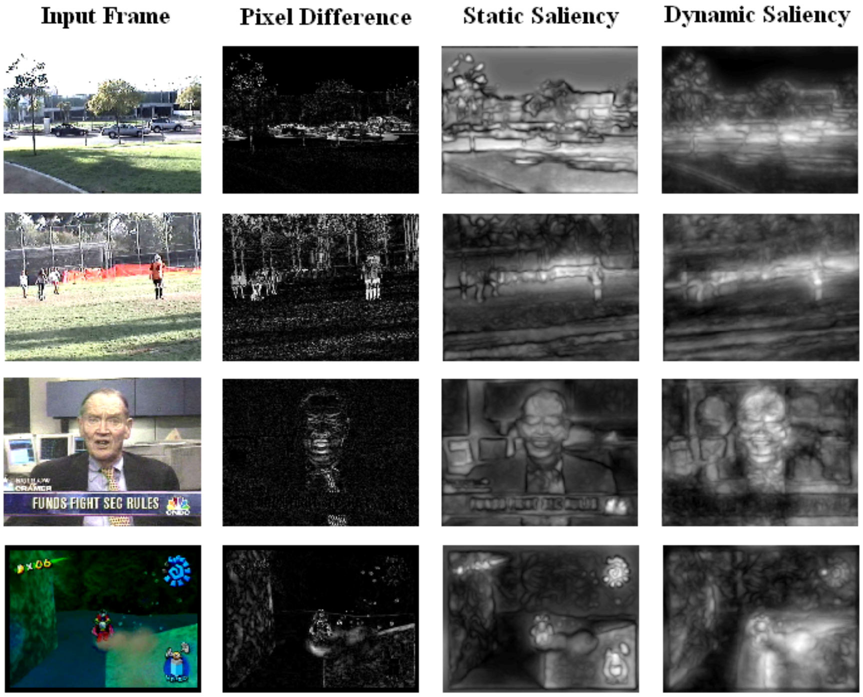

Figure IV.4 The saliency maps for several frames of video from [27]. The saliency maps generated from purely temporal components and spatial components are provided for comparison.

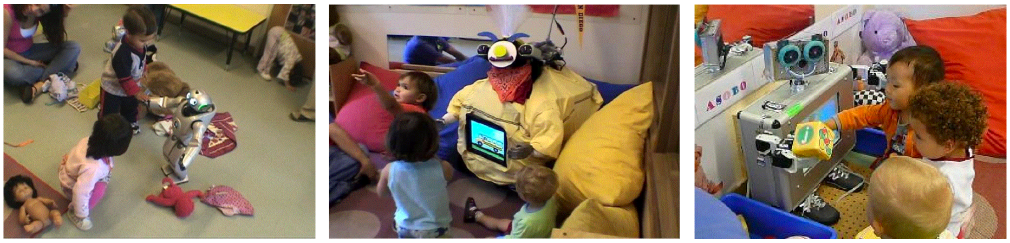

Figure IV.5 Three robot members of the RUBI project. **Left:** QRIO. **Center:** RUBI-1, the first prototype developed at UCSD. **Right:** RUBI-3 (Asobo) the third prototype developed at UCSD. It teaches children autonomously for weeks at a time

towards humans. While powerful algorithms for detecting the presence of humans using video have already been developed [15], they tend to be computationally expensive and thus best suited for scanning a small foveal region of a scene. As such we were interested in investigating whether a lightweight saliency model could be used on peripheral regions to help orient the fovea towards the most promising regions of the visual scene.

In this section, we modify the algorithm for real time implementation, and further evaluate whether the bottom up saliency algorithm is useful as a preprocess to facilitate higher level more specific tasks, namely, looking for humans.

IV.C.1 Simplified Algorithm

In the earlier session, we used three channels but we will only use the gray channel in the simplified algorithm. Furthermore, we assumed earlier that the filter responses follow generalized Gaussian distributions with different shape and scale parameters and the saliency is calculated by equation IV.8. In this simplified algorithm, we assume filter responses follow Laplacian distributions (generalized Gaussian distribution with shape parameter set to 1), with the same scale parameters. Equation IV.8 thus simplifies to:

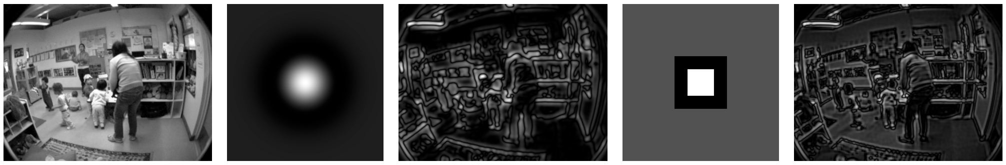

Figure IV.6 An example of DoB (Difference of Boxes) filter and its comparison to a DoG filter. From left to right: a camera frame input, a DoG filter, the DoG's filter response, a DoB filter, the DoB's filter response.

$$-\log p(F_z = f) = const. \cdot \left(\sum_{j=1}^{5} \sum_{i=1}^{5} \left| f^{i,j} \right| \right) + const. \qquad \text{(IV.9)}$$

This formulation is simply the sum of the absolute value of filter responses and no parameters need to be trained from natural videos. We further modify the original spacial filters from DoG to DoB (Difference of Boxes) for faster implementation. Figure IV.6 shows an example of the filter and compares its response to that of DoG. In the results reported below, a bank of 5 DoB filters are used with center widths $\{3, 5, 9, 17, 33\}$. The corresponding surround widths were $\{5, 9, 17, 33, 65\}$. Five temporal scales are used with temporal parameters τ taken values of $\{1, 2, 4, 8, 16\}$ [3].

In order to ensure that the simplifications in our approach still maintain the important properties of the original algorithm, we evaluated it with the same method used in section IV.B. The results are shown in table IV.3. Compare this to earlier results reported in table IV.2, the performance of this algorithm was very close to that of the original algorithm and that of Itti & Baldi [28].

IV.C.2 Robot Camera

A two degree of freedom (pan and tilt) robot camera was constructed using an iSight IEEE1394 640x480 camera with a fisheye lens (160° FOV), 2 Hitech

[3] The temporal scales were set to be faster here than in section IV.A so that the camera motion does not affect the saliency results.

Table IV.3 Performance of the simplified algorithm evaluated by predicting people's eye fixations when watching videos. This result is very close to the original algorithm and the surprise model (table IV.2).

Method	**KL**	**ROC area**	**% above median**
RUBI Saliency	0.0366	0.5797	61.25%

HS-322HD servo motors, and a Phidgets servo control card operated by a Mac Mini (1.87 GHz Intel Core Duo). The robot camera was placed in Room 1 of the UCSD's Early Childhood Education Center (ECEC), where the RUBI project is taking place. The camera was located on a bookshelf above the reach of the children (≈ 18 months). The system collected data continuously for 9 hours during one day's operation of ECEC, from 7:30am–4:30pm.

Images were processed in real-time. They were received from the camera at 640 × 480 resolution at approximately 15 FPS (i.e. every 66 msec). For the purpose of computing saliency, they were downsampled to a 160 × 120 pixel resolution. A saliency map was then computed in six-times-faster-than-real-time for all the pixels at the speed of 11 msec per frame. It is potentially important for a saliency algorithm to be faster than real time to be useful for robotic applications, so that there is time left for post-processes such as face detection.

The camera was controlled in two ways. One is saliency contingent and another is to repeat the motion that has been used in the other condition.

Experimental Camera – Saliency Track At the start of each experiment, the camera was moved to a central location. Starting 30 frames after any camera movement, on each successive frame, if the maximum saliency pixel exceeded threshold and the desired motion was more than 10 degrees in either the pan or tilt direction, the servos would reposition the camera so that the maximum saliency pixel in the saliency map was now at approximately the center of the image plane.

15 frames after a movement was initiated (to allow for the movement's

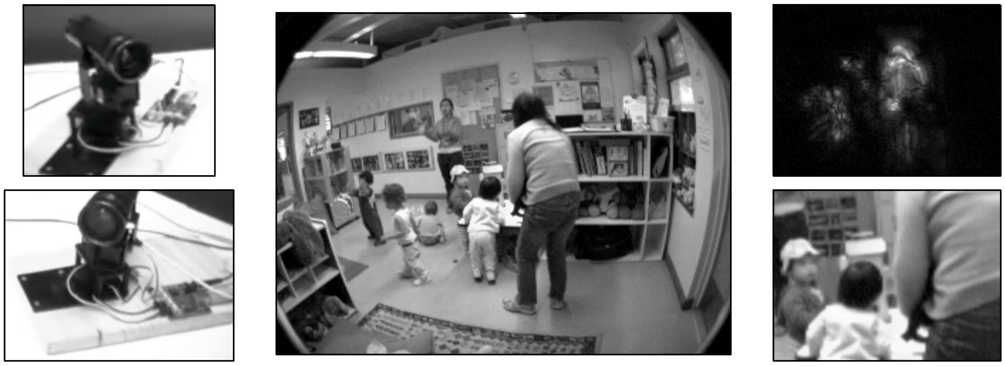

Figure IV.7 A simple robotic camera (left) collected 160° images at 640 × 480 resolution (center) and compute saliency maps (top right). The camera then orient towards the most salient pixel. After movement, a snapshot is taken (bottom right).

completion), an image of the camera's view was saved. Additionally, a fovea view containing the center 160 × 120 pixels of the high resolution 640 × 480 image was saved, simulating the foveal region over which high level but computationally expensive perceptual primitives could operate (*e.g.*, person detection, expression recognition). Figure IV.7 shows a couple pictures of the camera, an example of the input, its saliency map and its fovea.

Control Camera – Playback An additional camera control condition was implemented. In this condition the camera played back in open-loop the exact the same movements as in the previous salience-directed movement condition. This served as a control with the same motion statistics as the salience condition, but the movements were not caused directly by current events in the world.

Each condition ran sequentially for 3 minutes at a time. A pair of salience and playback conditions would take about 6 minutes. There was an additional 3 minute break between cycles. In all, 64 cycles were completed and 4964 images were collected.

