SALIENCY, ATTENTION AND VISUAL SEARCH: AN INFORMATION THEORETIC APPROACH

NEIL D. B. BRUCE

A DISSERTATION SUBMITTED TO THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN COMPUTER SCIENCE AND ENGINEERING
YORK UNIVERSITY
TORONTO, ONTARIO
JULY 2008

SALIENCY, ATTENTION AND VISUAL SEARCH:
AN INFORMATION THEORETIC APPROACH

by **Neil D. B. Bruce**

a dissertation submitted to the Faculty of Graduate Studies of York University in partial fulfilment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

SALIENCY, ATTENTION AND VISUAL SEARCH: AN INFORMATION THEORETIC APPROACH

by **Neil D. B. Bruce**

Examination Committee Members:

1. John K. Tsotsos
2. Richard P. Wildes
3. Robert S. Allison
4. J. Douglas Crawford
5. Andrew W. Eckford
6. Ronald A. Rensink

Abstract

This dissertation explores the concept of visual saliency as it pertains to attentional selection, visual search, and machine vision. A novel framework for visual saliency is put forth derived from consideration of the problem in the context of information theory. The proposed definition is distinguished from previous efforts on this front and is demonstrated to be a natural principled definition for salient visual content. Specifically, the proposal deemed Attention by Information Maximization (AIM) seeks to select visual content that is most informative in a formal sense in the context of a specific scene, and is put forth in a form that is amenable to considering more general definitions of context. Efficacy in predicting human gaze patterns is demonstrated and the proposal is revealed to outperform existing models in the prediction of fixation points. With regard to biological plausibility, an important consideration is the extent to which the model behavior agrees with the psychophysics and neurophysiology literature. To this end it is revealed that AIM is able to account for an unprecedented range of classic psychophysics

results including some subtle and counterintuitive results and may be achieved via a neural implementation that is consistent with observations concerning surround modulation in the cortex. More general modeling considerations are also addressed including compatibility with descriptions of how attention as a whole is achieved and constraints on possible architectures for achieving attentional selection in light of recent psychophysics and neural imaging results. The applicability of this definition within a machine vision context is also discussed revealing some interesting properties as emergent from the basic framework.

Acknowledgements

I wish to express my sincere gratitude to my supervisor, John Tsotsos, for his support and guidance during the completion of this dissertation. For allowing me the freedom to explore my interests, with understanding and continual encouragement and intellectual support. For providing a context in which to learn matters that transcend any specific subject knowledge.

To my parents and family, for their continued and unwavering effort in helping and encouraging me to realize my goals. For their love, understanding, support, and advice without which this work would not have been possible.

To my wife, Chrissie, for her patience, caring, companionship, compassion and love. For enriching my life and being a constant source of motivation and happiness.

To my lab mates and friends for providing necessary distractions from work and bringing balance and richness to my doctoral experience.

To my committee, Dr. Richard Wildes, Dr. Robert Allison, Dr. Douglas Crawford, Dr. Andrew Eckford and to my external examiner Dr. Ronald Rensink,

for providing fruitful suggestions and invaluable feedback that helped to better this dissertation.

To Dr. Laurent Itti, for sharing his spatiotemporal eye tracking data allowing a more thorough assessment of the modeling work.

Finally, I wish to acknowledge the financial support received from the Ontario Graduate Scholarship Program and the National Sciences and Engineering Research Council.

Table of Contents

List of Tables

List of Figures

Abbreviations

AIM - Attention based on Information Maximization: The model put forth in this dissertation.

DCT - Discrete Cosine Transform: A sequence of finitely many data points expressed as a sum of cosine functions at different frequencies.

FIT - Feature Integration Theory: An early influential attention model first put forth in [222].

fMRI - Functional Magnetic Resonance Imaging: A specialized form of brain imaging that measures blood flow and blood oxygenation related to neural activity.

ICA - Independent Component Analysis: A computational means of separating a set of mixed multivariate signals into additive subcomponents assumed to be mutually independent.

infomax - An optimization principle for neural information processing systems that seeks a function that maps a set of input values to a set of outputs such that the Shannon mutual information between input and output values is maximized

(see [131]).

IT - Inferotemporal Cortex: A higher cortical area that forms part of the ventral visual processing stream and is thought to represent shape and complex object representations.

Jade - An algorithm that learns independent components based on higher order cumulants by Jean-François Cardoso [32].

LIP - The Lateral Intraparietal area: Part of the parietal lobe of the brain and thought to be involved in targeting of saccadic eye movements and possibly the representation of visual salience.

LGN - The Lateral Geniculate Nucleus: Part of the thalamus, and an early visual area that receives direct input from the retina.

MST - Medial Superior Temporal area: A late visual processing region along the dorsal pathway implicated in visual navigation, and optic flow.

MT - Medial Temporal area: A key area in motion processing along the dorsal visual processing stream.

PCA - Principal Component Analysis: A vector space transform, often used to reduce multidimensional data sets to a lower dimensional space for analysis.

RF - Receptive Field: The region of space in which a neuron will elicit a response to stimuli.

ROC curve - Receiver Operating Characteristic Curve: A graphical representa-

tion of the number of true positives versus false positives for a binary classifier as its decision threshold is varied.

ROI - Region of Interest.

V1 - The primary visual cortex: The simplest earliest cortical visual area located at the posterior pole of the occipital cortex.

V2 - Visual area V2: The second major area of the visual cortex, with many properties similar to primary visual cortex.

V4 - Visual area V4: An intermediate area on the ventral processing stream, thought to represent intermediate complexity object features such as simple geometric shapes.

WTA - Winner-Take-All: A competitive strategy in recurrent neural networks. Output nodes inhibit each other and in some instances activate themselves reflexively. The result is typically that a single node in the output remains active.

1 General Introduction

Visual attention is a problem that has been studied intensely, and encompasses a wide range of different subareas, each of which are important in understanding the overall system. In this dissertation, the focus is on the notion of visual saliency and the bottom-up processes that give rise to its computation. That is, what is it about certain visual patterns that result in attention being directed in an automatic and rapid fashion such as a bright light, a colored sign or rapid movement. Alternatively, why is it that certain visual patterns are difficult to detect, such as an animal camouflaged by its natural environment. These considerations motivate the subject matter of this dissertation which considers the cortical basis for the computation that underlies these behavioral observations. The basis for such computation is explored in a principled manner, through consideration of information coding within the cortex. A variety of interesting observations emerge from this formulation including a system that is better able to capture eye movement patterns than its predecessors, that explains a wide range of psychophysics behavior, and may be

achieved by way of an implementation reminiscent of circuitry observed within the visual cortex. The resultant system also has important applications in a machine vision context, and may be of interest to those carrying out research in machine vision and image and signal processing. The fundamental principal by which the proposed model operates is that of selecting visual content that is *informative*, and the model is referred to as Attention by Information Maximization (AIM) in light of this.

1.1 Motivation

There are a variety of considerations relating to the notion of saliency that leave much to be desired. The majority of models that provide an explanation of saliency related computation assume that saliency is represented in a single hierarchical topographical representation, and that selection of regions of interest proceeds on the basis of this representation (e.g. [222] shows the basic structure that motivates existing approaches). There are a variety of observations concerning attentional selection derived from the behavioral and imaging literature that simply do not support this sort of architecture. That being said, one ambition of this dissertation is to disentangle the notion of saliency from the associated attentional selection strategy. An additional shortcoming of existing saliency based models, is that they may be viewed in some cases as simply an approximate implementation of general

observations concerning cortical computation, but say little concerning the underlying impetus for such computation and are vague with respect to some important details [97; 98]. In this dissertation, consideration of saliency computation proceeds on the basis of a principled assumption concerning the role of saliency related computation and the relation of this computation to cortical and behavioral manifestations is secondary. This is important in that the end result has greater value in the sense that it offers insight on the possible *raison d'être* for saliency based computation. A secondary approach to consideration of saliency in the past has been the examination of the relationship between somewhat *ad hoc* feature measures and fixational eye-tracking data. This is an avenue that has not been fruitful as there is no single measure that appears to correlate strongly with fixated content, and this approach informs little on the motivation underlying cortical saliency related computation. As a whole, there exist many aspects of saliency related computation that leave much to be desired, and these are highlighted in further detail in chapter 3.

1.2 Contributions

The central contribution of this dissertation concerns the claim that bottom-up saliency related computation should serve to effectively maximize information sampled from one's environment insofar as this determination is based on stimulus

properties. A formal infrastructure is derived on the basis of this premise and the plausibility of this proposal is evaluated via a variety of different avenues. Foremost, it is demonstrated that the proposed model offers greater agreement in predicting fixational eye movements than any existing effort. This is achieved using two separate data sets with very different properties.

An additional important aspect of the proposed computation concerns a possible neural analogue. To this end, it is demonstrated that the proposed model requires circuitry with many properties exhibiting substantial agreement with local surround computation in the cortex.

Certain prior efforts at characterizing saliency computation, in particular those derived from the psychophysics literature examine the extent to which model predictions agree with the large body of psychophysics results that exist concerning visual search associated with various target-distractor paradigms [197; 222; 248; 251]. To this end, it is demonstrated that the proposed model agrees with a greater range of observations in this domain than existing efforts and additionally offers insight on some experimental domains that previously lacked any satisfactory explanation. This arrives in part through emphasizing the role that the statistics of the natural world play in the formation of visual circuitry. Specifically, construction of a representation optimal for representing said statistics produces a variety of behaviors as emergent properties.

In the domain of attention modeling in general, there lacks agreement as to the specific mechanism underlying attentional selection and the neural representation upon which attentional selection operates. To this end, it is demonstrated that the proposed model is compatible with a variety of claims concerning how attentional selection operates and in particular, it is established that the proposed model is amenable to operation within a distributed hierarchical representation. It is argued that this sort of representation has the most favorable supporting experimental evidence based on recent work.

Finally, a central role of attention in the domain of machine vision, concerns the selection of smaller subsets of the visual field for directed processing. The importance of this derives from its use as a pre-cursor to object recognition and localization among other machine vision tasks. It is demonstrated within this dissertation, that the saliency proposal lends itself to selection of stable points and regions of interest, emerging as a potentially important contribution to the machine vision literature.

As a whole the dissertation seeks to provide the foundation for a variety of interesting potential avenues for future research efforts. In addition, the work affords a definition of saliency computation that is principled and demonstrates greater agreement with eye tracking, psychophysics, and anatomical data and is a better fit with recent thinking concerning general cortical attentional architectures than

its predecessors.

1.3 Organization

The content of the dissertation and its organization is as follows:

As an ambition of this dissertation lies in placing the notion of saliency within a broader context of how attention is achieved in general, chapter 2 reviews the history of modeling work surrounding attention, describing the varying opinions on the macroscopic structure of the system that orchestrates attention within the cortex. A basic understanding of the structure and properties of the human visual system is assumed. The emphasis is on providing a computational description, and the elements included are intended to provide the basic background necessary to understand the dissertation as a whole.

A more specific aspect of attention concerns the notion of visual saliency, and this concept is explored in detail in chapter 3. In particular, an attempt is made to precisely define what the term *saliency* is intended to capture and to review in a critical fashion, prior definitions for what constitutes salient content. This chapter serves a dual role in giving a more precise history of saliency than that offered by chapter 2, while additionally providing motivation for the balance of the dissertation by way of pointing out shortcomings of prior work.

The core of the dissertation lies in a principled definition for what comprises

salient visual content, and this definition is motivated in chapter 4. Cast as a problem of information coding, a basic framework is described including its ties to cortical machinery and preliminary results on eye tracking data are presented.

A detailed analysis of performance in predicting eye movements appears in chapter 5 including an exhaustive exploration of the parameter space and a detailed look at implementation issues.

An important aspect in the validation of the model, is the extent to which it may be achieved within a neural implementation, and moreover, the extent to which this implementation is consistent with observed neuroanatomy. This consideration and the associated details are considered in chapter 6. A proposal for a simple neural circuit that is sufficient to achieve the desired implementation is presented and shown to have a striking relationship to certain circuitry observed within the visual cortex.

An additional *litmus test* on the plausibility of the model, lies in the extent to which it agrees with behavioral results. Chapter 7 explores model behavior in the context of a variety of classic psychophysics paradigms, demonstrating agreement of model behavior across a wide range of classic visual search experiments. Insight is also gained on the impetus for certain visual search behaviors when viewed in the context of AIM.

As mentioned earlier in this chapter, an ambition of this dissertation lies in

connecting the notion of saliency to the broader body of work concerning how attentional selection in general is achieved within the cortex. As there exist many different proposals for how attention is achieved, chapter 8 discusses the compatibility of the model with these various accounts. Furthermore, this chapter makes the case that one class of models seems to have more support than others in light of recent research, and the implications of AIM in the context of this class of models is discussed. The generality of the model is demonstrated via application to spatiotemporal data and validated on a very different eye tracking data set than that considered in chapters 4 and 5.

In addition to the neuroscience of attention, there is also great potential for applications of attention within a machine vision context. Currently, the bulk of work in this area is restricted to the use of attention as a front-end to object recognition systems, or more generally, for the selection of relevant landmarks. Chapter 9 considers the applicability of the model in the domain of machine vision in a few different contexts. It is revealed that the model output is stable across certain transformations, and may be of utility in image enhancement, and the search for certain visual objects.

As a whole, the body of work put forth in this dissertation includes a novel proposal for the role of saliency related computation in the cortex. In contrast to existing models that employ crude approximations, or normalization operations

with weak biological relevance, the dissertation puts forth a very specific proposal for the form of rapid local context related gain control in visual processing. Evidence in favor of this proposal is put forth in the form of a few very different eye tracking experiments, comparison with important results from decades of behavioral data from the psychophysics literature, and in establishing a neural analogue for the implementation of such a mechanism.

2 Computational Modeling of Attention

Despite the perception that we "see everything around us", there is significant disparity between the amount of visual information that is received at the retinae, and the proportion of this visual information that reaches later processing or impacts on conscious awareness. Attention is crucial in determining visual experience. The spirit of attention is perhaps best captured by William James [100]:

> Everyone knows what attention is. It is the taking possession by the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought. Focalization, concentration, of consciousness are of its essence. It implies withdrawal from some things in order to deal effectively with others.

Attention provides a mechanism for selection of particular aspects of a scene for subsequent processing while eliminating interference from competing visual events. A common misperception is that attention and ocular fixation are one and the same phenomenon. Attention focuses processing on a selected region of the visual field that needn't coincide with the centre of fixation. The first observation of this phenomenon is often credited to Helmholtz who had observers fixate a pinhole sized

light target prior to a very brief illumination of a printed sheet of large clear letters [240]. Following the observation of this experiment, Helmholtz noted:

> The electrical discharge illuminated the printed page for a brief moment during which the image of the sheet became visible and persisted as a positive after-image for a short while. Hence, perception of the image was limited to the duration of the after-image. Eye movements of measurable size could not be performed during the duration of the flash and even those performed during the short persistence of the after-image could not shift its location on the retina. Nonetheless, I found myself able to choose in advance which part of the dark field off to the side of the constantly fixated pinhole I wanted to perceive by indirect vision. Consequently, during the electrical illumination, I in fact perceived several groups of letters in that region of the field....The letters in most of the remaining part of the field, however, had not reached perception, not even those that were close to the point of fixation.

In the following sections, these issues are touched on in more detail with the hope that the reader will attain a greater appreciation of what exactly attention entails, and what attention provides with regard to our ability to interpret and act upon visual stimuli. This chapter affords an overview of certain aspects of the current understanding of visual attention with consideration to neurobiology, the role of attention in computer vision, and with an emphasis on computational modeling of attention.

2.1 The Need for Attention

A question that frequently arises with regards to attention, particularly in the context of computer vision, is that of *why* attention is a necessary component of

such a system. Computer vision systems are frequently developed with the intention of operating in real time but often fall short of this mark at the time of their inception. It is not unusual to read statements such as "the algorithm should allow for real time performance given future hardware developments", or "this algorithm may operate in real-time given massively parallel computing hardware". Attention is typically thought of as a means of focusing processing on some subset of the incoming stimuli thus reducing the computational load. On the basis of this view, the importance of attention is sometimes downplayed with regards to its role in computer vision under the assumption that advances in hardware will eventually overcome such limitations. The intention of this section is to motivate *why* this is an invalid assumption, since the issue at hand is greater in scope than simply reducing computational running times. One of the primary goals of attention, unrelated to complexity, concerns interference between signals generated by unrelated image events and selecting between possible outcomes. In a feedforward network, crossover between signals and blurring may result in a response at the output level that is highly confused. This issue is elaborated on in the following discussion. Tsotsos examined the problem of visual search as derived from first principles [226] within a well defined framework including images, a model base of objects and events, and an objective function that affords a metric of closeness between an image subset and an element of the model base. On the basis of this formulation, it may be shown

that visual search in the general case (i.e. when no explicit target is given) is NP-complete. One conclusion that emerges on the basis of this analysis and other complexity arguments [7; 27; 154; 230], is that the computational complexity of vision demands a pyramidal processing architecture. Such an architecture is observed in the primate brain on the basis of increasing receptive field size and the observed connectivity between neurons as one ascends the visual pathway [174]. Pyramidal processing may greatly reduce the computation required to accomplish a particular task by reducing the number of instances to be processed. Tsotsos et al. outline four major issues that arise in a pyramid processing architecture, all of which result in corruption of information as input flows from the earliest to later layers [228]. The four cases are depicted in figure 2.1. The pyramid depicted in the top left (Fig. 2.1a) demonstrates the context effect. The response of any given unit at the top of the pyramid results from input from a very large portion of the input image. As such, the response of a given unit at the top of the pyramid may result from a variety of different objects or events in the image. On the basis of this observation, one may observe that the response at the output layer with regards to a particular event depends significantly on the context of that event. The top right pyramid (Fig. 2.1b) demonstrates the blurring effect. A small localized event in the input layer eventually impacts on the response of a large number of neurons at the output layer. This may result in issues in localizing the source of the response at

the output layer, as a localized event may be represented by a large portion of the highest layer. The pyramid on the bottom left (Fig. 2.1c) displays the cross-talk effect. Cross-talk refers to the overlap of two image events in the pyramid which results in interference between signals in higher layers of the pyramid. Finally, the pyramid on the bottom right (Fig. 2.1d) displays the boundary effect. Units at the outer edges connect to fewer units in higher layers of the pyramid. As a result, a significantly stronger response may result from the same stimulus centered in the visual field relative to near the boundaries. Means of overcoming this difficulty are discussed in detail in [44].

An additional argument concerning the role of attentional processing is motivated by consideration of the action domain. That is, one can only perform actions upon a few or even a single item at a time. This is evidently another motivating factor with respect to the role of attention and is advocated by studies such as that of Rizzolatti et al. [193]

At this point, the rationale of the preceding discussion may not yet be apparent. The motive for addressing such issues is that an appropriate attentional mechanism may overcome the aforementioned issues inherent in pyramid processing. In particular, the Selective Tuning Model [228] was designed with these issues in mind. Deactivation of appropriate connections in the network allows each of the aforesaid issues to be overcome. The exact mechanism by which such issues are handled

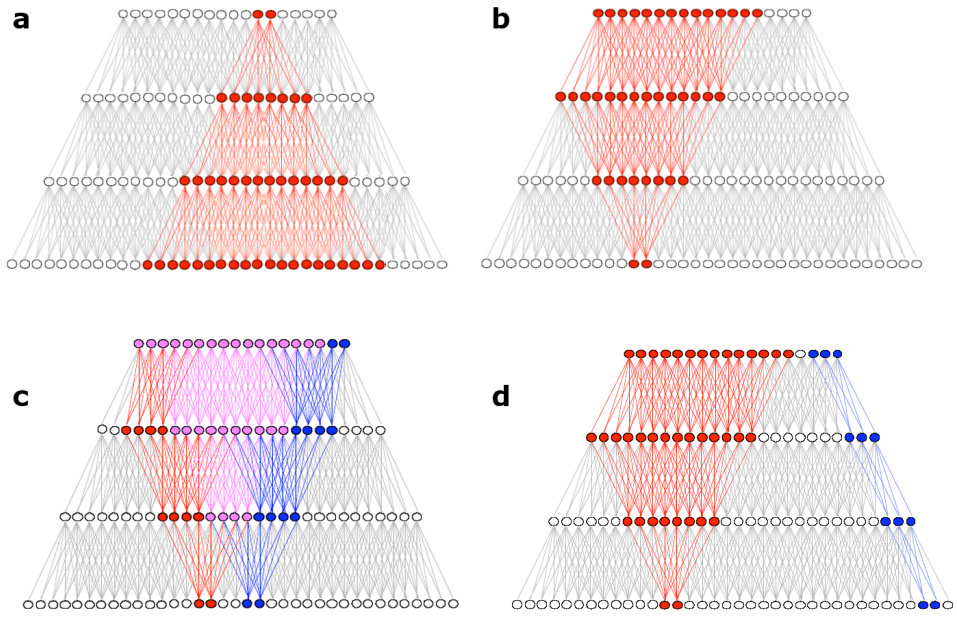

Figure 2.1: Four major issues in pyramid information flow: a. Context: Responses of units at the highest layer of a processing hierarchy are dependent on a large region at the input layer., b. Blurring: A stimulus at the input layer impacts on the response of a large number of units at the output layer, c. Cross-talk: Two unrelated visual stimuli result in overlap in the processing hierarchy resulting in a response in some units that corresponds to mutual interference of the two stimuli., d. Boundary-effect: Units at the boundary on the input layer connect to fewer units at the highest layer of the pyramid than those at a central location. Adapted from [227].

becomes evident in the description of the Selective Tuning Model presented in section 2.4.

2.2 Attention in the Human Brain

The neurobiology of attention has been a controversial subject for decades with much of our understanding coming from behavioral studies often drawing ambiguous or conflicting conclusions. That said, recent advances in imaging technology (fMRI in particular) have given rise to a significant number of imaging studies affording a more definite picture of the neuroanatomy and processes involved in attention. The following discussion offers a picture of the current understanding of the effects of top-down biasing on neural activity and in particular, some thoughts on where signals that initiate top-down bias originate.

2.2.1 Top-down Bias in the Visual Cortex

A key property of attention is the top-down modulation of visually evoked activity. Such modulation is believed to facilitate bias for a particular location of the visual field, or a particular stimulus property [51]. Modulation is thought to come in one of the following three forms: response enhancement, information filtering, and increase of response sensitivity [106].

There is evidence in favor of two distinct types of response enhancement result-

ing from top-down bias. The first of these types comes in the form of a presumed enhancement of cell response brought about by top-down attentional bias. Such enhancement has been detected in virtually every area of the primate visual system including LGN [163], V1 [148], V2 [132; 148], V4 [41; 79; 132; 137; 148; 210], MT [224; 225], and the lateral intraparietal area (LIP) [28; 40]. Top-down attentional bias then seems to impact on neural activity in virtually every area of the visual cortex and even down to the level of a single cell [146]. It has been demonstrated that bias may act in favor of location in the visual field [50], or particular stimulus properties such as luminance and color [149], line orientation [79; 136], and direction of motion [68]. There also exists some evidence that biasing for location occurs at a faster rate than bias for stimulus type [88; 235]. Explanations for this observation range from a hierarchical model of attention wherein selection of stimulus attributes requires prior selection of location [88], to parallel channels mediating each of space and attribute modulation with the spatial channel faster [51].

Secondly, modulation of cell responses also seems to come in the form of an increase of baseline activity in the absence of any visual stimulus. Increase in baseline firing rates of 30-40% have been observed in response to covert deployment of attention in areas V1 [109], V2 and V4 [132], and LIP [40].

In a scene containing multiple competing stimuli, interaction between cells responding to competing stimului may exhibit mutual suppression interaction [107;

108; 191]. There is evidence that top-down attention may result in modulation of suppressive interaction. Spatially directed attention to a stimulus in the receptive field of a particular neuron may eliminate suppressive activity of a non-attended stimulus falling in the same receptive field [191]. Attentional effects tend to be less pronounced when competing stimuli fall outside of the receptive field of the neuron in question. A further discussion of this interaction including a more detailed account of suppressive surround effects appears in chapter 6.

2.2.2 Where are Attentional Signals Generated?

Although the effects of attention may be seen in all areas of the visual cortex, there is evidence that top-down bias signals are generated outside the visual cortex and transmitted via feedback connections to the visual cortex. In particular, current evidence seems to favor selection achieved by way of competitive interaction in the visual cortex with bias signals originating within parietal and frontal cortices [106].

Unilateral lesions in a variety of brain areas give rise to unilateral neglect of the visual field contralateral to the lesion [16; 87; 183]. Cases range from mild, where patients have difficulties directing attention to competing stimuli in the affected visual hemifield, to severe resulting in a total lack of awareness of anything in the affected hemifield. Areas in which lesions result in the described effects include the parietal lobe [232], parts of the frontal lobe [46; 86], the anterior cingulate [101],

the basal ganglia [46], and the pulvinar [242]. It has been hypothesized that such areas form a network for directing attention to visual areas [139; 180]. In the case of patients exhibiting neglect, it has been shown that bottom-up computation in the affected hemifield proceeds as normal and may have an impact on behaviour [55; 56; 57; 76; 134] further reinforcing the importance of the previously mentioned areas in top-down modulation.

Imaging studies have further reinforced some conclusions drawn from lesion studies while casting doubt on others. fMRI studies have suggested that activity in frontal and parietal areas seems to correspond to attentional operations, and not merely the attentional modulation of visual responses [187].

One phenomenon that remains poorly understood is a particular asymmetry that is observed in neglect cases. Right sided parietal lesions result in hemispatial neglect much more frequently and with greater severity than is the case with left side parietal lesions [231]. Based on the observed asymmetry, it has been suggested that the right hemisphere directs attention to stimuli in both hemifields while the left hemisphere directs attention in the right hemifield [139]. A variety of fMRI studies have explored this asymmetry with some reporting a significantly stronger activation of the right parietal lobe [42; 159; 234] and others reporting symmetric activation of parietal lobes [70; 108] with some observing such symmetric interaction regardless of the visual field in which stimulus is presented [108].

2.3 Selected Computational Models of Attention

There exist an abundance of "theories" of attention ranging in specificity from very general conceptual descriptions to highly detailed computational mechanisms. This section is intended to serve as a review of computational models of attention that are much better classified as belonging to the latter of these categories. The discussion that follows is intended to serve as a historical review of biologically plausible computational models of attention, or, neurobiological models of attention that include concepts that might impact on the computational modeling of attention. Any models that are purely computational, purely descriptive, purely speculative, or only propose small variations on existing models, are excluded. Following this review, Tsotsos' Selective Tuning Model [226] is described in some detail as the Selective Tuning Model contains a variety of properties consistent with the views put forth in chapter 8.

Early Ideas

Perhaps the earliest mention of attention that borders on modeling of attention comes in the form of Broadbent's theory of early selection [24]. Early selection posits that rudimentary visual processing occurs preattentively and that focused attention is then required to facilitate higher visual processes such as object recog-

nition. Shortly after this time Deutsch, Deutsch and Norman proposed an opposing view known as late selection [52] that requires preattentive processing of the entire scene to a high-level with attention selecting a subset of this highly processed information. Since that time, a number of more detailed accounts of attention have been suggested mostly in agreement with ideas of early selection.

Adaptive Resonance Theory (Grossberg, 1976)

Adaptive Resonance Theory (ART) [78] falls under the classification of a model of attention based on its original concern for human cognitive information processing and filtering. Grossberg developed a theory of information processing based on a number of principles derived from experiments involving cognitive development, reinforcement learning and attentional blocking. Central ideas of ART with respect to attention are as follows: i. The magnitude of the response of a cell may be modulated by top-down priming. ii. Sufficiently large bottom up activation drives a cell. iii. A cell becomes active if it receives sufficient top-down and bottom-up activation. iv. top-down attentional modulation should exist in all cortical areas that learn. The basis for such rules revolves around the general principle of guarding stored memories against transient changes while maintaining plasticity in learning. ART has since evolved into a series of real-time neural network models for pattern recognition, unsupervised learning and classification([123], [58]). ART provides

an interesting early account of the utility of attentive behavior in learning. Grossberg's description satisfies its ambitions in explaining the aforementioned tradeoff in learning behavior and outlines specific circuitry that accounts for the dynamics of attentive behavior in early visual areas. The chief contribution to the attention literature, is a specific scheme for the modulation of signals in early sensory coding.

Feature Integration Theory (Treisman and Gelade, 1980)

Treisman and Gelade's Feature Integration Theory (FIT) [222], inspired by experimental work pertaining to visual search, proved to be a very influential early view of selective visual attention. FIT proposes that visual information is represented in a number of topographically organized feature maps. For example, a particular feature map might correspond to a topographic representation of local edge strength, or "blueness" over the image in question. Attention is then deployed on the basis of shifting an attentional spotlight over a "master map" constructed on the basis of information transfer from the various feature maps. In their proposed model, only information that falls under the attentional spotlight is said to reach the level of conscious awareness. The mechanism for information transfer between feature maps and the master map in the general case is not described with any specificity. A number of key predictions arise from FIT: First, visual search for a unique element is constant with regard to display size and number of distractors

since activity in an appropriate feature map can guide the attentional spotlight directly to the unique element. Secondly, search for an element that is only unique on the basis of a conjunction of features increases linearly as a function of the number of distracting elements. This second effect is explained in the context of their model by the observation that no single feature map can directly shift the attentional spotlight and as such, a serial visit of each of the distracting elements is required. Ideas inherent in feature integration theory have had a profound influence on many models of attention that have been proposed in the quarter century since its inception. In particular, models that include a saliency map as a means of guiding attention typically share many attributes with FIT. When introduced, FIT provided a convincing computational explanation for trends observed in slopes derived from visual search experiments. That said, more recent psychophysical results suggest that search tasks are perhaps not simply divisible into two distinct search paradigms.

Correlation Theory (von der Malsburg, 1981)

von der Malsburg's Correlation Theory of Brain Function [239] marks a very early account of the binding problem and perhaps the first account of this problem in a computational neuroscience context. The binding problem refers to problem of creating a unified percept from the responses of many separate neurons distributed

throughout the visual system. Correlation Theory is motivated by the necessity of responses of cells corresponding to different parts or properties of a single object to be integrated to arrive at a single unified percept of the object. von der Malsburg proposes that this task is accomplished by way of synaptic modulation in which cells switch between conducting and non-conducting states. Such modulation is governed by correlation in the temporal structure of cellular signals. Correlation in the timing of cell responses, signals that such responses correspond to a single object. This paradigm allows connections corresponding to irrelevant stimulus responses to be deactivated momentarily to reduce interference between different memory traces affording an increase in memory capacity. Correlation theory is important from the perspective of its introduction of the binding problem to the attention literature, and satisfies its aim of highlighting the importance of, and describing a mechanism for, allowing active cells to express relationships amongst themselves.

Koch and Ullman (1985)

In 1985, Koch and Ullman introduced a model of selective attention [110] based on a number of ideas inherent in Treisman's proposal [222]. Koch and Ullman suggested that, as was the case in Treisman's approach, attention is directed on the basis of a master feature map (called a saliency map by Koch and Ullman) derived from a variety of elementary feature maps. Feature maps are assumed to be computed in

parallel over the entire image, and afford a topographical representation of image content with regard to a particular elementary feature (e.g. color, edge orientation, movement direction, etc.). Koch and Ullman's model is characterized by a second step wherein the early topographic representation (saliency map) is projected into a more central non topographic representation which contains properties of only the selected location. Items are chosen for the central representation on the basis of a winner-take-all network. Winning locations are successively inhibited so that attention continually shifts to the next most conspicuous position. It is worth noting that no single cell modulation occurs in the context of this model. Koch and Ullman suggest that the model might be implemented in "neural hardware" with a saliency map located in the striate cortex (V1) or lateral geniculate nucleus (LGN). Although the model of Koch and Ullman provides a more complete description of attention than some of the earlier models discussed, that focus on more specific concepts, it relies on a number of assumptions that one might question. Perhaps the biggest question, is that of how to put into effect modulation at the level of a single cell. This is a consideration that is fundamentally important in any computational model of attention. Also, one might express concern with the claim that their exists a single unique topographical salience map in the brain that guides the focus of attention, since there is no significant evidence in favor of this hypothesis.

The Shifter Circuit Model (Anderson and Van Essen, 1987)

Anderson and Van Essen's Shifter Circuit Model [7] is based on an infrastructure in which a set of control neurons dynamically modify synaptic strength of intracortical neurons. The result is that information from within a windowed region of V1 is selectively routed to higher cortical areas. In neurobiological terms, Anderson and Van Essen suggest that control neurons might reside in the pulvinar, with intracortical connection strength modified by way of multiplicative interactions on the dendrites. As was the case in many of the aforementioned models, Shifter Circuits rely on a master/saliency map which drives the responses of the control neurons. The representation of global saliency is suggested to be represented in the superior colliculus or, the parietal cortex. In addition to the contentious assumption of a single localized representation of salience, the suggestion that the routing of information relies on a simple switching mechanism among early visual areas is highly questionable, and further, fails to explain attentional modulation observed in extrastriate areas.

Sandon (1989)

The model of attention proposed by Sandon [201], marks the first complete implementation of a computational model of attention. The aim of Sandon's proposal

is the selective routing of information to an object recognition processing step. An image processed by their model is first represented at three spatial scales. Scaled images are passed through a number of attention modules, which select features to be attended. The finest scale image is passed through the two attention modules, the intermediate scale image is passed through one attention module, and the coarsest scale image is not processed by any attention module. A scale arbitrator determines which of the three resulting streams is to be processed by the object recognition processor. The attention module consists of a hierarchical multi-scale network, in which features are computed in parallel, feature maps are transformed into feature contrast maps, and finally winner-take-all selection as described in Koch and Ullman's approach determines attended features. Given the criteria for models included in this review, one might question the inclusion of Sandon's model. The inclusion is based solely on the fact that Sandon's effort marks the first complete implementation of an attention model. That said, it offers little from a theoretical perspective in the context of attention, and lacks plausibility from the perspective of neurobiology.

Guided Search (Wolfe, 1989)

Wolfe and colleagues propose a computational model for visual search [251], that emphasizes a distinction between preattentive massively parallel computation of

feature information, and a secondary stage that performs more complex operations over a selected portion of the visual field. An activation map is generated on the basis of a weighted sum of feature maps computed in parallel. Feature maps are also adjusted on the basis of top-down demands. That is, top-down task demands may bias processing of specific categorical attributes (e.g. bias for vertical lines). The activation map is transformed into a saccade map through convolution with an averaging operator. The peak of the saccade map then determines the next saccade position. Guided search is very similar to Koch and Ullmans model with the slight distinction that Wolfe stresses thinking of attention as a limited capacity resource that is distributed in order of strength in the activation map. As such, guided search suffers from the same difficulties attributed to the model of Koch and Ullman.

MORSEL (Mozer, 1991)

Mozer proposes a connectionist model of object perception that includes an attentional mechanism that limits input to a network (BLIRNET) responsible for building location invariant representations of letters and words [150]. Attention is directed on the basis of an attentional map produced by bottom up input from a number of primitive feature maps, and top-down task bias (e.g. a temporal ordering imposed by higher levels of cognition when reading). Primitive feature information

is gated to BLIRNET by associating a probability with transmitted input based on the attentional map. From the perspective of attention theory, the sole distinction between MORSEL and its predecessors, is that a lower probability of transmission takes the place of inhibition. Some of the benefits of MORSEL have been its use in demonstrating the effects of deficits such as virtual lesions on attention. This last consideration is elaborated on in chapter 7.

VISIT (Ahmed, 1991)

Ahmed proposes a model for covert visual attention that predicts specific roles for a number of cortical areas [4]. The design proposed by Ahmed includes the following components: i. A set of basic feature maps assumed to correspond to the representation of the visual world on the retina, and in early areas of the brain including LGN, V1 and V2. ii. A gating system mediated by the pulvinar, which produces gated feature maps within higher visual areas such as MT, IT and V4 based on responses of early visual areas. iii. Bottom-up feature maps derived from early visual areas and represented in the superior colliculus. iv. A control center located in the posterior parietal cortex that controls access to working memory. The complexity of Ahmed's proposal is linear in the number of image pixels, and is successful in modeling aspects of visual search and spatial relationships. The primary contribution of VISIT to the literature might be considered the explicit

inclusion of various areas of the brain, and specific predictions surrounding the role of such areas in attention.

Olshausen et al. (1993)

Olshausen and colleagues suggest an attentional model [166] based on an implementation of Anderson and Van Essen's shifter circuit model [7]. The intention of the model is to form representations of input stimului that are invariant with regard to position, orientation, and scale at the output layer. Inputs from the retinal reference frame are selected based on modulation of synaptic weights in the network to route the desired input coordinates to an object centred reference frame. Such modification of synapic weights is achieved by control neurons originating in the pulvinar. Selection is based on Koch and Ullmans WTA mechanism [110]. Although the model achieves its aim of producing a position and scale invariant representation, the precise relation of aspects of the model to cortical computation remains unclear, including the existence of a neural analogue that achieves scale invariance by way of the proposed mechanism.

Niebur, Koch, et al. (1993)

Niebur et al. [158] and Niebur and Koch [156], propose a model for the neuronal implementation of selective visual attention that is based on correlation in the

temporal structure of a group of neurons. V1 neurons respond with a stimulus dependent mean firing rate according to a Poisson distributed spike train. Spike-trains of neurons within the receptive field of the focus of attention are distributed according to a time-dependent Poisson process while those outside the focus of attention respond with no correlation between action potentials. Temporal correlations in spike trains are detected by V4 inhibitory integrate-and-fire neurons that act as coincidence detectors inhibiting the response of unattended stimuli. Attentional effects on the temporal structure of firing rates affect cells in all visual areas. In contrast, attentional effects on neuron mean firing rates are limited to neurons in V4 and higher visual areas. Selection of attended regions is achieved by way of Koch and Ullman's attentional mechanism [110]. The model of Niebur et al. provides a plausible means of achieving synchronization, but only predicts modulation in V4 and later areas.

The Biased Competition Model (Desimone and Duncan, 1995)

Desimone and Duncan propose a Biased Competition model [51] wherein mutually suppressive interaction between competing stimuli facilitates selection. The model includes top-down bias for spatial location or feature type on the basis of a model of working memory. Strength of interactions is a function of the proximity of competing stimuli. Their initial proposal was largely descriptive but an updated

account establishes a more detailed picture of biased competition in the context of a neural model. Reynolds, Chelazzi and Desimone [191] describe a number of single cell recording experiments focusing on areas V2 and V4 of the visual cortex. They demonstrate that when two stimuli fall in a single receptive field, the neuron may be biased to elicit a response similar to that resulting from each one of the stimuli presented alone, through the influence of attention. The neural circuit describing such interaction consists of an output neuron, relying on two input neurons having both excitatory and inhibitory influences on the output. It is demonstrated that the equilibrium response is proportional to the relative contributions of the excitatory input and inhibitory input to the output neuron. A distinction between the mechanism described by Desimone and Duncan is the presence of both excitatory and inhibitory connections from the input neurons to the output neuron.

VAM (Schneider, 1995)

Following von der Malsberg's correlation theory [239], Schneider proposed a model based on attentional selection for object recognition and space based motor action called VAM [204]. The main distinction between the two proposals is that in VAM, computation is divided into distinct "what" and "where" pathways. Low level features are represented at the level of the primary visual cortex, such as colour and contrast information. The "what" pathway encompasses V4, the inferior tempo-

ral cortex, and the superior temporal sulcus in VAM. Shape primitives and object templates are represented among these regions. The "where" pathway is characterized by parietal areas that represent locations involved in various actions such as grasping or saccade execution. VAM includes an inhibition of return mechanism to avoid revisiting recently visited areas of space. The division of VAM into two distinct "what" and "where" pathways offers little with regard to advancing the understanding of attentive behavior, and this distinction is inherent in Ahmed's VISIT. The division into streams for recognition and motor-action perhaps offers a more useful division from a control-systems perspective.

SCAN (Postma et al., 1997)

The Signal Channeling Attentional Network (SCAN) proposed by Postma et al. [181] consists of a scalable neural network model, intended to simulate attentional scanning. SCAN consists of a hierarchy of gating networks that select an output pattern from the input image (by way of a bottom-up WTA process) that best matches an expectation pattern. The primary contribution of SCAN might be considered the description of a network architecture capable of explicitly routing information based on expectation.

Itti and Koch (1998)

Itti and Koch [98] extend the earlier work of Koch and Ullman [110] with an updated implementation including a number of minor modifications. The steps involved in the newer proposal are as follows: i. 42 feature maps affording topographic representations of orientation, intensity and color information are produced from the input image. This stage involves linear filtering at 6 spatial scales followed by convolution with a centre-surround kernel and a normalization operator. Feature maps are combined across scale to produce a single conspicuity map corresponding to each of intensity, color and orientation information. A linear sum of the resulting conspicuity maps give rise to a saliency map which, combined with a WTA network and inhibition facilitates successive shifts in attention based on the topographical saliency map. The model of Itti and Koch might be viewed as a description of the model of Koch and Ullman expressed at a more specific level of abstraction. The basic properties remain the same, with specific details of the feature maps fleshed out in an implemented model. As such, the concerns expressed in regard to the Koch and Ullman model also apply to the Itti and Koch model.

Cave (1999)

Cave's feature gate model [34] is based on a hierarchical structure whereby elements at each level compete for selection for the next level. Responses at a given level are gated so that as one ascends the hierarchy, the proportion of the visual field represented is reduced. Inhibition takes place at a number of levels to prevent interference with competing stimuli. Selection of open-gate elements is made on the basis of local differences in a winner-take-all network with top-down biases impacting on selection. Bias in Cave's model is achieved within the gating of stimuli, so that no modulation of the signals at the input takes place, only selective routing. Further, the proposed mechanism only allows for gating from discrete non-overlapping regions. The mechanisms included in Cave's model lack neurobiological analogues, and as such the model is not predictive of the behavior of attentional mechanisms in the primate brain.

Deco and Zihl (2001)

Deco and Zihl propose a model that includes a number of modules each consisting of populations of neurons [49]. The model exhibits parallel computation across the entire visual field with a serial focal mode by nature of the dynamics of the system. That is, neither saliency maps nor a focal search mode are explicitly included but

emerge as a result of intrinsic properties of the system. The model consists of a number of feature maps produced on the basis of sensory input and top-down influences. Each feature map is associated with an inhibitory pool that mediates inhibition among competing elements. A high-level map used to guide the focus of attention is derived from the aforementioned feature maps. One element of the Deco and Zihl model that distinguishes it from others is the idea that attention serves to control the spatial resolution at which processing occurs. Deco and Zihl have also considered the effects of lesioning the model with reference to studies of unilateral neglect.

2.4 The Selective Tuning Model (1995)

This section describes key details involved in Tsotsos' Selective Tuning Model [228]. In the most detailed account of Selective Tuning, Tsotsos and colleagues outline a number of problems that arise in the context of pyramid computation (described in section 2.1). Many design choices in the Selective Tuning Model are formed on the basis of overcoming such limitations. The following describes details of the model with commentary on how various components overcome issues of complexity and problems inherent in pyramid processing. Selective Tuning simultaneously handles the issues of spatial selection of relevant stimulus *and* features. Spatial selection is accomplished by way of inhibition of appropriate connections in the

network. Feature selection is accomplished through bias units which allow inhibition of responses to irrelevant features. The Selective Tuning Model is characterized by a multi-scale pyramid architecture with feedforward and feedback connections between units of each layer. A high level schematic of the model is depicted in figure 2.2. Details concerning the connectivity between adjacent layers are displayed in figure 2.3. Variables shown in figure 2.3 are as follows (Also refer to [228] for a more detailed description):

- $\hat{I}_{l,k}$: interpretive unit in layer l and assembly k
- $\hat{G}_l, k, j$: jth gating unit in the WTA network in layer l, assembly k which links $\hat{I}_{l,k}$ to $\hat{I}_{l-1,j}$
- $\hat{g}_{l,k}$: gating control unit for the WTA over inputs to $\hat{I}_{l,k}$
- $\hat{b}_{l,k}$: bias unit for $\hat{I}_{l,k}$
- $q_{l.j.i}$: weight corresponding to $\hat{I}_{l-1,i}$ in computing $\hat{I}_{l,j}$
- $n_{l,x}$: scale normalization factor
- $M_{l,k}$: set of gating units corresponding to $\hat{I}_{l,k}$
- $U_{l+1,k}$: set of gating units in layer l+1 connected to $\hat{g}_{l,k}$
- $B_{l+1,k}$: set of bias units in layer l+1 connected to $\hat{b}_{l,k}$

Selection is accomplished through two traversals of the pyramid. First, the responses of interpretive units are computed from the lowest level to the highest level of the pyramid in a feedforward manner. Next, WTA competition takes place between all units at the highest layer to select a single winning unit. In subsequent

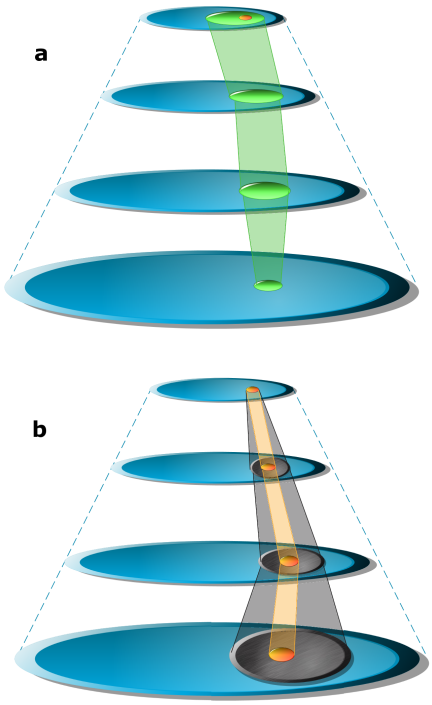

Figure 2.2: A high-level schematic of the selective tuning model. a. Bottom up feedforward computation. Stimulus at the input level (green) causes a spread in activity in successively higher layers. Winner selected at the highest layer is shown by the orange oval. b. Top-down WTA selection. WTA selection happens in a top-down-manner with the winning unit at each level indicated by the orange region. A suppressive annulus around the attended item caused by inhibition of connections is depicted by the greyed region.