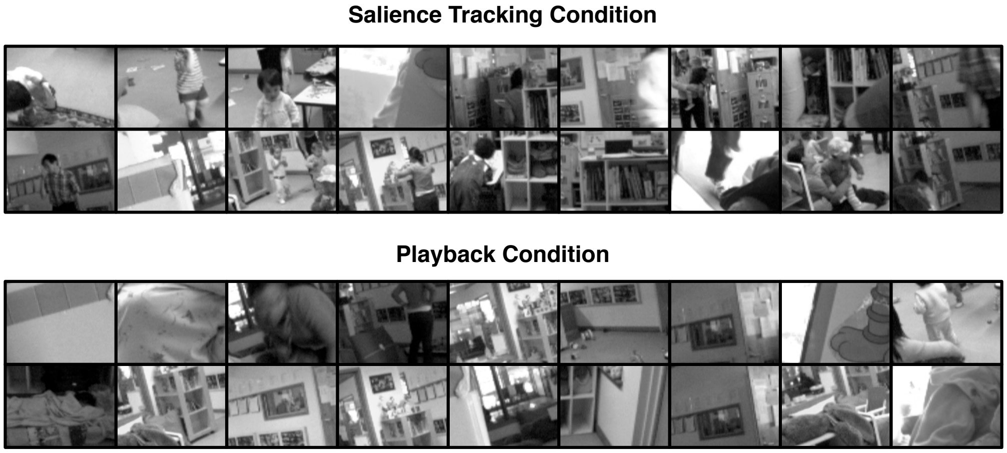

Figure IV.8 Center of snapshots in saliency tracking and playback conditions. In each case, 18 images were chosen randomly, and so the sample is representative. Many more people are attended in the saliency condition than the playback condition.

IV.C.3 Analysis of results

After the experiment a subset of the foveal center-images was chosen randomly and uniformly from the entire set. Some examples of the images from both conditions is shown in Figure IV.8. These images were coded by 4 coders. Two of the coders were authors of this paper and two were naive third parties. The coders were instructed to label the number of people they could see in each 160×120 foveal image. The coding was done in a double-blind fashion: the images were ordered randomly across labels and time collected. All coders, including the authors, were given no extra information to indicate which images came from which condition. All coders labeled 1050 images (510 saliency condition, 540 playback condition) in the same order.

The average Pearson correlation between the four coders across the 1050 labels was 0.8723. We marked a foveal snapshot as “containing a person” if two or more coders agreed that there was a person in the snapshot.

It should be noted that the control condition in our experiment was designed to be much smarter random merely random motion. In the control condition, the camera oriented toward regions of space that had been salient in the experimental condition. These regions are places that are tended to have people, such as the play area and the door way. In spite of this, the experimental camera (Saliency Tracking) performed much better than the control camera (Playback). In the Salience Tracking condition, 68.04% of images contained people. In the Playback condition, only 34.81% of images contained people. Thus by orienting toward salient events in the image plane, the camera attended to people twice as often as just looking in the places where people are likely to appear. This is remarkable given that the algorithm is not designed for people detection and the saliency map is calculated on bottom-up features with no knowledge of people's appearance. Note that with a detection rate of 68% per saccade, after 3 saccades, we are 96.8% likely to have seen at least one person. A post processing algorithm operating over these saccades would review $(3 * 160 \times 120)$ pixels, or 3/16 of the full image size. Thus, by using this algorithm to orient a robot camera, we can increase the chance of finding people while reducing the time needed for detection algorithms.

IV.D Discussion

We designed a feature space that can be calculated very efficiently, which leads to a simple, fast algorithm. A real time version of this algorithm has been applied to social interactive robot.

Our findings also agree with [56,69] in pointing out some disadvantages of using some of the previously proposed evaluation metrics. As the evaluation data sets are often collected by recording people's eye movement when viewing images and videos in a lab setting, stimuli are presented on a monitor and the head is often not allowed to move, introducing a strong central bias that confounds proper evaluation of the results. By shuffling the frames but maintaining the patterns of fixations, we effectively remove the effects of this bias. However, as [8] points out, there is also a central bias introduced by having humans center the camera on interesting parts of the scene - the center is inherently more likely to be salient.

Overall, our results show comparable performance with Itti and Baldi's surprise model [28, 32] in predicting human fixations. The simplified, real time implementation also works almost as well. And we show that a robot camera controlled to orient toward most salient region achieves higher probability of finding humans.

IV.E Acknowledgment

Chapter IV, in part, is a reprint of the paper in preparation "A Bayesian Framework for Dynamic Scenes", co-authored with Matthew H. Tong, Nicholas J. Butko, Javier R. Movellan and Garrison W. Cottresll. The dissertation author is the primary investigator and author of these two papers.

V Probabilistic Search

V.A Where to attend given a saliency map

The calculation of saliency, as described in the last several chapters, can be regarded as *pre-attentive* processing. The information of visual features is estimated over the entire visual field in parallel without the involvement of attention. In this section, we concentrate on the *attentive* processing; given a saliency map, how is attention allocated?

We propose that attention is allocated stochastically in proportion to the saliency map or as the probability modified by any monotonically increasing function. Since saliency is proportional to the probability of a target given the features, we will use the terms saliency map and probability map interchangeably. The term probability map, depending on the context, refers to the probability of a target at each location, or the probability of directing attention to that location, which are essentially the same after normalization over the visual field.

We will discuss our hypothesis in the context of single target paradigm where subjects perform speeded searches and respond whether a target is present in a display of distractors. In the visual search literature, there has been controversy that whether processing is parallel or serial, e.g. [70, 90]. As Townsend has pointed out decades ago, the average response time in the single target experiment paradigm is not sufficient to discriminate these two processes [72, 73]. For any serial model that accounts for the data, there exists a parallel model, likely limited capacity model, that will predict the exactly same data. In this paper, we stay neutral on whether the underlying information is processed in a parallel manner or a serial manner. Although our hypothesis about attention can not work independently from the underlying information process, it can work with both models.

In a serial model, attention is directed to one item at a time. We assume that the probability of an item being attended to is proportional to its saliency. A salient item is more likely to be attended to and be processed next. In a limited parallel model, many objects are processed at a time. We assume that attention is allocated proportional to the probability map. The processing resources is not equally distributed to the items that salient items enjoy more computational power. In this chapter, we will mainly develop our theory with a serial model. But the qualitative conclusions we make can be generalized to a parallel model.

In the context of a serial model for single target visual search tasks, our hypothesis is that the probability of an item being attended to is proportional to its saliency. This hypothesis, at first look, is not rational. If the goal is to maximize the probability of finding the target, the Bayesian optimal choice is to allocate attention to the item with the highest probability of the target, i.e. the most salient item. Why not concentrate attention on the most promising place and be optimal?

If we regard where to direct attention given the saliency map a decision that the visual system has to make. We can look into how people make decisions and hope that the underlying principles share some similarities. Strikingly, people seem to be irrational in the simplest binary choice tasks, often known as bandit problems by analogy with slot machines. Now try for yourself. If a city rains 60% of the days, will it rain tomorrow? Will you always say it will rain, as the rational choice theory would predict? Or will you hesitate, and possibly mix up your answers if being asked several times? Numerous experiments have shown that human and non-human tend to do the latter in these tasks (e.g. [40,45]). An interesting feature of this phenomenon is that people's choices tend to match the underlying probability of the outcomes, i.e. they probability match. Depending on the scenario, there also sees over-match where the portion of people's choice for the more likely outcome is bigger than its probability, and other times under-match. Overall, the decision is stochastic and the outcome with higher probability is more likely to be chosen. This also generalizes to multiple choice tasks. Probability match in decision making has been taking as evidence that people are not rational. Some later experimental works (e.g. [65]) and theories developed from the view of adaptive learning (see [80] for a review) have suggested otherwise. Whether people are rational or not in these decision making situations is still under controversy. If they are rational, as more researchers tend to believe nowadays, what is being optimized by the strategy of probability match is still unclear.

Here we will only briefly mention two situations where mixing the decisions stochastically could be preferable than simply choosing the most likely outcome when feedback is available only to the choice that is being made. The first one is that the organism does not have the perfect model of the environment yet. Thus, the estimated probability of the outcomes might deviate from the true probabilities. In this case, the observer wants to keep learning about the environment while trying to get as much reward as possible. It is then preferable to mix the choices up depending how confident the observer is on his model of the environment. The other situation is that the environment is dynamic. Thus, the knowledge of the environment collected in the past could become outdated in the future. It is to the observer's benefit to mix up the choices depending on how fast the environment is changing to keep his model of the environment updated. In both these scenarios, the uncertainty of the environment could drive an organism to make their decisions stochastically, trading off the current reward and learning of the environment for more future reward. In the reinforcement learning literature, it has been shown that probability match provides a good balance for the tradeoff of exploitation (pick the option with the highest expected reward) and the exploration (try something random) [21, 63].

Coming back to our own problem of where to direct attention, it mimics a multi-choice decision making scenario where the estimated probability of the outcomes is given by the saliency map. It is reasonable to assume that the organism do not have a perfect model of its environment and the environment itself could be changing. Furthermore, only attended items will be processed in further detail, i.e. feedback is only available for the choice being made. Thus, the same driving force that cause people to probability match, even though we do not understand fully, could result in probability match when directing attention.

V.A.1 Average Time to Find the Target

In this session, we infer that in a single target search task, if the target is present in the display, how long it takes in average for the subject to find the target. We assume for simplicity that the objects outside the stimuli display do not compete for attention, and that salience of each item do not change overtime due to eye movements etc.

Let s_{targ} denote saliency of the target, s_{dist} denote that of a distractor and there is n distractors in the display [1]. We define the term *distractor strength* as:

$$x = s_{dist}/s_{targ}. \tag{V.1}$$

This can be thought of as the relative salience of the distractors versus the target. For a classical feature search, the target is highly salient compared to the distractors, so the distractor strength x is very small. For classical inefficient conjunction target search, the distractors are as salient as the target, so the distractor strength x is approximately 1. As we will show later, distractor strength x is one of the key variables that decides how difficult a search task is.

When n distractors are present, the probability of the target being the

[1] For notation simplicity, we use s_{dist} for the salience of all distractors. The qualitative conclusions still hold if the salience of the distractors vary but the average salience is s_{dist}

first item to be attended to is:

$$\frac{s_{targ}}{s_{targ} + n \cdot s_{dist}} = \frac{1}{1 + nx} \tag{V.2}$$

The intuition is straightforward: the probability of attending to the target is small if the distractor is very distracting or the number of distractors is large.

Now the question is where to look next. An important question here is that whether previous attended items will be attended again. It is of controversy that whether subjects remember what items have been processed. Posner and Cohen [57] noticed an inhibitory effect which reduces the likelihood of attending to the previous attended locations. They named it inhibition of return. Klein [37] found this effect in visual search and suggested that it may function to improve efficiency. To avoid attention residing on one salient item, many earlier attention models have employed this mechanism that just attended items are not to be attended again, e.g. [31,74,92]. It could be implemented strictly that attended items will never be attended again or not so strictly that the probability of attending an item is zero right after it is attended but slowly raises over time. Some eye movement models went further to infer inhibition of return by maximizing information over time, e.g. [39]. Some recent findings confirmed that inhibition of return is indeed involved in visual search paradigms [49, 68]. However, it stays somewhat controversial that to what extent the attended distractors are inhibited [23–25] (see [90] for a review).

We will not investigate how much an attended item is inhibited, but allow the flexibility in our framework. We assume that after an item is attended to, it will be marked and not be attended to again with probability γ, where $0 < \gamma < 1$ and will referred to as inhibition rate. This says γ portion of the attended items will not be attended again, which can also be understand as that after an item is attended to, the probability of attending to it again is $1 - \gamma$.[2] When γ approaches 1, it is the scenario of strict inhibition of return where attended items will never be attended to again. When γ approaches 0, it reduces to no inhibition of return at all.

[2] Strictly, these two assumptions are different. But they will lead to same inference in our case.

Note that the mechanism of inhibition of return is much more complicated than to be summarized by a single parameter. For example, what items are inhibited and to what extent are dependent on the sequence of the attended items. However, our purpose here is to show that our framework could work with different assumptions about inhibition of return as our qualitative conclusions are independent of the exact value of inhibition rate.

Let $E(n, x)$ denote the expected number of distractors being attended before the target is attended to. It is zero if there are no distractors:

$$E(0, x) = 0 \tag{V.3}$$

When $n > 0$, by equation V.2, a distractor will be attended to with probability $1 - \frac{1}{1+nx} = \frac{nx}{1+nx}$. If a distractor is attended to, with probability γ, it is marked and will not be attended to again and the search procedure from then on is searching for the target in $n-1$ distractors. With probability $1-\gamma$, it is not marked and can be attended to again as any other distractors and the search procedure from then on is again searching for the target in n distractors. This gives us the recursive formula:

$$E(n, x) = \frac{nx}{1+nx}(\gamma(E_{n-1,x} + 1) + (1-\gamma)(E_{n,x} + 1)) \tag{V.4}$$

Taking together the initial condition given by equation V.3 and the recursive property given by equation V.4, we have:

$$E(n, x) = \frac{x}{1+\gamma x} n \tag{V.5}$$

Assume that the subject does not misclassify any distractor as the target (false alarm) nor misclassify the target as a distractor, and he correctly responded the presence of the target upon attending to it. Let t_{dist} denote the average time to process a distractor and t_{prst} denote the average additional cost in the response time when the target is present, including the time to process the target, the time needed to press the response button, etc. The expected response time is given by:

$$RT_{prst}(n, x) = E(n, x) \cdot t_{dist} + t_{prst} = \frac{x}{1+\gamma x} t_{dist} \cdot n + t_{prst} \tag{V.6}$$

This formula says that the expected response time when the target is present increases linearly with the number of items, with a slope of $\frac{x}{1+\gamma x}t_{dist}$.

V.A.2 Similarity to Former Models

When the distractor strength x is very small ($x \approx 0$), i.e. the target is highly salient relative to the distractor, $\frac{x}{1+\gamma x}t_{dist}$ is very small and generates a flat slope for the response time as the number of distractors increases. This generates a scenario of "parallel search" proposed in FIT (feature integration theory) [74]. The average number of distractors to be attended to before the target $\frac{x}{1+\gamma x}n$ is very small that the target is quite likely to be the first item attended; the target pops out.

When $x = 1$ and $\gamma = 1$, i.e. the target is equally likely to attract attention as any distractor and the inhibition of return is strict, the target is equally likely to be attended to as the 1st, $(n+1)$th, or any one in between. The average number of distractors being processed is $\frac{n}{2}$ and the slope of the response time is likely to be steep. This is equivalent to "serial search" in FIT.

When x moves away from 0, the slope $\frac{x}{1+\gamma x}t_{dist}$ increases smoothly with x. The search continuously changes from very efficient to less efficient to inefficient, which potentially account for Wolfe's classification of efficiency in search tasks [91]. For example, increased target/distractor similarity and distractor heterogeneity raise the RT slope. As the target contrasts less with the distractors, its salience decreases, leading to an increased distractor strength x. The slope of the expected time to find the target thus also increases. Similarly, as each distractor contrasts more from others distractors, their saliency increases, leading to an increased distractor strength x and a larger slope.

Moreover, recall that:

$$\begin{aligned} x &= \frac{s_{dist}}{s_{targ}} & \text{(V.7)} \\ &= \frac{\frac{1}{p(F=f_{dist})} \cdot p(F = f_{dist}|C = 1) \cdot p(C = 1|L = l_{dist})}{\frac{1}{p(F=f_{targ})} \cdot p(F = f_{targ}|C = 1) \cdot p(C = 1|L = l_{targ})} & \text{(V.8)} \\ &= \frac{p(F = f_{targ})}{p(F = f_{dist})} \cdot \frac{p(F = f_{dist}|C = 1)}{p(F = f_{targ}|C = 1)} \cdot \frac{p(C = 1|L = l_{dist})}{p(C = 1|L = l_{dist})} & \text{(V.9)} \end{aligned}$$

The first term $\frac{p(F=f_{targ})}{p(F=f_{dist})}$ is the relative rareness of the target and the distractors, and we showed earlier that it provides consistent account for many search asymmetries. The term on the right $\frac{p(C=1|L=l_{dist})}{p(C=1|L=l_{dist})}$ calculates whether the distractors or the target are on some preferred locations. The term in the middle $\frac{p(F=f_{dist}|C=1)}{p(F=f_{targ}|C=1)}$ is closely related to Wolfe's guidance in Guided Search model [92]. This term reflects the subject's knowledge of the target. If the subject knows what the target is and can estimate the likelihood terms, there will be strong guidance toward items that are consistent with subject's knowledge about the target.

Our distractor strength x can also find its correspondence in the selection ratio in Bundesen's partial report model, denoted as α [6]. It is the ratio of the selection strength v of a distractor and that of a target. If we re-interpretate the selection strength as the probability of the item being the target, then our distractor strength x is equivalent to selection ratio α. Then how can we interpretate the selection strength v in our probability terms? In his more recent visual attention models [4, 5], v is elaborated as the hazard function of classifying an item k into a category i, corresponding to our framework, this is the condition probability that given z has not been classified so far, what the probability it is classified as c at this moment. It is assumed that v has the formula:

$$v(k, i) = \eta(k, i)\beta_i \frac{w_k}{\Sigma_k w_k} \tag{V.10}$$

$v(k, i)$ has two components: The first is the categorization component $\eta(k, i)\beta_i$, where $\eta(k, i)$ notions the strength sensory evidence that item k belongs to category i, and β_i notions the perceptual decision bias associated with category i; The second is the selection component $\frac{w_k}{\Sigma_k w_k}$ where $w_k = \Sigma_i \eta(k,i)\pi_i$, and π_i notions the pertinence value of category i. The selection component is closely related to our saliency term in equation II.13 when considering multiple targets associated with different reward. Note that when there is no inhibition of return, the probability of an item being attended to anytime is proportional to its salience. Rearrange the terms, the utility or salience in equation II.13 associated with each element can be rewritten as:

$$\Sigma_i \; p(F_k = f|C_k = i) \; \left(\frac{1}{p(F_k = f)} p(C_k = i|L_k = l) \; r_i\right) \tag{V.11}$$

In particular, if we assume that the features of all elements are of equal probability in natural statistics (constant bottom-up saliency), and that there is no prior knowledge about the location of the targets. It simplifies to:

$$\Sigma_i \; p(F_k = f|C_k = i) \; r_i \tag{V.12}$$

If we draw an equal sign between $\eta(k,i)$ and $p(F_k = f|C_k = i)$, i.e. assume sensory evidence that item k belongs to category i is the likelihood term, and another equal sign between π_i (the pertinence value of category i), and r_k (the reward/utility of correctly identifying an element in category k). Given our hypothesis of probability match in allocating attention, Bundesen's selection component corresponds exactly to our probability of attending an element.[3]

The other component of categorization corresponds to the process in our model that once an item is attended to, it has to be classified as one of the categories, e.g. whether the item is a target or a distractor. This again involves calculating the probability of classifying the element belongs to each category[4].

$$p(C_k = i|F_k = f, L_k = l) = \frac{p(F_k = f|C_k = i)p(C_k = i|L_k = l)}{p(F_k = f)} \tag{V.13}$$

[3] The simplification of formula V.11 is not necessary. The correspondence still holds if we let π_i absorbs $p(F_k = f|C_k = i)p(C_k = i|L_k = l)$ besides r_i. However, the conceptual interpretation of π_i being the pertinence of category i will be compromised.

[4] The inference follows that in section II.A, adapting the independence assumption of the feature and location.

Although the probability to be estimated almost the same as for saliency, the decision to be made is very different. When this probability is calculated for the saliency purpose, it is to compare features across locations to direct attention, i.e. compare across k. Now it is to compare categories to classify the element given its features and location, i.e. compare across i. Thus $p(F_k = f)$ is a constant across categories and can be dropped from the formula:

$$p(C_k = i|F_k = f, C_k = i) \propto p(F_k = f|C_k = i)p(C_k = i|L_k = l) \tag{V.14}$$

Again, if each category is associated with reward r_i, the expected reward/gain to classify item k as category i is:[5]

$$g(k, i) \propto p(F_k = f|C_k = i)p(C_k = i|L_k = l)\ r_i \tag{V.15}$$

If we assume that the classification decision is made probabilistically proportional to the expected reward of each category (a reasonable assumption given our discussion about probability match in earlier sessions), the probability of categorizing item k as category i is:

$$\frac{p(F_k = f|C_k = i)p(C_k = i|L_k = l)\ r_i}{\Sigma_i p(F_k = f|C_k = i)p(C_k = i|L_k = l)\ r_i}. \tag{V.16}$$

We now draw an equal sign between β_i and $p(C_k = i|L_k = l) \cdot r_i$, i.e. assuming that the perceptual decision bias associated with category i is equivalent to the product of the prior probability of seeing an item of category k at the element's location and the reward of correctly classifying the element as category i. Recall that a equal sign was drawn earlier between sensory evidence, $\eta(k, i)$, that item k belongs to category i and the likelihood term $p(F_k = f|C_k = i)$. Bundesen's categorization component matches directly to equation V.15. When there is only one item on in the display, the probability of classifying item k as category i inferred in Bundesen's work is:

$$\frac{\eta(k, i)\beta_i}{\Sigma_i \eta(k, i)\beta_i}, \tag{V.17}$$

which is equivalent to equation V.16 in our framework.