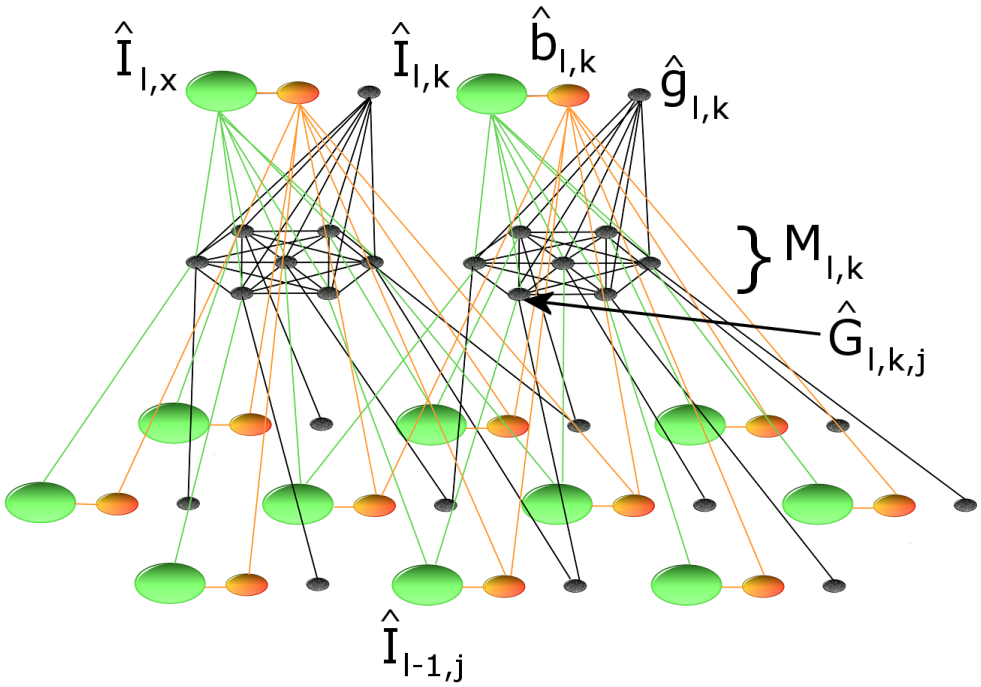

Figure 2.3: A detailed depiction of connectivity between units and layers in the Selective Tuning model. (From [228]).

layers, units in layer l that connect to the winning unit in layer l+1 compete for selection. This ultimately leads to selection of a localized response in the input layer. Figure 2.4 depicts a series of stages in the selection process. Note that interference between competing elements is eliminated by way of selection. Bias is handled through a connected network of bias units that impact on the response of interpretive units they are tied to in a multiplicative manner. Bias units may be used to modify the response of interpretive units to a particular stimulus type, such as blue items. Bias values less than one might be assigned to the response of non-blue units to bias selection in favor of blue pixels. Bias values are entered at the top level of the pyramid and propagate downwards through inter-layer connections between bias units. The value of a bias unit at layer l is given by the minimum of all bias values of units at layer l+1 to which the layer l unit is connected. The WTA process employed in Selective Tuning differs from that of Koch and Ullman [110] in a number of aspects. The effect of unit i in the WTA network on unit j is quantified by the following expression:

$$y = \begin{cases} q_{l,k,i}G^{t-1}_{l,k,i} - q_{l,k,j}G^{t-1}_{l,k,j}, & \text{if } 0 < \theta < q_{l,k,i}G^{t-1}_{l,k,i} - q_{l,k,j}G^{t-1}_{l,k,j} \\ 0, & \text{otherwise.} \end{cases} \tag{2.1}$$

with $\theta = \frac{Z}{2^{\gamma}+1}$, γ a parameter that controls the convergence rate of the WTA network (converges within γ iterations) and $G^{t_0}_{l,k,j} = b_{l-1}n_{l-1}I_{l-1,j}$. A more detailed version of the preceding description concerning the WTA scheme may be found in

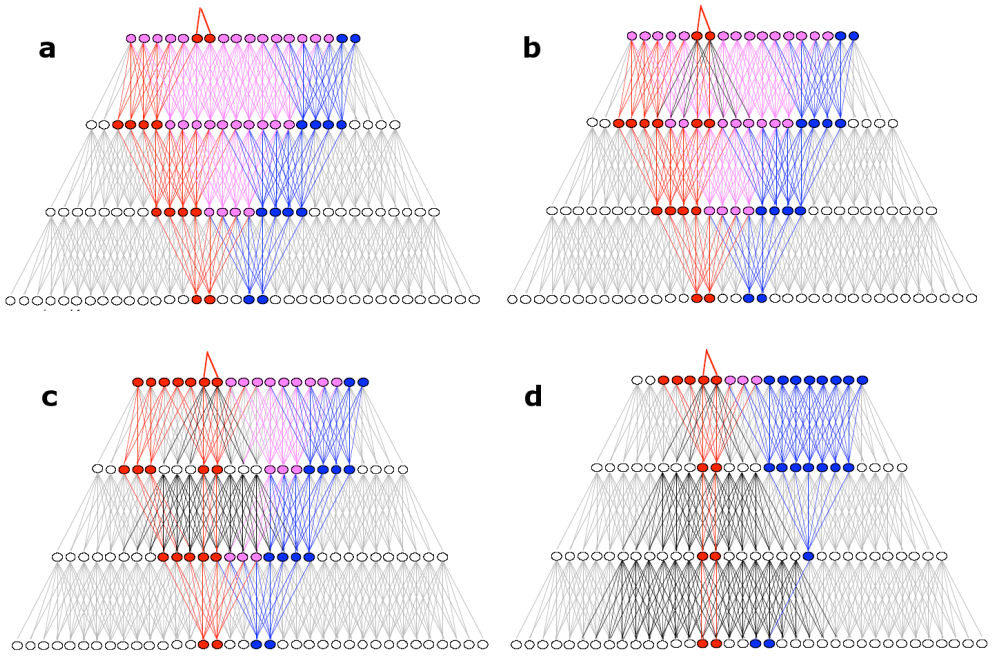

Figure 2.4: A series of stages in top-down winner take all selection depicting 4 hypothetical layers of a visual processing hierarchy. Note that attentional selection eliminates interference between the competing elements. a. Two winning units are selected at the highest level, no attentional effects are yet exhibited. b-d. Connections to winning units at layer (4,3,2 respectively) are inhibited and winners are selected at layer (3,2,1 respectively). (Adapted from [227]).

[228].

2.5 Visual Search

One experimental paradigm that has proved particularly useful for studying visual attention, is the task of visual search. Subjects are typically required to localize a stimulus element with particular characteristics among a number of distracting elements, with a response of some kind to indicate when the element is found, or in some cases, whether no stimulus element with the desired characteristics is present. The timing of such decisions has provided insights concerning the attentional mechanisms underlying visual processing, since performance differences may be telling in terms of the ability to narrow down processing to some subset of the visual stimulus presented.

Visual search tasks are typically conducted under conditions in which the target item is present 50% of the time. The number of distracting elements (called the set size) is varied, and reaction time to indicate that the target has been located, or, whether no such target is present is measured and may be observed as a function of set size. Slopes and intercepts of reaction time versus set size are then used to infer the role of attention in searching among the stimulus elements. There are of course many variations of the setup described including searching for one of several target items, or, setups in which accuracy is measured in lieu of reaction time.

In demonstrating why reaction time is a useful measure in studying attention, it may be instructive to provide a few examples of extreme cases that demonstrate the utility of such results. Consider first the task of searching for a red item among green distractors. The trend observed in this task, is near zero slope in the number distracting elements, which suggests a search achieved by processing the entire visual field in parallel [151]. In contrast, the task of searching for 2 among 5's exhibits a slope that is linear in the number of distractors with a cost of 20-30ms per item [250] with twice as much time required for target absent than target present trials on average. The obvious conclusion that emerges from the trends observed under such conditions is that search requires a serial search in which elements are visited one at a time. For a number of years, following Treisman's Feature Integration Theory [222], search tasks were described in terms of a strict dichotomy of serial versus parallel searches. Since this time, it has emerged that visual search tasks fall in a continuum of slopes ranging from near zero to greater than 30 ms per distractor [250].

Drawing inferences from reaction time slopes is a practice that should be carried out with some caution. This is a view that is advocated particularly strongly by Townsend [219; 220; 221]. Wolfe notes that various limited capacity parallel models give rise to patterns that appear to correspond to serial searches. Further, serial search in which processing time associated with each item is sufficiently low

might be misclassified as parallel searches [250]. The most important fact to put forth is that data drawn from visual search studies do not support a strict dichotomy between parallel and serial searches. Visual search slopes from different tasks range from shallow to steep. Although this does not preclude the possibility of parallel and serial processes acting together, it is important that visual search tasks not be grouped into parallel or serial categories. Feature searches may be made to exhibit linear slopes by narrowing the feature contrast between target and distractors. Wolfe proposes that search tasks may be more practically described using terms such as efficient or inefficient, owing to the obvious differences that do appear between tasks in which targets pop-out, versus those that do clearly require visiting a number of elements in series. Searches that are very efficient tend to be supported by a set of basic elements that may be computed in parallel including color, orientation, spatial frequency, curvature, motion, form, and depth.

3 Saliency: A Brief History and Critique

The central theme of this dissertation concerns the notion of *saliency*. A few passing references to the term were made in the preceding chapters, but the importance of this notion with respect to the subject matter of this dissertation warrants further discussion.

The term saliency describes the state of being *salient* which is defined according to the Merriam-Webster dictionary as

> standing out conspicuously, prominent, of notable significance.

This gives some sense of what a definition of saliency should capture, and also agrees in an intuitive sense with what tends to draw our gaze ignoring any task directives.

The inception of this term in the context of attention modeling derives from models that posit that a *spotlight* of attention is deployed on the basis of a unique topographical representation of saliency: a *saliency map*. Some models fitting this category were mentioned briefly in chapter 3 and are described in more detail

in section 3.2 in particular with respect to the relation of the computation they perform to the incoming retinal input.

In practice the idea of saliency is very often described in the context of saliency maps in the domain of attention. One important ambition of this work is to disentangle the notion of saliency from the idea of a saliency map. As described in the previous chapter there exist a variety of different models, some concerned heavily with the routing of signals through a hierarchy and other architectures quite distinct from those based on a saliency map.

A significant component of this chapter is to place the discussion of saliency outside of the discussion of models based on a unique topographical saliency map since as mentioned, some models include nothing resembling a saliency map but nevertheless discussion of the relation of the behavior of such models to the visual input is useful. For example, if one posits that the network underlying attention consists of a distributed hierarchy of winner-take-all networks, this says much concerning the flow of activity through the network. However, without knowledge of the nature of the underlying interpretive units, nothing can be said with respect to how the selection relates to the input to the network, the photoreceptor array that is the retina.

If one begins to consider the nature and properties of the millions of tiny computational units that comprise the visual cortex, a definition for what is salient

may emerge directly from architectures whose focus is a detailed account of the selection mechanism. In the case of a distributed hierarchical selection mechanism, saliency would correspond to distributed, localized activity in visual neurons in lieu of a single topographical representation of perceptual importance. Moreover, it is possible to consider definitions for what is salient in the absence of a specific selection mechanism. In this manner the definition of saliency is distinct from the selection of visual content and may or may not be compatible with a wide array of attention models, but its specification lies outside of the description of *how* content is selected and routed through the visual cortex, and the focus is on *what* is selected. The later chapters of this dissertation focus on *how* content is selected, but the emphasis for the remainder of this chapter and the next is on the *what*.

The intention of this chapter is to provide a historical overview of saliency. The focus is on efforts that include some description of how visual input translates into a representation of the perceptual importance of the content involved with a focus on the representation of salience rather than the mechanism implicated in attentional selection. As discussed in chapter 2 attention may be discussed in terms of covert and overt attention. Owing to the fact that there is as of yet no simple means of tracking the focus of covert attention, many saliency studies focus their attention on the prediction of eye movement patterns rather than discussing the relation to covert attention. It is worth noting that one possible means of observing the

covert focus of attention appears in [190] and such a paradigm might be adapted to consider tracking the covert focus of attention with the goal of assessing model performance.

The implicit assumption in many existing efforts is that overt attention equates to covert attention. As is discussed in chapter 7, this is not necessarily the case. Furthermore, the relation between overt and covert attention may be more complex than it appears on first inspection, and certainly more involved than is assumed in prior work. An additional contribution of the work in this dissertation is to describe precisely what claims may be made about a model on the basis of various overt and covert test cases.

For the time being, discussion is restricted to overt attention, but in chapter 7 discussion of how such results may generalize to conclusions concerning covert attention is included. The structure of each section in this chapter is a brief historical overview followed by some critical commentary. Passing references to more general modeling considerations are made but this discussion is largely relegated to chapter 8.

3.1 Correlates of Fixation Points

An important consideration with regard to the subject matter at hand, is the relationship between visual stimuli and eye movements. Early influential studies

by Buswell into reading and picture viewing marked the first experiments using non-intrusive eye trackers [29]. Several decades later, Yarbus conducted some eye tracking experiments involving picture viewing with different task directives, demonstrating the importance of task on the resultant eye movements made [254]. In the context of this dissertation, we are interested specifically in the relationship between eye movements made and the properties of the visual stimulus itself. It has recently been demonstrated that this remains an important consideration in the determination of where saccades are directed [61].

There are many basic measures on visual content such as edge strength or curvature that have been shown to correlate to varying degrees with the deployment of fixations. This section outlines a set of such contributions intended to reflect those that have been the most influential, while making the discussion as representative as possible of past efforts considering the relation of image content to saccadic eye movements.

In a somewhat exploratory effort, Privitera and Stark consider the relation of 10 different measures of local image content to the deployment of fixations [182]:

1. A measure based on Canny edge detection [31] which quantifies edges per unit area.

2. Masks selective for high curvature.

3. A 7x7 positive centre - negative surround operator.

4. Gabor masks measuring grey level orientation differences as in [157].

5. A pyramidal discrete wavelet transform based on Daubechies and Symlet bases.

6. A measure of local symmetry.

7. Michaelson contrast defined as $C = \|(L_m - O_m)/(L_m + O_m\|$ where L_m is the mean luminance of a local 7x7 region and O_m is the overall mean luminance of the image.

8. An entropy measure of the type often used to measure texture variance given by: $N = \sum_{i \in G} f_i log f_i$ where f_i is the number of times the i^{th} grey level appears in the neighborhood G.

9. Coefficients of the Discrete Cosine Transform with high frequency components indicating areas of interest.

10. The Laplacian of Gaussian operator.

The set of operators showed varying degrees of correlation with human fixation points with the exception of the DCT based operator which was uniformly poor. The most important result from this work, and the one result that was consistent

across all of the operators, was their inconsistency. That is, each of them performed well for some subset of the images and very poor for others. The conclusion to be drawn from this is that none of the operators could be classed as a universal predictor of fixations and also that this problem seems inherently hard. It also raises the question of whether local image content alone is sufficient to predict fixations to a high degree of accuracy.

Another important study of this type was conducted by Tatler, Baddeley and Gilchrist [213]. Tatler et al. considered the extent to which contrast, edge and chromaticity content at various spatial scales is predictive of fixation points. There are a few important aspects of their work. Perhaps the most important element of this work, is that the various operators tuned to combinations of angular and radial frequency, contrast and chromaticity are based on models of neurons in the visual cortex of primates. In this manner the various feature measures reflect (to some degree) the extent to which various types of neurons, akin to those appearing in the primary visual cortex of primates, are predictive of fixation patterns. Their findings indicate primarily that edge content and contrast are most predictive of fixation points, especially at high spatial frequencies. An additional important aspect of this work concerns the methodology in assessing agreement between features and eye tracking data and is discussed in further detail in chapter 4.

A related piece of work was presented by Pomplun [179]. Pomplun considers

the role of task related information in the context of fixations. Subjects were given a particular pattern to search for in complex natural scenes and fixated regions were compared with properties of the pattern with respect to their intensity, contrast, spatial frequency and orientation. For each of these categories, correlation between these measures on the target item, and fixated regions was determined to be statistically significant. The amount of guidance attributable to the features varied with intensity and contrast having the strongest correlation with fixated regions. This is an important reminder that there is an inherent cap on how well basic image features may predict fixation content since the goals of an observer exert influence on the selection of visual content. That said, fixated regions are not overwhelmingly biased with respect to the relation of their features to the target features suggesting there is still a significant role of bottom-up feature based selection.

Although work on visual correlates of fixations has given some insight on the nature of stimuli that tends to draw an observers gaze, this approach is inherently limited. One serious drawback of this approach, is that the relation of the features considered to the neural representation of such content is unclear. The fact that some choice of image operator produces output that is in some cases predictive of fixation patterns says little of what is happening in the brain. That said, operators based on either cortical processing or a principled decomposition of the visual signal may lead to hypotheses that are testable by psychologists or neuroscientists

based on their apparent role in guiding gaze. The fact that there is as of yet no single operator that is a reliable predictor of gaze patterns makes this line of work especially disappointing since we are left essentially with the knowledge that a variety of different image operators correlate with fixation patterns some of the time. This does not allow even the possibility of considering the relation of a particular feature or image operator to the underlying neural representation since none are universally good classifiers.

There are also some problems with this type of evaluation that lie with difficulties associated with experimental design. In some cases, the set of test images is sufficiently small that it is unclear how well the results presented generalize to visual sampling in general. An even bigger problem lies in disentangling top-down and bottom-up influences on selection. For example, knowledge that the image set contains mostly outdoor natural scenes taken in the summer time provides strong priors on the content of the scene and is liable to play a significant role in the deployment of fixations. In some cases, images consist of such stimuli as paintings, terrain photographs, and landscapes. Ignoring the effect of familiarity in this case, there is also the issue of the extent to which such stimuli are representative of *natural* scenes. For example, a painting will in many cases have a very different spectral profile than a natural scene. It follows that the observed behavior may not generalize to natural viewing conditions owing to an unusual neural representation of the

content. In later chapters, some of these issues pertaining to experimental design are explored further, and the problem of producing *natural* stimuli that minimize prior knowledge is considered.

An additional issue that pertains to all of the results in this chapter is that of assessing algorithm performance. In most cases a mechanism is assumed that selects a number of discrete coordinates for fixation points and such points are compared with human observers' data according to some metric. The selection procedures are wildly variable ranging from hierarchical Winner-Take-All networks to clustering algorithms. A resultant difficulty is gauging the quality of the underlying representation of saliency versus the performance associated with the selection procedure. In general, this sort of analysis will determine the extent to which the largest few peaks in the saliency representation correspond to human fixation points, but says little regarding the assessment of saliency away from these peaks. There is also an issue with determining the number of fixation points to consider, and depending on the cutoff this may produce artificially good or poor results. In general if one wishes to characterize the relative saliency of different targets or locations in space it would make sense to consider a metric that characterizes the quality of the representation of saliency. One might assume that the potential yields from the selection algorithm are proportional to and limited by the quality of the underlying saliency representation.

There is one more recent means of assessing the quality of a saliency measure that is favourable in regards to addressing some of the above issues. Tatler et al. [213] propose a means of assessing algorithm performance in a manner which produces an ROC curve as a performance measure. Their approach involves the following: A very large number of thresholds is chosen and each threshold applied to the representation of saliency. This yields a large number of binary maps, which are each treated as a classifier for fixated and non-fixated locations. Each of these classifiers is then applied to the fixation data set with regards to its hit rate (correctly classified fixated locations) and its false alarm rate (locations labeled fixation which were not). The range of classifiers produces hit rates and false alarm rates ranging from {0,0} ({hit rate, false alarm rate} respectively) for the lowest threshold to {1,1} for the highest threshold. In between the relative scores of hits versus false alarms can take on any range of values in [0,1] with the values increasing monotonically. Higher hit rates for lower false alarm rates means a better classifier typically. Additionally the overall performance may be quantified by considering the area under the ROC curve. This method provides a complete assessment of saliency over all of space, avoids the problems with choosing a single threshold, avoids the problems of assuming a specific selection mechanism, and offers a good sense of how well a measure of saliency might correlate with human data given the selection of a discrete number of points. An additional advantage of this means

of assessing saliency measures that will become apparent in chapters that follow is that it does not assume that attention must be directed to a single discrete point. In chaper 8, we argue that attention is more naturally modeled as deployed over a region of space with variable size and shape as opposed to a discrete point. Owing to the intimate relationship between saccadic eye movements and attention, this distinction is important.

3.2 Saliency in Saliency Models

As described in chapter 3, the introduction of saliency maps came with Treisman and Gelade's Feature Integration Theory [222]. To briefly reiterate, the basic structure of the model is that various basic features are extracted from the scene. Subsequently the distinct feature representations are merged into a single topographical representation of saliency. (In later work this single topographical representation has been deemed a saliency map.) A selection process then takes place which in vague terms selects the largest peak in this representation, and the *spotlight* of attention moves to the location of this peak. In this context, the combined pooling of the basic feature maps is referred to as the saliency map. Saliency in this context then refers to the output of an operation that combines some basic set of features.

Since there is some contention regarding the nature of the mechanism responsible for attentional selection in primates, we will focus for the time being on the

definition of saliency in this chapter. Further discussion of issues pertaining to the selection mechanism appear in chapter 8. In Treisman and Gelade's original description, there is little discussion with respect to the nature of features, or how they might combine.

Greater specificity in this regard was introduced by Koch and Ullman [110], who described selection on the saliency map in the form of a winner-take-all network. Such WTA behavior was thereafter applied to the problem of modal control of a robot head by Clark and Ferrier [38].

An implementation with greater specificity was produced by Itti and Koch [98], including specific feature definitions and claims concerning neurobiology and visual search performance. Since its introduction, this particular approach has seen considerable use and is generally the metric against which many models are compared.

Many additional descriptions of saliency in the domain of saliency based models constitute the addition of features to the basic structure described by Itti and Koch. The saliency map then becomes the amalgamation of some subset (not necessarily proper) of the original features along with an additional feature or features such as depth [172], motion [20; 186], symmetry [85], faces [20], or skin [185].

Some of the drawbacks already described in the previous section also apply to maps that pool features to create a single master saliency map. The biggest gain in systems that pool a number of features to derive a representation of saliency, is

that they have greater consistency in predicting fixation patterns across a larger set of images. This is perhaps unsurprising since we already detailed the fact that a variety of features that may have some intuitive relation to salient image content performed reasonably well on some images, and poorly on others. It follows then that a system that combines many of these types of features may produce more consistent results.

One element worth mentioning at this point pertaining to overt versus covert attention, is that the history of these models describes a story that is highly confused. The description of Treisman and Gelade, as well as Koch and Ullman is a description of how covert attention is achieved and as such the motivating experimental evidence derived from foveal displays with controlled fixation. However, in the subsequent work of Itti and Koch and the corresponding implementation, tests are carried out involving eye movements from which claims concerning covert (as well as overt) attention are drawn. As mentioned, it would be desirable to consider how these two issues interrelate as this is an issue neglected in previous contributions to the saliency literature.

One difficulty with a system that pools the responses of a variety of features into a single topographical representation of saliency, is that knowledge of the extent to which the underlying features are predictive of fixation patterns is lost. That is, it is no longer possible to measure the individual contribution of orientation content

versus spatial frequencies versus color. Nevertheless, it may be said that a variety of features combined appropriately may perform reasonably well in predicting fixation patterns, and certainly better than any single local operator. It is also worth noting that it may be possible to consider *disection* of the model, or anatomically in animal experiments to assess the relative contribution of the various features.

There is a significant criticism that may be levied against the definition of saliency captured by these saliency map models. The interpretive units in these models are hand-crafted on the basis of observations on primate neurophysiology. Observation in V1 of cells coding for different orientations, color opponency and intensity characterized by centre-surround receptive fields at various spatial scales have prompted their inclusion as basic building blocks of the saliency map paradigm. It is then perhaps not overly surprising that a selection mechanism acting on units whose response resembles those of cells appearing in the visual cortex should produce behavior that correlates with that observed in primates. The bigger issue however, is that although the models do give a definition of computational units that act on the incoming retinal image, they offer little in explaining *why* the neurons involved have the structure they do, and what this model translates into with respect to its relationship to the incoming stimulus.

One possible connection of the types of features typically found in early cortical processing to a quantitative principle is expressed by Koenderink and van Doorn

[111]. In this account, one may view the function of early visual circuitry as capturing the geometry of visual patterns according to a principled definition. There is therefore an explicit relationship between the V1 style features often employed by these models, and those that derive from a principled geometric framework. That being said, the specific relationship between the features employed in some studies of saliency [97; 98] and the geometric motivation put forth by Koenderink et al. exists only at an abstract level (i.e. there are some units with similar receptive field profiles across the two sets of features) with the specific relationship of features chosen, to those that are emergent from a principled geometric decomposition not a principle consideration in the work.

An additional account of the impetus behind the properties of features in early visual processing is found in the work of Linsker [128; 129; 130; 131]. Linsker's infomax principle [131] prescribes that a function that maps a set of input units to a set of outputs should be learned in a manner that maximizes the Shannon mutual information between the inputs and outputs. In performing experiments of this type based on Hebbian style learning on a neural network, one arrives at units that have properties reminiscent of those observed in early cortical processing including centre-surround configurations [128], responses to specific combinations of angular and radial frequency [129], and organization into cortical columns [130]. In this case, the learned receptive field profiles once again bear a resemblance to

those employed as basic building blocks of some saliency based models, but the connection is not made explicit.

The model of Itti and Koch offers at least a detailed quantitative description of the units, and hypothesized circuitry involved in attentional selection. The contributions of extensions to this model consist almost exclusively of the addition of various features, some lacking any neurobiological motivation. For this reason, these additional contributions offer little in contributing to what constitutes salient visual content in that their applicability is often restricted to specific types of stimuli. An additional consideration is that the saliency based models fail to address the issue of modulation throughout the cortex. It was mentioned that the pooling of features into a featureless representation precludes determination of what content gave rise to the resulting activation. This consideration in fact poses a deeper problem when attentional selection is considered. This issue is explored in greater detail in chapter 8.

3.3 Neuroanatomy

If one assumes that there is a unique topographical representation of saliency in the brain, the question follows of where this representation might reside.

Several areas of the brain have been proposed as possible loci for the saliency map. Proposals for the site of the saliency map have included V1 [124], areas

along the ventral pathway [138], one area on the dorsal pathway [39], and within the oculomotor network, specifically in the lateral intraparietal area [116]. Fecteau describes some important criteria that should be considered constraints on the possible candidates for the neural site of the saliency map [66]. It is worth stating in advance that the LIP area is the only region that satisfies all of these conditions and references follow where appropriate.

1. The neurons within the area should be featureless and retinotopic [17]. It is worth stating that although the local grouping of receptive field organization within LIP is retinotopic, the overall organization of LIP is highly complex [6].

2. Lesions within the area should produce deficits in orienting attention [140]. It is worth noting that LIP tends to be associated with overt rather than covert eye movements. Regions of 7a are monosynaptically connected with LIP and and yield far fewer presaccadic responses. Covert shifts of attention appear to be signalled by both LIP and 7a activation; however, activation in 7a only appears when a target excites neurons at a location different than the cue. Mesulam suggests that this may indicate that such neurons are more involved in shifting the covert or overt focus of attention rather than an explicit representation of saliency.

3. Facilitation should be possible through electrical stimulation [33; 45; 144; 145].

4. There should be input from the ventral pathway that codes for content necessary to compute relative saliency [67; 72; 202].

Additional elements that make the LIP area a suitable candidate for the locus of the saliency map include its connectivity to regions involved in generating saccadic eye movements including the frontal eye fields and the superior colliculus. LIP neurons have been demonstrated to respond to the abrupt onset of motion in the absence of selectivity for direction, and also to briefly flashed stimuli. Given all of the aforementioned properties, it is likely safe to say that the LIP plays some role in the neural representation of saliency, but the notion that it contains the saliency map remains a moot point. It is worth mentioning that LIP has considerable interconnection with various areas of the brain which poses questions for models that presume that some basic set of features, for example solely those of the type V1 computes, inform this representation of saliency.

The discussion of the location of the saliency map reveals an important consideration. Despite the variety of brain imaging technologies we now have at our disposal, the neural locus of the saliency map, or even the claim that such a representation exists remains open to debate. This consideration is explored in more

detail in chapter 8 within discussion focusing on more general modeling considerations.

3.4 Saliency in Computer Vision

An additional domain of interest relating to saliency comes from the computer vision literature. Although formal arguments have been made [226] on the necessity of attention for the problem of visual search, the computer vision community at large has yet to universally acknowledge the necessity of attention in computer vision systems. That said, there do appear to be more mentions of attention in the literature with attention being described as a front end to object recognition in particular for appearance based systems [54]. As chapter 9 explores this area in detail, a detailed review of these approaches is relegated to chapter 9.

One point perhaps worth noting pertaining to saliency in a computer vision context is that the current trend in the design of algorithms that select local features might be regarded as the complement of saliency algorithms. While operators that extract local features are often designed on the basis of invariance properties, the distinctiveness or saliency of such points is often an afterthought. This raises the question of whether success may be had in designing an operator to select salient or distinctive features, and adapt this paradigm to include desirable invariance properties. It is this consideration that forms the subject matter of chapter 9.

In the most ideal case, one might simultaneously consider constraints pertaining to invariance and distinctiveness and produce an interest operator that forms an optimal trade-off between these elements.

3.5 Conclusions

The preceding discussion reviews current thinking regarding saliency. The major issue with the field is that models either focus exclusively on image properties while saying nothing about the brain, or focus exclusively on implementing anecdotal observations concerning the brain, while yielding little in the way of insight concerning why such content is salient, or why the brain is organized as it is. Further, the current trend in this area is either simply consideration of *new* features, or the addition of features to the basic saliency map model, offering little in the way of progress or additional insight. This provides the motivation for the subject matter of the chapter that follows. The ultimate aim is to start with nothing except an appropriate *principle* regarding what is salient in terms of stimulus content in light of the role of attention and then to explore this definition in terms of stimulus properties and neural hardware. Ultimately it would be desirable to have a model with considerable generality, with significant success in predicting eye movement patterns, that extends to explain covert attention, with some insight on visual neurophysiology, and consistent with basic psychophysics derived from visual search

literature or other attention related areas. These issues are explored in some detail in the chapters that follow.

4 Towards a Principled Definition of Saliency

Chapter 3 reviewed the current state of the notion of *saliency* in the attention literature, while highlighting the shortcomings. In this chapter, the aim shifts to the solution, that is, an attempt to form a definition of saliency based on some criterion for optimality with a theoretical basis describing both what the definition means in terms of visual input, and how it might manifest itself neurally. It is demonstrated through the discussion in this chapter that ideas borrowed from information theory provide a natural interpretation of the problem.

4.1 Revisiting Attneave, 1954

The central core of the proposal is built on computational constraints derived from efficient coding and information theory. The intuition behind the role of these elements in saliency computation is perhaps best introduced by considering an example from an early influential paper by Attneave that considers aspects of information theory as they pertain to visual processing [8]. Within this work, Attneave provides

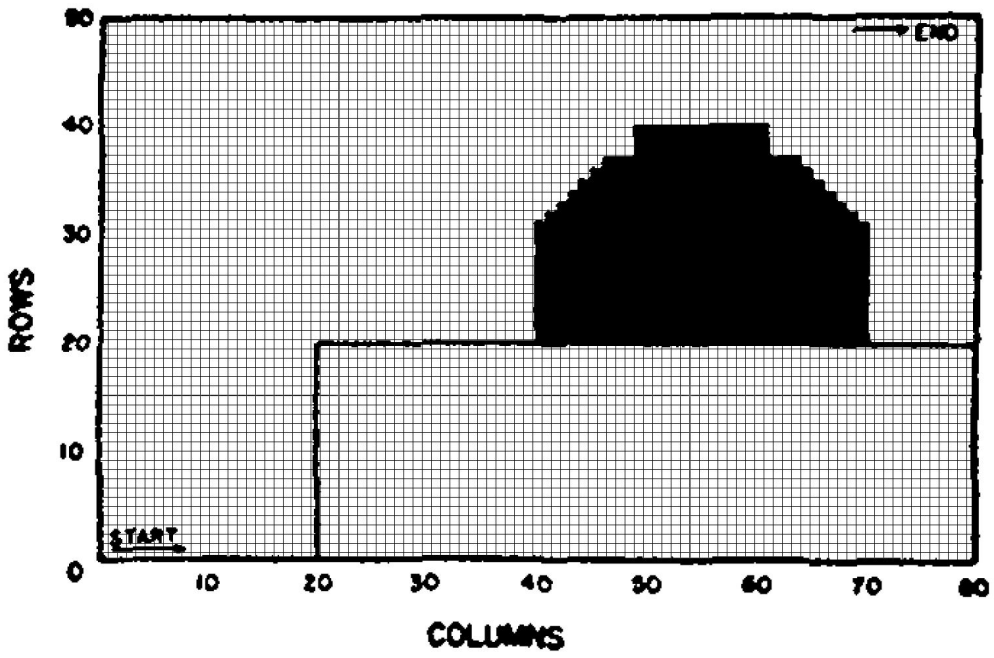

Figure 4.1: A crude depiction of an ink bottle on a desk corner from [8].

the following description and associated figure (labeled figure 4.1):

Consider the very simple situation presented in Fig. 1. With a modicum of effort, the reader may be able to see this as an ink bottle on the corner of a desk. Let us suppose that the background is a uniformly white wall, that the desk is a uniform brown, and that the bottle is completely black. The visual stimulation from these objects is highly redundant in the sense that portions of the field are highly predictable from other portions. In order to demonstrate this fact and its perceptual significance, we may employ a variant of the "guessing game" technique with which

Shannon has studied the redundancy of printed English. We may divide the picture into arbitrarily small elements which we "transmit" to a subject (S) in a cumulative sequence, having him guess at the color of each successive element until he is correct. ... If the picture is divided into 5O rows and 80 columns, as indicated, our S will guess at each of 4,000 cells as many times as necessary to determine which of the three colors it has. If his error score is significantly less than chance [2/3 X 4,000 + 1/2(2/3 X 4,000) = 4,000], it is evident that the picture is to some degree redundant. Actually, he may be expected to guess his way through Fig. 1 with only 15 or 20 errors.

The intent of Attneave's example is to demonstrate that there exists significant redundancy in natural visual stimuli, and that human subjects appear to have some degree of an internal model of this redundancy. A second observation that is not made in the original description, but that is fundamental to the subject matter of this dissertation, is that one might also suggest that the areas of the scene where subjects make the greatest number of errors on average in guessing, are those that contain content of interest. A comparison to some of Shannon's work on modeling redundancy in the English language may also be made in observing that words that occur less frequently within a particular context will tend to provide more information. In the sentence "The man went to the store to buy some bread", most would agree that selecting the subset of words "man went store buy bread" provides more

information than "The to the to some", the former being the 5 words that occur least frequently in English writing of the 10, and the latter being those words that occur most frequently. In the same manner, those visual patterns that are less likely or predictable are more informative in a Shannon sense [205]. This is the intuition that underlies the proposed saliency computation, that the saliency of visual signals may be equated to a measure of information or the degree of surprisal that they carry. The possibility of considering information in a principled sense associated with ensembles of neurons is expressed in the work of Palm [173] who outlines the possibility of characterizing the surprise associated with the temporal structure of spike trains generated by a neuron. The generality of Shannon's proposal makes it amenable to considering the surprise or information associated with various aspects of neural information processing, of the sort that forms the motivation in this chapter.

Imagine a hypothetical generalization of the game described by Attneave in which a human participant is required to describe the contents of a region of a scene containing arbitrary structure, lightness, contrast and colors. Although it is not practical to carry out an experiment of this type, most would agree that there exists some general intuition concerning what a certain portion of a scene is expected to contain on the basis of its context. Consider for example figure 4.1: Under the blacked out regions (left) labeled A,B and C, one would likely claim to

have some general intuition concerning the contents of each hidden region on the basis of the surround and the contents of the scene. It is also evident from the frame on the right that the image content hidden within regions A and B come very close to this intuition whereas the content that lies beneath region C is very far from our expectation and would almost certainly require the greatest number of guesses within the hypothetical guessing game. This region is also the most informative in a Shannon sense on this basis. The intuition this example provides is that salient content corresponds to content that is surprising and fundamentally *informative* in a Shannon sense. In the latter part of this chapter, this computation is described formally and examples of its output are given.

4.2 A Mathematical Theory of Communication

It has been discussed formally why the complexity of certain visual tasks demands attention [226]. The solution to this problem, is that some subset of visual content is selected at the expense of other content for high-level processing such as object recognition. In a dynamic environment, it is inevitable that the time available to evaluate one's current surroundings and act upon the resultant representation restricts the content that receives this deeper analysis to only a subset of all visual input. This raises the question of what content is selected, and why.

Certainly an important element is the current goals of the observer. If a monkey

Figure 4.2: An example of how context shapes our expectation of scene content. The content hidden behind regions labeled A and B come close to one's expectation while that hidden by C is arguably quite different from what one would guess. C carries the most surprisal or carries the greatest self-information in a Shannon sense.

is hungry and is traveling in search of fruit, the tendency to attend to and fixate more vivid colors is inevitable. That said, there is also an important influence of the nature of the visual content itself. That is, in the search for fruit a bright light flashing, an unusual object in the forest, or sudden movement suggestive of a predator is almost certain (hopefully for the monkeys sake!) to warrant the deployment of attention and/or fixation. For now the discussion is restricted to what properties of the stimulus itself results in selective attention either overt or covert independent of task related information.

As discussed, in order to consider this problem, it would be useful to establish a formal context in which to consider the elements involved. Some previous efforts have suggested the possibility of information theoretic measures as a means of guiding selection. Previous efforts fall short of establishing why this may be an appropriate domain in which to consider the problem of visual saliency. Furthermore, we will establish in the following discussion that the formulation that previous approaches consider, is perhaps not the most intuitive interpretation of the problem of attentional selection from an information theoretic perspective. Attneave's example suggests that one line of reasoning derived from information theoretic ideals agrees anecdotally with our intuition of what constitutes salient content. Specifically, the saliency of a region of a scene corresponds to the likelihood of error, or expected number of guesses to describe the region on the basis of its context or more formally

it's Shannon surprisal or self-information.

4.2.1 Some Previous Efforts

Previous work pertaining to attention that appeals to information theoretic considerations focuses almost exclusively on the notion of entropy. Shannon entropy is defined as follows: Given a discrete random variable X with possible outcomes $x_1, ..., x_n$, the Shannon entropy $H(x)$ is given by

$$H(X) = -\sum_{i=1}^{n} p(x_i) log(p(x_i))$$

The continuous case is analogous with

$$H(f) = -\int_{-\infty}^{\infty} f(x) log(f(x)) dx$$

with f a probability density function on $\mathbb{R}$.

There are numerous interpretations of this definition all of which are equivalent. Entropy might be described as a measure of uncertainty. Intuitively the definition of entropy should require that the measure of entropy is at a maximum for a uniform distribution since this implies total uncertainty. Consider the example of a coin toss. For a standard coin the value of $H(x)$ is given by $-0.5 * log(0.5) - 0.5 * log(0.5) = 1$. Now imagine we know ahead of time that the coin is biased and comes up heads 90 percent of the time. In that case, $H(x) = -0.9 * log(0.9) - 0.1 * log(0.1) = 0.469$. In the latter situation since we can predict to a greater extent the outcome of the

coin-toss the associated uncertainty or entropy is reduced. The specific nature of Shannon's definition is uniquely defined in satisfying 3 criteria. These include that H is at a maximum for a uniform distribution, that H is a continuous function of the probabilities in question, and that regrouping the various outcomes should not change the measure of entropy.

A slightly different interpretation of this quantity is related to the idea of surprise. The value of $-log(p(x_i))$ is often referred to as self-information and also sometimes described as the surprise associated with outcome i. Defined in this manner, entropy measures the average surprise over all possible outcomes. For example, in the case of the coins, it may be unsurprising to learn that the biased coin came up heads. That said, this outcome is far more frequent than the coin coming up tails and thus the average surprise is consequently lower than the unbiased coin case.

Another useful interpretation, particularly with respect to neural representation is that of coding. In short, coding in this context refers to the representation of some concept or event by another. Consider a hypothetical language that consists of characters from a 4 letter alphabet. Let us assume that the likelihood of each of these letters occurring in a message written in this language are $p(x_1) = 0.5, p(x_2) = 0.25, p(x_3) = 0.125, p(x_4) = 0.125$. Consider first the possibility of encoding this language based on a simple binary representation. That is, let $x_1 = 00, x_2 =$

$01, x_3 = 10, x_4 = 11$. Since the alphabet consists of 2 symbols which are always required to transmit a letter, the average number of bits required for a letter is 2 based on this encoding. Now consider a slightly different encoding given by $x_1 = 0, x_2 = 10, x_3 = 110, x_4 = 111$. In this case the number of bits is 1 to convey x_1 and 3 to convey x_3 and x_4. However, one must consider how frequently such letters occur. The average number of bits in this case is given by $0.5 * 1 + 0.25 * 2 + 0.125 * 3 + 0.125 * 3 = 1.75$. So we have a more efficient code for the language in this case and an example which serves to show the relationship between coding and entropy. This becomes an important point in considering neural representation. $H(x)$ captures the average number of bits required to store the random variable X. The precise mathematical details of this discussion are outside of the scope of the necessary background, but additional details may be found in [205].

In the context of discussing visual stimuli, the relation of entropy to the visual content varies. In the context of image content, often entropy is regarded in relation to the grey values within a local patch of the image. Given some local neighborhood, one can consider the distribution of grey values and derive a measure of entropy on the local neighborhood. For example, let us assume that one is dealing with a binary image, with pixels either black or white. In some local region, one might have a mix of black and white pixels, or a patch that is predominantly white or predominantly black. The case of an equal number of black and white pixels corresponds to a

uniform distribution and hence maximum entropy. The entropy is much lower in instances where the patch is predominantly one color. One can do the same sort of analysis on a greyscale patch, the only difference being that the number of bins is greater. Note that the spatial arrangement of pixels has no effect on the resulting entropy in this case.

Recently, a variety of proposals based on information theory concerning attention and fixation behavior have emerged. Najemnik and Geisler consider fixation behavior predicted by a Bayesian ideal observer with the focus on predicting sequences of fixations [153]. They demonstrate that human observers appear to compute an accurate posterior probability map in the search for a target within 1/f noise, and that inhibition of return proceeds according to a very coarse representation of past fixations. An important element of this work lies in showing that target search in primates appears to operate according to maximizing the information about the location of the target in the selection of fixations.

Lee and Yu propose a model of information maximization as the basis for saccadic eye movements [119]. As with others, the approach of Lee and Yu is based on local entropy. A distinguishing factor of the approach, is that it is based on a local hypercolumn representation of Gabor filters rather than simpler features. A normalization procedure takes place on the basis of the response values of local hypercolumns and those in the surround. The choice of this operation comes

from prior observations that a transformation based on local contrast tends to yield greater correlation to fixation patterns than the raw features themselves. A step is also included which maintains a prior memory in the form of a mental mosaic which may factor into the choice of saccades by discounting the role of mutual information between any given target location and the existing mental mosaic. This model has a variety of nice properties, but fundamentally maintains a measure of local entropy as the basis for selection. The model does not consider performance based on eye tracking data.

The model of Lee and Yu was imported to an experimental setting by Renninger et al. to establish its plausibility as a description of primate saliency computation. [189]. The task involved determining whether the silhouette of a particular shape matched with a subsequently presented silhouette. Eye movements were tracked during the presentation to observe the strategy underlying the selection of fixations. Renninger et al. demonstrate that the selection of fixation points proceeds according to a strategy of minimizing local uncertainty. This is the same as a strategy of maximizing information assuming information equates to local entropy. This will typically correspond to regions of the shape silhouette which contain several edgelets of various orientations. In agreement with the work of Najemnik and Geisler, it was found that there is little benefit to the optimal integration of information across successive fixations. One critique that might be levied against an

entropy based definition is that a definition based on minimizing local uncertainty, or entropy is inherently local. A further commentary on this issue appears later in this chapter. Mechanisms for gain control at the level of a single neuron have been observed which have been shown to correspond to a strategy based on information maximization [23]. Although the proposal put forth in this paper is distinct from a description that involves sequences of fixations, the search for a specific target, or specific task conditions, it is nevertheless encouraging that there do appear to be mechanisms at play in visual search that serve to maximize some measure of information in sampling, and it is also the case that the findings of these studies may be viewed as complementary to our proposal rather than conflicting. It is also interesting to note that recent studies have observed quantitatively an increase in the information transmission of a cell associated with stimulation outside of the classical receptive field [238].