[5] The formula can potential become much more complicated if considering different punishment for different kind of misclassification, a common scenario in the literature of cost-sensitive classification, e.g. [12].

V.B When to Stop - a sequential decision making procedure

In the last two sessions, we discussed what attracts attention and how attention is directed given a saliency map. In visual search tasks, there is at least one more important aspect: when to stop and make a response.

Examine the process of a search trial, it fits very well into the sequential decision making scenario. Whenever an item is being processed, some information is gathered about the display, and the observer has to decide whether to stop and make a response or to keep gathering more information.

Assume that when the target is identified, the subject stops the trial and response that the target is present. Then the crucial question facing the subject is that if I have not see the target so far, shall I stop? And if I stop, what do I report?

Let $W_{R-}(k)$ and $W_{R+}(k)$ denote the cost/waste functions of stopping to report absence or presence after k distractors have been processed. For simplicity, assume that the target and the distractors in the display are distinguished enough that there is no misclassification, i.e. the misses and false alarms only occur when the subject stops and makes an incorrect response before a target is identified. The cost function involves the cost of errors including misses or false alarms and the cost of time.

Let $D_{1:k}$ denotes the event that the first k items processed are all distractors. Let T and $\neg T$ denote that the target is present and absent respectively. Let W_m and W_f denote the cost/waste function for misses and false alarms respectively, and W_t denotes the cost/waste of time. The expected total cost of stopping and reporting absence or presence when no target has been identified after processing

k distractors is:

$$W_{R+}(k) = W_m(P(T|D_{1:k})) + W_t(k \cdot t_{dist}) \tag{V.18}$$

$$W_{R+}(k) = W_f(P(\neg T|D_{1:k})) + W_t(k \cdot t_{dist}) \tag{V.19}$$

$$= W_f(1 - P(T|D_{1:k})) + W_t(k \cdot t_{dist}) \tag{V.20}$$

$P(T|D_{1:k})$ is the probability that the target is present after k distractors have been encountered and is the probability of miss if absence is reported at this moment. Note the difference between the probability of miss and miss rate. The former is the number of miss trials normalized by the total number of trials, including both target present trials and target absent trials. The latter is the number of misses normalized by the number of only the target present trials. Miss rate equals probability of miss divided by the portion of target present trials $\frac{P(T|D_{1:k})}{P(T)}$.

The probability of miss $P(T|D_{1:k})$ can be estimated with Bayes' rule and written as a function of $P(D_{1:k}|T)$, the probability of encountering k distractors consecutively from the beginning when the target is present:[6]

$$P(T|D_{1:k}) = \frac{P(D_{1:k}|T) \cdot P(T)}{P(D_{1:k})} \tag{V.21}$$

$$= \frac{P(D_{1:k}|T) \cdot P(T)}{P(D_{1:k}|T) \cdot P(T) + P(D_{1:k}|\neg T) \cdot P(\neg T)} \tag{V.22}$$

$$= \frac{P(D_{1:k}|T) \cdot P(T)}{P(D_{1:k}|T) \cdot P(T) + [1 - P(T)]} \tag{V.23}$$

The probability of not attending to the target after processing k items when the target is present equals to the product of the probability of not attending to the target every step till k, i.e. $P(D_{1:k}|T) = \Pi_{i=1}^{k} P(D_i|T)$. When there are n distractors with distractor strength x, with inhibition rate γ, it is given by:

$$P(D_{1:k}|T) = \Pi_{i=1}^{k} \frac{nx - (i-1)\gamma x}{1 + nx - (i-1)\gamma x} \tag{V.24}$$

We do not specify the formula of cost functions W_m, W_f and W_t because we are trying to keep the framework as general as possible. We will discuss their

[6] Note that $P(T) + P(\neg T) = 1$ and that $P(D_{1:k}|\neg T) = 1$ because when the target is absent, only distractors will be attended to.

qualitative properties which leads to the common phenomenon known as the trade off between mistakes and time. Then we will develop further on a tractable special case to illustrate that the framework account for some other interesting phenomena.

From equation V.24 and V.23, when k increases, $P(T|D_{1:k})$ decreases. The intuition is straightforward: the more distractors you have encountered, the less likely the target is there. Assuming all cost functions of misses, false alarms and time are monotonically nondecreasing, then the total cost of reporting presence $WR{+}(k)$, given by equation V.20, is also monotonically nondecreasing. That says, if you are going to report presence of the target before actually seeing it, you might as well do so before the trial even started. This sounds pretty ridiculous. But imagine the scenario that false alarms do not cost anything ($W_f \equiv 0$) but a miss is lethal ($W_m \to \infty$), the optimal strategy is probably just to always say presence regardless of the display.

A normal visual search task, however, is not so drastic. Thus the optimal strategy is always to report absence if the target is not encountered. The total cost of reporting absence, given by equation V.18, has two components. The first term decreases with k but the second increases with k. In a reasonable setting, $k = 0$ does not provide least cost that the observers will not report absence regardless of the display, although this could happen in some drastic scenario. For example, if time is super precious ($W_t \to \infty$), and misses cost much less than false alarms ($W_m \ll W_f$). On the other hand, $W_{R-}(k)$ will not always decrease with an increasing k because when $k \to \infty$, the cost of possible misses $W_m(P(T|D_{1:k}))$ floors at $W_m(0)$, but the cost of time $W_t(k \cdot t_{dist})$ keeps growing, often super linearly[7].

Thus, the total cost of $W_{R-}(k)$ is minimized by an intermediate k, which provides a good trade off between time and error. The subject stops the trial and report absence, if k items are processed without seeing the target. k is dependent on many factors, including number of distractors n, inhibition rate γ, average time to process a distractor and of course the cost function of misses and time.

[7] Most people would rather do 1 hour of search task in a lab every day for 24 days rather than doing it consecutively for 24 hours.

In particular, in a reasonable setting, when n increases, if k stays the same, the cost of misses increases but that of time stays the same. On the other hand, if k increases to the point where the probability of miss holds still, the cost of misses stays the same, but that of time increases. Neither of these is likely to minimized $W_{R-}(k)$, the sum of the two component. The optimal k often lies in between, where both components increases by some amount, showing the trade off of time and error. The probability of miss goes up as well as the time spent when the number of distractors increase.

What we just discussed is that if n is given, what is the best k to minimize the cost. There could be another complication. If many search trials with various number of distractors are mixed, the global optimal strategy might trade off misses among trials and yield less cost than locally optimizing the cost function for each n. We are not going to discuss this in detail but will illustrate this effect with a special case. Consider the situation that the probability of miss must be kept below a certain threshold h[8]; as long as it is less than h, it does not matter what value it takes. In this case, the cost function of miss is: $W_m(y) = 0$ if $y \leq h$; $W_m(y) \to \infty$ if $y > h$. The best strategy is then to keep the probability of miss at h – spending least time while not violating the miss criteria. Note that this special case is equivalent to the stop criteria proposed in [4] where the subject rejects a trial when the probability of missing a target arrives a certain constant threshold. We will refer to this special case as “miss thresholding”.

If the observer is keeping the probability of every single trial at h, the optimal number of items to process before reporting absence k can be found by setting $P(T|D_{1:k}) = h$. From equation V.23, $P(D_{1:k}|T)$ can be solved as a function of $P(T|D_{1:k})$:

$$\begin{aligned} P(D_{1:k}|T) &= \frac{1-P(T)}{P(T)} \cdot \frac{P(T|D_{1:k})}{1-P(T|D_{1:k})} && \text{(V.25)} \\ &= \frac{1-P(T)}{P(T)} \cdot \frac{h}{1-h} && \text{(V.26)} \end{aligned}$$

[8] False alarms are punished so that simply reporting present regardless of the display is not acceptable.

For inference simplicity, assume distractor strength $x = 1$ that the target is as saliency as distractors, $\gamma = 1$ that the inhibition is strict. Equation V.24 simplifies to:

$$\begin{aligned} P(D_{1:k}|T) &= \Pi_{i=1}^{k} \frac{n-(i-1)}{1+n-(i-1)} && \text{(V.27)} \\ &= \frac{n-k+1}{n+1} && \text{(V.28)} \end{aligned}$$

Taking equation V.26 and V.28 together, we can solve k:

$$k = \left(1 - \frac{1-P(T)}{P(T)} \cdot \frac{h}{1-h}\right)(n+1) \tag{V.29}$$

Consider this experiment setup: the target is present in half of the trials, i.e. $P(T) = 0.5$; the probability of miss must be kept below 20%; and the number of distractors are either $n = 12$, $n = 24$ or $n = 36$, and each takes one third of the trials. If the trials are optimized individually, the best k for different n's are 9, 18 and 27 respectively, and the average processed items for rejected trials are 18 items. However, if the subject uses the strategy of keeping the probability of miss of $n = 12$ to 0, that of $n = 18$ to 20%, and that of $n = 24$ to 40%, the k is now 12, 18 and 12 respectively, resulting in the average number of processed items being 14. The strict inference could be much more complicated than what we just did in this special case. But the point is clear that optimizing for each n does not necessarily maximize the total cost over the entire trials. That being said, if a change of one probability term affect all local optimal strategy the same qualitative way despite of the value of n, the effect will likely applies to the global optimal strategy as well. This is the case for our following discussion and we will carry out our inference based on local strategies.

Now we will look at the effect of $P(T)$ on the decision making of when to stop. In our special case of miss thresholding where the probability of miss $P(D_{1:k}|T)$ is held still, the miss rate which is given by $\frac{P(D_{1:k}|T)}{P(T)}$ increases when $P(T)$ decreases. However, you may wonder, why to hold probability of misses still, but not to hold the miss rate still? This of course is dependent on the experimental setup. But often, the subjects are told to make as few mistakes as possible besides other things. The error rate is normalized by the total trial numbers. Thus, the miss rate is not as relevant because if the present trial takes up only 1% of all trials, the error from misses can not exceed 1% no matter how high the miss rate is.