An effort found in the computer vision literature is that of Kadir and Brady [104]. The approach of Kadir and Brady examines the local Shannon entropy based on intensity or color values in a local neighborhood. This operation is done across several spatial scales. Various scales are chosen based on peaks in the entropy versus scale curve. Subsequently the magnitude change of the probability density function as a function of scale is measured for each peak. The final saliency map is then given by the product of the local entropy and the magnitude change in

the local probability density function. This typically results in the selection of a variety of circular regions at several spatial scales. This approach is of interest in how it handles the scale variance of features of interest. In the class of entropy based approaches, this model is unique in this regard.

As alluded to in the previous chapter, there is currently interest in the selection of features as a precursor to object recognition. One entropy based approach in this regard is presented by Fritz et al. [73]. In their model, local features are selected on the basis of entropy with PCA employed as a precursor to reduce the dimensionality locally. The focus of this work is on recognition performance, but the work nevertheless employs an entropy based approach to select a subset of available visual content.

All of the aforementioned approaches that equate visual saliency to a measure of information are largely advocating the same theory, that being that a measure of how much local entropy is present is the basis for selecting gaze points or features. A slightly different way of saying this would be that fixation or feature selection chooses the region/features that requires the greatest number of bits to encode the content of the region and thus it is a region that contains the most information. It is worth pointing out that in the latter way of stating this idea there is one element that one might take issue with. The local content will give rise to a complex neural representation. That is, a stimulus with a relatively uniform distribution

with respect to a particular feature measure (e.g. luminance) may actually give rise to a highly peaked distribution of firing rates in terms of a neural representation. This implies the feature under consideration (e.g. grey values or orientation) in measuring entropy influences heavily the choice of gaze points and there is no particular formula for what the feature or set of features should look like except that this behavior should be governed by the specific nature of the neural representation involved. Additionally, a particular distribution within a chosen feature space says little about the resulting neural representation. That being said, the features employed in some studies do bear a resemblance to those that are emergent from principled approaches to deriving features such as those selective for specific combinations of angular and radial frequency [175]. That said, even under the assumption of such a representation, there are instances under which an entropy based definition seems to fail to predict observed behavior. In addition to these issues, there are also some more general philosophical considerations that render this interpretation questionable. These issues are discussed in the latter sections of this chapter.

4.2.2 A More Natural Interpretation

The entropy based interpretation of the problem lies essentially in the claim that the local region that contains the greatest number of bits of information, or un-

certainty is worth attending to. At first glance this appears to be a nice intuitive interpretation of the problem. The aim of this section is to introduce a slightly different interpretation of the problem of attention in an information theoretic context and contrast this against the entropy definition. The core of the problem with the entropy based definition is that the notion of uncertainty is misleading and does not always correspond with the content that informs most upon the content of the scene. The problem of selection of visual content can be stated as follows: At any given time, an observer is faced with an array of input covering the entire visual field and is faced with the decision of what to attend to or fixate in terms of some subset of the overall visual field. When phrased in this manner, the problem is very reminiscent of a basic problem addressed by information theory, that is, *given a limited capacity channel, which messages should be sent to convey the maximum amount of information?*. In this regard, the incoming visual input may be regarded as the *message space* and the local regions that are attended to or fixated as the *messages* being transmitted, in this case to other areas of the brain for more detailed analysis. As in the classic case, it is the nature of the message space that determines what messages are appropriate to send. The implication of this is that observation of a single message, or local region is insufficient in itself to determine what should be selected. This gives some early insight into what is lacking in the definition of some of the local entropy based approaches and any

local feature measure for that matter. Further discussion in this chapter elaborates on this consideration providing details on cases where the traditional measures fail spectacularly.

For example, consider a scenario wherein the visual field consists of a portrait hanging on an otherwise blank white wall. Most would agree that having a variety of local snapshots of the portrait will provide a greater sense of the content of the scene than a handful of local regions of the empty wall. Of course in this case the local activity and richness of content (and entropy on this basis) in the portrait is sufficient to predict this result, however, consider the case of a portrait hanging on a wall with highly textured and colorful wallpaper. The local activity may no longer be sufficient to define the portrait as a region of interest in the absence of some measure of context. It is this thinking that forms the basis for the proposed approach.

An element of this intuition is captured in the work of Topper [217]. Topper considered a variety of basic features such as intensity, contrast, edges etc. and their relation to eye movements. He concluded that a variety of basic features (many of those examined by Privitera et al. [182].) produce greater correlation to eye movements when remapped to a space representative of their information content based on Shannon's self-information measure. The definition and description of this information operation was mentioned briefly earlier in this chapter. A more

detailed discussion of this measure follows in the remainder of this section since it is an important element of the proposal put forth in this dissertation. Topper's work relies on a simple transformation based on local features and thus suffers from some of the methodological problems discussed in chapter 3 such as having a poor connection to human performance, and inconsistent performance for any of the features proposed. Nevertheless, as we will demonstrate there is something in this basic idea that makes it amenable to addressing the general problem of visual saliency. When viewed in a specific light the problem of visual saliency appears in a manner that suggests self-information as a transformation from some internal neural representation to one of visual saliency. Although this is far removed from the operation on basic features considered by Topper, as we are not interested in arbitrary feature sets, or any *ad hoc* choice of operators, it is interesting to note the performance gain resulting from an operation with some relationship to the eventual formulation derived in this chapter.

An additional consideration that has not yet been mentioned, is that of anisotropy in sampling. That is, content centered in the fovea has a more complex representation than that appearing in the periphery in regard to photoreceptor density. This is an issue that has been avoided by many previous efforts at understanding visual saliency. For the time being, the discussion assumes a representation in which all content exists in a relatively low resolution uniformly sampled representation.

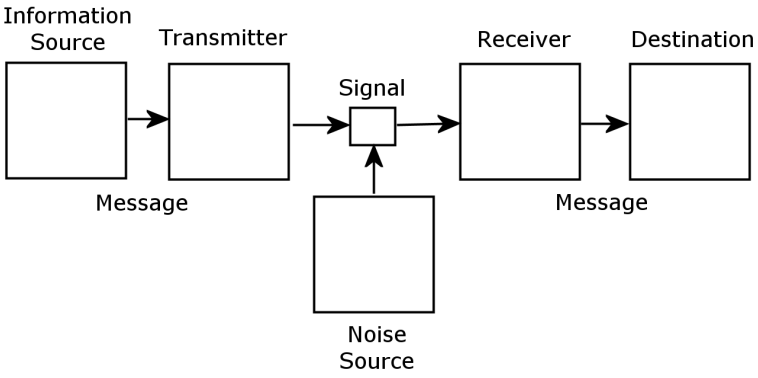

Figure 4.3: A depiction of the basic elements of a communication system. Adapted from [205].

Some deeper analysis of this consideration appears in the chapters that follow.

Let us consider this issue in a somewhat more formal context. Consider first figure 4.3. Figure 4.3 depicts the typical schematic of a general communication system. An information source is conveyed to a transmitter, from which messages are then passed on to a receiver and the messages arrive at their destination. In its intermediate form the message may or may not be subjected to the influence of some noise process. The nature of communication is such that the entire message cannot be transmitted in its entirety in parallel, but rather pieces of the message may be sent over time. The size of these pieces is governed by the capacity of the channel and given a limited time course or an ever changing information source, only a small subset of all possible messages may be sent.

In a similar manner, the complexity of content in the visual field prohibits a

deep analysis of all such content in parallel. This difficulty is solved by selecting some subset of visual content (a message) from the information source (incoming input comprising the contents of the visual field) and relaying this content to higher areas of the visual cortex (the receiver). The encoding of the incoming signals may be viewed as the actions that early neurons of the visual cortex perform on such signals.

In the classic problem, the messages that warrant transmission are governed by the properties of the message space. That is, in the example of the sentence *the man went to the store to buy some bread*, if the message space consists of our vocabulary of English words, the decision to send messages that are *improbable* yields a sentence that is highly informative.

A basic definition from information theory that encapsulates this idea is termed *self-information*. Self-information quantifies the amount of information that knowledge of the outcome of a certain event adds to one's overall knowledge. The self-information associated with the outcome A_n is given by $-log(p(A_n))$. Self-information is related to entropy in that entropy quantifies the weighted average self-information over the entire set of possible outcomes. So in the example of the biased coin toss, the self-information tells us that learning the coin came up tails is more informative than learning it came up heads.

Framed in the context of the communication problem, in the same manner as

self-information defines which packets to send to convey the maximum amount of information about the information source or message space, it may tell us which local regions of an image or scene inform most upon the contents of the image, and as such deserve the status of greater scrutiny by higher brain structures, or foveation.

The preceding formulation is expressed in figure 4.4. Figure 4.4 shows a scene S and some local neighbourhoods N_k with their local surround C_k. The question is that of the extent to which each N provides information about the contents of C, or S. Note that there is no claim being made as to the shape or size of N or C, but rather the figure merely uses circle shaped regions as an illustration of a generic neighborhood and its surrounding context.

4.2.3 Self-Information versus Entropy

The preceding discussion outlines the basic definition of entropy and its relation to self-information. The following discussion focuses specifically on the difference between these measures in the context of vision. The difference is subtle but important on two fronts. The first consideration lies in the expected behavior in *popout* paradigms and the second in the neural circuitry involved.

Let $X = [x_1, x_2, ..., x_n]$ denote a vector of RGB values corresponding to image patch X, and D a probability density function describing the distribution of

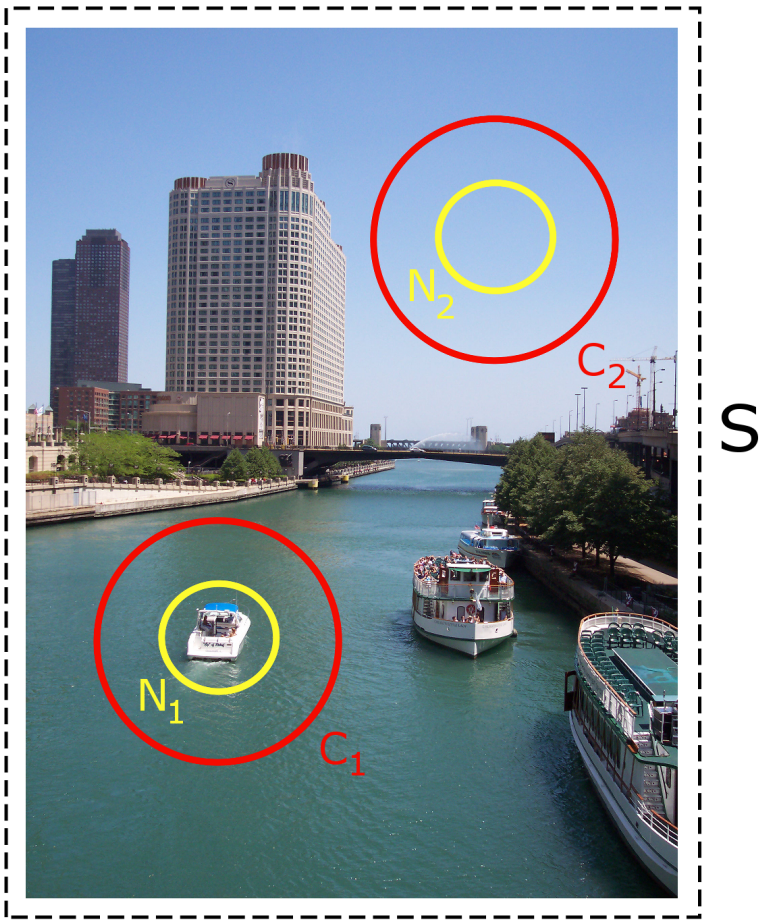

Figure 4.4: An illustration of the basic setup. The question is that of the extent to which each N informs on the associated C, or indeed S. Most would agree that N_1 provides more information about the contents of the scene than N_2.

some feature set over X. For example, D might correspond to a histogram estimate of intensity values within X or the relative contribution of different orientations within a local neighborhood situated on the boundary of an object silhouette [188]. Assuming an estimate of D based on N bins, the entropy of D is given by: $-\sum_{i=1}^{N} D_i log(D_i)$. In this example, entropy characterizes the extent to which the feature(s) characterized by D are *uniformly* distributed on X. An example for 2 feature domains (grey levels and orientations) is show in figure 4.5. The woman's eye, and the bottom region of the silhouette have relatively more uniform distributions and hence higher entropy. Note that the silhouette shown is of the type employed in the study of Renninger et al [188]. Self-information in the proposed saliency measure is given by $-log(p(X))$. That is, Self-information characterizes the raw likelihood of the specific n-dimensional vector of RGB values given by X. $p(X)$ in this case is based on observing a number of n-dimensional feature vectors based on patches drawn from the area surrounding X. Thus, $p(X)$ characterizes the raw likelihood of observing X based on its surround and $-log(p(X))$ becomes closer to a measure of local contrast whereas Entropy as defined in the usual manner is closer to a measure of local activity. The importance of this distinction is evident in considering figure 4.6. Figure 4.6a depicts a variety of candles of varying orientation, and color. There is a tendency to fixate the empty region on the left, which is the location of lowest entropy in the image in color and orientation space.

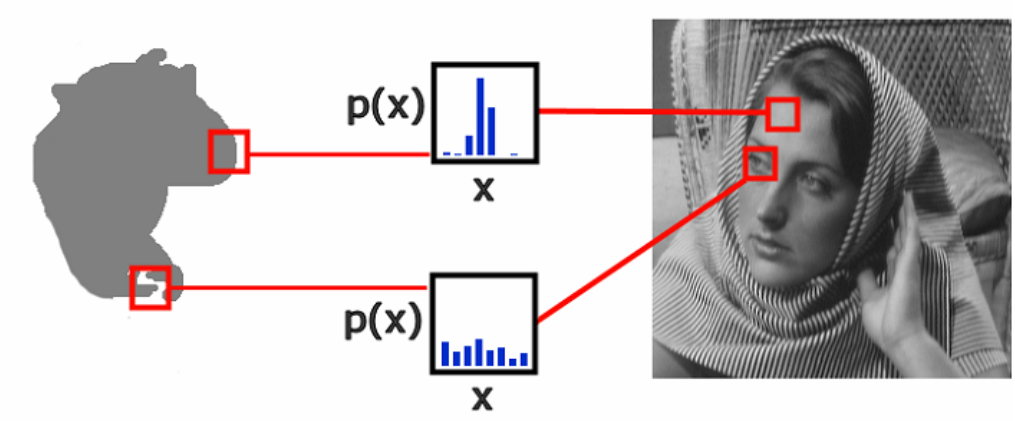

Figure 4.5: The distributions associated with two different features (greylevels and orientations). The eye region and the region on the lower part of the silhouette have more uniform distributions and hence higher entropy.

In contrast, this region receives the highest confidence from the quantity proposed in this work as it is highly informative in the context of this image. In figure 4.6b a silhouette of a figure is presented of the type considered by Renninger et al. In this case there is a tendancy to fixate the relatively flat region on the lower right of the figure, which corresponds to a highly peaked distribution in orientation space and is the locus of minimum entropy according to their measure.

It is worth briefly mentioning scale since evidently the local entropy and self-information associated with a more global context are influenced by the size of the region under consideration. For example, one might argue that for a sufficiently large window (or viewed from sufficiently far away), an entropy based operator

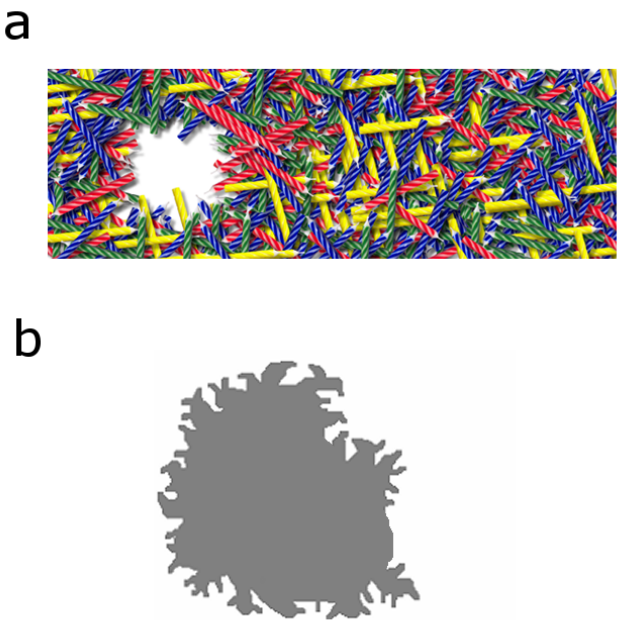

Figure 4.6: Examples that highlight the difference between entropy and self-information. a. Fixation invariably falls on the empty patch, the locus of minimum entropy in orientation and color but maximum in self-information when the surrounding context is considered. b. There is a tendency to fixate the relatively flat region on the lower right of the silhouette, a region of minimum entropy in orientation space. This is contrary to the predictions of the model of Renninger et al. [188].

might also predict the empty white region as most salient. It is less clear in the case of the object silhouette whether this is true. That said, the self-information operator provides the behavior one would expect regardless of the scale under consideration, whereas in this case an entropy based operator would offer an appropriate prediction only within a specific range of neighborhood sizes. The issue of scale also has a strong relationship to receptive fields. One would expect the receptive field size at any given layer to influence the region size over which entropy or self-information might be computed on the basis of the number of interneuron connections increasing as a function of this window. These are thoughts to bear in mind in the discussion of circuitry and coding in chapter 6.

4.3 A Computational Approach to Measuring Local Information

One problem that has not yet been discussed, is that of how the self-information of some local neighborhood (N in figure 4.4) can be computed in a general sense. Consider first the data carried by a single pixel of a color image; the content of the pixel is characterized by three numeric values corresponding to the red, green, and blue content of the pixel. If one wishes to characterize the distribution of pixel values in an entire image, this requires considering where all of the pixels

lie in a three dimensional space of RGB values. This is not difficult to estimate, but consider how the problem scales up to 2 pixels, or n pixels. In the case of 2 pixels, the distribution resides in a six dimensional space. In the case of n pixels, the estimate resides in a 3*n dimensional space. Convergence of the estimate is inversely proportional to the dimensionality of the space and grows exponentially with dimensionality. To appreciate this problem consider the following: Choosing 100 values randomly on the unit interval [0,1] provides a reasonable covering of the interval with a small distance on average between points. To achieve a similar covering of a 48 dimensional space (the dimensionality of a 4x4 RGB image patch) would require 10^{96} points, a quantity greater than estimates of the number of particles in the universe! Chapter 6 explores this issue in much greater detail than is presented here, including motivating this problem in a more formal setting, examining how the brain resolves this issue, and proposing specific circuitry in this regard. For now the discussion is restricted to a brief overview of the solution to this problem in a computational sense with only a very brief mention of neuroanatomy.

There exist a variety of statistical techniques that seek to discover structure in high-dimensional data. There also exists rather compelling evidence that the human brain encodes visual content in a manner closely resembling the statistical machinery underlying some of these techniques [13; 69]. One such approach that appears to result in an especially close match to the encoding appearing in the vi-

sual cortex of primates is Independent Component Analysis (ICA) [120]. ICA is a method for reducing a data set to subcomponents that are statistically independent and non-Gaussian. When applied to local patches drawn from natural images, ICA learns basis functions that resemble orientation selective cells at various angular and radial frequencies, color opponent cells, spatiotemporal receptive fields, and disparity selective cells all very similar to those found in the primary visual cortex of primates. These results strongly suggest that the primate visual system employs a sparse representation of the type learned by ICA algorithms, which takes advantage of redundancy in visual content. A few studies that consider the relationship of ICA to the specific properties of cortical cells indicate that the tuning of independent components learned from natural images is a good match with cortical neurons including the representation of color opponency [35], luminance/chromaticity preference [35], peak orientation and orientation bandwidth [35; 233], peak spatial frequency and spatial frequency bandwidth [35; 233], and aspect ratio [35; 233].

The most important property of a sparse representation, which will become apparent in the sections that follow, is that the activation of different types of cells are assumed to be mutually independent. For example, a cell that responds optimally to a vertical edge will not respond to a horizontal edge. Having such an independence assumption means that the distribution of each type of cell encoding an image can be modeled independently, greatly reducing the data required for any

estimate on the distribution of such variables. Specific details of the relation of this operation to the representation of local image content is presented in the section that follows.

4.4 The Model

In this section, the definition of saliency is given in more specific terms under the assumption of a learned V1 like independent representation as discussed. Saliency is determined by quantifying the self-information of each local image patch. To reiterate the problem briefly, even for a very small image patch, the probability distribution resides in a very high dimensional space. There is insufficient data in a single image to produce a reasonable estimate of the probability distribution. For this reason a representation based on independent components is employed for the independence assumption it affords as discussed. ICA is performed on a large sample of 11x11 RGB patches drawn from natural images to determine a suitable basis. For a given image, an estimate of the distribution of each basis coefficient is learned across the entire image through non-parametric density estimation. The probability of observing the RGB values corresponding to a patch centred at any image location may then be evaluated by independently considering the likelihood of each corresponding basis coefficient. The product of such likelihoods yields the joint likelihood of the entire set of basis coefficients. The overall architecture for this

computation is depicted in figure 4.7. Elements depicted in the figure are as follows: Each shaded rectangle depicts an operation involved in the overall computational framework (For a more detailed description, refer to appendix A):

Infomax ICA: A large number of local patches are randomly sampled from a set of approximately 4000 natural images. Based on these patches a sparse spatiochromatic basis is learned via infomax ICA. An example of a typical mixing matrix labeled as A is shown for 7 x 7 windows.

Matrix Pseudoinverse: The pseudoinverse of the mixing matrix provides the unmixing matrix which may be used to separate the content within any local region into independent components. The functions corresponding to the unmixing matrix resemble oriented Gabors and color opponent cells akin to those appearing in V1.

Matrix Multiplication: The matrix product of any local neighbourhood with the unmixing matrix yields for each local observation a set of independent coefficients corresponding to the relative contribution of various oriented Gabor-like filters and color opponent type cells.

Density Estimation: Producing a set of independent coefficients for every local neighborhood within the image yields a distribution of values for any single coefficient based on a probability density estimate in the form of a histogram or Kernel density estimate.

Joint Likelihood: Any given coefficient may be readily converted to a probability

by looking up its likelihood from the corresponding coefficient probability distribution. The product of all the individual likelihoods corresponding to a particular local region yields the joint likelihood.

Self-Information: The joint likelihood is translated into Shannon's measure of Self-Information by $-log(p(x)))$. The resulting information map depicts the Saliency attributed to each spatial location based on the aforementioned computation.

Details of each of the model components are as follows:

Projection into independent component space provides, for each local neighborhood of the image, a vector w consisting of N variables w_i with values v_i. Each w_i specifies the contribution of a particular basis function to the representation of the local neighborhood. As mentioned, these basis functions, learned from statistical regularities observed in a large set of natural images, show remarkable similarity to V1 cells [13; 69]. It is important to note that under certain conditions, ICA is equivalent to Linsker's infomax proposal [131], and certain algorithms [13; 120] may be seen as heuristics that achieve Linsker's original proposal. For a detailed discussion of this issue, readers may wish to refer to [162].

The ICA projection then allows a representation w, in which the components w_i are as independent as possible. For further details on the ICA projection of local image statistics see [26]. As discussed, salience may be defined based on a strategy for

maximum information sampling. In particular, Shannon's self-information measure [205], $-log(p(x))$, applied to the joint likelihood of statistics in a local neighborhood decribed by w, provides an appropriate transformation between probability and the degree of information inherent in the local statistics. It is in computing the observation likelihood that a sparse representation is instrumental: Consider the probability density function $p(w_1 = v_1, w_2 = v_2, ..., w_n = v_n)$ which quantifies the likelihood of observing the local statistics with values $v_1, ..., v_n$ within a particular context. An appropriate context may include a larger area encompassing the local neigbourhood described by w, or the entire scene in question. The presumed independence of the ICA decomposition means that $p(w_1 = v_1, w_2 = v_2, ..., w_n = v_n) = \prod_{i=1}^{n} p(w_i = v_i)$. Thus, a sparse representation allows the estimation of the n-dimensional space described by w to be derived from n one dimensional probability density functions. Evaluating $p(w_1 = v_1, w_2 = v_2, ..., w_n = v_n)$ requires considering the distribution of values taken on by each w_i in a more global context. In practice, this might be derived on the basis of a nonparametric or histogram density estimate. In chapter 6, we demonstrate that an operation equivalent to a non-parametric density estimate may be achieved using a suitable neural circuit.

One aspect lacking from the preceding description is that the saliency map fails to take into account the dropoff in visual acuity moving peripherally from the fovea. In some instances the maximum information accommodating for visual acuity may

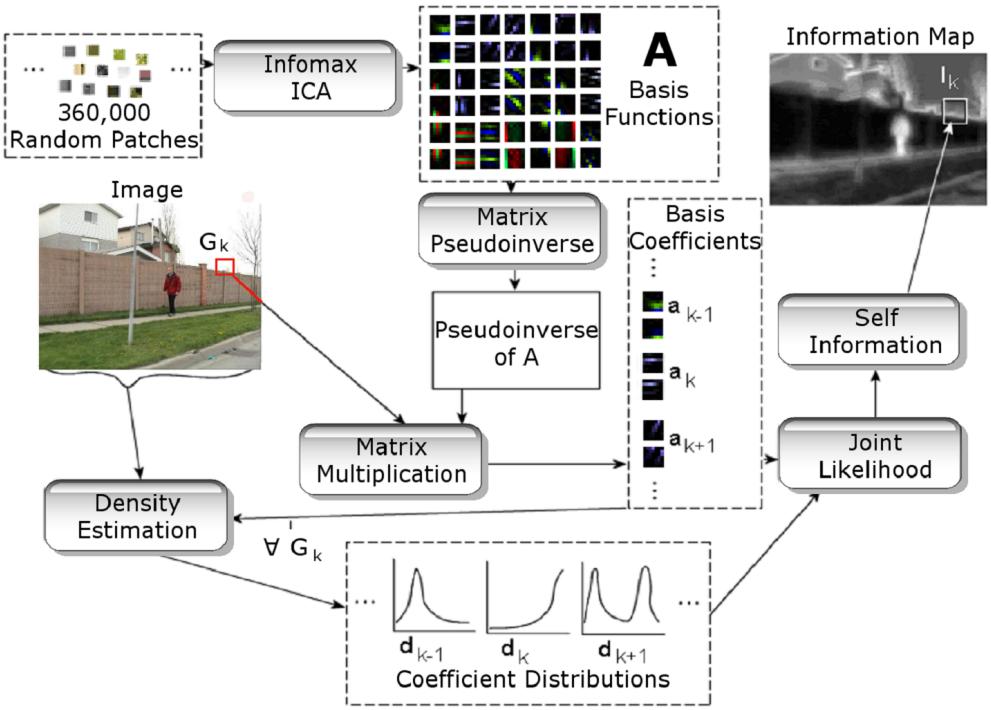

Figure 4.7: The proposed model. For additional details refer to appendix A.

correspond to the center of a cluster of salient items, rather than centered on one such item. For this reason, the resulting saliency map is convolved with a Gaussian with parameters chosen to correspond approximately to the dropoff in visual acuity observed in the human visual system. This simply means that the information gain in making a saccade reflects the dropoff in visual acuity thus resulting in an arguably more appropriate representation for comparison with eye movements. This is an issue that is revisited in more detail in the chapter that follows.

It is perhaps worth mentioning at this point some of the important contributions of the model to the literature, since the preceding setup offers a greater sense of how the model relates to the following subject matter relative to the more general description that appeared in chapter 1:

1. A bottom-up model of overt attention with selection based on the self-information of local image content.
2. A qualitative and quantitative comparison of predictions of the model with human eye tracking data, contrasted against the model of Itti and Koch [98]. Some preliminary results are presented in this chapter to give some sense of behavior on real data, and a more thorough analysis appears in chapter 5.
3. Demonstration that the model is neurally plausible via implementation based on a neural circuit with properties akin to the circuitry involved in early visual processing in primates. This is largely the subject matter of chapter 6. The

focus is on how the model relates to neuroanatomy and circuitry.

4. Discussion of how the proposal generalizes to address issues that defy explanation by existing saliency based attention models. This content is discussed in chapter 8.

5. There are a variety of behaviors observed in primate psychophysics that are nicely captured by the proposal of an information theoretic basis for attentional selection. These are described in detail in chapter 7.

6. A role of the proposed saliency descriptor in machine vision. Focus is on the selection of interest points and the possibility of the representation assumed by the model to select interest points. This discussion appears in chapter 9.

It is also important to note that there are many aspects of visual perception that are related to, but distinct from the proposal. For example, task and contextual factors impact significantly on what is targeted by selective attention. It is important to bear in mind that the focus of this dissertation is on modulation of cortical activity associated with the properties of the visual stimulus itself. This is a problem distinct from task and context in general, and also distinct from other more specific task related factors, such as the attenuation/alteration of neural signals to subserve object representation as one example. This is a problem distinct from the mechanism by which a subset of visual input is selected, and distinct from task or

context related bias, with the focus on retaining information as inputs ascend the visual hierarchy based solely on the properties of the stimulus itself.

4.5 A First Look at Model Behavior and Performance

The following section evaluates the output of the proposed algorithm as compared with the bottom-up model of Itti and Koch [98]. The model of Itti and Koch is perhaps the most popular model of saliency based attention and currently appears to be the yardstick against which other models are measured. The results at this stage are based on a first choice of parameters and are not intended to be an exhaustive or thorough depiction of model behavior and performance but rather are intended to provide some sense of how the model relates to the processing of real image data with respect to predicting saccadic behavior. In the chapter that follows, an exhaustive performance evaluation and exploration of the parameter space is presented.

4.5.1 Experimental Eye Tracking Data

The data that forms the basis for performance evaluation is derived from eye tracking experiments performed by the author. Subjects observed 120 different color images. Images were presented in random order for 4 seconds each with a mask between each pair of images. Subjects were positioned 0.75m from a 21 inch CRT

monitor and given no particular instructions except to observe the images. Images consist of a variety of indoor and outdoor scenes, some with very salient items, others with no particular regions of interest. The eye tracking apparatus consisted of an ERICA workstation including a Hitachi CCD camera with an IR emitting LED at the centre of the camera lens. The infrared light was reflected off two mirrors into the eye facilitating segmentation of the pupil. Proprietary software from ABB corporate research was used to analyze the data. The parameters of the setup are intended to quantify salience in a general sense based on stimuli that one might expect to encounter in a typical urban environment. Data was collected from 20 different subjects for the full set of 120 images.

The issue of comparing between the output of a particular algorithm, and the eye tracking data is non-trivial. Previous efforts have selected a number of fixation points based on the saliency map, and compared these with the experimental fixation points derived from a small number of subjects and images (7 subjects and 15 images in a recent effort [182]). As mentioned in chapter 3, there are a variety of methodological issues associated with such a representation. The most important such consideration is that the representation of perceptual importance is typically based on a saliency map. Observing the output of an algorithm that selects fixation points based on the underlying saliency map obscures observation of the degree to which the saliency maps predict important and unimportant content and in partic-

ular, ignores the determination of saliency levels away from highly salient regions. Secondly, it is not clear how many fixation points should be selected. Choosing this value based on the experimental data will bias output based on information pertaining to the content of the image and may produce artificially good results.

The preceding discussion is intended to motivate the fact that selecting discrete fixation coordinates based on the saliency map for comparison may not present the most appropriate representation to use for performance evaluation. In this effort, we consider two different measures of performance. Qualitative comparison is based on the representation proposed in [112]. In this representation, a fixation density map is produced for each image based on all fixation points, and subjects. Given a fixation point, one might consider how the image under consideration is sampled by the human visual system as photoreceptor density drops steeply moving peripherally from the centre of the fovea. This dropoff may be modeled based on a 2D Gaussian distribution with appropriately chosen parameters, and centred on the measured fixation point. A continuous fixation density map may be derived for a particular image based on the sum of all 2D Gaussians corresponding to each fixation point, from each subject. The density map then comprises a measure of the extent to which each pixel of the image is sampled on average by a human observer based on observed fixations. This affords a representation for which similarity to a saliency map may be considered at a glance. Quantitative performance evaluation

is achieved according to the procedure of Tatler et al. discussed previously [213]. The saliency maps produced by each algorithm are treated as binary classifiers for fixation versus non-fixation points. The choice of several different thresholds and assessment of performance in predicting fixated versus not fixated pixel locations allows an ROC curve to be produced for each algorithm.

4.5.2 Experimental Results

Figure 4.8 affords a qualitative comparison of the output of the proposed model with the experimental eye tracking data for a variety of images. Also depicted is the output of the Itti and Koch algorithm for comparison. Figure 4.9 demonstrates a quantitative comparison based on ROC curves which yields areas of 0.753 ± 0.00816 for self-information and 0.728 ± 0.00840 for Itti and Koch based on a very loose upper bound for the variance associated with the area under the curve as suggested in [43; 47] which is significant with $p < 0.0001$. Although in reality the error bounds are likely to be much tighter than those mentioned, the preceding serves to establish significance for the results shown without the need for computationally intensive analysis required to establish tighter bounds.

In the implementation results shown, the ICA basis set was learned from a set of 360,000 11x11x3 image patches from 3600 natural images drawn from the Corel Stock Photo Database and using the Lee et al. extended infomax algorithm [120].

Figure 4.8: Results for qualitative comparison. Within each boxed region defined by solid lines: (Top Left) Original Image (Top Right) Saliency map produced by Itti + Koch algorithm. (Bottom Left) Saliency map based on information maximization. (Bottom Right) Fixation density map based on experimental human eye tracking data.

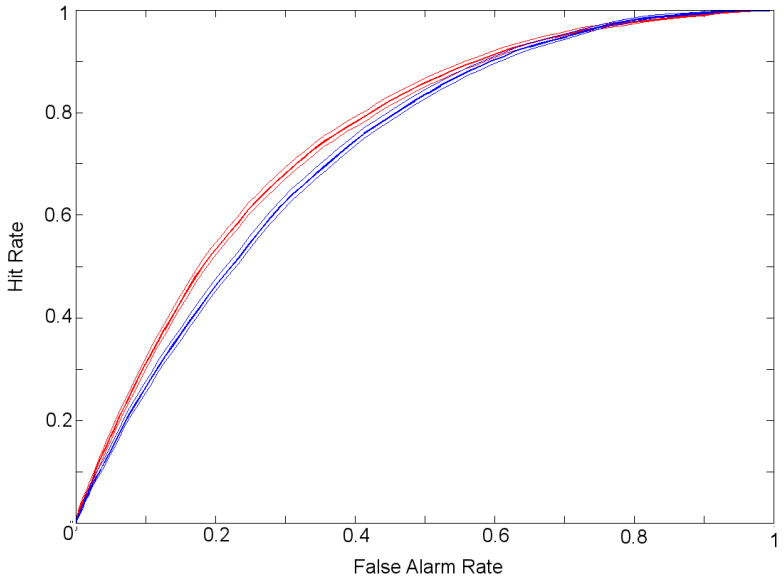

Figure 4.9: ROC curves for AIM (red) and Itti and Koch (blue) saliency maps along with 99% confidence intervals. Area under curves is 0.753 ± 0.00816 and 0.728 ± 0.00840 respectively based on a conservative upper bound for the variance associated with estimates of the area under the ROC curves as in [43; 47].

Processed images are 340 by 255 pixels. Ψ consists of the entire extent of the image and $w(s,t) = \frac{1}{p} \ \forall \ s,t$ with p the number of pixels in the image (340x255 in this example). One might make a variety of selections for these variables based on arguments related to the human visual system, or based on performance and these considerations are explored in detail in chapters that follow. In our case, the values have been chosen on the basis of simplicity in order to obtain an early sense of model performance. In particular, we wished to avoid tuning these parameters to the available data set. The ROC curves appearing in figure 4.9 give some sense of the efficacy of the model in predicting which regions of a scene human observers tend to fixate. As may be observed, at this stage the predictive capacity of the model is significantly better than the approach of Itti and Koch based on a more or less random first choice of parameters and encoding of image content. Encouraging is the fact that favorable performance is achieved using a method derived from a principled definition, and with no parameter tuning or *ad hoc* design choices. An independent comparison of these models has been conducted demonstrating the superiority of this proposal on an additional dataset for a completely ad hoc choice of parameters [99].

4.6 Conclusion

We have described a strategy that predicts human attentional deployment on the principle of maximizing information sampled from a scene. Although no computational machinery is included strictly on the basis of biological plausibility, nevertheless there are some interesting early parallels to the machinery that facilitates early visual processing in primates including computational units that resemble those implicated in early visual processing in primates and a model that has inhibition from the surround as a central element of salience based processing. Comparison with an existing attention model reveals the efficacy of the proposed model in predicting salient image content. There are many issues that arise out of the discussion of this chapter, and much more exploration required in order to evaluate the plausibility and generality of the model. That said, preliminary results in predicting eye movement patterns are very encouraging, as are early indicators of parallels between this proposal and the structure of the primate visual cortex. Later chapters reveal a deeper connection to psychophysics and neural circuitry as well as a more thorough investigation of the model in characterizing eye movement patterns.

5 Basis Functions, Context, and Overt selection

The ambition of chapter 5 is to validate the proposal put forth in chapter 4 for predicting eye movements and in addition, to explore aspects of the role of projection pursuit in the representation of natural image content. There are various parameters involved in the model, and various choices for such parameters make sense from biological or engineering vantage points. This chapter includes a systematic exploration of the parameter space with consideration for algorithm performance and biological plausibility. In each case, the model is compared with that of Itti and Koch [98] according to the performance metric of Tatler et al. [213] demonstrating the efficacy of the model (AIM) in predicting the locations of fixation points based on a fixed 2D image with a fixed head position.

The following sections consider a variety of implementation details associated with the basic framework put forth in AIM along with some related neurobiological considerations. It is worth first observing the qualitative and quantitative results associated with some of the better parameter choices to offer a sense of the efficacy

of the proposal in predicting fixations on spatiochromatic data. Figures 5.1 and 5.2 offer a sense of the qualitative similarity of the model output to the fixation density maps. The results shown correspond to the 31x31 receptive field size for the Jade ICA algorithm with PCA preprocessing retaining 95% of the variance. The Jade algorithm is discussed in the latter part of this chapter. The choice of this particular parameter set is intended to offer a sense of performance for one of the *better* choices of parameter sets and also to offer results that employ the same conditions as the results on psychophysics paradigms considered in chapter 7 in light of the issues pertaining to color opponency discussed in this chapter. The quantitative performance associated with this set of conditions appears in figure 5.3. Note the performance is significantly better than the Itti and Koch algorithm as described in [98].

5.1 Dimensionality and Receptive Field Size

In the preceding chapter, the utility of a sparse representation with respect to density estimation was demonstrated. In machine vision applications, the use of ICA to achieve a sparse representation is often preceded by Principal Component Analysis, a projection pursuit which selects axes that explain the most variance in the data. This is an important step since in some cases the space or time complexity of algorithms involved prohibits consideration of data sets above a certain dimen-

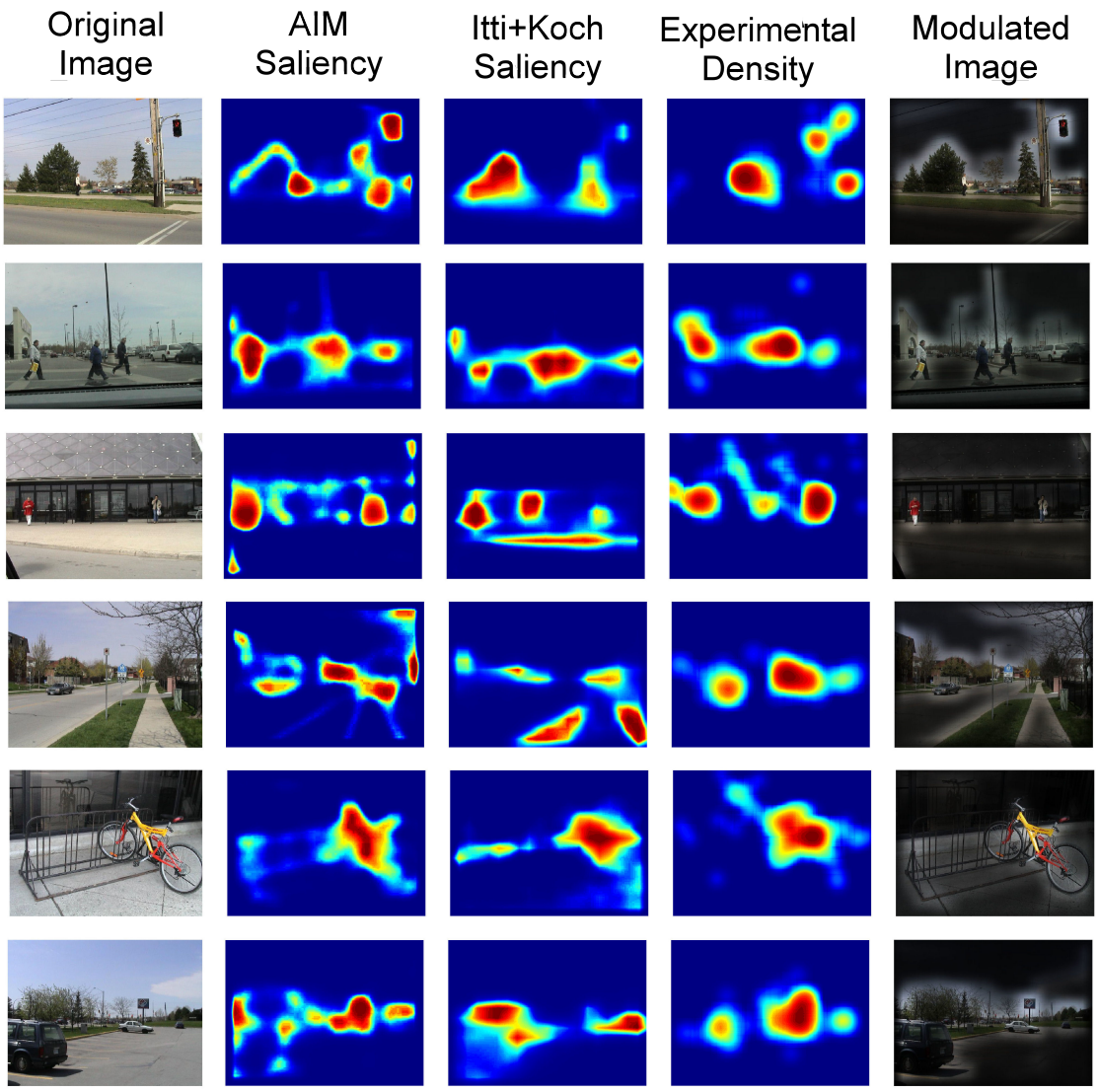

Figure 5.1: Qualitative comparison of the output of AIM, the algorithm of Itti and Koch and the experimental data. From left to right: Original Image, AIM output, Itti and Koch output, experimental density and the original image modulated by the output of AIM via a product to offer a sense of the localization of saliency related activation.

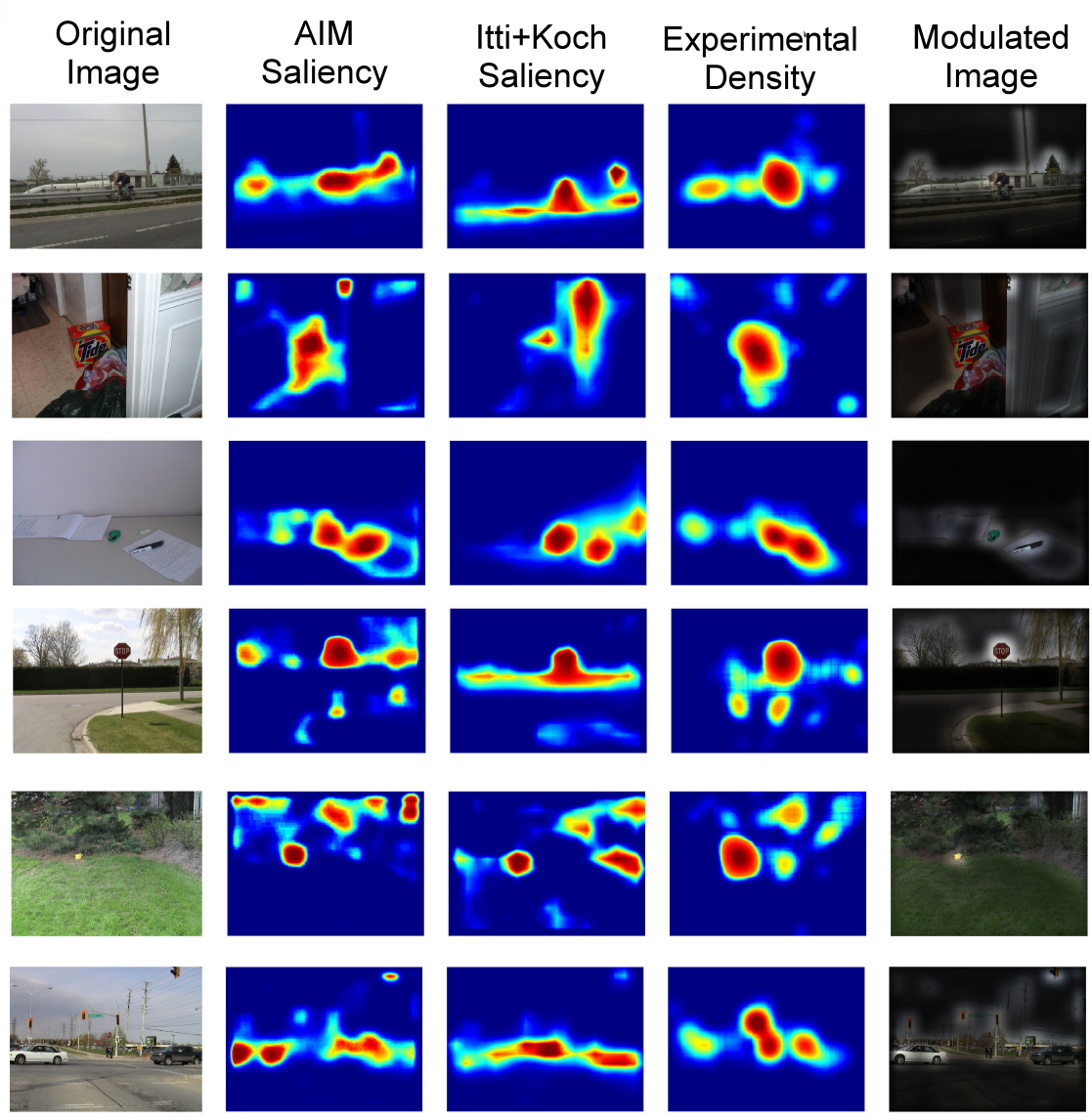

Figure 5.2: Additional examples of the sort offered by figure 5.1.

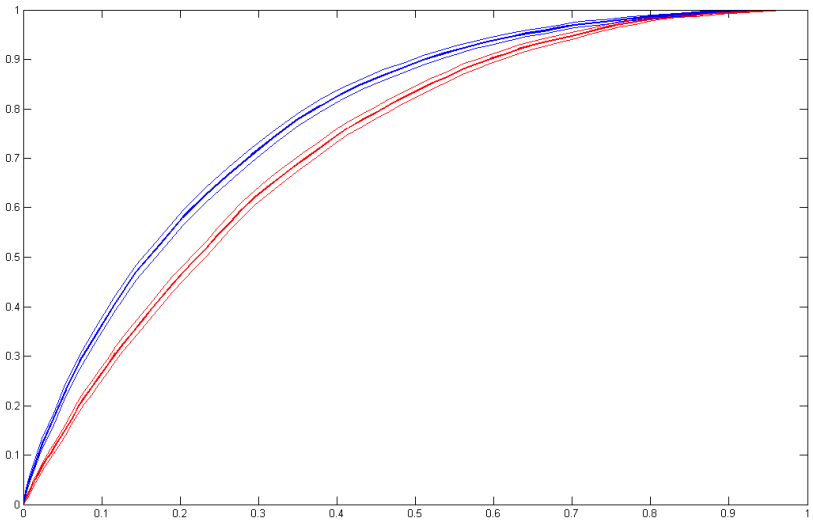

Figure 5.3: ROC curves for AIM (blue) and Itti and Koch (red). Area under the curves is 0.781 for AIM and 0.728 for Itti and Koch.

sionality. For this reason, it is sensible to consider dimensionality reduction as a precursor to learning a sparse representation since it is possible in many instances to capture an overwhelming proportion of the variance in the data using far fewer dimensions provided axes are chosen correctly. This also makes sense in a biological system in which energy and space are limiting factors. For example, if it requires 2000 neurons to adequately represent a particular state, it may be preferable to instead represent 99.99% of the variance associated with that state using only say 200 neurons. This will have the effect of greatly reducing energy requirements and boosting the representational capacity of the brain. It should be noted that there does exist the possibility that the small variance carried by these lower components may be important for certain tasks and that caution must be taken in reducing the data in this manner. That said, as the results that follow in this chapter demonstrate, it is possible to attain a significant savings in the data dimensionality at only a small cost with respect to perceptual quality.

In order to test the impact of dimensionality reduction on performance in predicting eye movements, we have considered the data from appendix C which includes the raw performance associated with various choices for receptive field size, dimensionality determined by variance captured, spatial scale at which processing is performed and ICA algorithm. Results include preprocessing via PCA retaining 90%, 95%, 97.5%, 99%, 99.5% and 99.9%. In each case, factors (e.g. choice of ICA

basis) are considered in isolation by averaging across all scores for dimensions not under consideration. (e.g. to consider the impact of the basis used, scores are averaged across variance retained, scale of processing and whether or not convolution is performed)

It is natural to consider these results in combination with considering receptive field size as dimensionality is a function of receptive field size. Furthermore, owing to constraints imposed by memory and time complexity, it was not possible to run all combinations of receptive field size and variance. Results include 11x11, 21x21 and 31x31 windows, a rather coarse sampling in space sufficient for determining the impact of receptive field size on the determination of visual saliency. A first step lies in determining the number of components that correspond to these choices of variance retained. The choice of a lower cap of 90% was based on the observation that below this level the visual reconstruction of the images in question became severely degraded. Number of components to retain corresponding to the three choices of window size are depicted in table 5.1. Note the diminishing returns as additional components are added.

A complete picture is offered by figure 5.4 depicting the variance explained versus number of components for each of the three choices of window size. Of note is the rapid diminishing returns in retaining more components in addition to the property of scale invariance. Of interest is the fact that scale invariance is preserved

Var\RF	11x11	21x21	31x31
0.9	4	7	13
0.95	9	25	54
0.975	21	69	150
0.99	48	163	349
0.995	72	251	536
0.999	135	478	1021

Table 5.1: For various receptive field sizes, the number of basis functions required to retain the desired variance.

for chromatic patches.

It is also interesting to consider how the perceptual quality of these representations varies with the number of principal components retained. Examples for a variety of choices for number of principle components is depicted in figure 5.5. Note once again the diminishing returns associated with the perceptual quality of these representations in increasing the number of components retained. The examples at the higher end of the spectrum are perceptually indistinguishable from the original unprocessed image.

The ROC evaluation described in chapter 5 was performed for all the various conditions described above and results are summarized in table 5.2 in the form of the area under ROC curve scores associated with the various conditions.

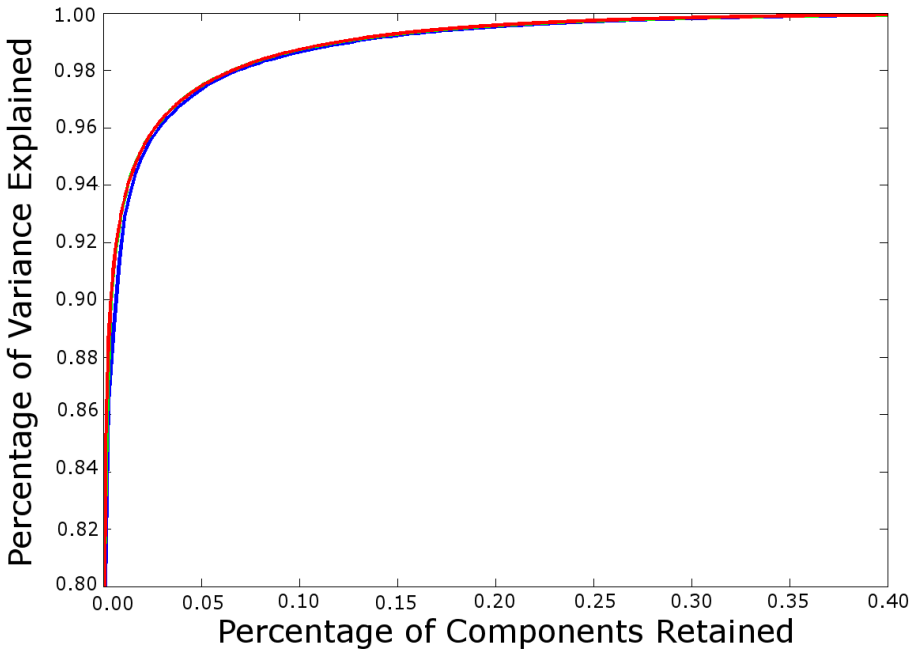

Figure 5.4: Variance explained versus number of principal components retained. Note the rapid diminishing returns and scale invariance of this relationship. Curves correspond to 11x11 windows (blue), 21x21 (green), 31x31 (red).

Figure 5.5: An image reconstructed with Principal Components discarded leaving (left to right, top to bottom) 99.9, 99.5, 99, 97.5, 95, and 90 percent of the local variance respectively.

Var\RF	11x11	21x21	31x31
0.9	0.719	0.732	0.749
0.95	0.732	0.755	0.769
0.975	0.738	0.758	N/A
0.99	0.735	0.755	N/A
0.995	0.731	N/A	N/A
0.999	0.728	N/A	N/A

Table 5.2: Demonstrates the effects of receptive field size and dimensionality reduction on area under ROC curve scores. N/A refers to conditions for which the computational requirements associated with the ICA learning prevented learning for the combination of variance retained and receptive field size in question.

The results shown in table 5.2 demonstrate some interesting aspects of the saliency computation. With regards to the role of dimensionality reduction, there is a large difference between a representation that retains 90% versus 95% of the variance for all RF sizes, but there is much less of a difference among all other categories. This suggests that above a certain level of variance retained, there may be little benefit of dealing with a higher dimensionality representation and also that 95% variance may be sufficient. This level of dimensionality reduction presents considerable savings as compared with a lossless representation and is worthy of inclusion in an implementation intended for machine vision applications. The story concerning receptive field size is very different. In this case there appears to be a considerable benefit of increasing the receptive field size. This no doubt is related to the fact that features shift from including localized edge content to larger blobs or patches of color. It is also evident from the unfilled locations in table 5.2 that the strategy of *learning* a basis has inherent limitations owing to the quadratic increase in dimensionality with receptive field size. One might expect further improvement from a 41x41 or larger patch size, but hardware limitations deny consideration of larger regions. For this reason, additional benefit may be had with regard to performance in constructing an analytic basis with the desired properties.

There are some important considerations with regard to any system that seeks

to emulate aspects of human behavior especially those that claim to be biologically motivated/plausible. It is possible to make design decisions in line with the macroscopic structure of a biological system that ignore some of the *smaller* details that have as emergent properties behavior that is misleading in the sense that introducing some of these ostensibly lesser constraints can change the behavior of the system entirely. One factor that may be suggested as falling into this category is that of receptive field size. Many systems that claim biological plausibility tend to overlook this consideration. In the model of Itti, Koch and Niebur, there is no discussion of how the receptive field sizes employed in their implementation relate to neurobiology. Owing to the pyramid representation employed in their implementation, some of the receptive fields employed are of a size more closely resembling those appearing in higher visual areas. There is no discussion of the implications of this from the perspective of how the model relates to a neurobiological analogue. It is also unclear why some set of *basic features* should be the only factors that contribute to the determination of saliency since the larger receptive fields seem to imply the involvement of higher visual areas for which neurons code for a variety of more complex features. These issues are further discussed in chapter 8 as more general issues pertaining to modeling are addressed.

5.2 Sparse Representation

The process of learning a sparse representation is not an exact science. Unlike a projection pursuit based on principal components, there is not a closed form solution to seeking components for which the optimality criterion is independence. To satisfy ourselves that the choice of ICA basis is not an overwhelming factor in the determination of saliency, we have chosen two of the more popular algorithms as a basis for comparison. It is worth noting that two basis representations that are both *truly* independent should yield the same result in the determination of saliency and thus similar results from the two algorithms should afford at least some confidence on the independence assumption inherent in the computation.

Independent component analysis proceeds according to the assumption that a multivariate signal is comprised of an additive combination of subcomponents that are mutually independent. The goal of such algorithms generally involves an iterative procedure that seeks to maximize the statistical independence of the constituent elements. There are a variety of criteria that are related to independence that are exploited by these algorithms including seeking non-Gaussianity, high kurtosis and minimizing mutual information.

Two commonly employed ICA algorithms, which are robust and produce stable solutions across multiple runs are the infomax algorithm [120], and Jade [32]

algorithm. A brief description of each algorithm and their differences follows:

The basic framework of ICA assumes the existence of n independent signals $s_1(t), ..., s_n(t)$ and n observations comprised of mixtures of the independent signals $x_1(t), ..., x_n(t)$ so that $x_i(t) = \sum_{j=1}^{n} a_{ij} s_j(t)$. For convenience this can be represented more compactly as $x(t) = As(t)$ and A is referred to as the mixing matrix. ICA seeks to recover $s(t)$ given only $x(t)$ and the problem may be formulated as finding a separating matrix B such that $y(t) = Bx(t)$ where $y(t)$ is an estimate of $s(t)$. Algorithms seeking independent components proceed by way of optimizing a contrast function on the output y. That is, by optimizing a scalar measure of some distributional property of the output y. Such contrast functions are abbreviated as $\Phi(y)$.

5.2.0.1 Infomax

The infomax method proceeds according to maximizing the contrast function $\Phi[y] \equiv -H[g(y)]$ where $H[.]$ denotes the Shannon entropy. This projection pursuit then seeks a form for y that is effectively as uniform as possible.

5.2.0.2 Jade

The Jade algorithm is based on higher order statistics and involves observing correlation beyond second order through consideration of cumulants. Although the

details of this procedure are beyond the scope of this dissertation, it is worth noting that the optimization is fundamentally tied to the kurtosis of the source distributions. In each case, optimization proceeds according to maximizing a function that is correlated with the independence of the constituent elements.

5.2.1 Sparsity and ROC scores

The ROC scores averaged across all conditions for the two different ICA algorithms are 0.7415 for Jade and 0.7423 for Infomax. Thus, there is no significant difference associated with the choice of ICA algorithm and a detailed look at the data indicates that for all choices of other parameters the ROC scores associated with the two algorithms is very similar. This is an encouraging result, as any truly independent basis should yield the same quantitative determination of visual saliency, and the metrics based on these two algorithms are nearly identical.

One issue that is not addressed in the literature concerning ICA specific to chromatic content, is the extent to which the resulting representation of color opponency is in agreement with that observed in the visual cortex. While a thorough analysis of this problem is beyond the scope of this dissertation, anecdotal evidence in the form of observations made in performing the experiments that appear in chapter 7 suggests that the representations learned via higher order cumulants given by the Jade algorithm tends to yield a representation of color opponency that is closer to

that which appears in the visual cortex.

5.3 Centre and Surround

Thus far we have considered only the case where the local context consists of the entire image in question. In the chapter that follows the focus shifts to the implementation including the plausibility of this account and possible cortical mechanisms for the proposed computation. It is worth however considering the scores associated with a model in which the surround statistics are determined locally. Unfortunately, in the absence of the sort of parallel hardware with which the brain is equipped, this is a consideration that proves to require an overwhelming degree of computation requiring a local density estimate to be performed for every local neighbourhood. In the chapter that follows, a means of computing this quantity via a simple circuit architecture is described which gives a sense of a possible cortical analogue of this behavior. That being said, although an exhaustive determination of performance based on local centre and surround regions proves computationally prohibitive it is at least worth carrying out this performance evaluation for a single sensible choice of these parameters motivated by biological observations.

The specific choice of parameters for this analysis are based on the data appearing in Figure 7 in [178] corresponding to a dropoff in surround modulation by a factor of approximately 200 over approximately 5 degrees visual angle. The

receptive field size relative to this surround varies with change in scale and was fixed at 21x21 pixels corresponding to the Jade algorithm preserving 95% variance. This yields an ROC score of 0.7472 for the image processed at full scale, and 0.7620 processed at half scale. In both cases these are once again significantly better than the algorithm of Itti, Koch and Niebur, but fall short of the scores associated with a more thorough optimization using a global context. In any case, it may be said that this result suggests promise in regard to model performance based on local center-surround interaction and a more exhaustive exploration of the parameter space might bring scores in line with the global context case. It is also interesting to note, that this particular form of computation would lend itself well to an implementation in hardware.

5.4 Visual Acuity

The role of visual acuity within saliency computation is an issue that has for the large part been ignored in computational modeling. There are many factors associated with the central proposal of this dissertation, and the issue of neuron density in the cortical representation is a deep issue in itself that this document does not seek to explore in detail. That being said, there are a few simple aspects of this issue that deserve discussion.

In the chapter that follows, it is clear in the discussion of how the proposal of

AIM relates to cortical circuitry that the proposal is amenable to an implementation in which photoreceptor density is non-uniform. Thus from a theoretical perspective the issue of non-uniformity in the underlying cortical representation is inherent in the basic proposal, but for the sake of parsimony in computation and in the complexity of the implementation, is not captured in the quantitative performance metrics presented.

A secondary issue of interest pertaining to visual acuity concerns the role of clustering of targets. Some behavioral experiments indicate that experimental participants may direct saccades to the centre of a group of targets as opposed to any single target suggesting that the determination of salient content may account for photoreceptor density being sufficient to sample a number of targets in a single saccade provided they are all within a localized region [83]. To determine whether this consideration has any impact on the proposed saliency computation, the resulting saliency landscape has been convolved with a Gaussian with parameters chosen to approximately correspond to the dropoff in visual acuity observed from the fovea to the periphery in order to determine the overall saliency associated with locations for which no targets are present but for which a number of salient targets are nearby. Average ROC scores associated with convolved versus non-convolved saliency maps according to this metric are 0.7525 and 0.7313 respectively. Although there does appear a small difference in the score for these two conditions computationally,

it is worth noting that some mechanism that accounts for visual acuity would be required to capture the results appearing in [83] and is a worthwhile consideration in addressing clustering effects.

As a whole, the specific relationship between visual acuity, photoreceptor density and cortical representation is an area that would benefit from further consideration on the side of computational modeling, but any further consideration is beyond the scope of this thesis.

5.5 Discussion

Performance metrics over all conditions vary from the low end of the spectrum with scores in the ballpark of the algorithm of Itti and Koch [98] to performance significantly greater at the high end. It is interesting to note that many of the parameters have little effect on the determination of saliency scores and qualitatively on the resulting saliency landscape. As a whole, the results demonstrate that the proposal has the following advantages over its predecessors:

1. A principled proposal for the specific nature of cortical saliency computation. This is in contrast to the model of Itti and Koch [98] in which the normalization operator (which is the closest analogue of the proposed computation within their model) is a crude approximation. This is arguably the most important portion of the computation specific to saliency when separated from

other more pejorative elements such as the selection mechanism.

2. A saliency metric that outperforms any of its competitors in predicting eye movement patterns in human observers.

3. A saliency metric with early indicators of biological plausibility owing to the similarity of elements involved to cortical cells despite the motivation underlying its construction being a means of local likelihood evaluation achieved by way of independence. It is revealed in the chapters that follow that this computation has an even stronger relationship to cortical circuitry and has as emergent properties, a variety of behaviors that agree well with a wide range of psychophysics results.

In the chapters that follow, issues surrounding this basic architecture, including its relationship to more general aspects of attention modeling are explored further, underscoring the plausibility of the proposal.

6 Efficient Coding and Density Estimation in the Brain

In chapter 4 we employed the assumption of a transformation to a sparse representation as a mathematical convenience in the context of the proposed model. In this chapter we demonstrate why such a representation is a necessity rather than a mere convenience. We highlight evidence for sparse coding in the primate brain and further document the apparent ubiquity of sparse coding within the animal kingdom. Following this, more formal discussion is put forth demonstrating the necessity of sparsity from the perspective of complexity arguments.

The notion of neural representation of likelihoods is a necessary component for any model of the brain that posits the involvement of statistical inference in decision making. Assuming a sparse representation, we demonstrate that a simple neural circuit may achieve neural density estimation to a great deal of precision. The key to this lies in the fact that the likelihood of any individual neurons firing rate can be estimated on the fly without an explicit representation of a probability density

function.

6.1 Why a Sparse Code?

The brain is required to encode incoming sensory patterns in addition to its internal states through patterns of neuron firing. If each neuron in a particular representation is described as being either active or passive, this provides a simple means of considering properties of the code. For example, how many units are active on average in representing a particular stimulus or state? In a code where only a single unit is active for any given stimulus or state, it is very easy to *read* the representation in the brain, but the number of states that can be represented is limited to the number of neurons. On the other hand if each possible stimulus or state is represented by a large number of active neurons, this allows the possibility for an enormous number of stimuli or internal states to be represented. That said, a result of this dense code is that states are hard to learn and hard to read, with a significant possibility of interference or producing an incorrect interpretation of the brain state by virtue of the fact that many neurons are shared between different representations. Foldiak provides a nice summary of the properties of these various types of coding schemes in [71] and these are depicted in table 6.1.

Immediately evident in glancing at table 6.1, is that a sparse coding scheme has some nice tradeoffs between local and dense coding schemes. The human brain

Property \ Coding Scheme	Local	Sparse	Dense
Representational Capacity	very low	high	very high
Memory Capacity	limited	high	low
Speed of Learning	very fast	fast	slow
Generalization	none	good	good
Interference	none	controlled	strong
Fault Tolerance	none	high	very high
Simultaneous Items	unlimited	several	one

Table 6.1: A summary of the properties of various coding schemes. From [71].

requires a representation that includes a reasonable capacity for representation, while being robust to interference and at the same time retaining the ability to learn quickly, in some instances based on only a small number of exemplars.

For a dense code, the representational capacity is very high (2^N in the binary case). This is not especially useful since the representational capacity of any practical system will never approach the requirement of this many states. It is sufficient to use an encoding in which a handful of units is active in representing a particular state. This allows sufficient representational capacity while maintaining many of the advantages that a sparse code affords. It is not worth considering all of these properties exhaustively, but it is worth stating that sparse codes may allow suffi-

cient representational capacity for visual representation, are fast to learn, and are reasonably tolerant to errors while maintaining strong generalization.

It is interesting to consider the precise criterion for the tradeoff between local and dense codes. A large part of the gain from a sparse code comes from the fact that representation takes advantage of the statistical properties of the patterns under consideration. This was a sentiment expressed by Barlow in an important piece of work in which he expressed the need for *suspicious coincidences* or *sensory cliches* to be observed in order to allow efficient representation of content [11]. For example, since there exists a great deal of redundancy in the representation that appears at the level of the retinal input, it makes sense to transform this data in such a way that the redundancy in the resultant representation is minimized. This is the reason why the visual system is very capable at distinguishing between seemingly complex stimuli with subtle differences such as faces, relative to simpler patterns such as random dots.

6.2 Physiological Evidence for Sparse Coding

With respect to efficient coding in the visual cortex, this is an area that has received a great deal of attention in the past decade. In early work, Linsker demonstrated that according to the infomax principle a layered network with Hebbian learning across layers may result in self-organization into opponent-type [128], and oriented

filters [129], as well as organization into cortical columns [130].

The mid 1990's introduced some additional important computational work, further demonstrating an explicit link between principles such as information transfer, sparsity and neural encoding. Olshausen and Field [69] and Bell and Sejnowski [13] published important work at this time demonstrating that learning an efficient code based on minimizing redundancy, or maximizing independence gives rise to a representation with properties very similar to neurons appearing in V1. That is, the application of an approach such as independent component analysis to a large set of local image patches gives rise to a basis that resembles a bank of gabor-like filters. It is worth mentioning that in some cases, the algorithms in question may be seen as an efficient means of learning an infomax relationship in accord with Linsker's proposal. [131]

Since this time, further efforts based on sparse coding have yielded models of spatiotemporal cells in visual cortex [233], color-opponent cells [121; 212], disparity selective cells [165] and also complex V1 cells [95] and cells coding for contours [91] as well as simple V2 type cells [94]. All of this lends credibility to the claim that the visual cortex appears to implement a sparse code of natural stimuli among early visual areas and suggests that this sparsity may only deepen as one ascends to higher visual areas.

On the side of human data, consideration of the extent to which sparse coding

appears in the cortex has proven nontrivial. One obstacle lies in the difficulty of physically recording across an ensemble of neurons over which sparseness might be measured.

The most obvious observation that may be cited as evidence in favor of sparse coding is that it is very difficult to find an effective stimulus for neurons appearing in even intermediate visual areas and exceedingly difficult for areas such as IT. This strongly suggests a sparse code in itself since it implies that only a small number of units is active for any given stimulus as the sparseness across stimuli is equivalent to the narrowness of tuning [71]. Even as early as V1 many units respond only to a localized edge or grating and only if the stimulus has a specific orientation, spatial frequency, direction of movement, or stereo disparity. Even at this stage such a configuration is indicative of the fact that only a small number of units will be active for a particular stimulus. In some interesting work by Vinje and Gallant [237; 238] it was demonstrated that V1 neuron responses were sparse and additionally stimulation of the non-classical receptive field resulted in decreased correlation in neighboring neurons indicating whitening of the responses.

Units in higher areas such as IT often respond only to specific complex geometrical patterns and are difficult to describe in terms of simple properties such as color, orientation or motion. Baddeley et al. demonstrated sparse responses within IT cells on par with the activity ratio observed in V1 [10].

Arguments favoring sparsity also appeal to notions of energy use. For example, it has been argued that the biochemical energy available for action potentials limits the average firing rate of neurons to an amount less than 1 Hz [9], and furthermore it has been argued that only 2 percent of any population of cortical neurons can exhibit high firing rates on the basis of energy related arguments [122]. In addition to the human visual system, there are a variety of examples in nature where sparse coding has been observed. Examples include sparse coding in auditory neurons in rats [53], olfactory neurons in insects [177], somatosensory neurons in rats [21] and recordings from rat hippocampus [214]. Sparsity has also been observed in prefrontal cortex in rhesus monkeys [1]. Sparsity is also observed among motor neurons. Examples of this include include rabbit [14] and rat [22] motor cortices. In the case of rats it was demonstrated that stimulating a single neuron was sufficient to cause movement of a whisker.

An additional aspect of coding is that while principles such as minimizing redundancy tend to produce a sparse code, sparsity in itself does not imply meaningful features. That is, it might be possible to construct a random code with the same sparsity as one designed according to some optimality criterion. The additional advantage that comes from the latter scheme, is that units correspond to real phenomena in the statistics and thus have intrinsic meaning. This allows consideration of similarity among codes as well as the kind of similarity present. This is an im-

portant consideration in a system in which simple primitive features are gradually combined to form elements that code for more complex patterns in a possibly even more sparse encoding. As a whole the evidence is suggestive of sparse coding as ubiquitous in the visual cortex, and with units that represent meaningful elements of visual content.

6.3 Sparse Coding and Computational Complexity

It is worth commenting on complexity as it pertains to coding in the context under consideration. We are concerned specifically with how the brain performs inference based on quantities that rely on probability density functions. The following demonstrates how the relationship between the data required for a probability density estimate and the dimensionality of the space is an exponential one. This implies that a sparse representation is of utility for a system that scales up to realistic problem sizes. Consider first the case of estimating a one dimensional probability density function. Let us assume that the distribution is adequately described by k points. Now consider the same estimate on a set of 2 variables that are pairwise dependent. To achieve a similar covering of the PDF, one requires k^2 points and in general, a D dimensional space will require k^D points where each point may be assumed to correspond to a training sample or observation that informs on the PDF. It is easy to appreciate how quickly this quantity grows. For example, let us

consider the case mentioned in chapter 4. For a $4x4$ RGB image patch ($D = 48$), with $k = 10^2$ one requires 10^{96} samples to achieve the same covering of the space as a 1D distribution. To put this number into perspective, estimates of the number of particles in the universe are on the order of 10^{80} and in order to achieve a covering of the desired precision for a 4x4 image patch one would have to experience 4.8 x 10^{87} training samples every second of their life assuming a lifespan of 66 years. The reason this is not a problem in practice, is that the entire (in this case 48 dimensional) space is not interesting. In fact, only a small proportion of this space corresponds to the sort of observations that exist in the real world. For this reason it makes sense to i. encode content in a manner that exploits this property and ii. have some degree of sparsity built in to the system since even for relatively small degrees of dependencies between features, the training required becomes unfeasible.

6.4 Density Estimation in an Ensemble of Neurons

This section introduces a simple neural circuit that is equivalent to a non-parametric density estimate based on any kernel function. The relationship between the resulting circuitry and neural circuitry is considered. In particular, the structure of the circuit is shown to be very similar to some well-established neural circuits, and some additional requirements (e.g. long range lateral excitation) are also consistent with its design. A corollary of the result is a circuit for entropy estimation.

It is worth noting that similar computation might be achieved in the context of a population coding scheme (e.g. [75; 102]) but that the proposed formulation has the advantages of similarity to simple cortical circuits and additionally does not require memory of the probability density distribution, but instead computes only likelihoods of observed firing rates locally and based on the current state of neurons involved in the estimate.

In this formulation, we demonstrate how a sparse representation lends itself to density estimation based on simple circuitry acting on an ensemble of neurons. The importance of such a circuit cannot be understated. The assumption that the brain is capable of building probability densities is fundamental to many schools of thought concerning how the brain functions (e.g. perception as Bayesian inference). The following demonstrates how a sparse representation allows the estimate of the likelihood of the firing rate associated with any given neuron to an arbitrary precision, and without the requirement of an explicit representation of the underlying probability density function.

Definition Let N denote an arbitrary ensemble of n neurons in the brain with the condition that N contains n distinct neurons that code for all types of features at a specific position in space-time.

Definition Let Ω denote the set of neurons that comprise the *context* of N. That is Ω codes for the same features as N but in the surrounding context. Returning

to the visual analogue, one might consider Ω to be a set of adjacent hypercolumns surrounding N. An additional possibility would be a non-spatial context for Ω. For example, one might consider Ω to describe previous exemplars for the elements of N drawn from a specific context. Experience with a tropical bird might result in a variety of exemplars in which auditory neurons that code for high temporal frequency are active, while those coding for low frequency signals remain relatively inactive. One might also consider the possibility that the elements of Ω comprise a *summary* of many exemplars stored in associative memory. The generality of Ω is unlimited with the sole constraints being that it describes a context in which N resides, and the $\Omega_k \in \Omega$ each contains the same n elements as N.

Definition Let Δ be a function that minimizes mutual dependence among the various $N_i \in N$ and $\Omega_{k,i} \in \Omega_k$ such that the assumption of independence between N_i and N_j and between $\Omega_{k,i}$ and $\Omega_{k,j}$ is valid $\forall i \neq j$ within some finite error limit. One might liken this to the operation that a hypercolumn in the visual cortex performs on incoming sensory input if searching for a neural analogue. It is worth noting that this transformation is not a mathematical requirement for the formulation that follows, but is necessary to overcome the combinatorial explosion of connections associated with representations of higher dimensionality.

Definition Let us denote $\Delta(N)$ as N' and $\Delta(\Omega)$ as Ω'. To avoid any confusion based on notation, it should be noted that $\Delta(N_i) \neq N_i'$ but rather each N_i' is a

mixture of all elements of N and $\Omega'_{k,i}$ is a mixture of all Ω_k.

Lemma 6.4.1 *For a non-parametric density estimate based on a symmetric kernel K, the contribution of $\Omega'_{k,i}$ to the estimate of N'_i is equivalent to the contribution of N'_i to the estimate of $\Omega'_{k,i}$.*

Proof A kernel density estimate in this case is a function of the difference in firing rate between N'_i and $\Omega'_{k,i}$. Let γ equal this distance. Since K is a symmetric kernel, $K(\gamma) \equiv K(-\gamma)$. ∎

Lemma 6.4.2 *The likelihood associated with the firing rate of N_i based on a kernel density estimate with symmetric kernel K is given by $\sum_{k=1}^{n} K(N'_i - \Omega'_{k,i})$.*

Proof Assuming the N_i are mutually independent, the contribution of $\Omega_{k,j}$ to the likelihood estimate of N_i is 0 $\forall j \neq i$. Then a kernel density estimate corresponding to the firing rate associated with N_i is given by $K(\Omega'_{1,i} - N'_i) + K(\Omega'_{2,i} - N'_i) + \ldots + K(\Omega'_{n,i} - N'_i)$. By Lemma 6.4.1 this expression can be written as $\sum_{k=1}^{n} K(N'_i - \Omega'_{k,i})$.

∎

The last step in the proof of lemma 6.4.2 may seem unnecessary, but facilitates translation to a neural circuit at a later stage.

Theorem 6.4.3

$$p(N) = \prod_{i=1}^{n} (\sum_{k=1}^{n} K(N'_i - \Omega'_{k,i})$$

Proof As Δ merely transforms N without modifying its constituent elements, $p(N) = p(N')$. Since $N' = (N'_1, N'_2, ..., N'_n)$, $(p(N') = p((N'_1, N'_2, ..., N'_n))$. Owing to the independence assumption on the N_k this expression may be rewritten as $\prod_{k=1}^{n}(p(N_k))$. Lemma 6.4.2 then gives $\prod_{i=1}^{n}(\sum_{k=1}^{n} K(N'_i - \Omega'_{k,i}))$ ∎

Based on the form given in this equation, it is easy to see possibilities for the layout of a neural circuit that evaluates $p(N)$. Recall that this is a necessary step in demonstrating the neural plausibility of the model proposed in chapter 4. The resultant formulation captures the likelihood of firing rates within some neuronal ensemble localized in time and/or space. Note that this allows an estimate to an arbitrary level of precision of the firing rate associated with any neuron, without requiring any explicit representation of the underlying probability density function.

6.5 An Example

To make the previous discussion more explicit, it may be instructive to provide an example of how the preceding applies to a real scenario.

Let N be a local neighborhood of an image, comprised of a number of RGB values on an nxn patch.

In the following formulation, we assume an estimate of the likelihood of the components of N based on a Gaussian kernel density estimate. Any other choice

of symmetric kernel may be substituted, with a Gaussian window chosen only for its common use in density estimation and without loss of generality.

An independent representation of N may be achieved through an algorithm that seeks such a transformation. In this example, we have employed the algorithm of Lee et al [119] which performs independent component analysis. The basis set is derived from a large number of natural image patches chosen at random from images in the Corel stock photo data set.

Let $N_{i,j,k}$ denote the set of independent coefficients based on the neighborhood centered at j, k (i.e. $\Delta(N')$ note the reversal in notation). That is N consists of a set of coefficients corresponding to the relative contribution of various oriented gabor-like filters and red-green or blue-yellow color opponency at various spatial scales. Note that in this example Ω is given by the set of all $N_{i,s,t}$ $\forall i$ and $\forall s \neq j$ or $t \neq k$ An estimate of $p(N_{i,j,k})$ based on a Gaussian kernel is given by:

$$\frac{1}{\sigma\sqrt{2\pi}} \sum_{\forall s,t \in \Omega} \Psi(s,t) e^{-(N_{i,j,k} - N_{i,s,t})^2 / 2\sigma^2} \tag{6.1}$$

with $\sum_{s,t} \Psi(s,t) = 1$. $\Psi(s,t)$ describes the degree to which the coefficient N' at coordinates s, t contributes to the probability density estimate. The inclusion of this parameter reflects the fact that neurons may have a greater impact on the likelihood estimate of neurons they are close to. The inclusion of such a parameter may be appropriate in some instances and less so in others. On the basis of the form given in equation 6.1 it is evident that this operation corresponds to the neural circuit

depicted in figure 6.1. Note that in this case a logarithmic nonlinearity follows the likelihood estimate yielding a measure of self-information. Figure 6.1 demonstrates only coefficients derived from a horizontal cross-section. The two dimensional case is analogous with parameters varying in i, j, and k dimensions. K consists of the Kernel function employed for density estimation. In our case this is a Gaussian of the form $\frac{1}{\sigma\sqrt{2\pi}}e^{-x^2/2\sigma^2}$. $\Psi(s,t)$ is encoded based on the weight of connections to K. As $x = N_{i,j,k} - N_{i,s,t}$ the output of this operation encodes the impact of the Kernel function with mean $N_{i,s,t}$ on the value of $p(N_{i,j,k})$. Coefficients at the input layer correspond to coefficients of N. The logarithmic operator at the final stage might also be placed before the product on each incoming connection, with the product then becoming a summation. The similarity between independent components and V1 cells, in conjunction with the plausible circuitry proposed here lends credibility to the claim that *information* may contribute to driving overt attentional selection.

It is interesting to note that the structure of this circuit at the level of within feature spatial competition is remarkably similar to the standard feedforward model of lateral inhibition, a ubiquitous operation along the visual pathways thought to play a role in attentional processing [30]. This consideration is revisited in the section that follows, demonstrating an explicit relationship to cortical surround suppression.

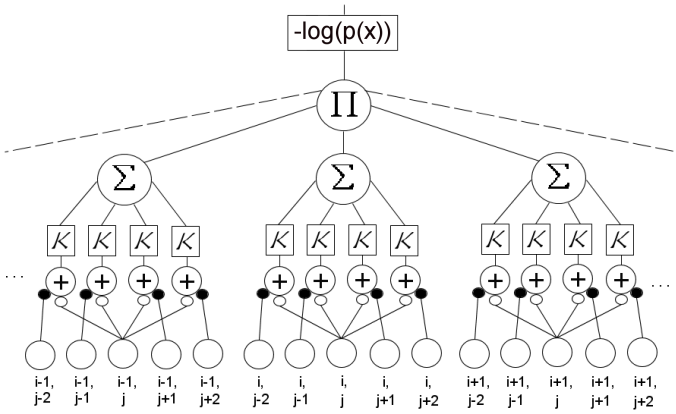

Figure 6.1: A 1D depiction of the neural architecture that computes the self-information of a set of local statistics. The operation is equivalent to a Kernel density estimate. Coefficients correspond to subscripts of $N'_{i,j,k}$. The small black circles indicate an inhibitory relationship and the small white circles an excitatory relationship. κ indicates a symmetric Kernel function.

With regard to the neural circuitry involved, we have demonstrated that self-information may be computed using a neural circuit in the absence of a representation of the entire probability distribution and an updated version of the overall model is shown in figure 6.2. Since entropy quantifies average self-information over some domain, a corollary of this result is a circuit that computes entropy, simply through the summation of many units that quantify self-information over some domain (e.g. the surround).

6.6 Surround Suppression, Gain Control and Redundancy

Perhaps the foremost consideration pertaining to neural circuitry, is the extent to which the proposal agrees with observations concerning cortical circuitry and neurophysiology. To this end, this section reviews a variety of classic and recent results derived from psychophysics and imaging experiments on the nature of surround suppression within the cortex. Necessary conditions on an architecture that seeks to maximize information in selection, are weighed against the experimental literature to determine a possible neural analogue for the implementation of AIM. As a whole, the discussion establishes that a variety of peculiar and very specific constraints imposed by the implementation show considerable agreement with the computation implicated in surround suppression further providing support for AIM, and also offering some insight on the nature of computation responsible

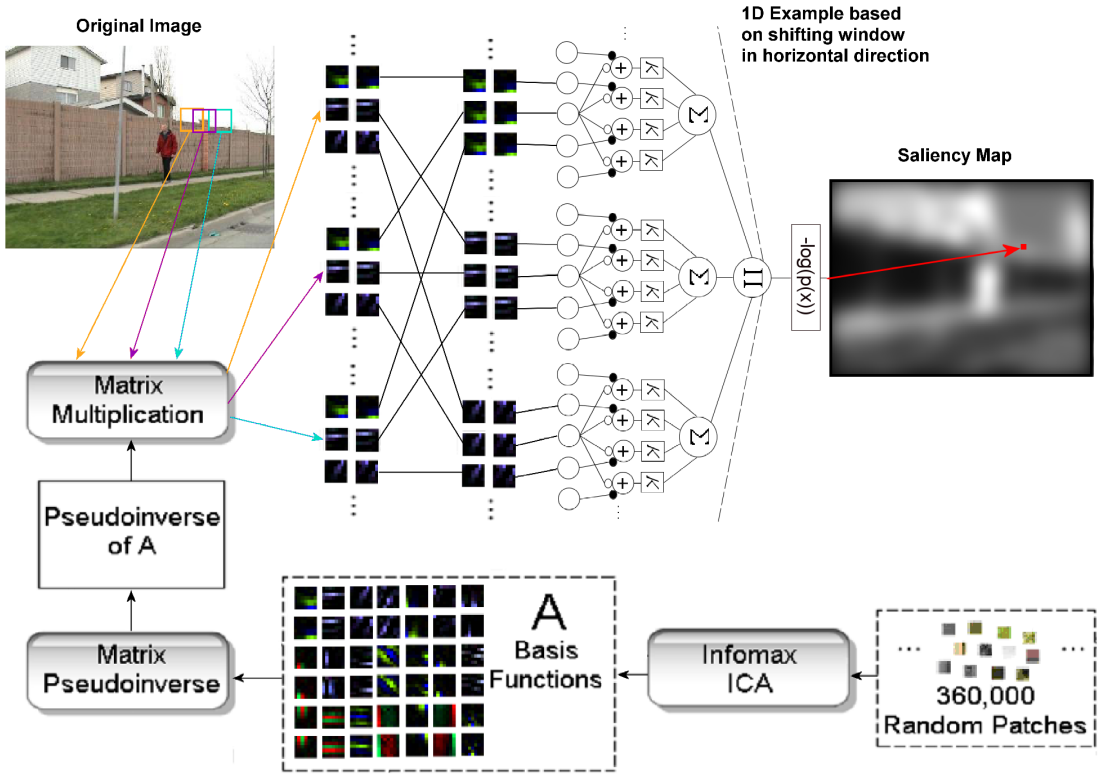

Figure 6.2: The proposed model. Shown is the computation corresponding to three horizontally adjacent neighbourhoods with flow through the network indicated by the orange, purple, and cyan windows and connections. The connections shown facilitate computation of the information measure corresponding to the pixel centered in the purple window. The network architecture produces this measure on the basis of evaluating the probability of these coefficients with consideration to the values of such coefficients in neighbouring regions.

for iso-orientation surround suppression in early visual cortex. Debate concerning the specific nature and form of surround suppression has rekindled in recent years, which has resulted in a large body of interesting results that further elucidate the details of this process. The following discussion reviews these results and offers further insight through a meta-analysis of recent studies. In each case, experimental findings are contrasted against the computational constraints on AIM to establish plausibility of the proposed computation.

6.6.1 Types of features

A great deal of research has focused specifically on the suppression that arises from introducing a stimulus in the surround of a localized oriented Gabor target. The specific nature of iso-orientation (iso-feature) surround suppression as dictated by the details of AIM includes two key considerations: 1. Suppression of a cell whose receptive field lies at the target location should occur only for a surround stimulus that is the effective stimulus for this cell. For example, for a vertically oriented Gabor target, suppression of a cell that elicits a response to the target will occur only by way of a similar stimulus appearing in the surround. Recall that a fundamental assumption is that the responses of different types of cells at a given location are such that the correlation between their responses is minimal and this is a phenomenon that is observed cortically. In the domain of studies pertaining to

surround suppression, the literature is undivided in its agreement with this assumption. When considering the cell response or psychometric threshold associated with a target patch, suppression from a surround stimulus is highly stimulus specific and is at a maximum for a surround matching the target orientation, with suppression observed only for a narrow orientation band centered around the target orientation [178; 207; 247; 253; 255; 257]. This is consistent with a local likelihood estimate in which the independence assumption is implicit. 2. Suppression should be observed for all feature types, and the nature of, and parameters associated with suppression should not differ across feature type. This is an important consideration since studies of this type have largely focused on oriented sinusoidal stimuli, but nevertheless similar suppression associated with color, or velocity of motion for example, should also be observed and the nature of such suppression should be consistent with that observed in studies involving oriented sinusoidal target and surrounds. One recent effort provides strong evidence that this is the case through single cell recording on macaque monkeys [207]. Shen et al. demonstrate that centre-surround fields defined by a variety of features including color, velocity and oriented gratings all elicit suppression and with suppression at a maximum for matching centre and surround stimuli.