This example provides some intuition about when the portion of target present trials goes down, the miss rate can go up without increasing the error rate. In general cases when $P(D_{1:k}|T)$ is not necessarily held constant, this intuition still applies. The effect of $P(T)$ can be seen qualitatively from equation V.25. When $P(T)$ decreases, to keep the left side and the right side of the equation the same, either $P(T|D_{1:k})$ has to decrease or k has to decrease (increase $P(D_{1:k}|T)$ on the left). These two changes, reflected in the cost function of equation V.18, increase the cost of miss or increasing the cost of time respectively. Generally, the optimum lies in between that both changes will be made. The decrease of k, however, will lead to an increase of miss rate, because when the target is present in the display, the subject is more likely to decide the target is absent before it is found and produces a miss.

That decreasing $P(T)$ will lead to earlier stop and higher miss rate has been reported by Wolfe et al. [93]. In that work, they increased the reward for hit and punishment for miss when $P(T)$ is lowered. This should bias subjects to avoid miss more in low $P(T)$ conditions. But since we do not know the cost functions of subjects[9], it is not very clear to what extend this manipulation biases the subjects' decisions. Nevertheless, if a miss is catastrophic and a false alarm is also costly, time becomes relatively cheap. The observer will take his time to make sure the target is present or absent regardless of $P(T)$. Thus we could take some relieve that in situations such as searching for a tumor in a film, a radiologist probably will not behave like subjects performing search tasks in labs who want to get out as fast as possible.

[9] It is often found to be sub-linear, e.g. losing $\$1,000$ is not feel 2 times as bad as losing $500.

V.B.1 Dynamic estimation of probability terms accounts for gambler's fallacy

Another interesting phenomenon is that subjects seem to dynamically adjusting their stopping criteria over trials. Chun and Wolfe showed that when the observer produces a miss, the next trial will take much longer time, as if the subject becomes more careful after the mistake [11]. This phenomenon is replicated by Wolfe et al. [93]. Furthermore, they observed "gambler's fallacy" that the search time decreases after a hit trial, as if the subject slack off after the correct response.

What do subjects adjust over trials that leads to the change of stopping criteria after a hit and a miss trial? Looking back at the cost function at equation V.18, the key players are the time spent and the probability of miss $P(T|D_{1:k})$, which is in turn decided by $P(T)$ and $P(D_{1:k}|T)$. The subject has the control of how much time to spend so there is not much uncertainty in regard to the cost of time. Furthermore, the probability of a target trial is often disclosed to the subject beforehand and held constant over the entire trial set. So there is not much uncertainty over $P(T)$ either. The term $P(D_{1:k}|T)$, however, is not so easy to calculate. The formula of $P(D_{1:k}|T)$ in our framework, given in equation V.24, is already complicated enough, not to mention this is under some simplification assumptions. An alternative way of calculating this probability for the subjects is to estimate it empirically. That is, of the target present trials, how often the target is not processed in the first k items.

Hit trials provide evidence toward $P(D_{1:k}|T) = 0$ as the target is attended to with k items, while miss trials provide evidence toward the opposite direction of $P(D_{1:k}|T) = 1$. A reasonable subject will take a history of trials into consideration and adjust his estimation of $P(D_{1:k}|T)$ with a small amount upon new evidence. Upon a hit trial, the estimation of $P(D_{1:k}|T)$ decreases. It is important to note that the underlying probability does not change but the empirical estimation of the subject decreases. Going back to our example of miss thresholding where the subject wants to hold $P(T|D_{1:k})$ still, the decrease of the estimation of $P(D_{1:k}|T)$ will lead to the decrease of the estimation of $P(T|D_{1:k})$ (equation V.23). To compensate for this decrease, k will be reduced, leading to less time spent before rejecting a trial. An intuitive way of reasoning from the subject's view is: since I successfully found the target with my current strategy, maybe my stopping criteria overkill and I can shorten the time a bit. The decrease of time in turn raises the miss rate right after a hit, because it was the estimation of $P(D_{1:k}|T)$ being adjusted, not the true underlying $P(D_{1:k}|T)$. The underestimation of $P(D_{1:k}|T)$ makes the subjects look less careful after a hit trial. A miss trial works the opposite way. The estimation of $P(D_{1:k}|T)$ increases after a miss trial, leading to the increase of time spent to reject a display. In general cases where the probability of miss does not need to be held still but trade off with time softly, the qualitative direction of the change of k still applies.

To further illustrate this effect, figure V.1 shows how the number of items to be processed k, varies near a miss trial and a hit trial, based on the example where the probability of miss is held still by simulation. In the simulation that generates this particular figure, the number of items in display is set to $n = 20$; the probability of target trials is 0.1; the probability of miss is set to $h = 0.01$ (resulting in a miss rate of 10%); and the learning rate of $P(D_{1:k}|T)$ is 0.5, i.e. $P(D_{1:k}|T) = 0.5 + 0.5P(D_{1:k}|T)$ after a miss trial and $P(D_{1:k}|T) = 0.5P(D_{1:k}|T)$ after a hit trial; 10,000 consecutive trials were simulated. We can compare this stimulation result to that reported in supplementary figure 1 of [93]. Although the exact shape of the curves differ from the human data (particularly that the processing time after a miss trial keeps high for many trials), our simulation of this special case of miss thresholding showed all three key observations made in [93]. First, the rejecting time is lower before a miss trial than that of a hit trial. Second, the rejecting time is raised by a significant amount after a miss trial. Third, the rejecting time is lowered after a hit trial. These three characteristic is present in our simulation for a wide range of parameter settings, i.e. the qualitative effect is intrinsic to our framework rather than a set of parameters.

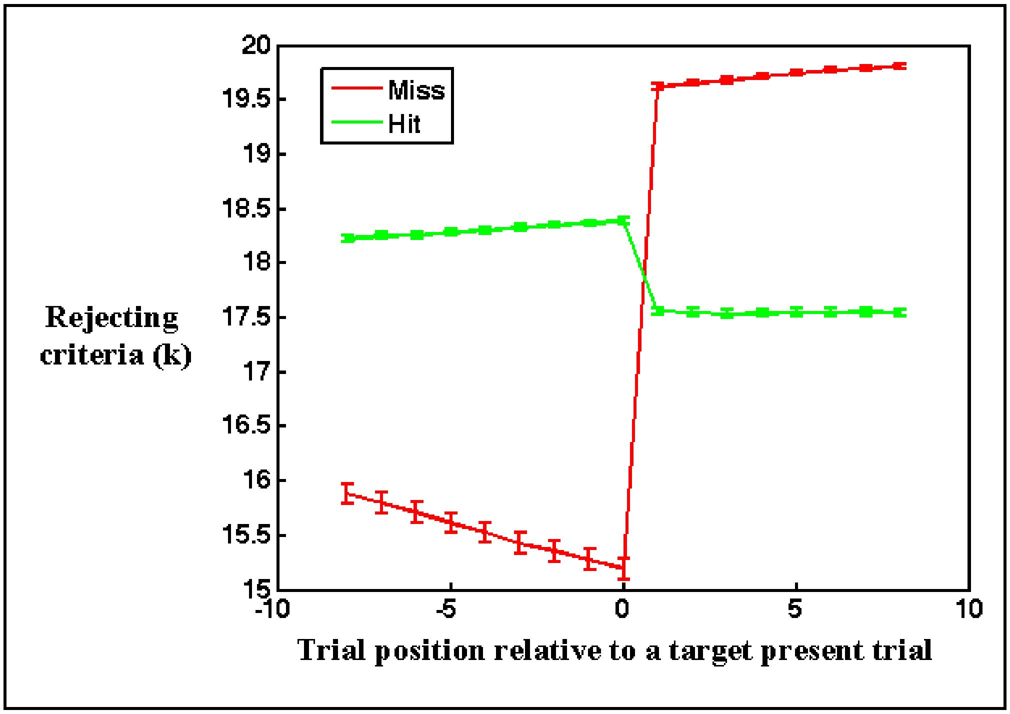

Figure V.1 The rejecting strategy changes as subjects dynamically adjusting the estimation of $P(D_{1:k}|T)$. The red curve shows the number of items to process before rejecting around a miss trial and the green curve shows that around a hit trial.

The above discussion is partly based on the assumption of the experimental setup that $P(T)$ is known to the subject and there is no need to estimate $P(T)$ on the fly. It is then interesting to speculate what if $P(T)$ is unknown and can change over time. In this case, subjects have to estimate $P(T)$ and adjust their strategy accordingly. Assuming $P(T)$ changes smoothly that the subject can estimate $P(T)$ over recent trials[10]. A target present trial provide evidence toward $P(T) = 1$ and a target absence trial provide evidence toward $P(T) = 0$. As discussed earlier, a larger $P(T)$ will lead to a larger k and a smaller $P(T)$ will lead to a smaller k. Target present trials will thus lead to longer rejecting time[11] while target absent trials will lead to shorter rejecting time.

[10] If $P(T)$ is random and not correlated in consecutive trials, there is not much the subject can do to predict $P(T)$.

[11] Less time does not necessarily lead to less miss in this case because $P(T)$ changes.

Imagine this scenario: you are going through a 1000-page-long book of essays to mark all the phrase "visual search". You went through the first 100 pages without seeing any such phrase. You figure you are probably at essays that do not deal with visual search very much. So you fastened your search and went through the next 400 pages with only rare occasions of seeing the phrase. Then you find your target phrase 10 times over the next 20 pages. You slow down your search and read the pages there after more carefully until the phrase stops showing up frequently around page 600. You then go through the rest 400 pages less carefully. While this task is not that realistic and is far from well controlled experimental setups, it conveys the idea that when the probability of the target frequency is unknown and changes dynamically, observers estimate it from recent experience and adjust their search strategy accordingly. When the target occurred frequently, it is likely to occur again in the immediate future and as the observer, you should keep your eyes open.

V.C Accounting for mean average response time

In earlier sections, we discussed what attracts attention, how attention is allocated and the stopping strategies in a search task. The previous qualitative accounts of visual search phenomena required no free parameters. As an exercise in showing that our model can account quantitatively for data as well, we present an example of fitting PS to response time data of target present and target absent trials in some visual search tasks. The human data presented here are recovered from figures in [74, 86, 91].