6.6.2 Relative contrast

Given a cell with firing rate $N_{i,j}$ that codes for a specific quantity at coordinates i,j in the visual field (e.g. a cell selective for a specific angular and radial frequency as part of a basis representation with its centre at location i,j), a density estimate on the observation likelihood of the firing rate associated with $N_{i,j}$ as discussed earlier in the chapter is given by:

$$p(N_{i,j}) = \sum_{\forall s,t \in \Omega} f(N_{i,j} - N_{s,t}) \tag{6.2}$$

Where f is a monotonic symmetric kernel with its maximum at $f(0)$ and Ω the region over which the surround has any significant impact. For further ease of exposition in observing the behavior of equation 6.2, assume without loss of generality that f comprises a Gaussian kernel. Then equation 6.2 becomes:

$$\frac{1}{\sigma\sqrt{2\pi}} \sum_{\forall s,t \in \Omega} e^{-(N_{j,k} - N_{s,t})^2 / 2\sigma^2} \tag{6.3}$$

As there also exists a spatial component to this estimate, it may be more appropriate to also include a parameter that reflects the effect of distance on the contribution of any given cell to the estimate of $N_{i,j}$ which might appear as follows:

$$\frac{1}{\sigma\sqrt{2\pi}} \sum_{\forall s,t \in \Omega} \Psi(s,t) e^{-(N_{j,k} - N_{s,t})^2 / 2\sigma^2} \tag{6.4}$$

Once again Ψ drops off according to the distance of any given cell from the target location, reflecting the decreasing correlation between responses. Assuming that

surround suppression is the basis for the computation involved in AIM, equation 6.4 demands a very specific form for the suppressive influence of a surrounding stimulus on the target item. According to the form of equation 6.4, suppression depends on the relative response of centre and surround stimuli and should be at a maximum for equal contrast centre and surround stimuli: Raising or lowering the contrast of a stimulus pattern will generally result in a concomitant increase in the response of a cell for which the pattern in question is the effective stimulus. There is therefore a direct monotonic (nonlinear) relationship between the firing rate attributed to centre or surround, and their respective contrasts. Support for suppression as a function of relative centre versus surround contrast is ubiquitous in the literature [3; 30; 171; 207; 252; 255; 256; 257] although there is as of yet no consensus on why this should be the specific form for the suppressive influence of a surround stimulus. There also exists a large body of prominent studies revealing that this suppression is indeed at a maximum for equal contrast centre and surround stimuli [3; 171; 207; 255; 257]. Note that this implies mathematical equivalence between surround suppression and a likelihood estimate on a given cell's response as defined by the response of neighboring cells and implies divisive modulation of a cells response by a function of its likelihood. This is an important consideration as it offers insight on the role of surround suppression which has recently become an issue of considerable dispute [178] and implicates surround suppression as the

machinery underlying the implementation of AIM. It is also worth noting that the suppressive impact of cells in the surround is observed to drop off exponentially with distance from the target giving the specific form of Ψ [178].

6.6.3 Spatial configuration

For the sake of exposition, let us assume that the computation under discussion is restricted to V1. From the perspective of efficient coding, no knowledge of structure is available at V1 beyond that which lies within a region the size of single V1 receptive field. A pure information theoretic interpretation of the surprise associated with a local observation as determined at the level of V1 should reflect this implying an isotropic contribution to any likelihood estimate in the vicinity of the target cell, regardless of the pattern that forms an effective stimulus for the cell in question. That is, for a unit whose effective stimulus is a horizontal Gabor pattern, equidistant patterns of the same type in the vicinity of the target should result in equal suppression regardless of where they appear with respect to the target and this is reflected in the implementation put forth in [25]. It is also expected that likelihoods associated with higher order structure over larger receptive fields are mediated by higher visual areas either implicitly at the single cell level or explicitly via recurrent connections. In line with the assumption that computation is on the observation likelihood of a pattern within a given region, and that structures are

limited to an aperture no larger than a V1 receptive field, it is indeed the case that suppression from the surround is isotropic with respect to the location of a pattern appearing in the surround independent of target and surround orientations [178]. By virtue of the same consideration, one would also expect the spatial extent of surround suppression to be invariant to the spatial frequency of a target item. This is also a consideration that is evident in the literature [178]. In consideration of observation likelihoods associated with more complex patterns, it is interesting to consider the nature of surround suppression among higher visual areas. Recent studies are discovering more and more examples of suppressive surround inhibition among higher visual areas with the same properties and divisive influence as those that are well established in V1. Extrastriate surround inhibition of this form has been observed at least among areas V2 [18; 60; 92; 113; 203; 208; 211; 258]. This is suggestive of the possibility that saliency is represented within a distributed hierarchy, with local saliency computation mediated by surround suppression at various layers of the visual cortex. This issue is revisited in chapter 8 and weighed against more general proposals concerning how attention is achieved within the primate brain.

6.6.4 Fovea versus Periphery

If the role of local surround suppression is in attenuating neural activation associated with unimportant visual input and/or redirecting the eyes via fixational eye movements one would expect the influence of such a mechanism to be prominent within the periphery of the visual field. Petrov and McKee demonstrated that surround suppression is in fact strong in the periphery and absent in the fovea [178]. This is consistent, as Petrov and McKee point out, with a role of this mechanism in the control of saccadic eye movements. Furthermore, there are additional points they highlight that support this possibility, including the fact that the extent of suppression is invariant to stimulus spatial frequency. Also of note, is the fact that the inaccuracy of a first saccade is proportional to target eccentricity and this correlates with the extent of surround suppression as a function of eccentricity [178]. Note that the cortical region over which surround suppression is observed does not vary with eccentricity implying that computationally, an equal number of neurons contribute to any given likelihood estimate of the form appearing in equation 6.4. All of these considerations are in line with a role of this mechanism in the deployment of saccades.

6.6.5 Summary

We have put forth the proposal that the implementation of AIM is achieved via local surround circuitry throughout the visual cortex. As a whole, there appears to be considerable agreement with the proposal and the specific form of surround suppression. The demonstration of equivalence of a likelihood estimate on the surround of a cell with the apparent form of suppressive inhibition implies modulation of cell responses at a single cell level through divisive gain as a function of the likelihood associated with that cell's response. This provides a more specific explanation for the nature of computation appearing in suppressive surround circuitry and further bolsters the claim that saliency computation proceeds according to a strategy of optimizing information transmission.

6.7 Ω and the Gist of a Scene

With respect to the context of a local neighbourhood, there is an additional school of thought concerning visual analysis that is worth commenting on. In the example given, the likelihood of N can be computed based on a context of cells that are spatially proximal to N. Although there exist long range lateral connections in the visual cortex, it is unclear whether the extent of such connections are sufficient to explain behavior in certain psychophysics paradigms. In light of this consideration,

we might consider a possibility that falls on the opposite extreme. Consider the possibility that the visual system is able to form a general *at a glance* model of the statistics of the scene. That is, the assumption that a preattentive representation, albeit coarse, of the spatiochromatic profile and spatial frequency content of the scene is available at a glance.

This is in fact a view that has become popular in recent years. Oliva et al. have conducted a variety of work concerning what they call the *gist* of a scene (See [167; 169]). This work is motivated by the observation that although the visual system encodes visual content on the basis of localized receptive fields that respond to specific stimulus patterns, some global properties of the scene may be observed at a glance. Further, among this work it is established that global estimates of scene properties are sufficient to describe context [218]. Among early information that is available, is properties of spatial layout (such as openness, expansion, mean depth) [168], surface properties (colors, textures and materials) and possibly functional properties such as those of use for navigation or camouflage [77]. In discrimination tasks between scene categories (desert vs. forest vs. waterfall etc.), most judgements on the context could be made within 35 ms and judgements on spatial layout and surface properties even more quickly (mostly on the order of 25 ms).

One interesting study that relates to this idea is that of Chong and Treisman

[36]. In this work, subjects were presented with a variety of circles of varying size, density and grouping. It was determined that some basic statistical properties could be determined reliably and quickly over the display. Changes in frequency of sizes or groupings did not have any significant effect on accurate estimation of mean size. When circles were colored, average sizes based on color were also estimated to a suprising degree of accuracy. This suggest that binding of color and size information that one might expect for this task is unnecessary, but rather possibly preattentive color based segmentation allows global statistics to be considered over a subset of the display. All of this supports the notion that statistical estimates of certain properties are computed in parallel over the display.

It is clear how the discussion of *gist* relates to discussion of the proposed model. In lieu of local circuitry that computes a density estimate over some local surround, the likelihood estimate might be based on a more general representation of the gist of the scene. That is, an Ω that quantifies basic properties of the scene in order to determine what is informative.

Of course a representation that is entirely global is lacking with respect to certain behaviors one might predict. For example, consider figure 6.3. In figure 6.3 there is a tendency to fixate the green and red patches with surround of the opposite color. That said, a global representation of the chromatic profile would be insufficient to predict this result. For this reason, some locality is required.

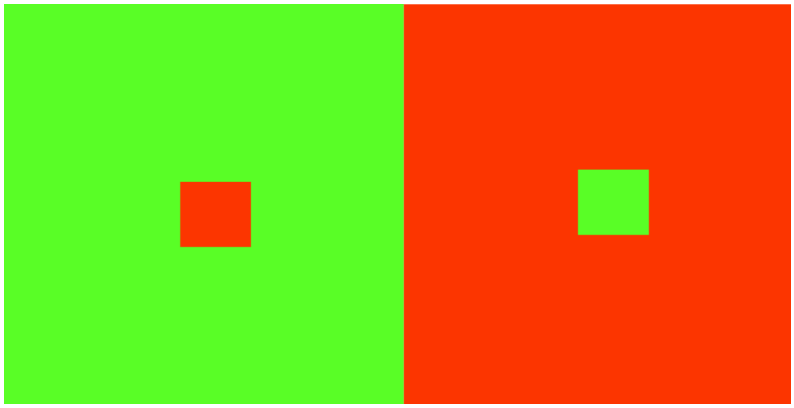

Figure 6.3: An example of why some locality is required. While the number of red and green pixels is equal, the local arrangement of such elements is important.

Whether this is achieved via more global detectors that represent basic content on larger neighborhoods while maintaining some locality, or through local circuitry in a visual hierarchy remains to be determined. An additional possibility with respect to gist is that it serves to provide an early determination of context so that visual units can be adapted to processing said context. One might view this possibility as selection through localized circuits but with a global tuning of thresholds on the basis of the gist.

6.8 Discussion

With respect to the claim of a model for which the basis of selection is that of maximizing a criterion based on information as signals flow through the visual

hierarchy, the discussion of this chapter highlights some interesting points. The self-information model began with a simple definition for what comprises salient content, and the resulting mathematical framework ends up with an encoding of local content based on a representation remarkably similar to V1. A close look at density estimation yields circuitry with properties very similar to circuits that implement a ubiquitous operation in visual processing. Additionally, physical requirements of the circuitry appear to have a neuroanatomical analogue.

7 Psychophysics and New Insights

7.1 Eye movements and Attention

A consideration that renders difficult any effort towards modeling attention, is that there is as of yet no simple means of monitoring the position and scope of the focus of covert attention. For this reason, many evaluations of attention models are carried out in observing fixational eye movements under the assumption that saccadic eye movements serve as an index for the focus of attention. This is an area that is rife with controversy.

It is worth briefly reviewing current thinking associated with this topic in order to establish the generality of claims that may be made based on eye tracking data and also to consider alternatives in establishing plausibility. In recent years, several researchers have proposed that fixational eye movements are related to shifts in the covert focus of attention. Most of these efforts posit that covert shifts in attention may be tracked through the observation of microsaccades [62; 80; 118]. The impetus for such claims derives from the premotor theory which claims that

shifts in attention are accompanied by a cancelled saccade plan [206]. This claim finds support in the work of Engbert and Kliegl who observed that microsaccades recorded in a cueing task with a central fixation tended to be biased in the cued direction [62]. That said, there are also a handful of studies in which no such correlation was observed [74; 80; 117; 196]. Hafed and Clark posit that this may be explained by a combination of microsaccades corresponding to shifts in attention and corrective saccades towards the point of fixation [80]. It has also been noted that the time course associated with saccades away from a cued location is consistent with inhibition of return [74]. An additional result that confuses matters demonstrates that the abrupt onset of a stimulus shows no correlation to the spatial orienting of microsaccades. A recent effort of Horowitz et al. [90] attempts to bring some lucidity to the discussion through an experiment involving cueing alongside a target presentation requiring a manual response. The conclusion from this study is that although many pathways are shared by fixation and the orienting of attention, that the two are functionally separate. A commentary on this work however reveals that there does exist correlation between microsaccades and attention albeit this relationship is weak. Clearly the existing body of evidence on this issue leaves much to be desired. As an alternative means of assessing model plausibility, it is perhaps worth considering predictions of AIM for visual search tasks requiring shifts in the covert focus of attention with a central fixation location and

comparing the behavior from such studies with model predictions. It is with this consideration in mind that we revisit a large body of classic psychophysics results many of which form the motivation for existing models to observe the extent to which behavior is consistent with the behavior of AIM.

7.2 Attention and Visual Search

The study of visual search has been influential in shaping the current understanding of computation related to attention and the determination of visual saliency. Owing to the large body of psychophysics work within this area, in addition to some of the peculiarities that are observed within the visual search paradigm, it is natural to consider how model predictions measure up against the wealth of psychophysics results in this area. It is with this in mind that we revisit a variety of classic results derived from the psychophysics literature revealing that AIM exhibits considerable explanatory power and offers some new insight on certain problem domains. Generally, models of attention assume that the focus of attention is directed according to a competitive Winner-take-all process acting on some neural representation in the cortex. An important element of this representation is the saliency of a target item relative to the saliency of the distractors since this is the determinant of search efficiency according to various selection mechanisms [51; 98; 110; 228]. It is assumed then throughout the discussion, that search efficiency is a function of

the ratio of target to distractor saliency in line with other similar efforts [124]. This assumption allows the consideration of saliency to be disentangled from the mechanisms that underlie attentional gating which remains a contentious issue.

7.3 Serial versus Parallel Search

An observation that has been influential in earlier models of attention, is that certain stimuli seem to be found effortlessly from within a display, while others require considerable effort to be spotted seemingly requiring elements of the display to be visited in turn. Consider for example figure 7.1. In the top left, the singleton item distinguished by its orientation is found with little effort seemingly drawing attention automatically. This phenomenon is sometimes referred to as "pop-out". Pop-out results in the immediate and automatic deployment of attention to items defined by feature contrast associated with many basic features including orientation, color and motion [161]. The same may be said of the singleton defined by color in the top-middle frame; however, the singleton in the top right frame requires examining the elements of the frame in turn to locate the target. These observations form the motivation for Treisman's Feature Integration Theory [222], a seminal work in attention modeling based on the observation that some targets are found effortlessly and seemingly in parallel while others seem to require a serial search of target items with the search time increasing as a linear function of the number of

distracting elements. In particular, the distinction between these two cases is when a target item is defined by a conjunction of features rather than a single feature. On the bottom row of figure 7.1 is the output of AIM with the saliency scale shown on the left hand side. Warmer colors are more salient, and this scale is used in all examples scaled between the maximum and minimum saliency values across all examples within an experiment. As can be seen in figure 7.1 the target relative to distractor saliency is very high for the first two cases, but the target saliency is indistinguishable from that of the distractors in the third case, suggesting no guidance towards the target item and hence requiring a visit of items in serial order. Thus, the distinction between a serial and parallel search is an emergent property of assuming a sparse representation, and saliency based on information maximization. Since the learned feature dimensions are mutually independent, the likelihood is computed independently for uncorrelated feature domains implying unlikely stimuli for singletons based on a single feature dimension, but equal likelihood in the case of a target defined by a conjunction. This behavior seen through the eyes of AIM is then a property of a system that seeks to model redundancy in natural visual content and overcome the computational complexity of probability density estimation in doing so. To make clear the reason for this behavior, consider figure 7.2 which shows a probability density representation of the response of a small number of hypothetical cells (idealized examples for the purpose of exposition) to

the stimuli appearing in figure 7.1. For the case shown in figure 7.1 (top left), a large number of units respond to the stimuli oriented 15 degrees from vertical, and only a small number to the bar 15 degrees from horizontal. On the basis of this, the likelihood of the response associated with the singleton is lower and thus it is more informative. Since an approximately equal number of units respond to both green and red stimuli, this stimulus dimension dictates that all of the stimuli are equally informative. The situation for the stimulus shown in figure 7.1 (top middle) is analogous except that color is the discriminating dimension and orientation dictates all stimuli are equally salient. In the case of figure 7.1 (top right), there is a singleton element, but the number of units responding to all four cell types is approximately equal and as such, a serial search of the elements is required.

An additional example of a conjunction search is featured in figure 7.3: The 5's that are small, rotated and red are easily spotted, but finding the 2 requires further effort. It is worth noting that this account of visual search has been revised to some extent with more recent experiments demonstrating an entire continuum of search slopes ranging from very inefficient to very efficient [249]. This is a consideration that is also supported by AIM as more complex stimuli that give rise to a distributed representation may yield very different ratios of target versus distractor saliency.

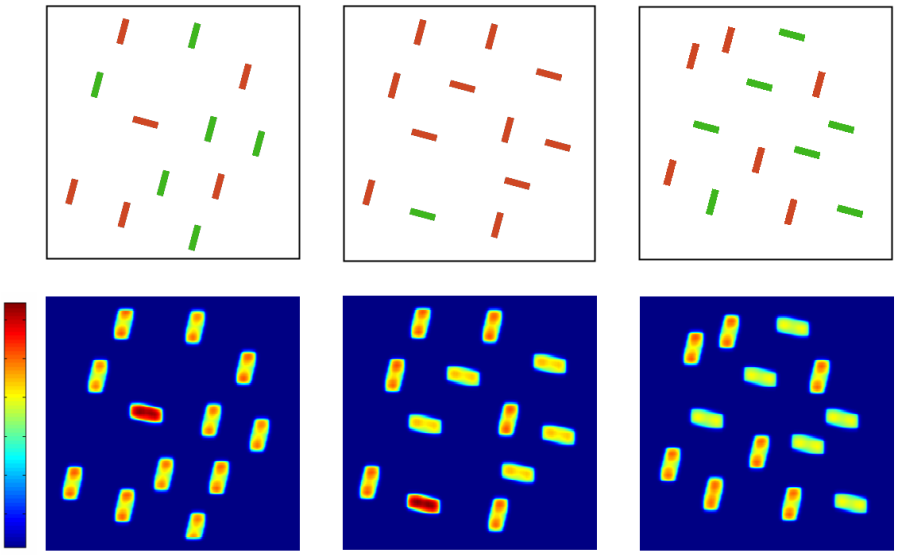

Figure 7.1: Three stimulus examples wherein a singleton element is present. In the top left case, defined by orientation, top middle by color and top right by a combination of the two. Associated saliency appears in the corresponding maps on the bottom. This result mimics the classic serial-parallel dichotomy that forms the basis for some classic attention models.

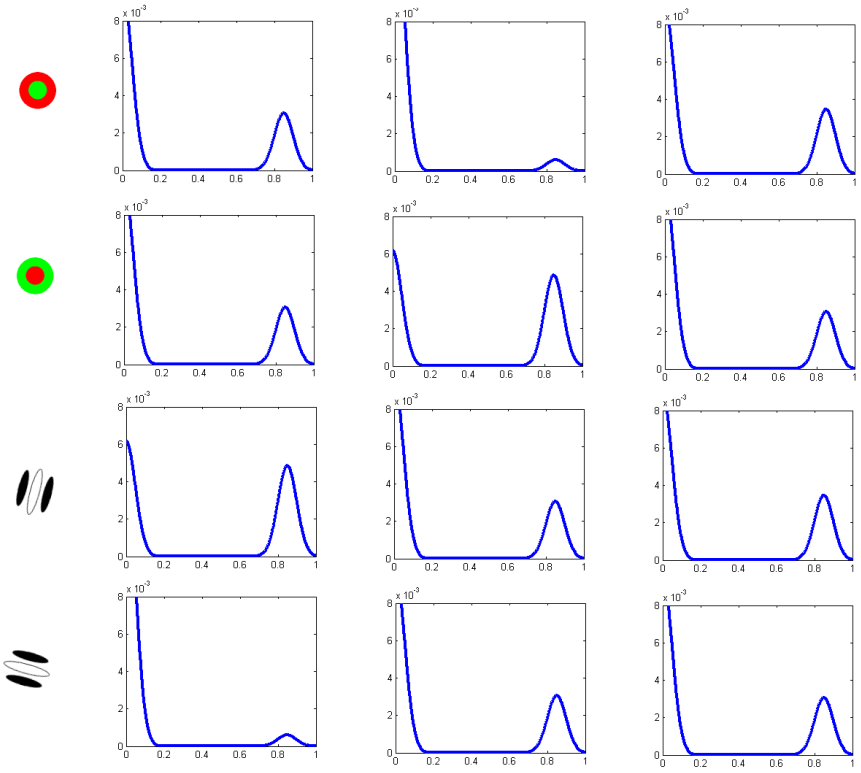

Figure 7.2: Hypothetical probability densities associated with the response of four types of units. Shown are examples based on idealized units for the stimulus in question and crafted to exemplify how the responses of the units in question give rise to the observed effects.)

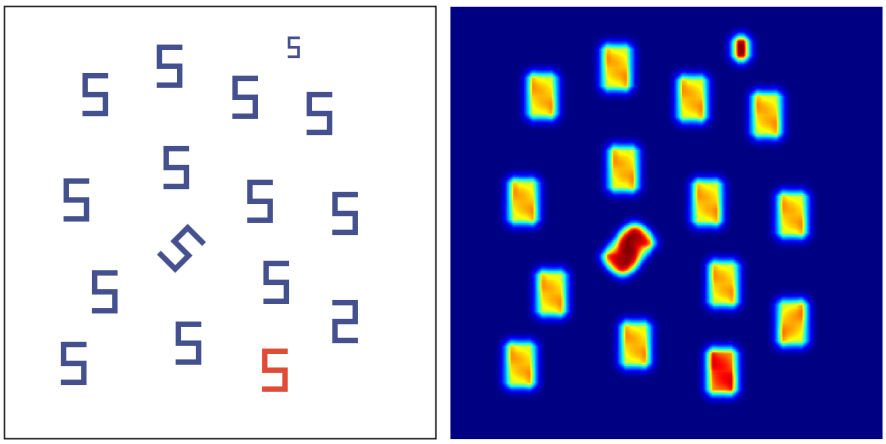

Figure 7.3: An additional example of a conjunction search. The 5's that are small, rotated and red are immediately spotted, however the blue 2 requires effort to spot. Right: Saliency associated with the stimulus pattern.

7.4 Target-Distractor Similarity

An additional area of psychophysics work that has been very influential is that of observing the effects of target-distractor similarity on difficulty of search tasks. Generally, as a target item becomes more similar in its properties to the distracting items, the search becomes more difficult [59; 176]. An example of this modeled on the experiment of Duncan and Humphreys appearing in [59] is shown in figure 7.4 (top). Moving from the top left to top right frame, a shift of the target away from the distractors in color space occurs. The resulting saliency appears below each example and the ratio of distractor to target saliency is 0.767, 0.637, 0.432

and 0.425 respectively. There is one important element appearing in this example that perfectly matches the data of Duncan and Humphreys: In the two rightmost stimulus examples, the distractor to target saliency ratio remains the same. This implies that beyond a certain distance for a particular feature dimension, a further shift along this feature dimension makes no difference in search efficiency. This is exactly the effect reported in [59]. In AIM, the effect emerges due to a single neuron type responding to both target and distractor items. Once the target item is far enough from distractors in feature space there is zero response in the unit tuned to target properties as a result of the distractors regardless of the absolute distance in feature space. Hence the specific effect observed in [59] also appears as an emergent property of modeling redundancy and with saliency equated to information. Interestingly, the resulting ratio of target to distractor saliency is almost identical to the experimental results despite the simplifying assumptions in learning the V1 like neural representation.

7.5 Distractor Heterogeneity

A question that follows naturally from consideration of the role of target-distractor similarity is that of whether distractor-distractor similarity has any effect on search performance. The most telling effect in this domain is that increasing the heterogeneity of the distractors yields a more difficult search [59; 152; 200]. Consider

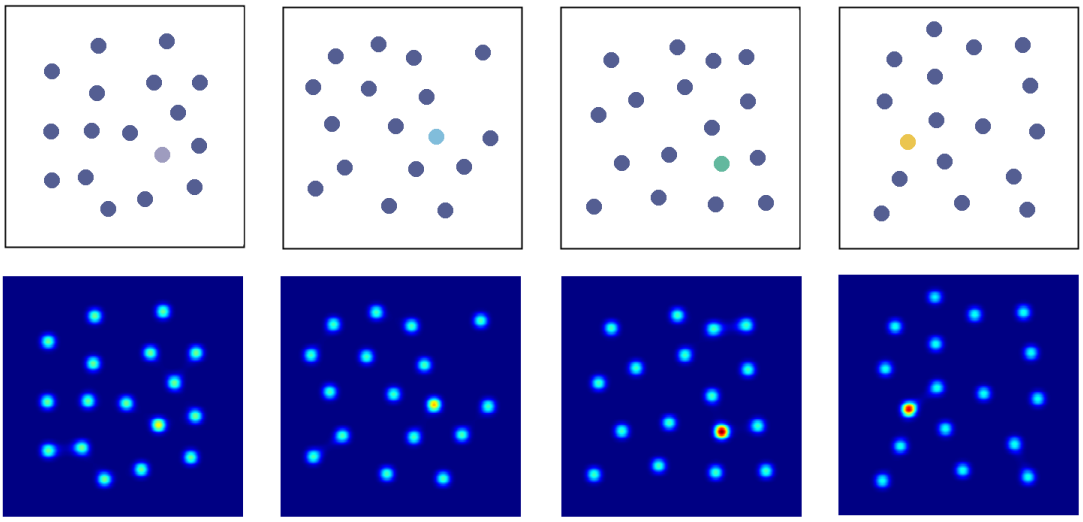

Figure 7.4: From left to right the distance in color space between target and distractors increases. Bottom: Resulting saliency from application of AIM to the stimulus examples. Of note is that the target saliency increases to an extent, but remains constant for the two rightmost conditions.

for example figure 7.5. In the top left case, the item 15 degrees from horizontal appears to pop-out. This effect is diminished in the top middle frame and severely diminished in the top right frame. The saliency attributed to each of these cases appears below each stimulus example. The finding that an increase of distractor heterogeneity results in a more difficult search is consistent with AIM behavior. Distributing the distractors over several different cell types rather than a single type of neuron means that the distractors are considered less probable and hence more informative thus decreasing the ratio of target to distractor saliency. There is also a secondary effect in the example given of target distractor similarity since broad tuning means that cells tuned to a particular orientation may respond weakly to a distractor type other than that for which they are tuned, or the target. This serves to highlight the importance of the specifics of a neural code in the determination of visual saliency and also offers insight on why the determination of efficiency in visual search tasks may be difficult to predict. It is worth noting that this basic effect captures behaviors that models based on signal detection theory [236] fail to. For example, a horizontally oriented bar among distractors at 30 degrees is much more salient than a horizontal bar among distractors 1/3 oriented at 30 degrees, 1/3 at 50 degrees and 1/3 at 70 degrees as observed in [198]. This is an important peculiarity of visual search that is inherent in an information seeking model, but absent from many competing models of saliency computation. These considerations

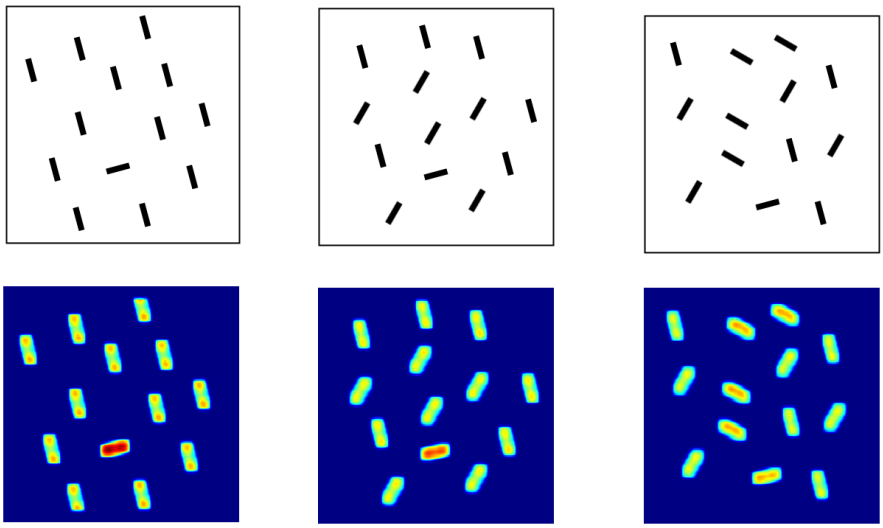

Figure 7.5: Top: An example of increasing distractor heterogeneity from left to right. The target at 15 degrees from horizontal becomes less salient in the presence of increasingly heterogeneous distractors. Bottom: Saliency associated with the stimulus examples. This effect demonstrates the somewhat curious effect of distractor heterogeneity in agreement with the results reported in [59].

are of course not limited to the orientation domain and may also be observed for other feature domains, for example for colored stimuli as observed in figure 7.6.

7.6 Search Asymmetries

A stimulus domain that has generated a great deal of interest involves so-called search asymmetries, due to their potential to reveal peculiarities in behavior that may further our understanding of visual search. One asymmetry that has received

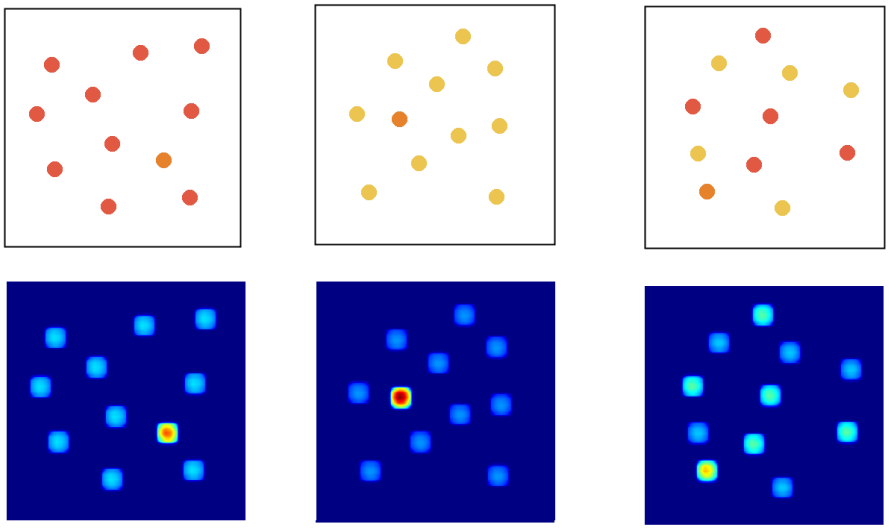

Figure 7.6: Increased distractor heterogeneity in color space (top) and corresponding saliency maps (bottom).

considerable attention is an asymmetry attributed to presence vs. absence of a feature as in figure 7.7 [223]. In this example, a search for a dash among plus signs is much more difficult than the converse. In examining the associated saliency maps as computed by AIM, it is evident that this behavior is also inherent in the information based definition. Note that this is simply a specific case of a more general phenomenon and the same might be observed of a Q among O's or any instance where a singleton is defined by a feature missing as opposed to its presence. This phenomenon can be explained by the fact that in the feature present case, the feature that distinguishes the target is judged to be improbable and hence informative. In the case of the feature absent, there is nothing about the location

that distinguishes it from background content in the context of the missing feature since the background regions also elicit a zero response to the "missing" feature. Rosenholtz reveals an additional class of asymmetries, which she points out are examples of poor experimental design as opposed to true asymmetries [199]. An example of such a stimulus appears in figure 7.8 (top). Rosenholtz points out that the asymmetry appearing in figure 7.8 which corresponds to the task of finding a red dot among pink being easier than the converse (top left and top second from left) may be attributed to the role of the background content [200]; a change in background color (top right and top second from right) causes a reversal in this asymmetry. From the resultant saliency maps, it is evident that AIM output also agrees with this consideration (figure 7.8 bottom). Reducing the contrast between the background and the target/distractors would also be expected to give rise to a more pronounced asymmetry as the response of a cell to target/distractors and background become less separable. This is indeed the behavior reported in [200]. An important point to note is the fact that viewed in the context of AIM; the color background asymmetry arises from the same cause as the feature presence-absence asymmetry, both a result of the role of the background in determining feature likelihood. In each case, it is the role of the background content in determining the likelihood associated with any particular firing rate. In the colored background examples, the background causes greater suppression of the target or distractors

depending on its color. One example Rosenholtz describes as an asymmetry in experimental design is that of a moving target among stationary distractors versus a stationary target among moving distractors, suggesting that the design be rectified by ensuring the motion of the distractors is coherent [197]. Under these conditions, the stationary search becomes more efficient but still remains significantly less efficient than the moving target case which it is suggested may be attributed to a basic asymmetry in processing motion. Viewed in the context of AIM, an additional possibility arises: If there exist units that elicit a response to non-target background locations and also to the stationary target this may have an effect of suppressing target saliency that will be absent in the moving target case. As such, a truly symmetric experiment would call for the case where the entire background elicits a response from the motion sensitive neurons in question with a localized region remaining stationary, along with the converse case. This might be tested using random textured stimuli rather than independent dots on a white background.

7.7 Basic Asymmetries

An additional aspect pertaining to asymmetric behavior in visual search paradigms corresponds to what are often described as "Basic Asymmetries" [197]. For example, a bar oriented at 15 degrees from vertical among vertical orientation distractors is a very easy search whereas the converse is more difficult. Additional examples of

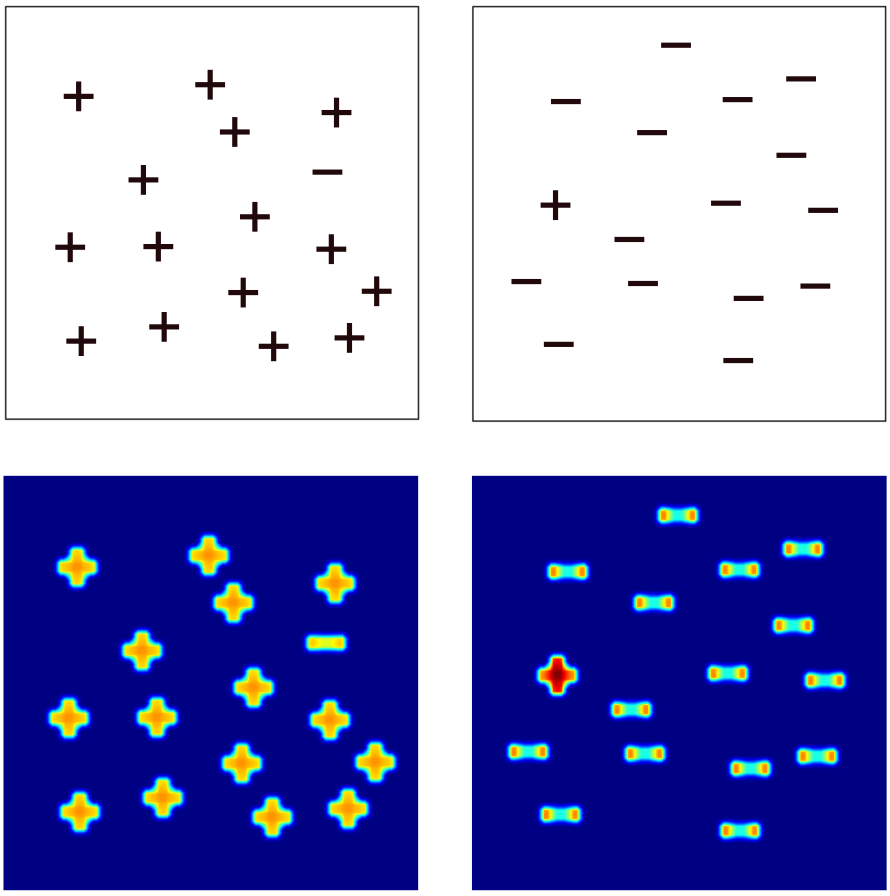

Figure 7.7: An experimental asymmetry. The task of locating the plus among dashes is easier than the dash among pluses. Bottom: Saliency associated with the two stimulus examples. This effect demonstrates a specific example of a general asymmetry related to feature presence versus absence as reported in [223].

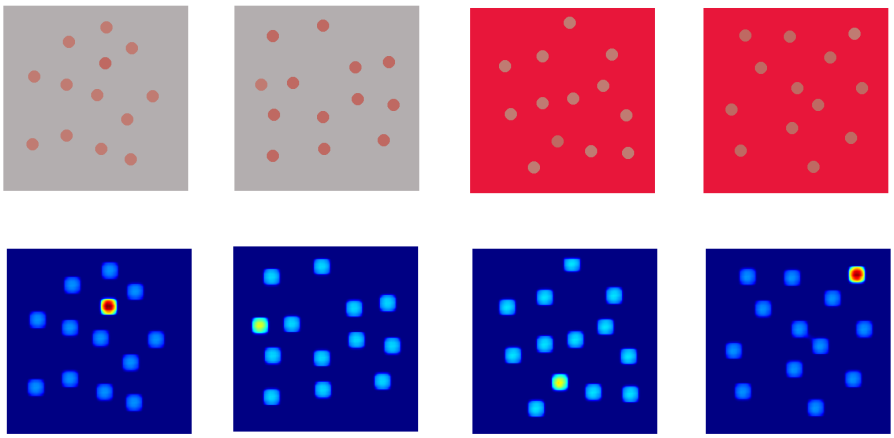

Figure 7.8: Top row: An asymmetry in experimental design as described in [200]. The red target pink distractor case is easier than the converse; a change in background color results in a reversal of the effect. Bottom row: Saliency based on model output.

this include a red target among orange distractors and a fast moving target versus slow moving distractors. Figure 7.9 demonstrates an example of this first case, and the corresponding saliency as determined by AIM. It may be somewhat surprising that the model output again agrees with what is observed behaviorally. This is an interesting result and demonstrates a behavior that heretofore has lacked a suitable explanation. Close inspection of the underlying basis reveals the explanation for this behavior: As vertically oriented structure is overrepresented in natural statistics, it is also given greater representation in the resulting neural encoding. As such, tuning for vertical orientation is stronger and additionally suppression associated with such cells is stronger. This presents an interesting avenue for further exploration, specifically to observe the extent to which these basic asymmetries may be explained by bias in the statistics of the natural environment and moreover, what predictions may be made with regard to searches that may fall in this category.

7.8 Visual Field Anisotropies and Neural Coding

The preceding explanation for so-called *Basic asymmetries* as being attributed to bias in coding warrants some further consideration. In particular, it is worth considering the extent to which bias in natural statistics exhibits behavioral correlates. A byproduct of this analysis may be additional support for the efficient coding hypothesis assuming any bias in statistics is reflected directly in coding. Prior work

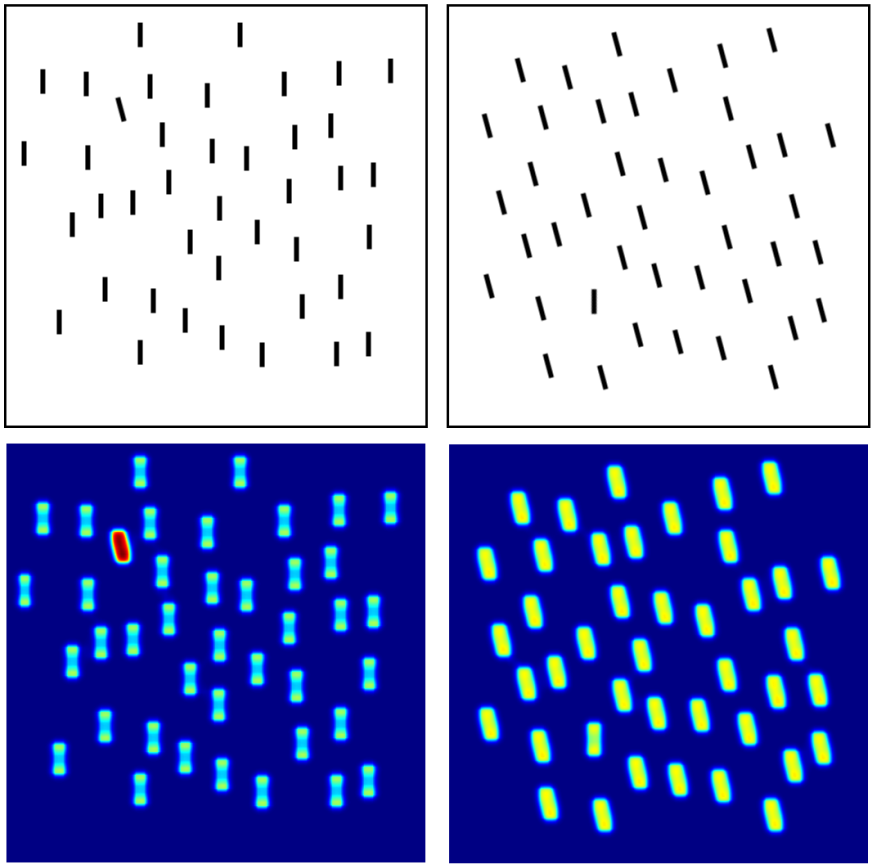

Figure 7.9: Top row: An example of a basic asymmetry in which the visual search for a target oriented 15 degrees from vertical is easily spotted among vertical distractors while the converse is not a pop-out task.

on natural image statistics has shown that a sparse code for local neighborhoods tends to yield cells with properties akin to those appearing in the visual cortex. If this claim is true, one would also expect that any variation in local statistics across the visual field would be reflected in the underlying neural hardware. In the following discussion this possibility is investigated at the level of whether there is support for this consideration since there are a variety of psychophysics paradigms for which performance varies with position in the visual field. Although the results can hardly be stated as conclusive from the perspective of causality, support for efficient coding does appear in the form of visual field anisotropies. For the purposes of supporting the efficient coding hypothesis, this is sufficient.

It has long been apparent that there exist anisotropies in human visual processing. For example, performance in various psychophysical tasks is much better for grating stimuli oriented horizontally or vertically than for the same stimuli presented at oblique orientations (See [37] for a review). This phenomenon has been termed the oblique effect. Previous efforts have considered a statistical basis for this effect and others, and in the case of the oblique effect, there does exist a bias in image content in favor of vertically and horizontally oriented edges [48]. In this work, we examine a different set of anisotropies, namely, domains for which performance varies as a function of position of stimulus in the visual field. Studies concerning laterality make up the bulk of psychophysical results fitting this

category, with performance differences for stimulus presentation in left and right visual field considered. Much of the literature considers the interaction between visual field and spatial frequency of stimuli, with a right visual field advantage for high spatial frequency content and a left visual field advantage for low spatial frequency content. There also exist upper-lower visual field asymmetries that have received relatively less attention in the literature. Articles describing upper-lower visual field asymmetries typically read very similar to a standard visual field laterality study with the exception that the visual world is rotated 90 degrees [64]. That is, left-right asymmetries manifest in very similar behavioral benefits/deficits to upper-lower asymmetries, with stimuli appearing in the upper and right visual fields showing similar performance benefits/deficits, and lower and left visual fields exhibiting similar benefits/deficits. Another interesting anisotropy concerns the so-called radial organization of the visual system. A variety of studies have found that judgments related to line orientation are best for lines oriented towards the centre of the visual field and worst for lines orthogonal to the centre. For example, in the lower left quadrant of the visual field performance for judgments on lines oriented at 45 degrees is best and lines oriented at 135 degrees is worst [81; 133]. It has been suggested that upper-lower visual field asymmetries might arise from the difference in statistics between sky and ground [64]. This claim has not been validated through consideration of actual scene statistics. Furthermore, there does

not exist a consensus on the origin of lateral asymmetries, or the so-called radial organization of the visual system. In the sections that follow, each of these effects is considered in the context of local statistics, with the aim of determining whether there might exist a statistical basis for such effects.

7.8.1 A Look At the Statistics

Explanations for the cause of visual field anisotropies are sparse in the literature, with the majority of work describing what is observed rather than why. The following effort aims at observing the manner in which angular and radial frequency statistics vary across the visual field, and evaluating these observations in the context of existing psychophysical results. In this light, we seek a local representation of statistics that allows observation of angular and radial frequency content as such content varies across the visual field (image), and lends itself well to qualitative analysis (i.e. orientation and spatial frequency statistics are directly observable). A common representation that adheres to these properties most often employed in the domain of signal processing, is the power spectrum of a signal (an image in this case). The power spectrum of an image I is given by $P(u) = |F(u)|^2$, the square of the magnitude of the Fourier transform of I. Although the frequency domain representation of an image affords observation of the appropriate orientation and spatial frequency content, the manner in which such content varies across space

is not directly observable. However, the power spectrum need not necessarily be computed over the entire image. If one divides the image into smaller subimages and computes local power spectra of each subimage, this affords direct observation of angular and radial frequency content as a function of position in the visual field without the need for discrete sampling in frequency. Although this offers only coarse discrete sampling over space, the resolution should be sufficient to make inferences concerning the plausibility of statistics as a basis for visual anisotropies. It is worth noting that one might also perform this analysis based on a locally windowed wavelet style representation which might allow more detailed analysis including consideration of quantitative aspects of the features present. That being said, the local representation is sufficient for the qualitative observation of the quantities of interest. Fig. 7.10 demonstrates a representation of a single image (left) using the proposed local power spectra representation (right). It is clear that the angular and radial frequency content are directly observable over various sections of the visual field. The first attempt in this work at observing statistics in general, was made through considering an average of the representation of the form depicted in fig. 7.10 obtained from 3600 different natural images, and normalized over each window to give a sense of the relative proportion of high and low spatial frequency content within each section of the visual field. Images were 1408x896 pixels, and were divided into 77 (11 by 7) 128x128 windows. Figure 7.11a. demon-

strates the local power spectra averages from the 3600 images. Qualitative biases visible in such a representation are marginal as demonstrated in figure 7.11a with a strong bias for low spatial frequency content over the entire visual field. However, given what is observed in figure 7.11a it is clear that an overall bias for low spatial frequency content may mask any subtle asymmetries existent in the local power spectra. We overcome this difficulty in figure 7.11b, by demonstrating the difference between each of the spectra depicted in figure 7.11a, and the average local power spectrum derived from every local window over the entire image. This offers an idea of the difference in shape of local power spectra across the visual field and the local statistics as compared with other regions of the visual field, and makes visible the subtleties present in the statistics. The spectra depicted in figure 7.11b are histogram equalized to make the relative presence of structure at various frequencies more evident. Since this does not affect the relative rank of such components, this is sufficient for our purposes.

7.8.1.1 The Oblique Effect

Previously we described the oblique effect, a performance deficit for orientation discrimination for oblique orientations relative to horizontal and vertical orientations. Recent efforts indicate that this effect persists over the entire visual field [81; 143]. As may be seen in figure 7.11a. there is a significant bias in the statistics in favor of

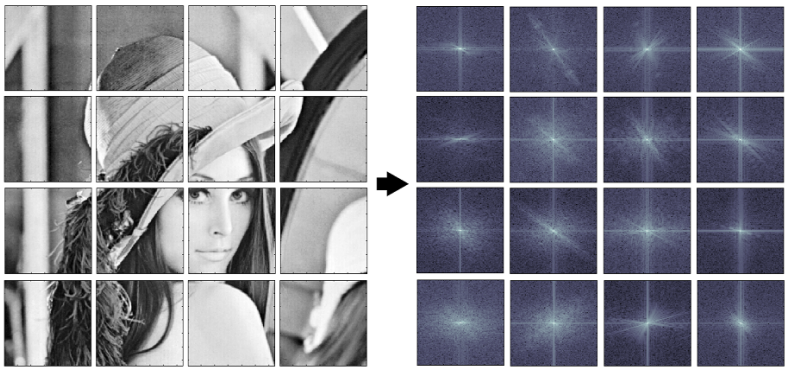

Figure 7.10: Representation of local orientation and spatial frequency content of an image based on the proposed local power spectra representation.

horizontally and vertically oriented structure across a wide range of spatial frequencies. This effect has been previously demonstrated through observation of power spectra obtained from entire images, but not locally as depicted in this case. The statistics suggest that there is a bias in favor of horizontally and vertically oriented edges over the entire image in agreement with psychophysical results [81; 143].

7.8.1.2 Lateralized and Upper/Lower Visual Field Asymmetries

As one might expect, there appears to be a significant difference in the statistics of upper versus lower visual field, with a bias in favor of high spatial frequency content in upper visual field. In contrast, there is no such bias across the vertical meridian. This calls into question the argument that basic sensory anisotropies arise from structure in the statistics since the upper-lower and left-right asymme-

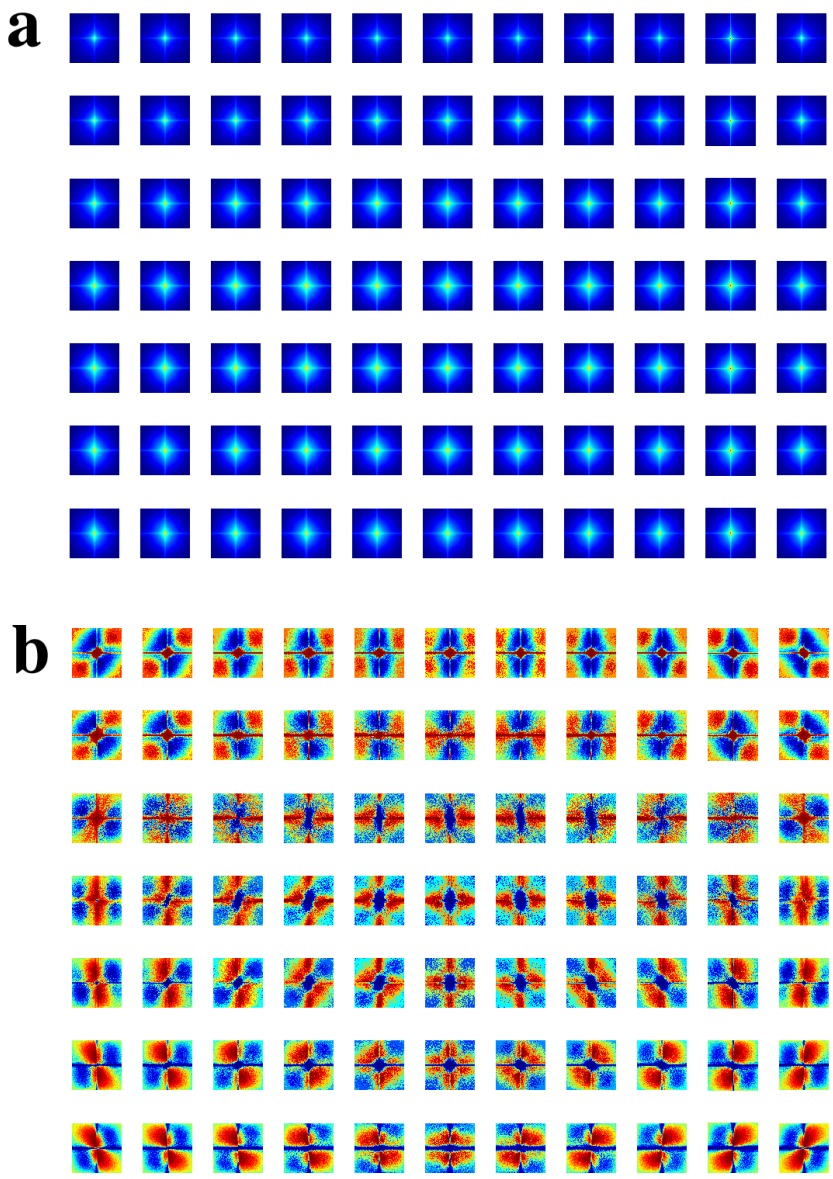

Figure 7.11: a. Average of local power spectra obtained from 3600 natural images. b. Difference between each spectrum depicted above, and average local power spectrum derived from every local neighbourhood of each image.

tries seem to manifest in very similar performance benefits/deficits. This conflict may be resolved by considering a few more recent experiments. Mondor and Bryden investigated the effect of varying SOA for a task requiring letter identification and lexical decisions for stimuli presented to left or right visual field [192]. A right visual field advantage was observed only in the case that the time between onset of cue and onset of stimulus (SOA) was sufficiently short. Rhodes and Robertson considered the effect of rotating the display during a typical laterality experiment. They found that left-right asymmetries persisted in the reference frame of the display rather than a retinal frame of reference [243]. One might conclude from these observations that lateral asymmetries arise from the manner in which the output of basic sensory channels is handled. In contrast, upper-lower asymmetries seem to manifest from more primitive factors such as the relative contribution of magnocellular and parvocellular pathways to the processing of stimuli in the upper and lower visual fields. There is evidence that the parvocellular pathway projects preferentially to visual areas corresponding to upper visual field and magnocellular layers to areas corresponding to lower visual field [64]. This effect has not been observed for left versus right visual cortices. Maehara et al. observed that a red background, thought to attenuate magnocellular pathways, relative to a green background gave rise to a greater deficit in detecting spot stimuli in the lower visual field than in the upper visual field [244]. Each of these results are suggestive of a primitive neural

basis for upper-lower visual field asymmetries existent in early visual areas. That said, more work is needed in comparing left-right to upper-lower asymmetries to resolve this issue.

7.8.1.3 Radial Organization of the Visual Field

We have described the apparent radial organization of the human visual system, wherein judgments on structure oriented towards the centre of the visual field tend to be best (of obliques) and structure orthogonal to the centre tend to give rise to the worst performance [81; 133]. Whether there exists a statistical basis for this effect may be determined in observing the orientation statistics as they vary across the image. As is seen in figure 7.11b, a somewhat surprising anisotropy is observed in the orientation statistics, with an apparent bias over the visual field for lines oriented toward the centre. One possibility for this observation, is that this effect arises as a result of geometric perspective, with edges in the visual field appearing to fade to the point of fixation on the visual horizon. Figure 7.12 demonstrates an image for which this effect is especially apparent along with its local power spectra. Whether this is truly the basis for the orientation anisotropy remains to be determined. That said, the proposed relationship marks the first possible explanation for this effect having an environmental and hence statistical backing.

We have demonstrated that in the case of basic sensory visual field anisotropies,

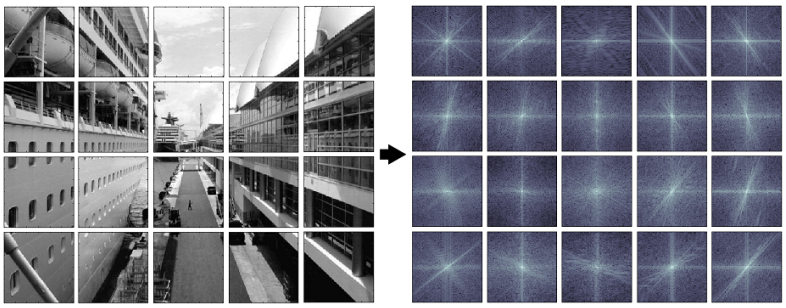

Figure 7.12: An image (left) for which the described perspective effect is particularly strong, along with its local power spectral representation (right).

there does appear to be a statistical basis for such effects. Further, we have presented an argument dissociating upper-lower visual field asymmetries from lateral asymmetries, in agreement with more recent psychophysical results. Finally, we have put forth a possible explanation for the apparent radial organization of the visual system, suggesting that geometric perspective may produce a sufficient statistical bias to account for this effect. These results speak to the role of anisotropic coding in regards to observed behavior in visual search paradigms. It appears that there does exist a basis for the basic orientation asymmetry and the preceding highlights an important additional avenue for assessing the nature of unexpected behaviors emergent from visual search psychophysics.

7.9 Discussion

In further support of the proposal put forth in AIM, it has been demonstrated that there exists a considerable range of basic behaviors observed in visual search that appear as emergent properties of the proposal. This incorporates an unprecedented range of effects including pop-out, the role of target distractory similarity, distractor heterogeneity, and various asymmetries that appear in the experimental literature. Finally, some novel results concerning efficient coding bolster the claim that certain basic asymmetries may occur by virtue of the fact that neural circuitry forms an efficient representation of natural image statistics, which includes anisotropic incidence of certain frequencies. Importantly all of these observations are emergent from the basic assumption of selection based on information and a detailed account of a specific component of saliency related processing.

8 Complex Features and a Hierarchical Representation of Saliency

Saliency based models assume that somewhere in the brain there must exist a topographical representation of the relative importance of different visual stimuli. This chapter also considers an alternative possibility in which saliency does not require an explicit topographical representation, since most areas of the brain seem to encode the strength of certain features (e.g. a horizontal edge, or a convex surface, or a face) and not their relative importance. A scheme is presented in which combining the circuitry that facilitates likelihood estimation, with traditional winner-take-all type selection produces information based selection without an explicit topographical representation of saliency. The contribution is perhaps not the proposed circuitry per se, but rather a demonstration that in combining the traditional school of thought concerning saliency, with a disjunctive class of models that consider only selection, the two may be achieved simultaneously without the requirement of a single topographical representation of saliency. That said,

the focus is on consolidating different schools of thought concerning attention, in a manner that is largely consistent with ideas derived from each.

As we have described in earlier chapters, there is evidence that suggests the possibility that the primate visual system may consist of a multi-layer sparse coding architecture [13; 69]. The proposed algorithm quantifies information on the basis of a neural circuit, on units with response properties corresponding to neurons appearing in the primary visual cortex. However, given an analogous representation corresponding to higher visual areas that encode form, depth, convexity etc. the proposed method may be employed without any modification. Since the *popout* of features can occur on the basis of more complex properties such as a convex surface among concave surfaces [93], this is perhaps the next stage in a system that encodes saliency in the same manner as primates. Given a multi-layer architecture, the mechanism for selecting the locus of attention becomes less clear. In the model of Itti, Koch and Niebur, a multi-layer winner-take-all network acts directly on the saliency map and there is no hierarchical representation of image content. There are however attention models that subscribe to a distributed representation of saliency (e.g. [228]) that may implement attentional selection with the proposed neural circuit encoding saliency at each layer. These issues are discussed in section 8.5.

A great deal of research effort is currently being placed on unsupervised learning of hierarchies of features towards representations amenable to application to object

recognition and visual representation in general. A claim made in this chapter is that the proposal put forth by AIM may be applied within a distributed hierarchical representation and section 8.5 offers a sketch of the details of this. That being said, as the research in this area remains in its infancy, it is difficult to offer an existence proof in the form of an implementation in this regard. As a compromise, the following discussion aims to show the generality of the proposal via consideration of the operation of AIM using a different basis set and employing very different eye tracking data than that presented in chapters 4 and 5. An additional example is provided by way of an analytic spatiotemporal basis, further establishing the generality of the approach. Although this falls short of a complete demonstration of hierarchical operation, it offers at least a demonstration of how the proposal may be applied to any sparse basis and also further establishes the efficacy of the proposal in explaining human eye tracking data through consideration of a very different eye tracking data set and also demonstrates that the application of the proposal is not limited to a learned basis but may have utility for analytic models also. It should be stated that the important component of this with respect to its extension to a hierarchical representation derives from the fact that saliency related modulation may occur at the level of a single cell in the absence of any explicit topographical representation of saliency. That being said, the reader is encouraged to bear this in mind in considering the various components that follow

in this chapter.

8.1 Spatiotemporal Saliency

The general nature of the original proposal implies that it may be applied to any set of neurons that constitute a sparse basis. For this reason, extension to space-time is straightforward assuming the early coding of spatiotemporal content observed in the cortex satisfies these criteria. There exist many efforts documenting the relationship between early visual cortical neurons and coding strategies that demonstrate that learning a sparse code for local grey-level image content yields V1 like receptive fields similar to oriented Gabor filters [13; 170]. Further efforts have demonstrated this same strategy yields color-opponent coding for spatiochromatic content [241] and also cells with properties akin to V1 for spatiotemporal data [233]. We have employed the same data and strategy put forth in [233] to learn a basis set of cells coding for spatiotemporal content. The data described in [233] was subsampled taking every second frame to yield data at 25 frames per second. The data set consists of a variety of natural spatiotemporal sequences taken from various angles of a moving vehicle traveling in a typical urban environment. Spatiotemporal volumes were then randomly sampled from the videos to yield 11x11x6 (x,y,t) localized spatiotemporal volumes that served as training data. Infomax ICA [120] was applied to the training set resulting in a spatiotemporal basis consisting

of cells that respond to various frequencies and velocities of motion. The basis resulting from dimensionality reduction via PCA retaining 95% variance followed by ICA yields a set of 60 spatiotemporal cells. A subsample of these (corresponding to 1st, 3rd and 6th frame of the volume) are shown in figure 8.1. Note the response to various angular and radial frequencies and selectivity for different velocities of motion. Aside from the application to spatiotemporal data and the different basis set, the saliency computation proceeds according to the description put forth in chapter 4.

An overall schematic of the model based on the learned spatiotemporal basis appears in figure 8.2. A localized region from adjacent frames (3 of 6 shown) are projected onto the learned basis. This yields a set of coefficients for the local region that describes the extent to which various types of motion are observed at the given location. The likelihood of each response is then evaluated by observing the response of cells of the same type in the surround or in this implementation, over the entire image. A sum of the negative log likelihood associated with all of the coefficients corresponding to the given coordinate (pixel) location yields a local measure of saliency.

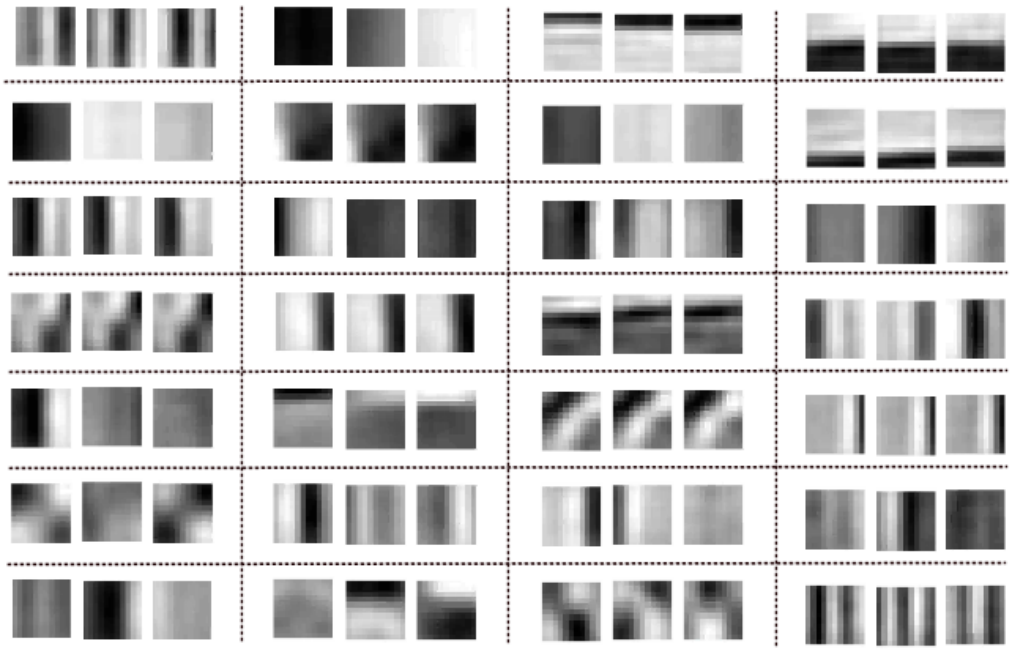

Figure 8.1: The receptive field profile of a subsample of the learned basis. Each dotted box depicts the receptive field in space corresponding to frames 1, 3 and 6 of the spatiotemporal basis volume associated with one basis function. Note the selectivity for various angular and radial frequencies and velocities and directions of motion.

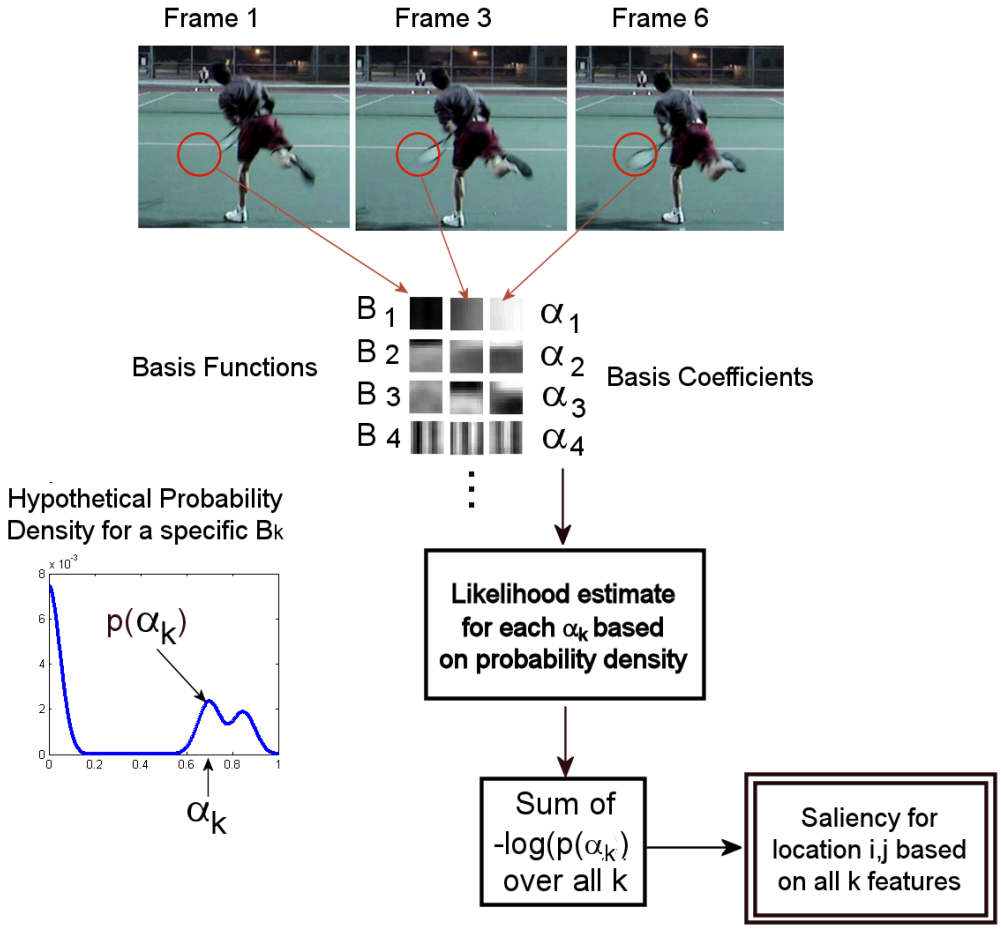

Figure 8.2: An overview of the computation performed by AIM. A spatiotemporal volume is projected onto a learned basis based on independent component analysis. The likelihood of any given cells firing rate may be estimated by observing the distribution of responses associated with cells of the same type in the surround or over the entire image. A summation of these likelihoods subjected to a log transform then yields a local measure of information.

8.2 Evaluation

An evaluation of the efficacy of the model in predicting spatiotemporal fixation patterns is achieved via comparison with eye tracking data collected for video stimuli. The eye tracking data employed for this study was that used in [96] and performance evaluation was carried out according to the same performance metric described in the aforementioned work.

The data consists of eye tracking data for a total of 50 video clips and from 8 subjects aged 22-32 with normal or corrected to normal vision. Videos consist of indoor and outdoor scenes, news and television clips and video games. Videos were presented at a resolution of 640x480 and at 60 Hz and consist of over 25 minutes of playtime. The total number of saccades included in the analysis is 12,211.

For any given algorithm, one may compare the saliency at fixated locations with randomly sampled locations. The Kullback-Leibler divergence of two distributions corresponding to these quantities is given by

$$D_{KL}(P,Q) = \sum P(i) log \frac{P(i)}{Q(i)}$$

where P and Q correspond to the distribution of randomly sampled and at-fixation sampled saliency values respectively based on 10 bin histogram estimates. The KL-divergence offers a performance metric allowing comparison of various algorithms. For more details on KL-Divergence, readers may refer to appendix B. Results

are compared against those put forth in [96] and proceeds according to the same performance evaluation strategy.

Figure 8.3 demonstrates the relative saliency of pixel locations for a variety of single frames from a number of videos. Note the inherent tradeoff between moving and stationary content as observed for the running tap as well as the ability to detect salient patterns on a relatively low contrast background.

Figure 8.4 demonstrates a histogram of the saliency associated with the fixated locations as compared with those from uniformly randomly sampled regions. Of note is the shift of the distribution towards higher saliency values for the distribution associated with fixated relative to random locations. The KL-divergence associated with this evaluation is 0.328. This is a 36 percent improvement over the Surprise model of Itti and Baldi with a KL score of 0.241 and a 60 percent improvement over the saliency model of Itti and Koch [98] with a KL score of 0.205. Importantly, this apparently strong performance comes from the same biologically plausible setup that yielded favorable performance for spatiochromatic data, without modification or any additional assumptions required for consideration of spatiotemporal neurons. This evaluation supports the claim of generality of information as a strategy in saliency computation and additionally offers a means of characterizing spatiotemporal saliency. Additionally, no prior model of scene content or memory is involved as in [96], but rather the prediction is based on the

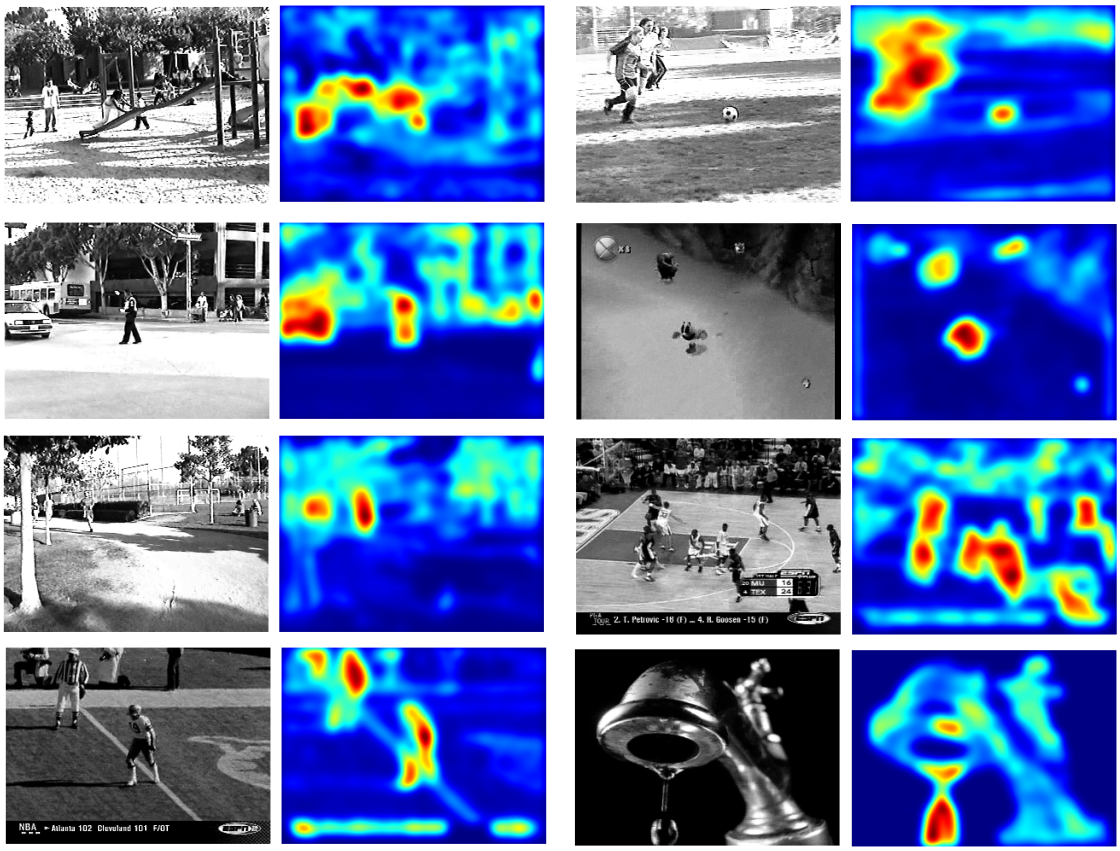

Figure 8.3: Sample frames from a number of different videos in the data set (left of each pair) and their associated saliency (right of each pair). Note the robustness to a variety of differing data and favorable performance even for low contrast structure, crowded scenes as well as the inherent weighting of moving versus stationary content.

current state of neurons that code for spatiotemporal content. Overall, the results provide further support of the generality of AIM in predicting fixation and visual search related behaviors and demonstrates the efficacy of the proposal in predicting fixation patterns on a qualitatively different data set than that comprised of still images.

8.3 Types of Motion Salience

It is also interesting to consider how the model responds to different categories of spatiotemporal stimuli, such as those described in [245] including a moving target on a moving background, scintillation and flicker. Figure 8.5 demonstrates frames from a variety of qualitatively different types of spatiotemporal stimuli and their associated salience. In figure 8.5a, a fast moving target is followed by the camera resulting in a fast moving structured background. Although the target is relatively more stationary than the background due to the panning of the camera, the running target is nevertheless considered salient. In figure 8.5b, scintillation on the surface of a lake results in a diffuse judgement of salience, which becomes localized upon emergence of a bubble on the same surface. In the case of figure 8.5c, a brief flash of lightning appears for only a few frames but is considered salient on the basis of this brief appearance. These examples serve to demonstrate the variety of possibilities in considering spatiotemporal patterns and that the salience associated with said

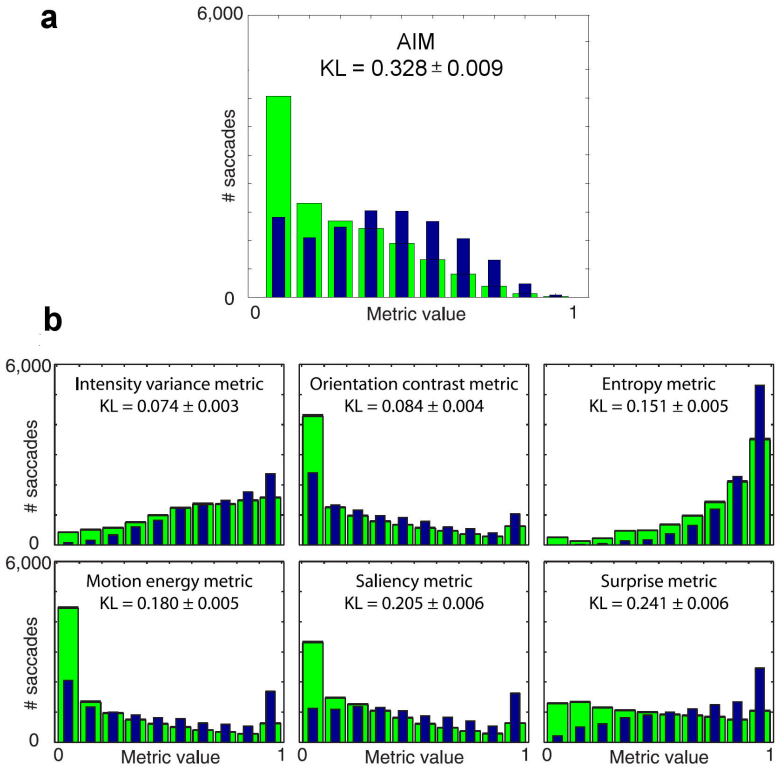

Figure 8.4: a. Saliency values corresponding to locations sampled randomly (green) and at fixations (blue) as produced by AIM. There is a tendency to fixate points that correspond to higher saliency values. The KL-divergence of the two distributions is 0.328 +/- 0.009 as compared with 0.241 +/- 0.006 for the Surprise metric [96] and 0.205 +/- 0.006 for the Saliency metric [98]. b. The same quantitative performance evaluation for the Saliency and Surprise metrics (reproduced from [96]).

categories appears to agree at an intuitive level with ones expectation.

8.4 An Analytic Basis

Perhaps the most influential parameter in the evaluation described in chapter 5 is the role of receptive field size on the resulting saliency determination. Recall that a limitation of an unsupervised learning approach is that the dimensionality of the problem prohibits receptive fields beyond a certain size. Fortunately unlike the spatiochromatic domain, there exists considerable prior work on analytic bases directed towards the representation of spatiotemporal content. For this reason, it is worth considering the extent to which such a representation lends itself towards saliency computation under the operation of AIM.

One such representation that is suitable for the purposes of this exercise and a natural choice owing to its widespread use and favorable properties is that of Adelson and Bergen [2]. A representation of this form has shown efficacy in the qualitative analysis of motion [246] and additionally has served a precursor to the determination of spatiotemporal saliency [245]. In the implementation results, a spatial Gaussian pyramid was constructed from the video stream yielding spatiotemporal volumes at 4 spatial scales. Subsequently, separable steerable filters were employed along x-t and y-t dimensions at each layer of the pyramid and saliency determined by AIM acting on the responses of spatiotemporal filters. An

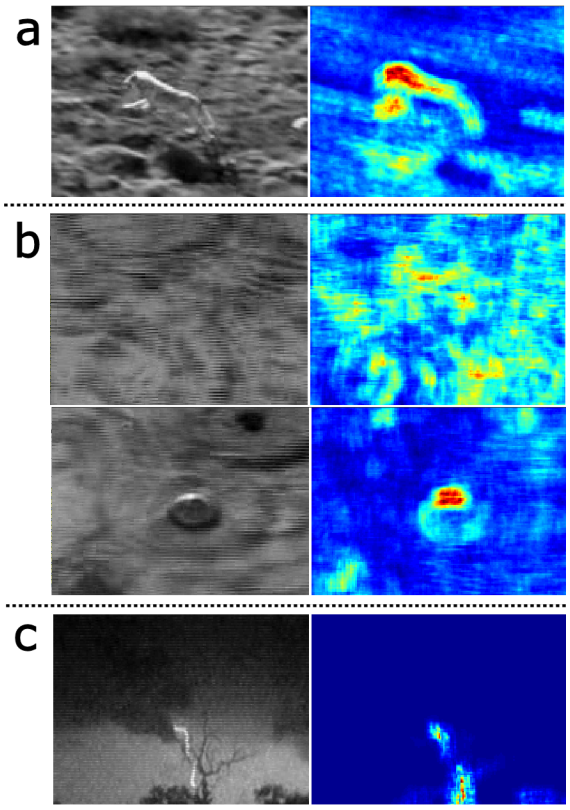

Figure 8.5: Examples of various qualitatively different categories of moving stimuli and associated salience for certain frames: a. A fast moving target is followed by a panning camera resulting in a structured and moving background. b. Scintillation on the surface of a lake results in a salience judgement that is diffuse. Emergence of a bubble on the surface results in the suppression of this diffuse salience in favor of the bubble. c. A very briefly flashed lightning bolt is judged as salient, despite its appearance for only a few frames and without any directional motion.

Levels of Gaussian Pyramid Considered	ROC Area
1	0.2217
2	0.2582
3	0.2711
4	0.2316
1-2	0.2505
1-3	0.2727
1-4	0.2867
2-3	0.2837
2-4	0.2926
3-4	0.2732

Table 8.1: Demonstrates the effects of receptive field size and dimensionality reduction on area under ROC curve scores.

overall information score was determined by considering a subset of scale space corresponding to some subset of the layers of the Gaussian pyramid. Layer 1 refers to the most detailed or highest frequency processing and layer 4 to the coarsest level of processing. Saliency scores corresponding to these various conditions are shown in table 8.1.

Table 8.1 reveals some interesting aspects pertaining to the saliency computation at hand. The first point of interest is that the saliency scores associated with

the mid-range levels of the Gaussian pyramid are superior to those at the very high and very low ends of the frequency range. An additional aspect of interest is that the saliency scores associated with representations that span scale space are superior to any of those that correspond to a narrow frequency band. This is a result of interest for saliency computation at large and lends credence to the use of a basis representation in the computation of visual saliency in lieu of an ad hoc set of arbitrary features. In further consideration of this observation it is interesting to consider possible differences in the magnitude spectra of fixated versus non-fixated locations. Eye tracking data was collected for the purpose of considering the extent to which local spatial frequency content informs on salient visual content. Data was collected for a set of 250 grayscale images, from 10 subjects each viewing 50 images (5 sets of 50 images, with 2 subjects viewing each set). Images were randomly chosen from the Corel stock photo database and presented in random order for 4 seconds each with a mask between each pair of images for 2 seconds. Analysis was based on the glint-pupil vector data obtained from an Arrington Research View-Point EyeTracker. Images were presented on a 21 inch CRT monitor at a resolution of 1024 x 768, with participants positioned at a distance of 70 cm. Participants were naive to the purpose of the study and were instructed simply to observe the images. A 120x120 window centred at each of the approximately 3000 fixation points was extracted and an average magnitude spectrum computed based on these

local regions. A template magnitude spectrum was derived from selecting 120x120 regions from 62,500 randomly sampled locations from the same image set. Figure 8.6 reveals the difference between local magnitude spectra sampled at fixation points and the average template local spectrum sampled from random locations. In confirmation of the scale space observations described here, it may be said that fixated points have a greater presence of mid-range frequencies with very low and very high frequency components relatively under-represented at fixation.

One final aspect of the spatiotemporal saliency computation that may be of interest in a machine vision context concerns camera movement or the movement of a robotic head. The nature of the computation performed results in movement associated with camera movement essentially being cancelled out. That is, a figure that is stationary in the scene being followed by a camera may be judged salient relative to the moving background. As a whole, these results further demonstrate the generality of the proposal and illustrate some interesting aspects of computation as it pertains to coding and scale space.

8.5 Towards a Hierarchical Representation of Saliency

It is interesting to consider how the content discussed in the previous sections fits in with the *big picture* as far as attention modeling is concerned. There are a variety of different schools of thought on the computational structure underlying

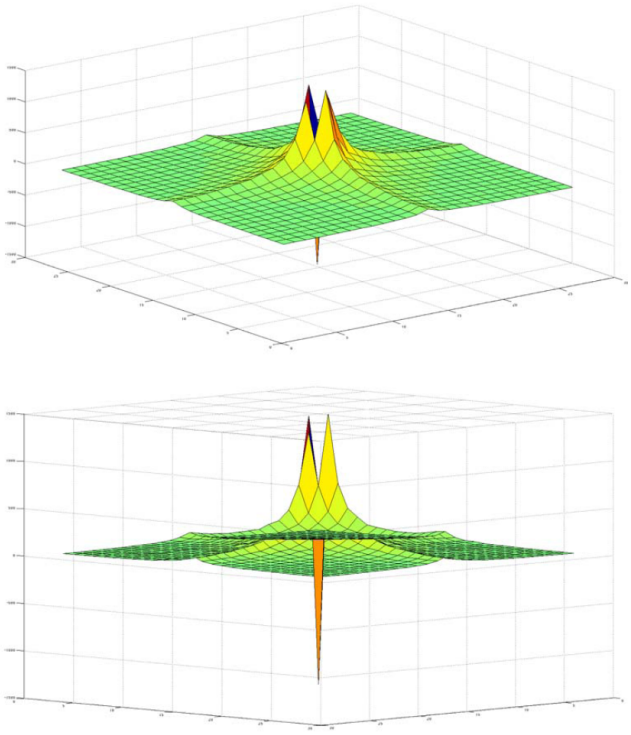

Figure 8.6: Two views of the difference between the average magnitude spectrum of fixated points versus the average of nonfixated regions. The centre hole corresponds to the origin and the elongated peaks moving to higher spatial frequencies correspond to vertical and horizontal structure.

attentional selection in primates ranging from those that posit the existence of a *saliency map* [98; 110; 124] in the cortex to those that claim a distributed representation over which winner-take-all behavior or competitive interaction facilitates attentional selection [51; 228]. Thus far we have depicted saliency in a manner more consistent with the former of these categories demonstrating the total information at any spatial location as the sum of information attributed to all cells that code for content at that location. The correspondence between the proposal and models based on a saliency map can then be thought of as considering the average gain across a cortical column corresponding to a particular location. What is perhaps more interesting is the relationship between the proposal and distributed models of attention. It is evident that as the observation likelihood is computed at the level of a single cell, it is possible that this signal is used to control its gain at the single cell level in accord with neurophysiological observations. It is evident that the proposal put forth is amenable to a saliency map style representation, but it is our opinion that recent results are more consistent with a distributed selection strategy in which gating is achieved via localized hierarchical winner-take-all competition and saliency related computation achieved via local modulation based on information. In this vein, the following discussion considers evidence in favor of a distributed representation for attentional selection as put forth in [228] and the relationship of such a representation to the proposal put forth by AIM. Visual pro-

cessing appears to constitute a dichotomy of rapid general perception on one hand versus slower detailed processing on the other. For example, it is possible to miss large changes in a scene when the changes are masked by an intermediate frame in the absence of focal attention. This phenomenon is referred to as change blindness [190]. Many studies demonstrate that certain quantities are readily available from a scene at a glance such as [65; 92] while other judgments require considerably more effort. This is evidently a product of a visual hierarchy in which receptive fields cover vast portions of the visual field and representations code for more abstract and invariant quantities within higher visual areas. Attentional selection within the model of Tsotsos et al. proceeds according to this assumption with attentional selection implemented via a hierarchy of winner-take-all processes that gradually recover specific information about an attended stimulus including the specific conjunction of features present and, in particular, the precise location of a target item. In line with this sort of architecture, recent studies have shown that a variety of judgements can be made on a visual stimulus with a time course shorter than that required for localization of a target item [65; 92]. The early access to general statistics associated with stimuli within the display is also encouraging, as this is a requirement for the computation performed in AIM. It should be noted that within the traditional saliency map paradigm, there is nothing inherent in the structure of the model that is consistent with this consideration as spatial selection forms

the basis for determining the locus of attention. Furthermore, the forest before trees priority in visual perception appears to be general to virtually any category of stimulus including the perception of words preceding that of letters [103] and scene categories more readily perceived than objects [15] in addition to a more general global precedence effect as demonstrated by Navon [155]. As a whole, the behavioral studies that observe early access to general abstract quantities prior to more specific simple properties such as location seem to support an attentional architecture that consists of a hierarchical selection mechanism with higher visual areas orchestrating the overall selection process. Further evidence of this arrives in the form of studies that observe pop-out of high-level features such as depth from shading [184], facial expressions [164], 3D features [63], perceptual groups [19], surface planes [84], and parts and wholes [248]. As mentioned, the important property that many of these features may share is an efficient cortical representation. Furthermore, pop-out of simple features may be observed for features that occupy regions far greater than the receptive field size of cells in early visual areas. It is unclear then, how a pooled representation in the form of a saliency map mediating spatial selection can explain these behaviors unless one assumes that it comprises a pooled representation of activity from virtually every visual area. The circuitry required to implement AIM is consistent with the behavior of local surround suppression with the implication that surround suppression may subserve the local

modulation involved in saliency computation in line with recent suggestions [178]. The only requirement on the neurons involved is sparsity and it may be assumed that such computation may act throughout the visual cortex with localized saliency computation observed at every layer of the visual hierarchy in line with more general models of visual attention. There also exists considerable neurophysiological support in favor of this type of selection architecture. In particular the response of cells among early visual areas appears to be affected by attention at a relatively late time course relative to higher visual areas [135; 160; 195] and furthermore the early involvement of higher visual areas in attention related processing is consistent with accounts of object based attention [209; 215]. In a recent influential result, it was shown that focused attention gives rise to an inhibitory region surrounding the focus of attention [89]. This result is a prediction of a hierarchical selection architecture [228] along with the ability to attend to arbitrarily sized and shaped spatial regions [142]; these considerations elude explanation within the traditional saliency map paradigm in its current form and are more consistent with a distributed hierarchical selection strategy [105]. The preceding discussion serves to establish the generality of the proposal put forth by AIM. The portion of saliency computation that is of interest is the normalization or local gain control observed as a product of the context of a stimulus. This is an aspect of computation that is only a minor consideration within other models and accounted for based on a

crude or general mechanism within a normalization operation with only loose ties to visual circuitry [98].

8.6 Discussion

We have put forth a proposal for saliency computation within the visual cortex that is broadly compatible with more general models concerning how attention is achieved. In particular, the proposal serves to provide the missing link in observing pop-out behaviors that appear within models that posit a distributed strategy for attentional selection; a subset of attention models for which favorable evidence is mounting. The proposal is shown to agree with a broad range of psychophysical results and allows the additional possibility of simulating apparent high-level pop-out behaviors. The model demonstrates considerable efficacy in explaining fixation data for a qualitatively different data set than that considered in earlier chapters demonstrating the plausibility of a sampling strategy based on information seeking as put forth in this dissertation. We have also demonstrated the potential utility of an analytic basis in exploring additional issues pertaining to saliency computation highlighting the importance of consideration of scale-space and coding. Finally, it is stressed that although results of prior chapters are displayed in a manner consistent with traditional *saliency map* style models, this may be thought of as a summary of the average local gain for each spatial location. The estimate of saliency resides

at the level of a single cell and this signal may be employed locally for gain control allowing saliency related computation while preserving a hierarchical representation on which attentional selection operates. This should be seen as an important step in moving forwards towards a definition of saliency that operates throughout the cortex and in addition allows hierarchical selection in space and features along the lines of that put forth in distributed models of attentional selection [228] that demonstrate greater accord with recent imaging and psychophysics results.

9 AIM in Machine Vision

In the computer vision literature, the selection of points, and regions of interest as a front-end to various machine vision tasks is a domain that has reached some maturity. The majority of techniques in this area are designed with invariance to various forms of deformations including zoom, rotation, viewpoint changes, blur, illumination and noise as their central design criterion. It is less clear to what extent the points selected correspond to content of interest to an observer. It is of interest then, given the proposal put forth in the earlier chapters, to consider the extent to which AIM may be applied in the domain of selecting points or regions of interest and its stability in this regard. Favorable results from a stability perspective imply an operator with desirable dual properties of selecting salient points, in addition to such points being stable across various deformations and hence amenable to machine vision applications. The current chapter evaluates AIM within this context, presenting a comprehensive comparison of the model across various forms of deformations as compared with the state of the art in this

area from the machine vision literature.

9.1 Interest Operators

The focus on selection in a machine vision context involves the selection of a variety of local regions of varying scale. This set of locations may then be used in object recognition, robot navigation, scene classification or a variety of other tasks. The use of interest points (or regions) in computer vision actually dates back to 1965 [194] with early influential work on interest operators by Moravec following in 1979 [147]. A surge in the popularity of this approach in a variety of machine vision tasks has been seen in recent years.

Owing to the purpose for which these interest points are employed, the selection of regions/locations often proceeds to satisfy a variety of criteria. The most obvious element that is important is that in looking at an object/landmark from a variety of different viewpoints, it is desirable to have the same interest points selected provided such locations are visible. For this reason, a primary goal in producing algorithms that select interest points is invariance to rotation, scaling or affine transformations. Other important elements of interest points include the distinctiveness of the features underlying selection, the accuracy of localization, and that the selection criteria allow a sufficient number of keypoints to be detected.

There is a long history of detecting interest points with some early efforts focused

on choosing interest points based on the detection of corners, blobs or edgels. The focus has now shifted in favor of scale and affine invariant local features. Of these early approaches, one that appears frequently in the modern literature is the Harris corner detector [82]. The Harris detector chooses points on the basis of two large eigenvalues in the autocorrelation matrix. Although the invariance of the basic operator to rotation, noise and changes in illumination is good, the approach is not robust to scale changes. It is interesting to note that in general, corners are also established as salient regions according to the definition of AIM.