Our model is developed in a somewhat nontraditional way. We did not reason the structure from the human data, but inferred what should be calculated from several basic assumptions. One consequence of this is that our model is more complicated than necessary when coming to fit the response time data. For example, the average response time in target present trials often changes linearly with the number of distractors, thus only one parameter is necessary to fit the slope, e.g. the unit processing time of each item in feature integration theory. Our model, on the other hand, has three parameters that affect the slope: distractor strength x, inhibition ratio γ and average processing time for a distractor (equation V.6).

In this session, we will hold the inhibition ratio γ still and change the others to fit the data. We show that our framework can quantitatively fit to the data and we will discuss the qualitative aspects of the values of the parameters. But since there are infinite combinations of possible fits, we do not claim significance in the particular values assigned to the parameters.

When $\gamma = 0$, there is no inhibition of return. In session V.A.1, we showed that in this case our model is equivalent to that of [4] in selection and classification. Thus the prediction of expected target present response time is also equivalent. In target absent trials, [4] assumed that subjects reject the trial when a constant probability of missing a target is arrived. This assumption is equivalent to our example of "miss thresholding". We will continue to use this example in this session for inference simplicity, while keeping in mind that it is a special case where there is a hard cut off of misses but no trade-off with time. When we assume no inhibition of return and the scenario of miss thresholding for stopping strategy, we have exactly the same account for the data as in [4] [12].

Bundesen's model has been shown to account well for the mean response time in both target present and target absent trials in some tasks [4]. However, as you can probably tell that the assumption of no inhibition of return is not that realistic in our framework, there seems to exist some difficulty for this model. The mean response time of target absent trials and target present trials are both linear with the number of distractors in the model (equation 13 and 18 in [4]), providing a good fit to the human data. The slope ratio SR can be easily inferred from the

[12] The exact formula inferred from our framework will look slightly different to those in [4] because our inference is based on a serial structure while his on a parallel structure. Our inference is the discrete equivalence to his continuous inference.

two equations and is given by:

$$SR = \frac{-\log h}{1 + \frac{h \log h}{1-h}} \ , \tag{V.30}$$

where h is the threshold of the probability of misses defined in section V.B (denoted as r in [4]). As h is a probability term, it can only vary in the range from 0 to 1. Figure V.2 shows how the slope ratio varies as the threshold of probability of miss changes. The slope ratio SR approaches infinity when h approaches 0 and SR approaches 2 when h approaches 1. Note that the probability of misses h equals the miss rate times the portion of target present trials $P(T)$. In a typical experimental setup, target present trials take up half of the total trials, i.e. $P(T) = 0.5$. If the miss rate is to be kept below, for example, 20%, the probability of miss h has to be kept below 10%, i.e. $h < 10\%$. Consequently $SR > 3$, i.e. the slope ratio will always be bigger than 3. However, classical conjunction search produces slope ratio of 2 [74]. Furthermore, Wolfe showed that the the majority of the search tasks have slope ratios around 2 with deviations on both sides, and that very large and very small slope ratios are also observed for some tasks [88]. Thus, Bundesen's model, as well as our example of setting the probability of misses to a constant and assuming no inhibition of return, can not account for the range of smaller slope ratios.

Now we will discuss the situation in the other extreme where the inhibition of return is strict, i.e. $\gamma = 1$. This assumption is, if not more, equally unrealistic as assuming no inhibition of return. However, as the reality may lie anywhere in between, it is informative to discuss the extremes besides the benefit of simplified inference. When $\gamma = 1$, the mean response time in target present trials given by equation V.6 simplifies to:

$$RT_{prst} = \frac{x}{1+x} t_{dist} \cdot n + t_{prst}. \tag{V.31}$$

When the target is absent, if the stopping strategy is to reject after k items are processed, the expected response time is given by:

$$RT_{abst} = k \cdot t_{dist} + t_{abst}. \tag{V.32}$$

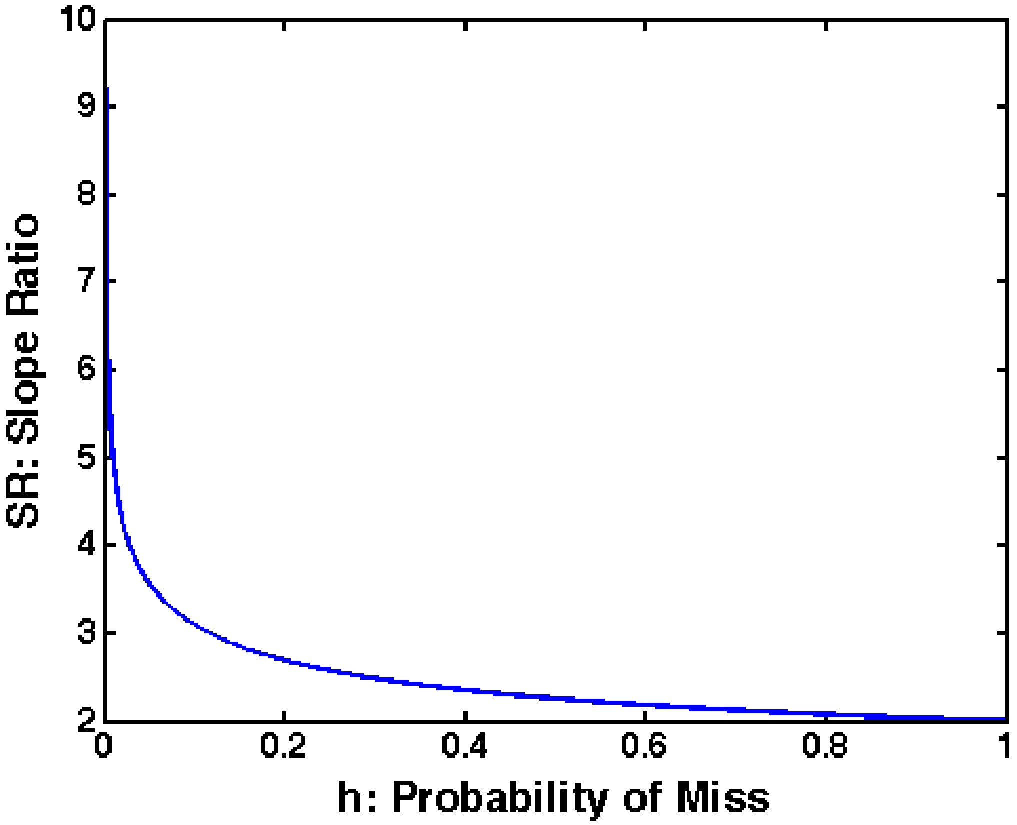

Figure V.2 The slope ratio is a function of probability of miss in Bundesen's model, equivalently in our special case with $\gamma = 0$. The slope ratio approaches infinity when the probability of miss approaches 0, and approaches 2 when that approaches 1.

where t_{abst} is the average extra time needed in an target absent trial.

We will again discuss target absent trials in the scenario of miss thresholding for simplicity, while the qualitative conclusions can be generalized as long as the cost functions do not go to extremes. With $\gamma = 1$, the probability of not attending to the target given the target is present, given by equation V.24, simplifies to:

$$P(D_{1:k}|T) = \Pi_{i=1}^{k} \frac{(n-i+1)x}{1+(n-i+1)x} \tag{V.33}$$

The optimal k can be inferred by considering the right side of equation V.33 and

V.26: k is the smallest of all k's that satisfy the following inequality:

$$\Pi_{i=1}^{k} \frac{(n-i+1)x}{1+(n-i+1)x} \leq \frac{1-P(T)}{P(T)} \cdot \frac{h}{1-h} \tag{V.34}$$

This does not have a closed form solution in general but simulation suggests that k is linear with n given x, h and $P(T)$. Three are three special cases where closed form solutions are available. We will discuss them each briefly to illustrate that our framework is capable to account for the human data in different scenarios.

The first situation is when the target and distractors are as salient as each other, i.e. $x = 1$. For example, the classical conjunction search for a red horizontal bar in a pool of green horizontal and red vertical bars fits into this scenario. We have discussed this case briefly in section V.B and k is given by equation V.29. In this case, k is linear in n, but the slope is dependent on $P(T)$ and h. The rejecting time grows when $P(T)$ increases, showing the effects of less presented target is likely to miss [93]; and it also grows when h decreases, showing the error-time trade off. In a standard experimental setup where $P(T) = 0.5$, the slope of target present trials is $\frac{x}{1+x}t_{dist} = 0.5t_{dist}$ and that of target absent trials is $(1 - \frac{h}{1-h})t_{dist}$. The slope ratio is then given by $2(1 - \frac{h}{1-h})$. When subjects are reasonably careful, h is small and the slope ratio is approximately 2. When $P(T)$ changes, the slope ratio will also change accordingly. Particularly, if $P(T)$ becomes very small, the slope of target absent trials may becomes smaller than the slope of target present trials $0.5t_{dist}$. That is, subjects spent less time before rejecting a trial than they need averagely to find a target when present. This is also observed in [93]. The intuition is straightforward in an extreme case, if $P(T) \to 0$ that the target almost never appears, the subjects can reject a trial without looking knowing the chance of miss is tiny although it will take them some effort (if they try) to find it when it is present.

The second case is when the target is very salient relatively to the distractors, i.e. $x \to 0$. The classical feature search where the target pops out fits into this scenario. In this case, as long as $h > 0$, inequality in V.34 can be arrived with $k = 1$. That is, when the cost of a miss is not infinity so that some trade off of misses and time is desirable, the trial can be rejected if the first items processed is not the target. Intuitively, the target is so salient that when it is present, the probability of it not being attended to the first is very small. Thus if the first item attended is not the target, the target is probably not present in the display and the subject can reject it confidently without processing a second item. This explains that the mean response time in very efficient search is flat not only for target present trials, but also for target absent trials [88, 91].

The third case is when $h = 0$. This assumption is not very realistic because it means that the subjects want to make perfect judgement no matter how long it takes, while trade off between mistake and time is commonly observed. However, if x is not too small and $P(T) \approx 0.5$, $h = 0$ is a good approximation to that when h takes a small value, which is often the case. Under this assumption, the subjects need to exhaustively scan every item before they reject a trial, i.e. $k = n$. Taken equation V.31 and V.32 which give the mean response time in target present and absent trials. We can fit parameters x, t_{dist}, t_{prst} and t_{abst} to human data in search tasks where the target does not pop out (when the target pops out, $x \rightarrow 0$, $h = 0$ is not a good approximation). Figure V.3 shows some example of fitting the model with this special case assumptions to some human data, and the corresponding parameters are shown in Table V.1.