Various extensions of the Harris detector have been proposed that incorporate different approaches to including scale invariance. An early blob based approach proposed by Lindeberg involves detecting blobs by applying a Laplacian-of-Gaussian operator at several spatial scales [125]. Selection of the maximum across scale provides scale invariance and the circular symmetry of the operator implies rotation invariance.

There has been considerable work devoted to affine invariant operators. Generally such approaches are based on similar ideas to the scale invariant operators but with some assumptions relaxed. As with the scale invariant operators the basis for affine invariant operators is typically based on detection algorithms that look for corners [5; 229] or blobs [12; 126] but in an algorithm robust to affine deformation.

In this chapter we compare the performance of AIM against a variety of affine

invariant region detectors comprised of the state of the art in this area as demonstrated by the evaluation of Mikolajczyk et al. [141]. Performance is evaluated according to the same evaluation described in [141] which is described in the section that follows. The remainder of this section is devoted to providing a brief description of the algorithms involved in the comparison.

9.1.1 Maximally Stable Extremal Region Detector (MSER)

MSER is based on connected components established in the intensity domain. That is, for a thresholded image, one may establish regions based on connected sets of pixels for which a collection of local adjacent pixels is either darker or brighter than all pixels in its surround. The algorithm includes a means of optimal threshold selection which forms the basis for its success as an affine region detector. Given thresholding, a monotonic change in image intensity means that the regions are perfectly preserved. Geometric changes should also imply that pixels belonging to a single connected component remain so. The threshold may be chosen so that the change in the area of connected components relative to the threshold is minimized. That is MSER implies that regions of interest correspond to those for which conversion to a binary image is locally stable over a wide range of thresholds.

9.1.2 Intensity Extrema-Based Region Detector (IBR)

IBR is based on detecting intensity extrema that persist across scale. Points determined in this process are then bound to an elliptical region by exploring the surrounding structure. Specifically, given a local intensity extremum, rays projecting from the locus of the extremum are extracted and the intensity profile along such rays is considered according to the function:

$$f_I(t) = \frac{|I(t) - I_0|}{\max(\frac{\int_0^t |I(t)-I_0|dt}{t}, d)}$$

where $I(t)$ is the intensity at position t, I_0 is the intensity at the extremum and d a small constant. Maxima on rays originating from the extremum are used to bound an affine covariant region.

9.1.3 Harris Affine (HarAff) and Hessian Affine (HesAff) Detectors

The Harris and Hessian Affine detectors both borrow ideas from earlier corner and blob based detectors encompassing the behavior of the Harris corner detector [82] and Lindeberg's blob based selection [127] in a single selection strategy. The premise of these operators lies in maximizing the stability of image regions to various deformations. The two algorithms are in essence corner detectors based on Harris' suggestion of corners as invariant features under various forms of image transformations. Scale selection in each case is based on the Laplacian and the shape of the

associated elliptical region determined by observing the second moment matrix of the intensity gradient: $M = \mu(\mathbf{x}, \sigma_I, \sigma_D) = \sigma_D^2 g(\sigma_1) * \begin{bmatrix} I_x^2(\mathbf{x}, \sigma_D) & I_x I_y(\mathbf{x}, \sigma_D) \\ I_y I_x(\mathbf{x}, \sigma_D) & I_y^2(\mathbf{x}, \sigma_D) \end{bmatrix}$.

Image derivatives are computed locally based on a Gaussian kernel of scale σ_D. An average of derivatives over the local neighborhood is computed via convolution with a Gaussian with scale σ_I. A strong response implies signal strength across two orthogonal directions which implies stability across certain transformations. The Harris detector is based on this principle. An additional idea explored in the paper of Mikolajczyk et al. [141] is based on the Hessian matrix: $H = H(x, \sigma_D) = \begin{bmatrix} I_{xx}(\mathbf{x}, \sigma_D) & I_{xy}(\mathbf{x}, \sigma_D) \\ I_{xy}(\mathbf{x}, \sigma_D) & I_{yy}(\mathbf{x}, \sigma_D) \end{bmatrix}$

Second derivatives in the matrix tend to correspond to blobs and ridges. Scale selection is incorporated via the characteristic scale as described by Lindeberg. This is the scale at which similarity between the feature detection operator and local image structure is at a maximum. Given selection of characteristic scale, iterative estimation of elliptical affine regions is achieved according to the algorithm of Lindeberg and Garding [127]. For a more detailed description of the details of these elements, readers should refer to [127] and [141].

9.1.4 AIM adapted to select ROIs

As results demonstrated in earlier chapters typically depict the saliency associated with AIM as a continuous surface, it is necessary to describe its extension to selection regions of interest. The first point of note is that the basis derived from ICA training yields a representation that is inherently invariant albeit over a small sampling of scale space. The basis employed in experimental results within this chapter was derived from the Jade algorithm with PCA preprocessing retaining 95% variance. While some range of scale space is captured in the raw basis, this falls well short of the size of regions selected by the abovementioned algorithms. For this reason, AIM is applied using the 21x21 sized ICA basis with the image resampled at several spatial scales. (From 100% down to 25% by increments of 5%). At each spatial scale, non-maximal suppression is applied over a radius of 4 pixels and remaining extrema are recorded to establish a raw set of interest points. The resulting point set is ranked on the basis of associated saliency values at extremum points, and the points corresponding to the top 4.5% are kept for each image as the final set of interest points associated with that image. In the absence of specific analysis on the features resulting from the ICA basis, there is no simple means of assigning a dominant orientation in the absence of further analysis of local image properties. For this reason, a circular region of support is established and the ra-

dius of this region set to correspond to the spatial extent of the window of analysis for the scale under consideration. For example, for the full scale image the radius of a local region of support extends 10 pixels in each direction from the extremum. For the lowest scale, the radius associated with an ROI is 40 pixels. The result of this process is a set of interest points at several spatial scales corresponding to the regions that are determined to be most salient at that scale. It is worth noting that for viewpoint changes that involve considerable affine transformations of scene content, there is an inherent cap on the amount of overlap among selected regions and also a base error associated with overlap scores depending on the extent of affine deformation over various regions of the image. Nevertheless, this analysis should serve to allow assessment of the stability of AIM as applied at multiple spatial scales and subject to various forms of deformation as compared with the best affine invariant operators defined for this purpose.

Some simple qualitative results reveal that this method is sufficient to provide region of interest selection with dominant scales ranking highest as depicted by the regions appearing in figures 9.1 and 9.2. Figure 9.1 depicts the 8 most salient regions and indicates appropriate selection of scale for the salient items. Figure 9.2 shows the regions associated with a more typical application which gives some sense of the preservation of selection of similar features across scale and deformations in a single image albeit with some differences as the dolls are hand painted and differ

Figure 9.1: An example of appropriate selection of scale for regions of interest selected by the modified AIM algorithm.

in exact detail. Nevertheless this suggests some qualitative evidence of agreement in feature selection across scale as similar sections of the dolls heads are selected across scale.

9.2 Evaluation Methodology and Results

The image set employed for evaluation purposes is that used in the study of Mikolajczyk et al. [141] and consists of examples of 5 different types of changes in imaging conditions over 8 image sets. Among these are two examples of viewpoint changes, two examples of scale changes consisting of combined rotation and zoom,

Figure 9.2: An example of selection of a large number of regions at several scales for an image of dolls with similar features but of different sizes.

two examples of image blur, one example of illumination variation, and one example of jpeg compression. Examples of these images are shown in figure 9.3.

Viewpoint changes vary from a fronto-parallel view to a view at 60 degrees resulting in significant foreshortening. Scale and blur are the result of systematic changes in the zoom and focus of the camera. JPEG compression corresponds to the xv image utility and results from varying the quality parameter from 40 down to 2%. Finally, illumination variation corresponds to varying the aperture of the camera.

The main evaluation criterion as put forth by Mikolajczyk et al., is that of repeatability. That is, the extent to which regions appearing within one image, are repeated in a corresponding image subject to the associated affine transformation. Regions are considered to be corresponding if the overlap associated with two regions is sufficiently large. That is, if:

$$1 - \frac{R_{\mu_\alpha} \bigcap R_{H^T \mu_\beta H}}{R_{\mu_\alpha} \bigcup R_{H^T \mu_\beta H}} < \epsilon_0$$

where R_μ corresponds to the region defined by $x^T \mu x = 1$, and H the homography between the two images. To avoid bias based on region size, all regions are first normalized to a radius of 30. In the results shown by Mikolajczyk et al., repeatability rates are shown for an overlap error of 40% with the authors noting that even regions with a 50% overlap error can be matched successfully with a robust descriptor. Note that in the figures shown, the labels haraff, hesaff, mseraff,

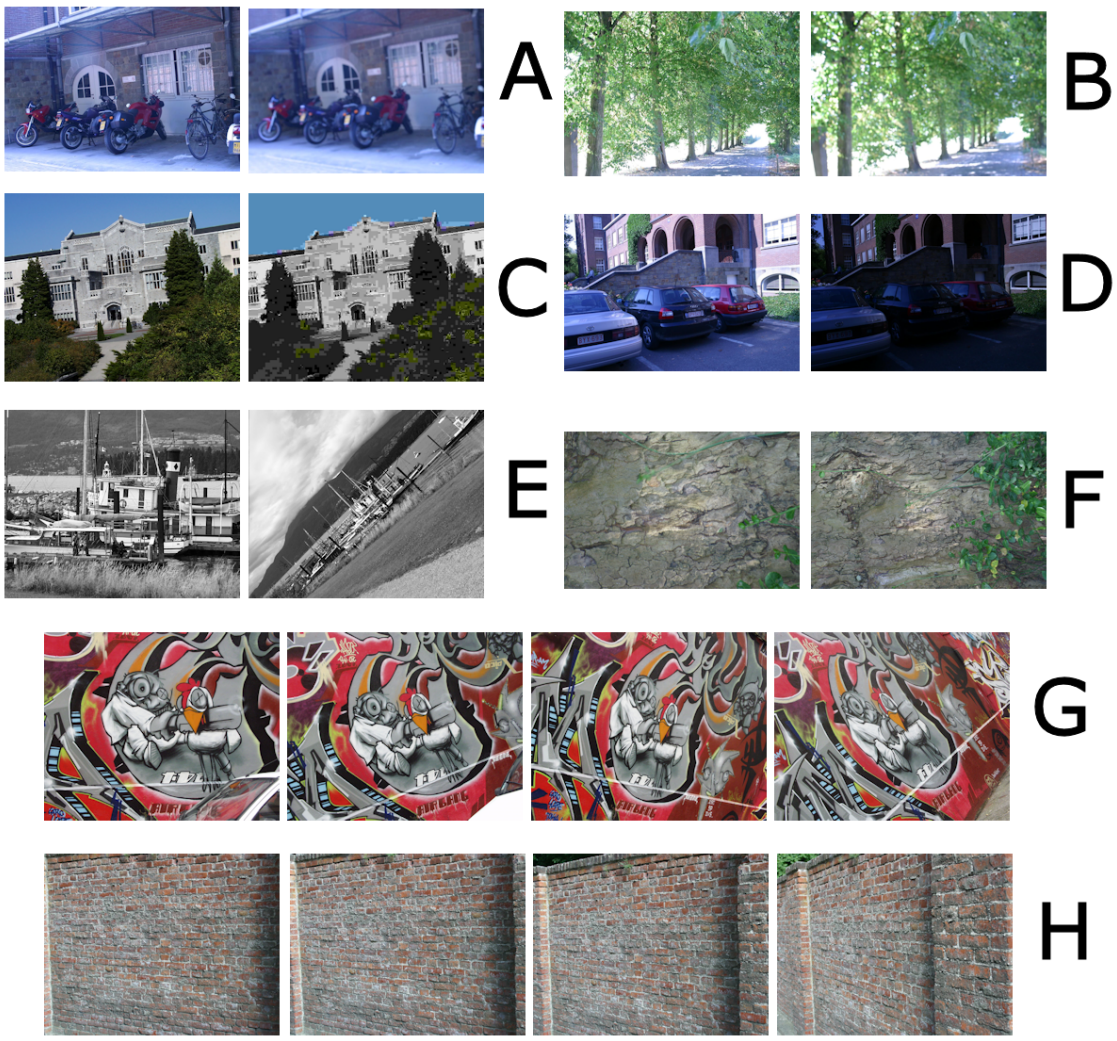

Figure 9.3: An example of the various transformations used for stability testing of the operators described in the previous section. Examples shown include blur (A: bikes, B: trees), JPEG compression (C: UBC), change in illumination (D: leuven), combined rotation and zoom (E: boat, F: bark), and change of viewpoint (G: graffiti, H: wall).

ebraff, and aimaff refer to the Harris Affine, Hessian Affine, MSER, EBR and AIM algorithms respectively.

Figures 9.4, 9.5, 9.6, 9.7, 9.8, 9.9, 9.10 and 9.11 demonstrate a comparison of AIM for the 40% overlap rate as compared with the algorithms described in the preceding section.

Figures 9.4, and 9.5 correspond to the bikes and trees images respectively and show the consistency of ROI selection across varying degrees of camera blur. As may be seen from these examples, the interest points selected by AIM exhibit a far greater degree of repeatability relative to all of the affine invariant operators.

Figure 9.6 corresponds to the UBC image and demonstrates repeatability scores associated with various degrees of jpeg compression. Once again, the AIM algorithm exhibits a very high degree of repeatability across all conditions of JPEG compression, retaining more than 95% of interest points for compression as extreme as 2% of original image size. In applications where image compression is common (e.g. search for trademark infringements on the web), the multiscale AIM selection algorithm may provide a natural means of obtaining interest points for matching.

Figure 9.7 corresponds to the leuven image and demonstrates scores associated with the change in illumination condition. In the case of varying illumination, the AIM algorithm is bested by the maximally stable extremal region algorithm

and to a small extent, by the Hessian-Affine algorithm. This is perhaps unsurprising as the MSER algorithm is by construction invariant to monotonic changes in image intensity. The AIM algorithm still elicits very reasonable performance in this domain, besting the oft used Harris-Affine algorithm. It is worth noting that the training set employed in constructing the ICA basis consists entirely of images captured during the daytime in plain light. It is possible that in training on a set that involves variation in illumination, that the AIM algorithm might become more robust to changes along this feature dimension. It is also worth noting, that one might achieve success via application of AIM for selection on an analytic basis that generalizes better under varying illumination conditions.

Figures 9.9 and 9.8 correspond to the boat and bark images that consist of examples of combined zoom and rotation. For the sequence involving the rotation and zoom of a texture (Figure 9.8) AIM proves superior for small deformations, while demonstrating slightly inferior repeatability to the Hessian-Affine algorithm across all other zoom conditions. This is a promising result as by construction AIM selects those regions that are most informative or salient while not explicitely seeking regions that are invariant to deformation. That being said, the nature of the basis itself has inherent invariance to such properties by virtue of its representation of orientation and scale space. In the case of figure 9.9 the results are very interesting. This image consists of a very large change in scale on a natural

scene. The repeatability rate is comparable to most of the other algorithms for the smallest zoom factor, but drops off severely in zooming out from the scene. This is almost certainly attributed to an important consideration that is currently absent from discussion of interest operators. At the closest scale, the boat fills much of the scene and many features on the boat itself are clearly visible and salient items. At the furthest degree of zoom, only the macroscopic features of the boat are clearly visible in the distance. While a persons face may be salient at a distance, a closer view may result in the eyes, nose or facial features becoming the salient items in the scene. This is an issue that is absent from the Affine-Invariant operators, but appears in the behavior of AIM by virtue of the premise underlying its construction. It is unclear whether this should be interpreted as a negative result since it is quite possible that while the overall repeatability of regions is reduced in the AIM result, there may actually be a greater correspondence of regions associated with items of interest at the current scale. In the case of recognizing objects, one may well be interested in only regions that correspond to salient items such as objects that are clearly visible at the scale under consideration. This line of reasoning receives support from the more robust repeatability associated with the textured (Bark) example, for which there is less change in the interpretation of the scene with change of zoom, and also a more uniform transformation of scale space.

Figures 9.10 and 9.11 correspond to the graffiti and wall images and comprise a

viewpoint change from frontoparallel to a view at approximately 60 degrees. The foreshortening in the graffiti example is sufficiently large that it is not possible for a circular region chosen in two views to match at a greater than 40% rate for the 50 and 60 degree angles. For this reason results in the graffiti condition are restricted to viewing angles of 10,20,30 and 40 degrees. The MSER algorithm exhibits the greatest performance in the case of the graffiti experiment, with the Hessian-Affine and AIM algorithms next in performance. It is worth noting that there is some baseline penalty associated with the selection of circular regions that increases with increases in viewpoint angle non-uniformly over the scene. It is foreseeable then that an AIM algorithm constructed in the manner described that includes further analysis to select elliptical regions may present the possiblity of trumping performance of the other algorithms across viewpoint. In the non-textured graffiti image case, there is once again the issue that the definition of what is salient changes across viewpoint: For example the appearance of grass or trees beyond the wall at oblique views might be interpreted as salient within the context in question. The important implication of this is that the raw repeatability scores should be taken with a grain of salt as these results may be misleading depending on the application in question. It is also perhaps worth mentioning that employing a likelihood estimate based on natural image statistics in general as opposed to those corresponding to a single scene may imply greater stability across these sorts of

deformations for the purposes of a task such as localization or viewpoint matching. In the case of the wall (textured example) the performance of AIM for smaller viewpoint changes is overwhelmingly better than its competitors. The inherent penalty associated with selection of circular regions is evident at larger viewpoint angles as demonstrated by the severe drop-off at the most severe deformation levels.

It is also instructive to demonstrate the nature of regions selected across deformation. To give some sense of the qualitative match among regions, figure 9.12 depicts a selection of regions for two different views of the graffiti data set for this purpose. Note that the parameters associated with these examples provide a greater spread and smaller number of regions than those depicted in the quantitative results for the purposes of exposition.

9.3 General Discussion

In either the case of a considerable zoom and rotation, or change of viewpoint, the AIM algorithm is average for the textured examples while dropping off significantly for more severe deformations in the case of structured natural scenes. This is as discussed a result of the change in profile of what is deemed salient at a particular scale across such deformations, which has little effect in the case of a textured scene. That being said, as discussed this is not necessarily a negative result as repeatability means little for a task such as object recognition if none of the regions correspond to

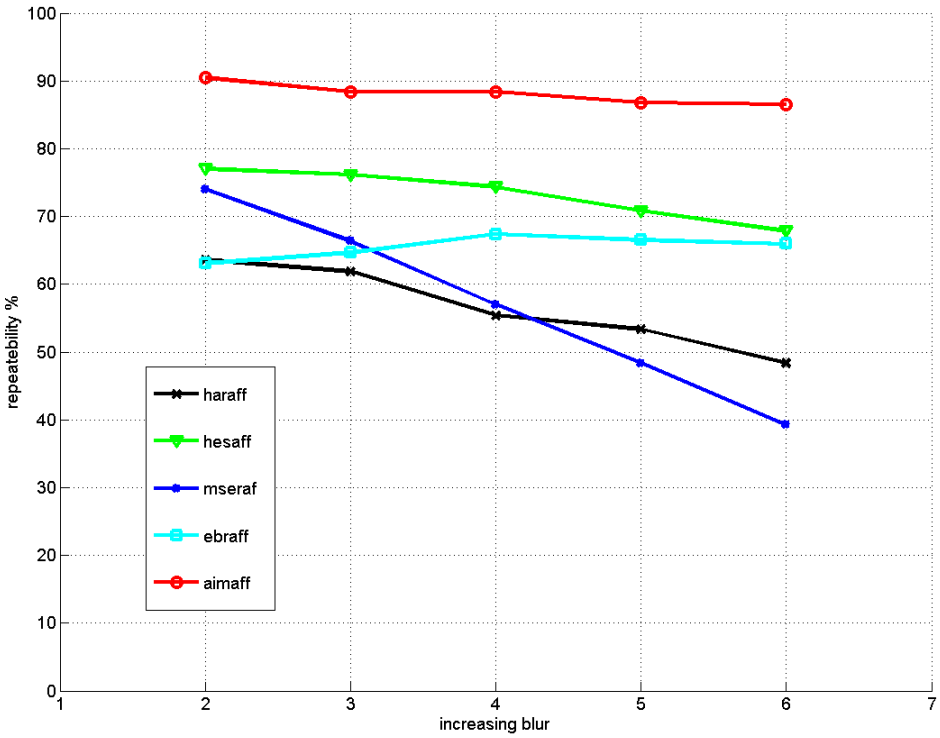

Figure 9.4: Repeatability scores for the bikes sequence which consists of varying degrees of blur.

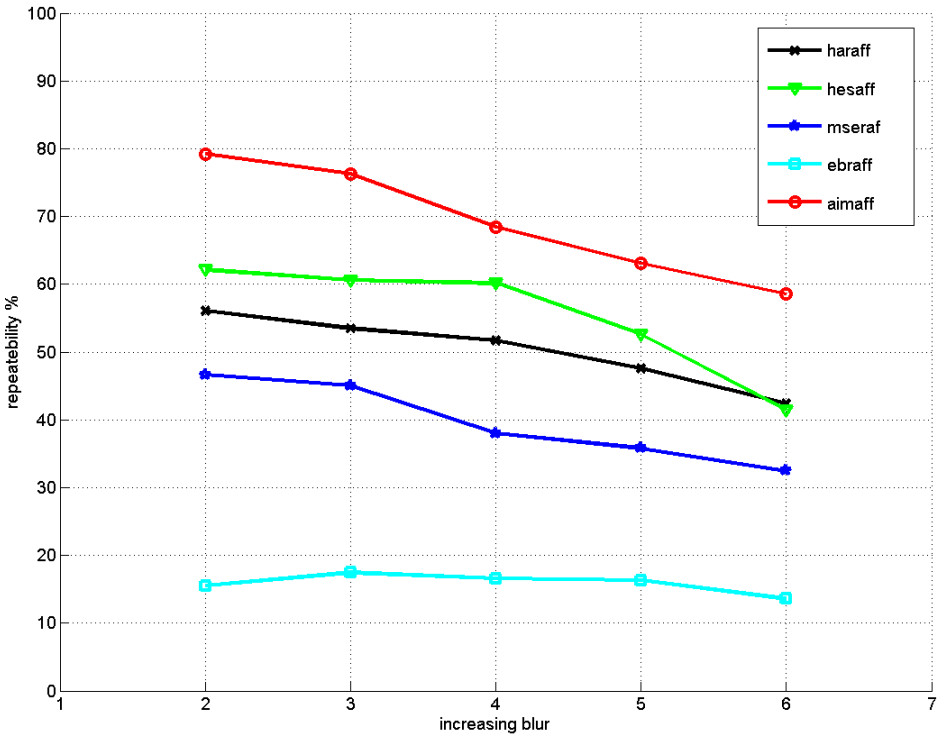

Figure 9.5: Repeatability scores for the trees sequence which consists of varying degrees of blur.

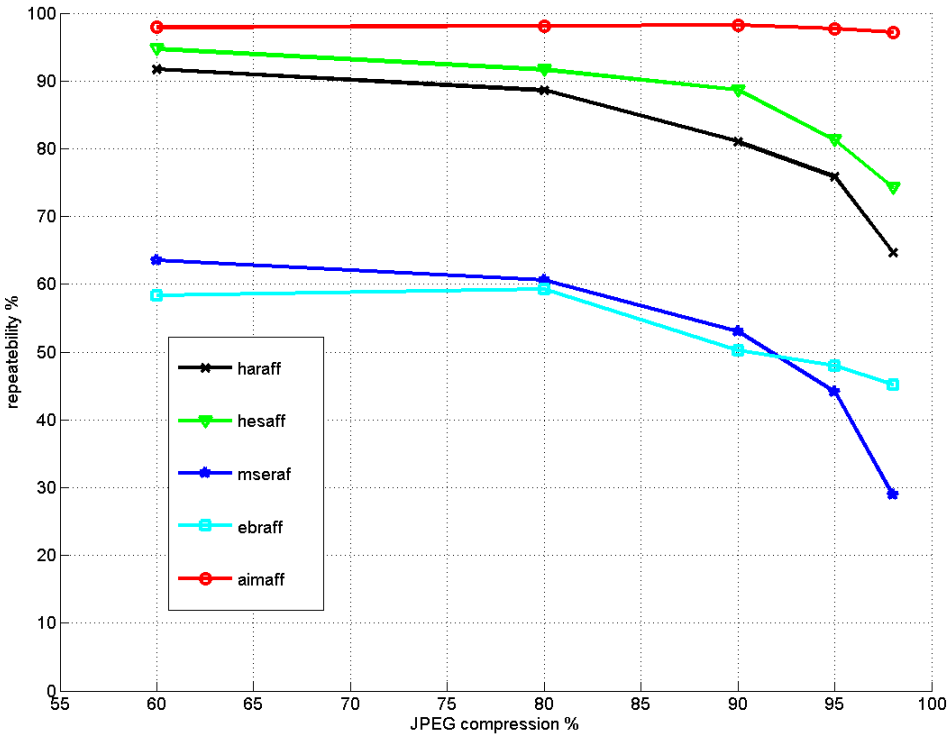

Figure 9.6: Repeatability scores for the UBC sequence which consists of varying degrees of JPEG compression.

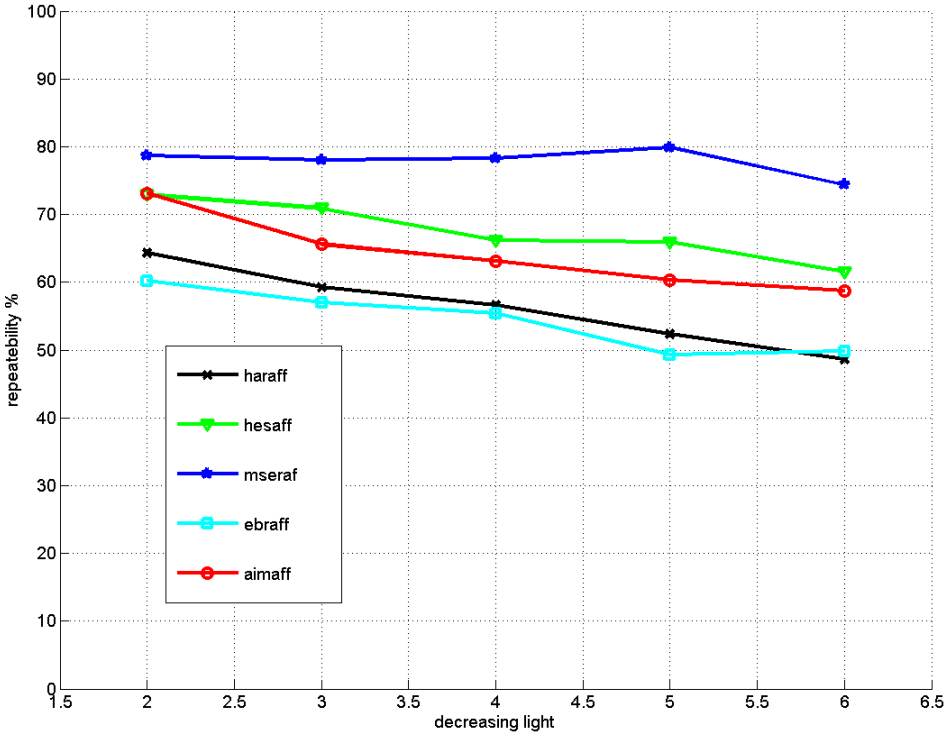

Figure 9.7: Repeatability scores for the leuven sequence which consists of varying illumination.

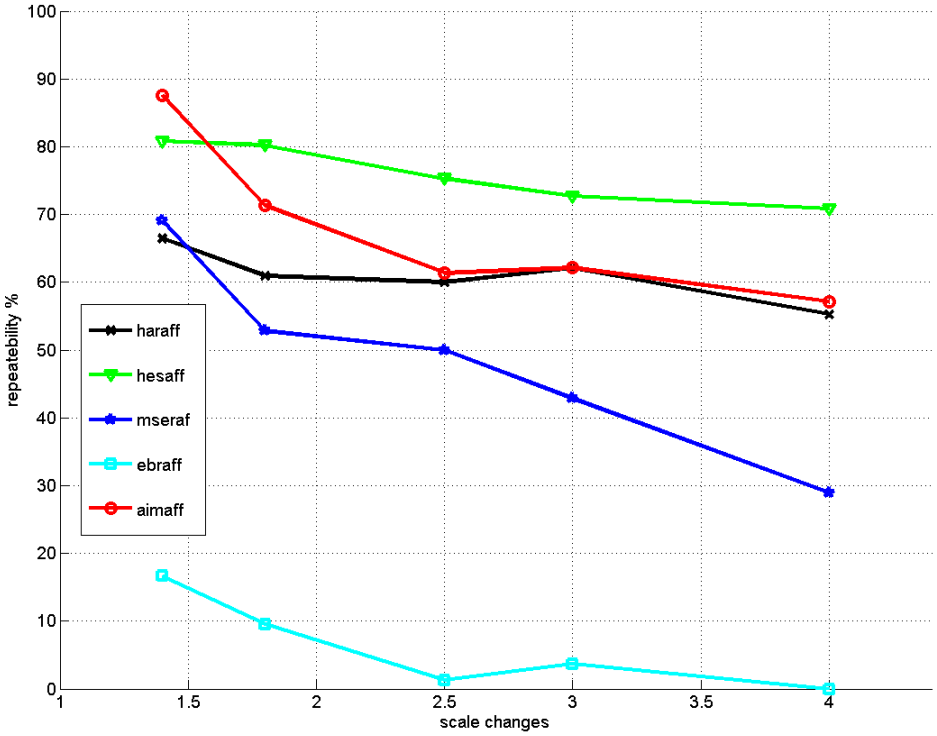

Figure 9.8: Repeatability scores for the bark sequence which consists of a combined rotation and zoom.

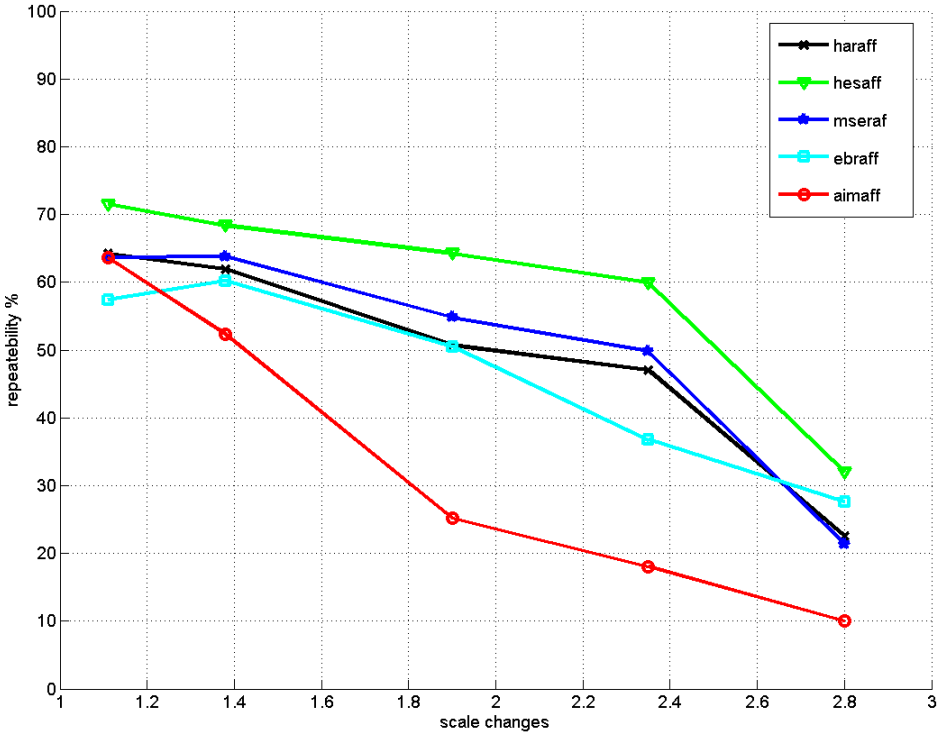

Figure 9.9: Repeatability scores for the boat sequence which consists of a combined rotation and zoom.

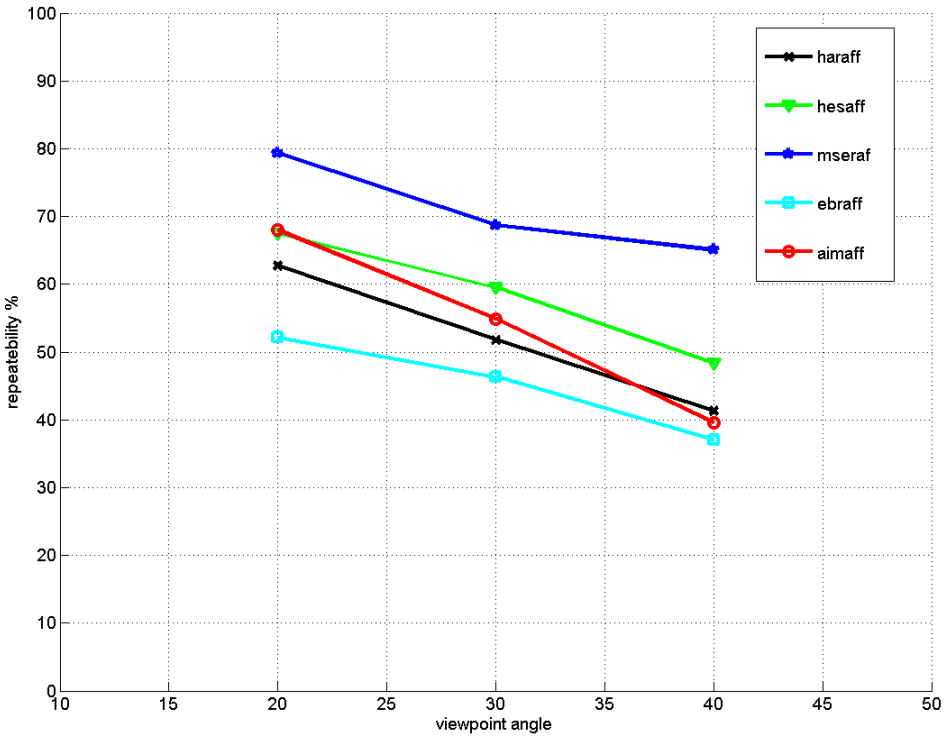

Figure 9.10: Repeatability scores for the graffiti sequence which consists of a large change in viewpoint angle.

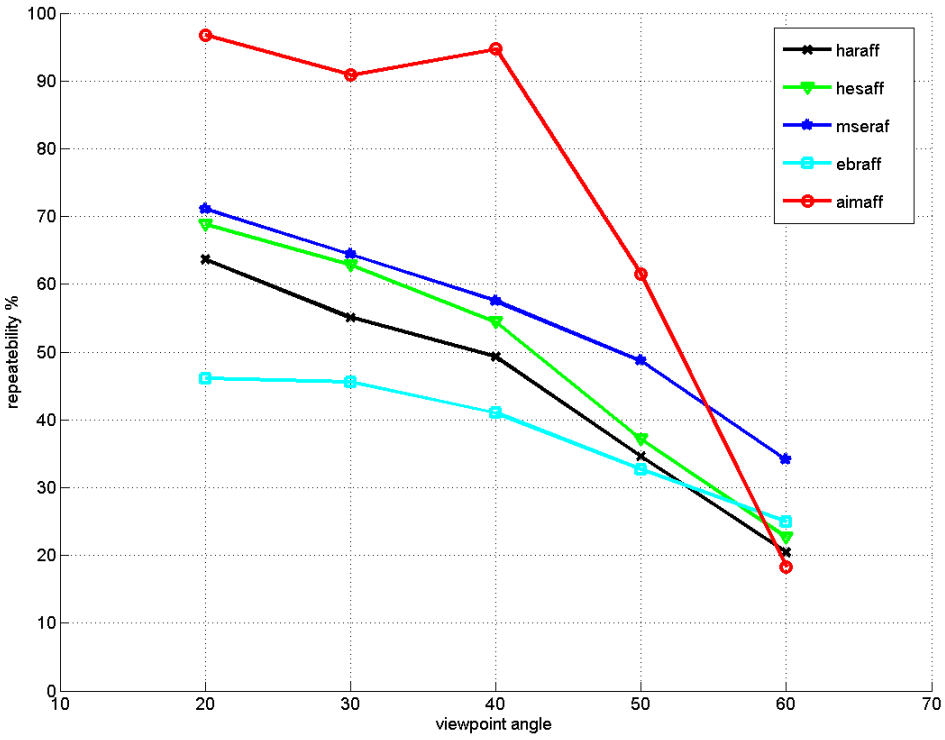

Figure 9.11: Repeatability scores for the wall sequence which consists of a large change in viewpoint angle.

Figure 9.12: Selection of regions associated with two different views of the graffiti example.

relevant objects at the current viewing scale. A remedy for this issue in the context of viewpoint changes may arrive by way of employing a general model of likelihoods based on the statistics of all natural images as opposed to those associated with a single scene for additional tasks such as viewpoint matching or localization. In all of the deformations that should not imply a significant change in what is salient or imply a strong penalty associated with circular as opposed to affine regions, AIM is a favorable choice. As a whole, the results demonstrate on the basis of the proposed algorithm, considerable applicability of the salience operator for the sake of selecting regions of interest, and also that with additional investigation, one may achieve superior performance across all experimental conditions including severe affine deformations. It would also be interesting to consider performance of the algorithms considered in the context of detection of objects for the sake of considering the applicability of regions selected to this task.

10 Conclusions and Future Directions

In this dissertation, various aspects of saliency computation were explored in the context of human vision and to a lesser extent, machine vision. The claim that saliency serves to maximize information sampled from one's environment in a bottom-up sense was put forth and evidence presented in this regard. In the section that follows, a summary of this investigation is put forth highlighting results and corresponding evidence.

10.1 Summary of Dissertation

The central premise of the body of work put forth in this dissertation, is that bottom-up saliency computation corresponds to a strategy that results in the propagation of informative signals up the visual hierarchy. Saliency is defined according to the Shannon self-information or surprisal associated with local content in a general sense. Comparison with eye tracking data demonstrates greater correlation with fixational eye movements than its predecessors, additionally offering the ad-

vantage of being built upon a principled definition. Analysis of various aspects of saliency computation highlight the role of receptive field size and scale space in the corresponding computation. It is also shown that the model may be implemented by way of neural circuitry exhibiting a close correspondence with observations concerning cortical surround inhibition. This applies to the nature of features involved, the spatial extent, the role of relative contrast and the emphasis on peripheral suppression observed in this computation. Further support for the model arrives in juxtaposing model behavior with a variety of classic visual search results from the behavioral psychophysics literature. It is demonstrated that a wide range of behaviors appear as emergent from the model, and the role of natural image statistics on this encoding is made clear. Discussion of more general architectures for attentional selection reveals that the model may be interpreted in a manner that allows its consideration in proposals that posit a distributed attentional selection strategy. It follows that various visual search behaviors that were heretofore lacking from the description of such distributed selection strategies, may be observed in a distributed paradigm along with other desirable behaviors that such models present. A further validation of the model in the context of predicting fixation behavior is put forth through consideration of eye tracking in a spatiotemporal context, demonstrating a greater agreement with the human data than existing models in this domain. Finally, the promise of this definition of saliency within a machine vision context

is considered, establishing the stability of the proposed saliency computation and offering a possibly promising component for various machine vision applications.

10.2 Future Directions

Many aspects of saliency computation have been considered for both biological and machine vision. That said, the basic framework put forth in this dissertation presents many promising avenues for future research efforts. The following outlines a variety of possible future directions of interest pertaining to this body of work.

10.2.1 Biological Vision

The most notable avenue for future research effort in the context of biological vision, is perhaps further consideration of the operation of saliency computation within a hierarchy comprised of simple features at early layers and more complex features associated with larger receptive fields and various forms of invariance at later layers. A strategy that consists of localized saliency related computation that results in the preservation of information above raw signal strength as visual input ascends the hierarchy may be of great benefit in establishing a general vision system capable of more complex vision tasks such as recognition.

A secondary consideration that is lacking from any existing biologically motivated vision system, is that of computation on a representation that is foveated.

This is a rather important consideration, and while the research presented within this dissertation employs relatively low resolution processing in an attempt to capture some element of this aspect of computation, it remains to be determined via a detailed analysis, what implications foveation has in regards to saliency related processing and the design of a vision system, either biological or for machine vision applications.

The claims made with regard to coding present a natural avenue for further investigation in order to establish the specific nature of surround inhibition in the cortex. Specifically it may be of interest to consider the role of infomax related local interactions at the level of the LGN and V1 as a means of distinguishing between cross-orientation and iso-orientation surround effects. An additional natural question in regards to the role of relative contrast lies in the determination of whether inhibition operates according to a sum of differences or a difference of sums according to the local neuronal responses.

It may also be interesting to consider the design of visual search paradigms on the basis of their predicted complexity as a means of further validating the model on the behavioral front.

10.2.2 Machine Vision

As the focus of this dissertation leans in favor of biological vision, there remains many possible avenues for future research efforts on the front of machine vision.

One possibly salient avenue in the case that visual content is coded within an invertible basis representation, is the use of AIM in a denoising or enhancement context. Modulation of cells within a basis representation based on their local context might allow a reconstruction that is perceptually superior, or less noisy than the image under consideration.

A second possible consideration in regards to machine vision, concerns the possibility of incorporating context into the evaluation of saliency related likelihoods. In chapter 6, it was noted that the *context* of a neuron need not be limited to its local spatial or spatiotemporal context, but might be considered on the basis of a more general definition incorporating natural statistics in general, or the statistics associated with a specific context such as a forest or urban environment. In regards to this possiblity, the work of Tishby et al. [216] may be especially relevant presenting a means of interpreting Shannon information in a manner that incorporates the notion of *relevance* of said information.

Lacking specific knowledge of feature properties in employing an ICA basis, detection of affine invariant regions is restricted to circular regions of support.

A natural extension of this work, would be consideration of features with known properties in order that a dominant orientation may be established. This may serve to further improve the results presented in chapter 9 and may also afford additional application specific advantages. It may also be possible to extend the methodology put forth in a manner specific to developing invariance to affine deformations, but the exact nature of such a process is less clear.

While the consideration of a hierarchical representation of visual content is of import, this consideration is also true in a machine vision context. An appropriate selection strategy combined with localized saliency computation within a visual hierarchy may be a considerable step towards the development of a system capable of achieving general machine vision.

Finally, there are many areas in which a more thorough analysis may be carried out in order to determine the impact of various properties on performance in the prediction of fixation points and also points of stability. A natural starting point may be consideration of different basis sets within this context.

As a whole, there are a great deal of starting points for promising research that emerge from the work put forth in this dissertation including avenues of interest to those involved in machine vision, computational neuroscience, psychophysics and neurobiology.

A Components of AIM

The following outlines the details of the various components depicted in figure 4.7.

Infomax ICA: A large number of local patches are randomly sampled from a set of 3600 natural images. Images are drawn from the Corel photo database and consist of a variety of photographs of outdoor natural scenes captured in a variety of countries. In total, 360,000 patches form the training set for ICA based on the random selection of 100 patches from each image. Infomax or Jade ICA is applied to the data in order to learn a sparse spatiochromatic basis. In the results shown in chapters 4 and 5 patches are learned based on 11x11, 21x21 and 31x31 windows.

Matrix Pseudoinverse: The pseudoinverse of the mixing matrix provides the unmixing matrix which may be used to separate the content within any local region into independent components. For each local neighbourhood one arrives at a set of coefficients that represents the relative contribution of the various basis functions in representing the local neighbourhood. The functions corresponding to the unmixing matrix resemble oriented Gabors and color opponent cells akin to those appearing

in V1.

Matrix Multiplication: The matrix product of any local neighbourhood with the unmixing matrix yields for each local observation a set of independent coefficients corresponding to the relative contribution of various oriented Gabor-like filters and color opponent type cells.

Density Estimation: For each local neighbourhood of the image, the product of a neighbourhood with the unmixing matrix yields a set of coefficients. Observing any individual coefficient independently in an area surrounding any given patch yields an estimate of the probability density function associated with the coefficient in question within the surrounding region. This allows an estimate of the likelihood associated with the coefficient value corresponding to the central patch.

Joint Likelihood: Any given coefficient may be readily converted to a probability by looking up its likelihood from the corresponding coefficient probability distribution. The product of all the individual likelihoods corresponding to a particular local region yields the joint likelihood.

Self-Information: The joint likelihood is translated into Shannon's measure of Self-Information by $-log(p(x)))$. The resulting information map depicts the Saliency attributed to each spatial location based on the aforementioned computation.

B Kullback-Leibler divergence

The Kullback-Leibler (KL) divergence is a principled measure of distance between two probability distributions P and Q [114; 115]. The term divergence rather than distance refers to the fact that the measure is not symmetric and thus $D_{KL}(P\|Q) \neq D_{KL}(Q\|P)$. Typically P is thought of as the true probability distribution and Q an approximation of P, with the KL-divergence affording a measure of the quality of the approximation. For probability densities P and Q, the Kullback-Leibler divergence of Q from P is given by:

$$D_{KL}(P\|Q) = \sum_i P(i) log(P(i)/Q(i))$$

for a distribution on a discrete random variable, and by

$$D_{KL}(P\|Q) = \int_{-\infty}^{\infty} P(x) log(P(x)/Q(x)) dx$$

for a distribution on a continuous random variable.