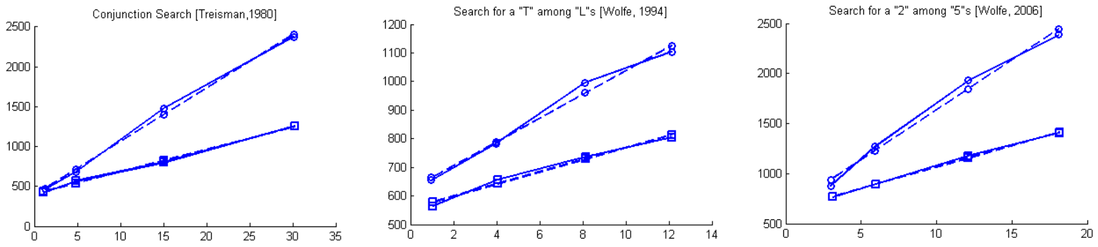

Figure V.3 PS is fit to four sets of human data. It provides the best linear fit to the data.

Before continue to the next section, it is of some interest to discuss the role of the parameters in search efficiency, in particular, the distractor strength x

Table V.1 The parameters used in Figure V.3

	x	t_{dist} (ms)	t_{prst} (ms)	t_{abst} (ms)
Conjunction	0.72	66.91	402.24	393.39
T vs. L	1.04	41.40	556.40	623.54
2 vs. 5	0.75	99.50	640.54	640.57

and the distractor processing time t_{dist}. Before deciding their roles, however, we should first examine what affects these two parameteres.

The distractor strength x, defined in equation V.1, is the relative rareness of the visual features of a distractor and a target. A smaller x means that the target, when present, is more likely to be attended to. This leads to a faster search by reducing the expected number of items to be processed before the target is attended to when present. The rareness of visual features is dependent on two major factors. One is the item's own features such as color, orientation, etc. Another is the item's contrast to its neighbor items, such as color contrast, orientation difference, etc. When the target and the distractor are similar, the second factor is small and the first dominates, and we observe search asymmetries when the target and the distractor switch roles, as discussed in section II.B.3. When the target and the distractor are very different, the second factor can override the first one and becomes dominant. The bigger the neighborhood difference, the smaller the probability as homogeneous surface and texture are frequent in natural statistics[13]. When the difference between the target and the distractors, or the difference among the distractors are manipulated, x changes accordingly, resulting in changes of search efficiency, as discussed in V.A.1.

t_{dist}, on the other hand, is time needed to classify a distractor as a distractor but not a target. It also depends on two factors. One is the complexity of the distractor. For example, it should be easier to classify color bars than classify faces. The other, again, is the similarity between the target and the distractor. For example, if the target is red, it is probably easier to classify a green distractor

[13] ICA algorithms have shown to learn features from natural images with sparse marginal distributions that respond to edges and texture borders [34, 53, 64].

as a non-target than classify an orange distractor as a non-target.

The efficiency of a search task often refers to the how fast a target can be found when present in a pool of distractors, e.g. the search is efficient if the slope is less than 5ms/item but very inefficient if it is more than 20ms/item [87]. The slope of mean response time when the target is present, given by $\frac{x}{1+\gamma x}t_{dist}$ (equation V.6) is relevant to both x and t_{dist}. Thus, efficiency of a search task is a one dimension projection of underlying multi-dimensional factors, including the difference between the target and the distractors, their own feature probability, and their complexity. Two search tasks often differ in more than one of these factors and the difference observed in efficiency is from the changed underlying factors combined.

V.D Acknowledgment

Chapter V, in part, is a reprint of the paper in preparation "Probabilistic Search: a New Theory on Visual Search", co-authored with Matthew H. Tong and Garrison W. Cottrell. The dissertation author is the primary investigator and author of these two papers.

VI Summary

In this dissertation, we fist developed a saliency framework in a principled way by investigating the goal of the visual system. Bottom-up saliency falls out from out framework as self-information and overall saliency as pointwise mutual information. Our definition of saliency, different from previous methods, is based on natural statistics. It accounts for feature and conjunction search, as well as many search asymmetries straightforwardly. We then implemented efficient bottom-up saliency algorithms on static images and dynamic scenes. They perform as well as state of art saliency algorithms in predicting human eye fixations. We further implemented a simplified, but real time bottom-up saliency algorithm on a robotic camera. The camera is orientated towards salient location in the space and it greatly improved the chance of seeing people.

We then investigated how attention is directed given a saliency map. We treat it as a decision making problem and hypothesized that it shares higher level decision making characteristics: probability matching. That is, attention is directed probabilistically according to the saliency map. Based on this hypothesis, we inferred that the response time in a visual search task when the target is present is linear with the number of distractors. Furthermore, the slope is positively correlated to the distractor strength, defined as the ratio of the distractor salience to the target salience. The two special cases of distractor strength near 0 and distractor strength being 1, corresponds to the feature search and conjunction search in feature integration theory. When the distractor strength varies, the difficulty of the visual search task varies, corresponding to different level of efficiency in guided search model.

We further treat visual search as a sequential decision making process and discussed when the subject will stop if the target is not found. Without specifying the cost functions nor fit any parameters, we were able to account for four phenomena observed: (1) miss rate increases with number of distractors; (2) miss rate increases when the portion of target present trials decreases; (3) search time decreases and miss rate increases after a hit trial, known as the "gambler's fallacy"; (4) search time increases and miss rate decreases after a miss trial.

Throughout this dissertation, we are trying to build a theoretic framework that account for human behavior as well as allow efficient implementation on complex images and videos so that it is also of value to computer vision and graphics. Most of work contained here, however, only deals with bottom-up saliency which is dependent on the stimuli but not take tasks into consideration. Also, our current work does not take visual acuity and temporal changes of saliency into account. How top down tasks affect visual attention, how saliency is updated over time, and how eye movements are planned sequentially are of great interest for future works.

Bibliography

[1] H. Barlow. What is the computational goal of the neocortex? In C. Koch, editor, *Large scale neuronal theories of the brain*, pages 1–22. MIT Press, Cambridge, MA, 1994.

[2] A.J. Bell and T.J. Sejnowski. The independent components of natural scenes are edge filters. *Vision Research*, 37(23):3327–3338, 1997.

[3] Neil Bruce and John Tsotsos. Saliency based on information maximization. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 155–162, Cambridge, MA, 2006. MIT Press.

[4] C. Bundesen. A theory of visual attention. *Psychological Review*, 97(4):523–547, 1990.

[5] C. Bundesen. A computational theory of visual attention. *Philosophical Transactions: Biological Sciences*, 353(1373):1271–1281, 1998.

[6] C. Bundesen, LF Pedersen, and A. Larsen. Measuring efficiency of selection from briefly exposed visual displays: a model for partial report. *Journal Experimental Psychology: Human Perception and Performance*, 10(3):329–39, 1984.

[7] D. Cai, G.C. Deangelis, and R.D. Freeman. Spatiotemporal receptive field organization in the lateral geniculate nucleus of cats and kittens. *Journal of Neurophysiology*, 78(2):1045–1061, 1997.

[8] R. Carmi and L. Itti. The role of memory in guiding attention during natural vision. *Journal of Vision*, 6(9):898–914, 2006.

[9] R. F. Caron and A. J. Caron. The effects of repeated exposure and stimulus complexity on visual fixation in infants. *Psychonomic Science*, 10:207–208, 1968.

[10] A. Chauvin, J. Herault, C. Marendaz, and C. Peyrin. Natural scene perception: visual attractors and image processing. *Progress in Neural Processing*, pages 236–248, 2002.

[11] M.M. Chun and J.M. Wolfe. Just say no: How are visual searches terminated when there is no target present? *Cognitive Psychology*, 30(1):39–78, 1996.

[12] C. Elkan. The foundations of cost-sensitive learning. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI'01)*, 2001.

[13] J. F. Fagan. Memory in the infant. *Journal of Experimental Child Psychology*, 9:217–226, 1970.

[14] R.L. Fantz. Visual experience in infants: Decreased attention to familiar patterns relative to novel ones. *Science*, 146(3644):668, 1964.

[15] I. Fasel, B. Fortenberry, and J.R. Movellan. A generative framework for real-time object detection and classification. *Computer Vision and Image Understanding*, 98:182–210, 2005.

[16] S. Friedman. Habituation and recovery of visual response in the alert human infant. *Journal of Experimental Child Psychology*, 13:339–349, 1972.

[17] U. Frith. A curious effect with reversed letters explained by a theory of schema. *Perception and Psychophysics*, 16:113–116, 1974.

[18] Dashan Gao and Nuno Vasconcelos. Discriminant saliency for visual recognition from cluttered scenes. In Lawrence K. Saul, Yair Weiss, and Léon Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 481–488, Cambridge, MA, 2004. MIT Press.

[19] Dashan Gao and Nuno Vasconcelos. Integrated learning of saliency, complex features, and object detectors from cluttered scenes. In *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2*, pages 282–287, Washington, DC, USA, 2005. IEEE Computer Society.

[20] Dashan Gao and Nuno Vasconcelos. Bottom-up saliency is a discriminant process. In *IEEE International Conference on Computer Vision*, Rio de Janeiro, Brazil, 2007.

[21] D.E. Goldberg. Probability matching, the magnitude of reinforcement, and classifier system bidding. *Machine Learning*, 5(4):407–425, 1990.

[22] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. In *Advances in Neural Information Processing Systems 19*, Cambridge, MA, 2007. MIT Press.

[23] T.S. Horowitz and J.M. Wolfe. Visual search has no memory. *Nature*, 394(6693):575–577, 1998.

[24] T.S. Horowitz and J.M. Wolfe. Search for multiple targets: Remember the targets, forget the search. *Perception & Psychophysics*, 63(2):272–285, 2001.

[25] T.S. Horowitz and J.M. Wolfe. Memory for rejected distractors in visual search? *Visual Cognition*, 10(3):257–298, 2003.

[26] Aapo Hyvarinen and Erkki Oja. A fast fixed-point algorithm for independent component analysis. *Neural Computation*, 9(7):1483–1492, 1997.

[27] L. Itti. Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes. *Visual Cognition*, 12(6):1093–1123, 2005.

[28] L. Itti and P. Baldi. Bayesian surprise attracts human attention. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 1–8, Cambridge, MA, 2006. MIT press.

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

[30] L. Itti and C. Koch. Computational modeling of visual attention. *Nature Reviews Neuroscience*, 2(3):194–203, 2001.

[31] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998.

[32] Laurent Itti and Pierre Baldi. A principled approach to detecting surprising events in video. In *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1*, pages 631–637, Washington, DC, USA, 2005. IEEE Computer Society.

[33] T. Kadir and M. Brady. Saliency, scale and image description. *International Journal of Computer Vision*, 45(2):83–105, 2001.

[34] Y. Karklin and M.S. Lewicki. Learning higher-order structures in natural images. *Network: Computation in Neural Systems*, 14(3):483–499, 2003.

[35] Wolf Kienzle, Felix A. Wichmann, Bernhard Schlkopf, and Matthias O. Franz. A nonparametric approach to bottom-up visual saliency. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 689–696, Cambridge, MA, 2007. MIT Press.

[36] R.A. Kinchla. Detecting targets in multi-element arrays: A confusability model. *Perception & Psychophysics*, 15:149–151, 1974.

[37] R. Klein. Inhibitory tagging system facilitates visual search. *Nature*, 334(6181):430–431, 1988.

[38] C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Hum Neurobiol*, 4(4):219–27, 1985.

[39] T.S. Lee and S.X. Yu. An information-theoretic framework for understanding saccadic eye movements. In *Advances in Neural Information Processing Systems 12*, Cambridge, MA, 2000. MIT Press.

[40] W. Lee. *Decision theory and human behavior.* Wiley New York, 1971.

[41] D.T. Levin. Classifying faces by race: The structure of face categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22:1364–1382, 1996.

[42] D.T. Levin. Race as a visual feature: Using visual search and perceptual discrimination tasks to understand face categories and the cross-race recognition deficit. *Journal of Experimental Psychology: General*, 129(4):559–574, 2000.

[43] Z. Li. A saliency map in primary visual cortex. *Trends in Cognitive Sciences*, 6(1):9–16, 2002.

[44] R. D. Luce. *Individual choice behavior.* New York: Wiley, 1959.

[45] N.J. Mackintosh. *The Psychology of Animal Learning.* Academic Press New York, NY, 1974.

[46] P. McLeod, J. Driver, and J. Crisp. Visual search for a conjunction of movement and form is parallel. *Nature*, 332:154–155, 1988.

[47] J. R. Movellan, F. Tanaka, B. Fortenberry, and K. Aisaka. The rubi project: Origins, principles and first steps. In *Proceedings of the International Conference on Development and Learning (ICDL05)*, Osaka, Japan, 2005.

[48] Javier R. Movellan, Fumihide Tanaka, Ian R. Fasel, Cynthia Taylor, Paul Ruvolo, and Micah Eckhardt. The rubi project: a progress report. In *Proceedings of the ACM/IEEE international conference on Human-robot interaction*, Virginia, USA, 2007.

[49] HJ Muller and A. von Muhlenen. Probing distractor inhibition in visual search: inhibition of return. *J Exp Psychol Hum Percept Perform*, 26(5):1591–605, 2000.

[50] HC Nothdurft. Faces and facial expressions do not pop out. *Perception*, 22(11):1287–1298, 1993.

[51] A. Oliva, A. Torralba, MS Castelhano, and JM Henderson. Top-down control of visual attention in object detection. In *Proceedings of International Conference on Image Processing*, pages 253–256, Barcelona, Catalonia, 2003. IEEE press.

[52] Bruno Olshausen and David Field. Sparse coding of natural images produces localized, oriented, bandpass receptive fields. *Technical Report CCN-110-95, Department of Psychology, Cornell University, Ithaca, New York 14853*, 1995.

[53] S. Osindero, M. Welling, and G.E. Hinton. Topographic product models applied to natural scene statistics. *Neural Computation*, 18(2):381–414, 2005.

[54] J. Palmer. Attention in visual search: Distinguishing four causes of a set-size effect. *Current Directions in Psychological Science*, 4(4):118–123, 1995.

[55] D. Parkhurst, K. Law, and E. Niebur. Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, 42(1):107–123, 2002.

[56] D.J. Parkhurst and E. Niebur. Scene content selected by active vision. *Spatial Vision*, 16(2):125–154, 2003.

[57] M. I. Posner and Y. Cohen. Components of attention. In H. Bouma and D. G. Bouwhuis, editors, *Attention and Performance X*, pages 55–66. Erlbaum, 1984.

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

[59] R. Ratcliff. A theory of memory retrieval. *Psychological Review*, 85(2):59–108, 1978.

[60] Laura Walker Renninger, James M. Coughlan, Preeti Verghese, and Jitendra Malik. An information maximization model of eye movements. In Lawrence K. Saul, Yair Weiss, and Léon Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 1121–1128, Cambridge, MA, 2004. MIT Press.

[61] R. Rosenholtz. A simple saliency model predicts a number of motion popout phenomena. *Vision Res*, 39(19):3157–63, 1999.

[62] D.L. Ruderman. The statistics of natural images. *Network: Computation in Neural Systems*, 5(4):517–548, 1994.

[63] P. N. Sabes and M. I. Jordan. Reinforcement learning by probability matching. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, editors, *Advances in Neural Information Processing Systems (NIPS) 8*. MIT Press, 1996.

[64] Honghao Shan, Lingyun Zhang, and Garrison W. Cottrell. Recursive ica. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 1273–1280, Cambridge, MA, 2007. MIT Press.

[65] D.R. Shanks, R.J. Tunney, and J.D. McCarthy. A re-examination of probability matching and rational choice. *Journal of Behavioral Decision Making*, 15(3):233–250, 2002.

[66] J. Shen and E.M. Reingold. Visual search asymmetry: The influence of stimulus familiarity and low-level features. *Perception & Psychophysics*, 63(3):464–75, 2001.

[67] Kaisheng Song. A globally convergent and consistent method for estimating the shape parameter of a generalized gaussian distribution. *IEEE Transactions on Information Theory*, 52(2):510–527, 2006.

[68] Y. Takeda and A. Yagi. Inhibitory tagging in visual search can be found if search stimuli remain visible. *Percept Psychophys*, 62(5):927–34, 2000.

[69] BW Tatler, RJ Baddeley, and ID Gilchrist. Visual correlates of fixation selection: effects of scale and time. *Vision Research*, 45(5):643–59, 2005.

[70] Thomas L. Thornton and David L. Gilden. Parallel and serial processes in visual search. *Psychological Review*, 114(1):71–103, 2007.

[71] A. Torralba, A. Oliva, M. S. Castelhano, and J. M. Henderson. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological Review*, 113(4):766–786, 2006.

[72] J. T. Townsend. A note on the identifiability of parallel and serial processes. *Perception & Psychophysics*, 10(3):161–163, 1971.

[73] J. T. Townsend. Some results concerning the identifiability of parallel and serial processes. *British Journal of Mathematical and Statistical Psychology*, 25:168–199, 1972.

[74] A. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12:97–136, 1980.

[75] A. Treisman and S. Gormican. Feature analysis in early vision: evidence from search asymmetries. *Psychological Review*, 95(1):15–48, 1988.

[76] A. Treisman and J. Souther. Search asymmetry: a diagnostic for preattentive processing of separable features. *Journal Experimental Psychology: General*, 114(3):285–310, 1985.

[77] Shimon Ullman, Michel Vidal-Naquet, and Erez Sali. Visual features of intermediate complexity and their use in classification. *Nature Neuroscience*, 5(7):682–687, 2002.

[78] A. van der Schaaf and J. H. van Hateren. Modelling the power spectra of natural images: statistics and information. *Vision Research*, 36(17):2759–70, 1996.

[79] P. Verghese. Visual search and attention a signal detection theory approach. *Neuron*, 31(4):523–535, 2001.

[80] N. Vulkan. An economist's perspective on probability matching. *Journal of Economic Surveys*, 14(1):101–118, 2000.

[81] K.-P. Hoffmann W. Einhauser, W. Kruse and P. Konig. Differences of monkey and human overt attention under natural conditions. *Vision Research*, 46(8-9):1194–1209, 2006.

[82] T. Wachtler, T. Doi E, Lee, and T.J. Sejnowski. Cone selectivity derived from the responses of the retinal cone mosaic to natural scenes. *Journal of Vision*, 7(8):1–14, 2007.

[83] M.J. Wainwright, O. Schwartz, and E.P. Simoncelli. Natural image statistics and divisive normalization: modeling nonlinearities and adaptation in cortical neurons. *Statistical Theories of the Brain*, pages 203–22, 2002.

[84] Q. Wang, P. Cavanagh, and M. Green. Familiarity and pop-out in visual search. *Perception & Psychophysics*, 56(5):495–500, 1994.

[85] R. Ward and J.L. McClelland. Conjunctive search for one and two identical targets. *Journal of Experimental Psychology: Human Perception and Performance*, 15(4):664–672, 1989.

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

[87] J.M. Wolfe. Visual search. In H. Pashler, editor, *Attention*, pages 13–73. University College London Press, London, UK, 1998.

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

[89] J.M. Wolfe. Asymmetries in visual search: An introduction. *Perception & Psychophysics*, 63(3):381–389, 2001.

[90] J.M. Wolfe. Moving towards solutions to some enduring controversies in visual search. *Trends in Cognitive Sciences*, 7(2):70–76, 2003.

[91] J.M. Wolfe. Guided search 4.0: Current progress with a model of visual search. In *Integrated Models of Cognitive Systems*. Oxford, New York, 2006.

[92] J.M. Wolfe, KR Cave, and SL Franzel. Guided search: an alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human Perception and Performance*, 15(3):419–33, 1989.

[93] J.M. Wolfe, TS Horowitz, and NM Kenner. Rare items often missed in visual searches. *Nature*, 435(7041):439–440, 2005.

[94] Keiji Yamada and Garrison W. Cottrell. A model of scan paths applied to face recognition. In *Proceedings of the Seventeenth Annual Cognitive Science Conference*, pages 55–60, Pittsburgh, PA, 1995. Mahwah: Lawrence Erlbaum.

[95] Lingyun Zhang, Matthew H. Tong, and Garrison W. Cottrell. Information attracts attention: a probabilistic account of the cross-race adavantage in visual search. In *Proceedings of the 29th Annual Cognitive Science Conference*, Nashville, Tennessee, 2007. Cognitive Science Society.

[96] L. Zhaoping and R.J. Snowden. A theory of a saliency map in primary visual cortex (v1) tested by psychophysics of colour–orientation interference in texture segmentation. *Visual Cognition*, 14(4):911–933, 2006.