KL-divergence is related to a variety of other information theoretic measures. For example, the KL-divergence of a distribution p_i from the Kronecker delta rep-

resenting the relationship $i = m$ corresponds to self-information. That is

$$D_{KL}(\delta_{im} \| (p_i))$$

dictates the number of additional bits needed to identify i assuming the receiver has knowledge of p_i.

Similar relations between KL-divergence and other information theoretic measures such as entropy, mutual information and conditional entropy also exist.

C ROC Areas for Different Parameters on Spatiochromatic Data

The following appendix contains the raw results corresponding to an exploration of the parameter state space in the determination of spatiochromatic saliency. The basis indices referred to in the raw data correspond those listed in table C.1, organized in two columns showing the index and corresponding basis respectively. The parameters of each basis are given by the width, the algorithm (Jade or infomax) used for ICA training, and the variance explained as a numerator out of 1000. Following this are the ROC areas associated with the various parameter sets.

In the raw data, quantities associated with each parameter set are as follows:

1. AREA: Area under the ROC curve for all images with the parameters that follow.
2. resize: Proportion of original image size (681x511) via bicubic downsampling.
3. convolve: Convolved with a Gaussian modeling the dropoff in visual acuity to

Basis Index	[wid][alg][var]	Basis Index	[wid][alg][var]
1	31infomax900	12	21infomax990
2	31infomax950	13	11jade990
3	31jade900	14	11jade995
4	31jade950	15	11infomax900
5	11infomax999	16	21jade900
6	21infomax900	17	11infomax950
7	11jade900	18	21jade950
8	21infomax950	19	11infomax975
9	11jade950	20	21jade975
10	21infomax975	21	11infomax990
11	11jade975	22	11infomax995

Table C.1: Indices associated with the various learned basis sets referenced in the raw data. Columns containing a string correspond to the width of the window size, the algorithm used and the variance captured (/1000) respectively.

simulate clustering behavior. This is a binary quantity.

4. scaling: 1 if density estimation is on an independent scale for each feature type, 2 if density estimate is on the same scale for all features.

5. basis: The specific basis employed for this condition. Refer to table C.1.

AREA: 0.789602 resize: 0.25 convolve: 0 scaling: 1 basis: 12
AREA: 0.788785 resize: 0.25 convolve: 0 scaling: 2 basis: 12
AREA: 0.787520 resize: 0.25 convolve: 1 scaling: 1 basis: 12
AREA: 0.787355 resize: 0.25 convolve: 1 scaling: 2 basis: 12
AREA: 0.785635 resize: 0.25 convolve: 1 scaling: 2 basis: 10
AREA: 0.785399 resize: 0.25 convolve: 1 scaling: 1 basis: 10
AREA: 0.782930 resize: 0.25 convolve: 1 scaling: 2 basis: 20
AREA: 0.782537 resize: 0.25 convolve: 0 scaling: 1 basis: 10
AREA: 0.782122 resize: 0.50 convolve: 1 scaling: 1 basis: 2
AREA: 0.782108 resize: 0.50 convolve: 1 scaling: 2 basis: 2
AREA: 0.782059 resize: 0.25 convolve: 1 scaling: 1 basis: 20
AREA: 0.781626 resize: 0.25 convolve: 0 scaling: 2 basis: 10
AREA: 0.780675 resize: 0.25 convolve: 1 scaling: 2 basis: 2
AREA: 0.780611 resize: 0.50 convolve: 1 scaling: 2 basis: 4
AREA: 0.780476 resize: 0.50 convolve: 1 scaling: 1 basis: 4
AREA: 0.779513 resize: 0.25 convolve: 1 scaling: 1 basis: 2
AREA: 0.778958 resize: 0.25 convolve: 1 scaling: 2 basis: 4
AREA: 0.777305 resize: 0.25 convolve: 1 scaling: 1 basis: 4
AREA: 0.777284 resize: 0.25 convolve: 0 scaling: 1 basis: 20
AREA: 0.777084 resize: 0.25 convolve: 0 scaling: 2 basis: 2
AREA: 0.776768 resize: 0.25 convolve: 0 scaling: 1 basis: 2
AREA: 0.776447 resize: 0.25 convolve: 0 scaling: 2 basis: 20
AREA: 0.774282 resize: 0.50 convolve: 0 scaling: 1 basis: 2
AREA: 0.773077 resize: 0.25 convolve: 1 scaling: 2 basis: 21
AREA: 0.772723 resize: 0.25 convolve: 0 scaling: 1 basis: 4
AREA: 0.772696 resize: 0.50 convolve: 0 scaling: 2 basis: 2
AREA: 0.772451 resize: 0.25 convolve: 1 scaling: 1 basis: 21
AREA: 0.772421 resize: 0.25 convolve: 0 scaling: 2 basis: 4
AREA: 0.772003 resize: 0.25 convolve: 1 scaling: 2 basis: 13
AREA: 0.771163 resize: 0.25 convolve: 1 scaling: 1 basis: 13
AREA: 0.770698 resize: 0.25 convolve: 1 scaling: 2 basis: 8
AREA: 0.770653 resize: 0.50 convolve: 0 scaling: 1 basis: 4
AREA: 0.770086 resize: 0.25 convolve: 1 scaling: 2 basis: 18
AREA: 0.769795 resize: 0.25 convolve: 1 scaling: 2 basis: 22
AREA: 0.769191 resize: 0.50 convolve: 0 scaling: 2 basis: 4
AREA: 0.769159 resize: 0.25 convolve: 1 scaling: 1 basis: 8
AREA: 0.769142 resize: 0.25 convolve: 1 scaling: 1 basis: 22
AREA: 0.768848 resize: 0.50 convolve: 1 scaling: 2 basis: 8

AREA: 0.768647 resize: 0.25 convolve: 1 scaling: 1 basis: 18
AREA: 0.768388 resize: 0.50 convolve: 1 scaling: 2 basis: 10
AREA: 0.768299 resize: 0.25 convolve: 1 scaling: 2 basis: 19
AREA: 0.768257 resize: 0.50 convolve: 1 scaling: 1 basis: 8
AREA: 0.768188 resize: 0.50 convolve: 1 scaling: 2 basis: 18
AREA: 0.768148 resize: 0.50 convolve: 1 scaling: 1 basis: 10
AREA: 0.767980 resize: 0.25 convolve: 1 scaling: 2 basis: 11
AREA: 0.767681 resize: 0.50 convolve: 1 scaling: 1 basis: 18
AREA: 0.767591 resize: 0.50 convolve: 1 scaling: 2 basis: 20
AREA: 0.767398 resize: 0.25 convolve: 1 scaling: 2 basis: 14
AREA: 0.767318 resize: 0.50 convolve: 1 scaling: 1 basis: 20
AREA: 0.767153 resize: 0.25 convolve: 1 scaling: 1 basis: 19
AREA: 0.767087 resize: 0.25 convolve: 1 scaling: 1 basis: 14
AREA: 0.766990 resize: 0.25 convolve: 1 scaling: 1 basis: 11
AREA: 0.764797 resize: 0.25 convolve: 1 scaling: 2 basis: 5
AREA: 0.764692 resize: 0.25 convolve: 1 scaling: 1 basis: 5
AREA: 0.763804 resize: 0.25 convolve: 0 scaling: 1 basis: 22
AREA: 0.763729 resize: 0.25 convolve: 0 scaling: 1 basis: 5
AREA: 0.762278 resize: 0.25 convolve: 0 scaling: 1 basis: 21
AREA: 0.762035 resize: 0.25 convolve: 0 scaling: 2 basis: 22
AREA: 0.761972 resize: 0.25 convolve: 0 scaling: 1 basis: 8
AREA: 0.761452 resize: 0.25 convolve: 0 scaling: 2 basis: 5
AREA: 0.760921 resize: 0.25 convolve: 0 scaling: 2 basis: 8
AREA: 0.760897 resize: 0.25 convolve: 0 scaling: 1 basis: 13
AREA: 0.760665 resize: 0.50 convolve: 0 scaling: 1 basis: 10
AREA: 0.760149 resize: 1.00 convolve: 1 scaling: 2 basis: 1
AREA: 0.760129 resize: 0.25 convolve: 0 scaling: 1 basis: 18
AREA: 0.760070 resize: 0.25 convolve: 0 scaling: 2 basis: 21
AREA: 0.759965 resize: 1.00 convolve: 1 scaling: 2 basis: 2
AREA: 0.759923 resize: 0.50 convolve: 1 scaling: 2 basis: 1
AREA: 0.759693 resize: 1.00 convolve: 1 scaling: 1 basis: 1
AREA: 0.759682 resize: 1.00 convolve: 1 scaling: 2 basis: 4
AREA: 0.759611 resize: 1.00 convolve: 1 scaling: 1 basis: 2
AREA: 0.759496 resize: 1.00 convolve: 1 scaling: 1 basis: 4
AREA: 0.759484 resize: 0.50 convolve: 1 scaling: 2 basis: 3
AREA: 0.759481 resize: 1.00 convolve: 1 scaling: 2 basis: 3
AREA: 0.759179 resize: 0.50 convolve: 1 scaling: 1 basis: 1
AREA: 0.759137 resize: 0.25 convolve: 0 scaling: 2 basis: 13
AREA: 0.758984 resize: 0.50 convolve: 0 scaling: 2 basis: 10
AREA: 0.758964 resize: 0.25 convolve: 0 scaling: 2 basis: 18
AREA: 0.758853 resize: 1.00 convolve: 1 scaling: 1 basis: 3
AREA: 0.758746 resize: 0.50 convolve: 1 scaling: 1 basis: 3
AREA: 0.758632 resize: 0.25 convolve: 0 scaling: 1 basis: 14
AREA: 0.758569 resize: 0.50 convolve: 1 scaling: 1 basis: 12
AREA: 0.758201 resize: 0.50 convolve: 0 scaling: 1 basis: 12
AREA: 0.758073 resize: 0.50 convolve: 0 scaling: 1 basis: 20
AREA: 0.758061 resize: 0.50 convolve: 1 scaling: 2 basis: 12

AREA: 0.757429 resize: 0.25 convolve: 0 scaling: 2 basis: 14
AREA: 0.756800 resize: 0.50 convolve: 1 scaling: 2 basis: 11
AREA: 0.756779 resize: 0.50 convolve: 1 scaling: 2 basis: 19
AREA: 0.756306 resize: 0.50 convolve: 1 scaling: 1 basis: 19
AREA: 0.756161 resize: 0.50 convolve: 1 scaling: 1 basis: 11
AREA: 0.756065 resize: 0.50 convolve: 0 scaling: 2 basis: 12
AREA: 0.756016 resize: 0.50 convolve: 0 scaling: 2 basis: 20
AREA: 0.755003 resize: 0.50 convolve: 1 scaling: 2 basis: 17
AREA: 0.754899 resize: 0.25 convolve: 1 scaling: 2 basis: 9
AREA: 0.754598 resize: 0.25 convolve: 1 scaling: 2 basis: 17
AREA: 0.754302 resize: 0.50 convolve: 1 scaling: 2 basis: 9
AREA: 0.753916 resize: 0.25 convolve: 1 scaling: 2 basis: 1
AREA: 0.753878 resize: 0.50 convolve: 1 scaling: 1 basis: 17
AREA: 0.753706 resize: 0.25 convolve: 1 scaling: 1 basis: 9
AREA: 0.753551 resize: 0.25 convolve: 1 scaling: 1 basis: 17
AREA: 0.753544 resize: 0.50 convolve: 1 scaling: 1 basis: 9
AREA: 0.753272 resize: 0.25 convolve: 1 scaling: 2 basis: 3
AREA: 0.752974 resize: 1.00 convolve: 1 scaling: 2 basis: 8
AREA: 0.752860 resize: 1.00 convolve: 1 scaling: 2 basis: 18
AREA: 0.752379 resize: 1.00 convolve: 1 scaling: 1 basis: 8
AREA: 0.752285 resize: 1.00 convolve: 1 scaling: 1 basis: 18
AREA: 0.752271 resize: 0.50 convolve: 0 scaling: 1 basis: 8
AREA: 0.752074 resize: 0.25 convolve: 0 scaling: 1 basis: 19
AREA: 0.751636 resize: 0.25 convolve: 1 scaling: 1 basis: 1
AREA: 0.751525 resize: 1.00 convolve: 0 scaling: 1 basis: 2
AREA: 0.751416 resize: 0.50 convolve: 0 scaling: 1 basis: 18
AREA: 0.751391 resize: 0.25 convolve: 0 scaling: 1 basis: 11
AREA: 0.750890 resize: 0.25 convolve: 1 scaling: 1 basis: 3
AREA: 0.750137 resize: 1.00 convolve: 1 scaling: 2 basis: 6
AREA: 0.749987 resize: 0.50 convolve: 0 scaling: 2 basis: 8
AREA: 0.749927 resize: 1.00 convolve: 0 scaling: 1 basis: 4
AREA: 0.749819 resize: 1.00 convolve: 1 scaling: 2 basis: 16
AREA: 0.749647 resize: 0.25 convolve: 0 scaling: 2 basis: 11
AREA: 0.749562 resize: 1.00 convolve: 0 scaling: 2 basis: 2
AREA: 0.749501 resize: 0.25 convolve: 0 scaling: 2 basis: 19
AREA: 0.749257 resize: 1.00 convolve: 1 scaling: 1 basis: 6
AREA: 0.749237 resize: 0.50 convolve: 0 scaling: 2 basis: 18
AREA: 0.748950 resize: 1.00 convolve: 1 scaling: 1 basis: 16
AREA: 0.747708 resize: 0.50 convolve: 1 scaling: 2 basis: 13
AREA: 0.747624 resize: 1.00 convolve: 1 scaling: 2 basis: 7
AREA: 0.747595 resize: 1.00 convolve: 0 scaling: 2 basis: 4
AREA: 0.747565 resize: 0.25 convolve: 0 scaling: 1 basis: 1
AREA: 0.747422 resize: 0.50 convolve: 1 scaling: 2 basis: 21
AREA: 0.747411 resize: 1.00 convolve: 1 scaling: 2 basis: 15
AREA: 0.747373 resize: 0.50 convolve: 1 scaling: 1 basis: 21
AREA: 0.747317 resize: 0.50 convolve: 1 scaling: 2 basis: 6
AREA: 0.747253 resize: 0.25 convolve: 0 scaling: 2 basis: 1

AREA: 0.747233 resize: 1.00 convolve: 1 scaling: 1 basis: 7
AREA: 0.747198 resize: 0.50 convolve: 1 scaling: 2 basis: 16
AREA: 0.747131 resize: 0.50 convolve: 1 scaling: 1 basis: 13
AREA: 0.746830 resize: 1.00 convolve: 1 scaling: 1 basis: 15
AREA: 0.746205 resize: 0.50 convolve: 1 scaling: 1 basis: 16
AREA: 0.746076 resize: 0.25 convolve: 0 scaling: 1 basis: 3
AREA: 0.745946 resize: 0.50 convolve: 1 scaling: 1 basis: 6
AREA: 0.745531 resize: 0.25 convolve: 0 scaling: 2 basis: 3
AREA: 0.743623 resize: 0.50 convolve: 0 scaling: 1 basis: 1
AREA: 0.743243 resize: 0.50 convolve: 1 scaling: 2 basis: 7
AREA: 0.742547 resize: 0.50 convolve: 0 scaling: 1 basis: 3
AREA: 0.742524 resize: 0.50 convolve: 1 scaling: 2 basis: 15
AREA: 0.742186 resize: 0.50 convolve: 1 scaling: 1 basis: 7
AREA: 0.741991 resize: 0.50 convolve: 1 scaling: 1 basis: 15
AREA: 0.741609 resize: 0.50 convolve: 0 scaling: 2 basis: 1
AREA: 0.741163 resize: 0.25 convolve: 1 scaling: 2 basis: 16
AREA: 0.741116 resize: 0.25 convolve: 1 scaling: 2 basis: 6
AREA: 0.741024 resize: 1.00 convolve: 1 scaling: 2 basis: 17
AREA: 0.740714 resize: 0.50 convolve: 0 scaling: 2 basis: 3
AREA: 0.740661 resize: 1.00 convolve: 1 scaling: 2 basis: 9
AREA: 0.740318 resize: 1.00 convolve: 1 scaling: 1 basis: 17
AREA: 0.740274 resize: 1.00 convolve: 1 scaling: 2 basis: 20
AREA: 0.739804 resize: 1.00 convolve: 1 scaling: 1 basis: 20
AREA: 0.739648 resize: 1.00 convolve: 1 scaling: 1 basis: 9
AREA: 0.739553 resize: 0.25 convolve: 1 scaling: 1 basis: 6
AREA: 0.739535 resize: 0.25 convolve: 1 scaling: 1 basis: 16
AREA: 0.739211 resize: 0.50 convolve: 1 scaling: 2 basis: 14
AREA: 0.739062 resize: 0.50 convolve: 1 scaling: 2 basis: 22
AREA: 0.739026 resize: 1.00 convolve: 1 scaling: 2 basis: 10
AREA: 0.738744 resize: 0.50 convolve: 1 scaling: 1 basis: 14
AREA: 0.738521 resize: 1.00 convolve: 1 scaling: 1 basis: 10
AREA: 0.738442 resize: 0.50 convolve: 1 scaling: 1 basis: 22
AREA: 0.737737 resize: 1.00 convolve: 0 scaling: 1 basis: 1
AREA: 0.736418 resize: 1.00 convolve: 0 scaling: 1 basis: 3
AREA: 0.734786 resize: 1.00 convolve: 0 scaling: 2 basis: 1
AREA: 0.733983 resize: 1.00 convolve: 0 scaling: 2 basis: 3
AREA: 0.733076 resize: 0.25 convolve: 1 scaling: 2 basis: 7
AREA: 0.732781 resize: 1.00 convolve: 1 scaling: 2 basis: 11
AREA: 0.732482 resize: 1.00 convolve: 1 scaling: 2 basis: 19
AREA: 0.732369 resize: 0.25 convolve: 1 scaling: 2 basis: 15
AREA: 0.732296 resize: 0.25 convolve: 1 scaling: 1 basis: 7
AREA: 0.732036 resize: 1.00 convolve: 1 scaling: 1 basis: 19
AREA: 0.731884 resize: 1.00 convolve: 1 scaling: 1 basis: 11
AREA: 0.731806 resize: 0.25 convolve: 0 scaling: 1 basis: 17
AREA: 0.731330 resize: 0.50 convolve: 1 scaling: 2 basis: 5
AREA: 0.731244 resize: 0.25 convolve: 1 scaling: 1 basis: 15
AREA: 0.731110 resize: 0.25 convolve: 0 scaling: 1 basis: 9

AREA: 0.730543 resize: 0.50 convolve: 1 scaling: 1 basis: 5
AREA: 0.730322 resize: 0.50 convolve: 0 scaling: 1 basis: 19
AREA: 0.729843 resize: 0.50 convolve: 0 scaling: 1 basis: 11
AREA: 0.729649 resize: 0.50 convolve: 0 scaling: 1 basis: 21
AREA: 0.728732 resize: 0.50 convolve: 0 scaling: 1 basis: 13
AREA: 0.727833 resize: 0.25 convolve: 0 scaling: 1 basis: 6
AREA: 0.727802 resize: 0.25 convolve: 0 scaling: 2 basis: 9
AREA: 0.727771 resize: 1.00 convolve: 0 scaling: 1 basis: 18
AREA: 0.727715 resize: 0.50 convolve: 0 scaling: 2 basis: 19
AREA: 0.727700 resize: 1.00 convolve: 0 scaling: 1 basis: 8
AREA: 0.727277 resize: 0.25 convolve: 0 scaling: 2 basis: 17
AREA: 0.727019 resize: 0.50 convolve: 0 scaling: 2 basis: 11
AREA: 0.727005 resize: 0.50 convolve: 0 scaling: 2 basis: 21
AREA: 0.726861 resize: 0.25 convolve: 0 scaling: 1 basis: 16
AREA: 0.726467 resize: 0.50 convolve: 0 scaling: 2 basis: 13
AREA: 0.725943 resize: 0.25 convolve: 0 scaling: 2 basis: 6
AREA: 0.725210 resize: 1.00 convolve: 0 scaling: 2 basis: 18
AREA: 0.725146 resize: 1.00 convolve: 0 scaling: 2 basis: 8
AREA: 0.725023 resize: 0.50 convolve: 0 scaling: 1 basis: 22
AREA: 0.724569 resize: 1.00 convolve: 1 scaling: 2 basis: 12
AREA: 0.724495 resize: 0.25 convolve: 0 scaling: 2 basis: 16
AREA: 0.724102 resize: 1.00 convolve: 1 scaling: 1 basis: 12
AREA: 0.723031 resize: 0.50 convolve: 0 scaling: 1 basis: 14
AREA: 0.722990 resize: 0.50 convolve: 0 scaling: 1 basis: 6
AREA: 0.722738 resize: 1.00 convolve: 0 scaling: 1 basis: 10
AREA: 0.722629 resize: 0.50 convolve: 0 scaling: 2 basis: 22
AREA: 0.722430 resize: 1.00 convolve: 0 scaling: 1 basis: 20
AREA: 0.722307 resize: 0.50 convolve: 0 scaling: 1 basis: 16
AREA: 0.720982 resize: 0.50 convolve: 0 scaling: 1 basis: 5
AREA: 0.720625 resize: 0.50 convolve: 0 scaling: 2 basis: 14
AREA: 0.720126 resize: 1.00 convolve: 0 scaling: 2 basis: 20
AREA: 0.720091 resize: 1.00 convolve: 0 scaling: 2 basis: 10
AREA: 0.719642 resize: 0.50 convolve: 0 scaling: 1 basis: 17
AREA: 0.719415 resize: 0.50 convolve: 0 scaling: 1 basis: 9
AREA: 0.719200 resize: 1.00 convolve: 1 scaling: 2 basis: 13
AREA: 0.718839 resize: 0.50 convolve: 0 scaling: 2 basis: 6
AREA: 0.718788 resize: 1.00 convolve: 1 scaling: 2 basis: 21
AREA: 0.718418 resize: 1.00 convolve: 1 scaling: 1 basis: 13
AREA: 0.718347 resize: 0.50 convolve: 0 scaling: 2 basis: 16
AREA: 0.718334 resize: 0.50 convolve: 0 scaling: 2 basis: 5
AREA: 0.717762 resize: 1.00 convolve: 1 scaling: 1 basis: 21
AREA: 0.715829 resize: 0.50 convolve: 0 scaling: 2 basis: 9
AREA: 0.715761 resize: 0.50 convolve: 0 scaling: 2 basis: 17
AREA: 0.715082 resize: 1.00 convolve: 0 scaling: 1 basis: 12
AREA: 0.713733 resize: 1.00 convolve: 1 scaling: 2 basis: 14
AREA: 0.712555 resize: 1.00 convolve: 1 scaling: 2 basis: 22
AREA: 0.712544 resize: 1.00 convolve: 1 scaling: 1 basis: 14

AREA: 0.712476 resize: 1.00 convolve: 0 scaling: 2 basis: 12
AREA: 0.712016 resize: 1.00 convolve: 0 scaling: 1 basis: 16
AREA: 0.711860 resize: 1.00 convolve: 0 scaling: 1 basis: 6
AREA: 0.711359 resize: 1.00 convolve: 1 scaling: 1 basis: 22
AREA: 0.710016 resize: 1.00 convolve: 1 scaling: 2 basis: 5
AREA: 0.709067 resize: 0.25 convolve: 0 scaling: 1 basis: 7
AREA: 0.708838 resize: 1.00 convolve: 1 scaling: 1 basis: 5
AREA: 0.708805 resize: 0.25 convolve: 0 scaling: 1 basis: 15
AREA: 0.707771 resize: 1.00 convolve: 0 scaling: 2 basis: 16
AREA: 0.707229 resize: 1.00 convolve: 0 scaling: 2 basis: 6
AREA: 0.704829 resize: 0.25 convolve: 0 scaling: 2 basis: 7
AREA: 0.703827 resize: 0.25 convolve: 0 scaling: 2 basis: 15
AREA: 0.702021 resize: 0.50 convolve: 0 scaling: 1 basis: 15
AREA: 0.701501 resize: 0.50 convolve: 0 scaling: 1 basis: 7
AREA: 0.697054 resize: 0.50 convolve: 0 scaling: 2 basis: 7
AREA: 0.696877 resize: 0.50 convolve: 0 scaling: 2 basis: 15
AREA: 0.696188 resize: 1.00 convolve: 0 scaling: 1 basis: 17
AREA: 0.695392 resize: 1.00 convolve: 0 scaling: 1 basis: 9
AREA: 0.694133 resize: 1.00 convolve: 0 scaling: 1 basis: 19
AREA: 0.693837 resize: 1.00 convolve: 0 scaling: 1 basis: 11
AREA: 0.691657 resize: 1.00 convolve: 0 scaling: 2 basis: 17
AREA: 0.691156 resize: 1.00 convolve: 0 scaling: 2 basis: 9
AREA: 0.691145 resize: 1.00 convolve: 0 scaling: 2 basis: 19
AREA: 0.690956 resize: 1.00 convolve: 0 scaling: 2 basis: 11
AREA: 0.689389 resize: 1.00 convolve: 0 scaling: 1 basis: 15
AREA: 0.688779 resize: 1.00 convolve: 0 scaling: 1 basis: 7
AREA: 0.687758 resize: 1.00 convolve: 0 scaling: 1 basis: 21
AREA: 0.686888 resize: 1.00 convolve: 0 scaling: 1 basis: 13
AREA: 0.685376 resize: 1.00 convolve: 0 scaling: 1 basis: 22
AREA: 0.684926 resize: 1.00 convolve: 0 scaling: 2 basis: 21
AREA: 0.684657 resize: 1.00 convolve: 0 scaling: 1 basis: 14
AREA: 0.684451 resize: 1.00 convolve: 0 scaling: 2 basis: 13
AREA: 0.684367 resize: 1.00 convolve: 0 scaling: 2 basis: 7
AREA: 0.684259 resize: 1.00 convolve: 0 scaling: 1 basis: 5
AREA: 0.683811 resize: 1.00 convolve: 0 scaling: 2 basis: 15
AREA: 0.682738 resize: 1.00 convolve: 0 scaling: 2 basis: 22
AREA: 0.682295 resize: 1.00 convolve: 0 scaling: 2 basis: 5
AREA: 0.681919 resize: 1.00 convolve: 0 scaling: 2 basis: 14
AREA: 0.751636 resize: 0.25 convolve: 1 scaling: 1 basis: 1
AREA: 0.779513 resize: 0.25 convolve: 1 scaling: 1 basis: 2
AREA: 0.750890 resize: 0.25 convolve: 1 scaling: 1 basis: 3
AREA: 0.777305 resize: 0.25 convolve: 1 scaling: 1 basis: 4
AREA: 0.764692 resize: 0.25 convolve: 1 scaling: 1 basis: 5
AREA: 0.739553 resize: 0.25 convolve: 1 scaling: 1 basis: 6
AREA: 0.732296 resize: 0.25 convolve: 1 scaling: 1 basis: 7
AREA: 0.769159 resize: 0.25 convolve: 1 scaling: 1 basis: 8
AREA: 0.753706 resize: 0.25 convolve: 1 scaling: 1 basis: 9

AREA: 0.785399 resize: 0.25 convolve: 1 scaling: 1 basis: 10
AREA: 0.766990 resize: 0.25 convolve: 1 scaling: 1 basis: 11
AREA: 0.787520 resize: 0.25 convolve: 1 scaling: 1 basis: 12
AREA: 0.771163 resize: 0.25 convolve: 1 scaling: 1 basis: 13
AREA: 0.767087 resize: 0.25 convolve: 1 scaling: 1 basis: 14
AREA: 0.731244 resize: 0.25 convolve: 1 scaling: 1 basis: 15
AREA: 0.739535 resize: 0.25 convolve: 1 scaling: 1 basis: 16
AREA: 0.753551 resize: 0.25 convolve: 1 scaling: 1 basis: 17
AREA: 0.768647 resize: 0.25 convolve: 1 scaling: 1 basis: 18
AREA: 0.767153 resize: 0.25 convolve: 1 scaling: 1 basis: 19
AREA: 0.782059 resize: 0.25 convolve: 1 scaling: 1 basis: 20
AREA: 0.772451 resize: 0.25 convolve: 1 scaling: 1 basis: 21
AREA: 0.769142 resize: 0.25 convolve: 1 scaling: 1 basis: 22
AREA: 0.753916 resize: 0.25 convolve: 1 scaling: 2 basis: 1
AREA: 0.780675 resize: 0.25 convolve: 1 scaling: 2 basis: 2
AREA: 0.753272 resize: 0.25 convolve: 1 scaling: 2 basis: 3
AREA: 0.778958 resize: 0.25 convolve: 1 scaling: 2 basis: 4
AREA: 0.764797 resize: 0.25 convolve: 1 scaling: 2 basis: 5
AREA: 0.741116 resize: 0.25 convolve: 1 scaling: 2 basis: 6
AREA: 0.733076 resize: 0.25 convolve: 1 scaling: 2 basis: 7
AREA: 0.770698 resize: 0.25 convolve: 1 scaling: 2 basis: 8
AREA: 0.754899 resize: 0.25 convolve: 1 scaling: 2 basis: 9
AREA: 0.785635 resize: 0.25 convolve: 1 scaling: 2 basis: 10
AREA: 0.767980 resize: 0.25 convolve: 1 scaling: 2 basis: 11
AREA: 0.787355 resize: 0.25 convolve: 1 scaling: 2 basis: 12
AREA: 0.772003 resize: 0.25 convolve: 1 scaling: 2 basis: 13
AREA: 0.767398 resize: 0.25 convolve: 1 scaling: 2 basis: 14
AREA: 0.732369 resize: 0.25 convolve: 1 scaling: 2 basis: 15
AREA: 0.741163 resize: 0.25 convolve: 1 scaling: 2 basis: 16
AREA: 0.754598 resize: 0.25 convolve: 1 scaling: 2 basis: 17
AREA: 0.770086 resize: 0.25 convolve: 1 scaling: 2 basis: 18
AREA: 0.768299 resize: 0.25 convolve: 1 scaling: 2 basis: 19
AREA: 0.782930 resize: 0.25 convolve: 1 scaling: 2 basis: 20
AREA: 0.773077 resize: 0.25 convolve: 1 scaling: 2 basis: 21
AREA: 0.769795 resize: 0.25 convolve: 1 scaling: 2 basis: 22
AREA: 0.747565 resize: 0.25 convolve: 0 scaling: 1 basis: 1
AREA: 0.776768 resize: 0.25 convolve: 0 scaling: 1 basis: 2
AREA: 0.746076 resize: 0.25 convolve: 0 scaling: 1 basis: 3
AREA: 0.772723 resize: 0.25 convolve: 0 scaling: 1 basis: 4
AREA: 0.763729 resize: 0.25 convolve: 0 scaling: 1 basis: 5
AREA: 0.727833 resize: 0.25 convolve: 0 scaling: 1 basis: 6
AREA: 0.709067 resize: 0.25 convolve: 0 scaling: 1 basis: 7
AREA: 0.761972 resize: 0.25 convolve: 0 scaling: 1 basis: 8
AREA: 0.731110 resize: 0.25 convolve: 0 scaling: 1 basis: 9
AREA: 0.782537 resize: 0.25 convolve: 0 scaling: 1 basis: 10
AREA: 0.751391 resize: 0.25 convolve: 0 scaling: 1 basis: 11
AREA: 0.789602 resize: 0.25 convolve: 0 scaling: 1 basis: 12

AREA: 0.760897 resize: 0.25 convolve: 0 scaling: 1 basis: 13
AREA: 0.758632 resize: 0.25 convolve: 0 scaling: 1 basis: 14
AREA: 0.708805 resize: 0.25 convolve: 0 scaling: 1 basis: 15
AREA: 0.726861 resize: 0.25 convolve: 0 scaling: 1 basis: 16
AREA: 0.731806 resize: 0.25 convolve: 0 scaling: 1 basis: 17
AREA: 0.760129 resize: 0.25 convolve: 0 scaling: 1 basis: 18
AREA: 0.752074 resize: 0.25 convolve: 0 scaling: 1 basis: 19
AREA: 0.777284 resize: 0.25 convolve: 0 scaling: 1 basis: 20
AREA: 0.762278 resize: 0.25 convolve: 0 scaling: 1 basis: 21
AREA: 0.763804 resize: 0.25 convolve: 0 scaling: 1 basis: 22
AREA: 0.747253 resize: 0.25 convolve: 0 scaling: 2 basis: 1
AREA: 0.777084 resize: 0.25 convolve: 0 scaling: 2 basis: 2
AREA: 0.745531 resize: 0.25 convolve: 0 scaling: 2 basis: 3
AREA: 0.772421 resize: 0.25 convolve: 0 scaling: 2 basis: 4
AREA: 0.761452 resize: 0.25 convolve: 0 scaling: 2 basis: 5
AREA: 0.725943 resize: 0.25 convolve: 0 scaling: 2 basis: 6
AREA: 0.704829 resize: 0.25 convolve: 0 scaling: 2 basis: 7
AREA: 0.760921 resize: 0.25 convolve: 0 scaling: 2 basis: 8
AREA: 0.727802 resize: 0.25 convolve: 0 scaling: 2 basis: 9
AREA: 0.781626 resize: 0.25 convolve: 0 scaling: 2 basis: 10
AREA: 0.749647 resize: 0.25 convolve: 0 scaling: 2 basis: 11
AREA: 0.788785 resize: 0.25 convolve: 0 scaling: 2 basis: 12
AREA: 0.759137 resize: 0.25 convolve: 0 scaling: 2 basis: 13
AREA: 0.757429 resize: 0.25 convolve: 0 scaling: 2 basis: 14
AREA: 0.703827 resize: 0.25 convolve: 0 scaling: 2 basis: 15
AREA: 0.724495 resize: 0.25 convolve: 0 scaling: 2 basis: 16
AREA: 0.727277 resize: 0.25 convolve: 0 scaling: 2 basis: 17
AREA: 0.758964 resize: 0.25 convolve: 0 scaling: 2 basis: 18
AREA: 0.749501 resize: 0.25 convolve: 0 scaling: 2 basis: 19
AREA: 0.776447 resize: 0.25 convolve: 0 scaling: 2 basis: 20
AREA: 0.760070 resize: 0.25 convolve: 0 scaling: 2 basis: 21
AREA: 0.762035 resize: 0.25 convolve: 0 scaling: 2 basis: 22
AREA: 0.759179 resize: 0.50 convolve: 1 scaling: 1 basis: 1
AREA: 0.782122 resize: 0.50 convolve: 1 scaling: 1 basis: 2
AREA: 0.758746 resize: 0.50 convolve: 1 scaling: 1 basis: 3
AREA: 0.780476 resize: 0.50 convolve: 1 scaling: 1 basis: 4
AREA: 0.730543 resize: 0.50 convolve: 1 scaling: 1 basis: 5
AREA: 0.745946 resize: 0.50 convolve: 1 scaling: 1 basis: 6
AREA: 0.742186 resize: 0.50 convolve: 1 scaling: 1 basis: 7
AREA: 0.768257 resize: 0.50 convolve: 1 scaling: 1 basis: 8
AREA: 0.753544 resize: 0.50 convolve: 1 scaling: 1 basis: 9
AREA: 0.768148 resize: 0.50 convolve: 1 scaling: 1 basis: 10
AREA: 0.756161 resize: 0.50 convolve: 1 scaling: 1 basis: 11
AREA: 0.758569 resize: 0.50 convolve: 1 scaling: 1 basis: 12
AREA: 0.747131 resize: 0.50 convolve: 1 scaling: 1 basis: 13
AREA: 0.738744 resize: 0.50 convolve: 1 scaling: 1 basis: 14
AREA: 0.741991 resize: 0.50 convolve: 1 scaling: 1 basis: 15

AREA: 0.746205 resize: 0.50 convolve: 1 scaling: 1 basis: 16
AREA: 0.753878 resize: 0.50 convolve: 1 scaling: 1 basis: 17
AREA: 0.767681 resize: 0.50 convolve: 1 scaling: 1 basis: 18
AREA: 0.756306 resize: 0.50 convolve: 1 scaling: 1 basis: 19
AREA: 0.767318 resize: 0.50 convolve: 1 scaling: 1 basis: 20
AREA: 0.747373 resize: 0.50 convolve: 1 scaling: 1 basis: 21
AREA: 0.738442 resize: 0.50 convolve: 1 scaling: 1 basis: 22
AREA: 0.759923 resize: 0.50 convolve: 1 scaling: 2 basis: 1
AREA: 0.782108 resize: 0.50 convolve: 1 scaling: 2 basis: 2
AREA: 0.759484 resize: 0.50 convolve: 1 scaling: 2 basis: 3
AREA: 0.780611 resize: 0.50 convolve: 1 scaling: 2 basis: 4
AREA: 0.731330 resize: 0.50 convolve: 1 scaling: 2 basis: 5
AREA: 0.747317 resize: 0.50 convolve: 1 scaling: 2 basis: 6
AREA: 0.743243 resize: 0.50 convolve: 1 scaling: 2 basis: 7
AREA: 0.768848 resize: 0.50 convolve: 1 scaling: 2 basis: 8
AREA: 0.754302 resize: 0.50 convolve: 1 scaling: 2 basis: 9
AREA: 0.768388 resize: 0.50 convolve: 1 scaling: 2 basis: 10
AREA: 0.756800 resize: 0.50 convolve: 1 scaling: 2 basis: 11
AREA: 0.758061 resize: 0.50 convolve: 1 scaling: 2 basis: 12
AREA: 0.747708 resize: 0.50 convolve: 1 scaling: 2 basis: 13
AREA: 0.739211 resize: 0.50 convolve: 1 scaling: 2 basis: 14
AREA: 0.742524 resize: 0.50 convolve: 1 scaling: 2 basis: 15
AREA: 0.747198 resize: 0.50 convolve: 1 scaling: 2 basis: 16
AREA: 0.755003 resize: 0.50 convolve: 1 scaling: 2 basis: 17
AREA: 0.768188 resize: 0.50 convolve: 1 scaling: 2 basis: 18
AREA: 0.756779 resize: 0.50 convolve: 1 scaling: 2 basis: 19
AREA: 0.767591 resize: 0.50 convolve: 1 scaling: 2 basis: 20
AREA: 0.747422 resize: 0.50 convolve: 1 scaling: 2 basis: 21
AREA: 0.739062 resize: 0.50 convolve: 1 scaling: 2 basis: 22
AREA: 0.743623 resize: 0.50 convolve: 0 scaling: 1 basis: 1
AREA: 0.774282 resize: 0.50 convolve: 0 scaling: 1 basis: 2
AREA: 0.742547 resize: 0.50 convolve: 0 scaling: 1 basis: 3
AREA: 0.770653 resize: 0.50 convolve: 0 scaling: 1 basis: 4
AREA: 0.720982 resize: 0.50 convolve: 0 scaling: 1 basis: 5
AREA: 0.722990 resize: 0.50 convolve: 0 scaling: 1 basis: 6
AREA: 0.701501 resize: 0.50 convolve: 0 scaling: 1 basis: 7
AREA: 0.752271 resize: 0.50 convolve: 0 scaling: 1 basis: 8
AREA: 0.719415 resize: 0.50 convolve: 0 scaling: 1 basis: 9
AREA: 0.760665 resize: 0.50 convolve: 0 scaling: 1 basis: 10
AREA: 0.729843 resize: 0.50 convolve: 0 scaling: 1 basis: 11
AREA: 0.758201 resize: 0.50 convolve: 0 scaling: 1 basis: 12
AREA: 0.728732 resize: 0.50 convolve: 0 scaling: 1 basis: 13
AREA: 0.723031 resize: 0.50 convolve: 0 scaling: 1 basis: 14
AREA: 0.702021 resize: 0.50 convolve: 0 scaling: 1 basis: 15
AREA: 0.722307 resize: 0.50 convolve: 0 scaling: 1 basis: 16
AREA: 0.719642 resize: 0.50 convolve: 0 scaling: 1 basis: 17
AREA: 0.751416 resize: 0.50 convolve: 0 scaling: 1 basis: 18

AREA: 0.730322 resize: 0.50 convolve: 0 scaling: 1 basis: 19
AREA: 0.758073 resize: 0.50 convolve: 0 scaling: 1 basis: 20
AREA: 0.729649 resize: 0.50 convolve: 0 scaling: 1 basis: 21
AREA: 0.725023 resize: 0.50 convolve: 0 scaling: 1 basis: 22
AREA: 0.741609 resize: 0.50 convolve: 0 scaling: 2 basis: 1
AREA: 0.772696 resize: 0.50 convolve: 0 scaling: 2 basis: 2
AREA: 0.740714 resize: 0.50 convolve: 0 scaling: 2 basis: 3
AREA: 0.769191 resize: 0.50 convolve: 0 scaling: 2 basis: 4
AREA: 0.718334 resize: 0.50 convolve: 0 scaling: 2 basis: 5
AREA: 0.718839 resize: 0.50 convolve: 0 scaling: 2 basis: 6
AREA: 0.697054 resize: 0.50 convolve: 0 scaling: 2 basis: 7
AREA: 0.749987 resize: 0.50 convolve: 0 scaling: 2 basis: 8
AREA: 0.715829 resize: 0.50 convolve: 0 scaling: 2 basis: 9
AREA: 0.758984 resize: 0.50 convolve: 0 scaling: 2 basis: 10
AREA: 0.727019 resize: 0.50 convolve: 0 scaling: 2 basis: 11
AREA: 0.756065 resize: 0.50 convolve: 0 scaling: 2 basis: 12
AREA: 0.726467 resize: 0.50 convolve: 0 scaling: 2 basis: 13
AREA: 0.720625 resize: 0.50 convolve: 0 scaling: 2 basis: 14
AREA: 0.696877 resize: 0.50 convolve: 0 scaling: 2 basis: 15
AREA: 0.718347 resize: 0.50 convolve: 0 scaling: 2 basis: 16
AREA: 0.715761 resize: 0.50 convolve: 0 scaling: 2 basis: 17
AREA: 0.749237 resize: 0.50 convolve: 0 scaling: 2 basis: 18
AREA: 0.727715 resize: 0.50 convolve: 0 scaling: 2 basis: 19
AREA: 0.756016 resize: 0.50 convolve: 0 scaling: 2 basis: 20
AREA: 0.727005 resize: 0.50 convolve: 0 scaling: 2 basis: 21
AREA: 0.722629 resize: 0.50 convolve: 0 scaling: 2 basis: 22
AREA: 0.759693 resize: 1.00 convolve: 1 scaling: 1 basis: 1
AREA: 0.759611 resize: 1.00 convolve: 1 scaling: 1 basis: 2
AREA: 0.758853 resize: 1.00 convolve: 1 scaling: 1 basis: 3
AREA: 0.759496 resize: 1.00 convolve: 1 scaling: 1 basis: 4
AREA: 0.708838 resize: 1.00 convolve: 1 scaling: 1 basis: 5
AREA: 0.749257 resize: 1.00 convolve: 1 scaling: 1 basis: 6
AREA: 0.747233 resize: 1.00 convolve: 1 scaling: 1 basis: 7
AREA: 0.752379 resize: 1.00 convolve: 1 scaling: 1 basis: 8
AREA: 0.739648 resize: 1.00 convolve: 1 scaling: 1 basis: 9
AREA: 0.738521 resize: 1.00 convolve: 1 scaling: 1 basis: 10
AREA: 0.731884 resize: 1.00 convolve: 1 scaling: 1 basis: 11
AREA: 0.724102 resize: 1.00 convolve: 1 scaling: 1 basis: 12
AREA: 0.718418 resize: 1.00 convolve: 1 scaling: 1 basis: 13
AREA: 0.712544 resize: 1.00 convolve: 1 scaling: 1 basis: 14
AREA: 0.746830 resize: 1.00 convolve: 1 scaling: 1 basis: 15
AREA: 0.748950 resize: 1.00 convolve: 1 scaling: 1 basis: 16
AREA: 0.740318 resize: 1.00 convolve: 1 scaling: 1 basis: 17
AREA: 0.752285 resize: 1.00 convolve: 1 scaling: 1 basis: 18
AREA: 0.732036 resize: 1.00 convolve: 1 scaling: 1 basis: 19
AREA: 0.739804 resize: 1.00 convolve: 1 scaling: 1 basis: 20
AREA: 0.717762 resize: 1.00 convolve: 1 scaling: 1 basis: 21

AREA: 0.711359 resize: 1.00 convolve: 1 scaling: 1 basis: 22
AREA: 0.760149 resize: 1.00 convolve: 1 scaling: 2 basis: 1
AREA: 0.759965 resize: 1.00 convolve: 1 scaling: 2 basis: 2
AREA: 0.759481 resize: 1.00 convolve: 1 scaling: 2 basis: 3
AREA: 0.759682 resize: 1.00 convolve: 1 scaling: 2 basis: 4
AREA: 0.710016 resize: 1.00 convolve: 1 scaling: 2 basis: 5
AREA: 0.750137 resize: 1.00 convolve: 1 scaling: 2 basis: 6
AREA: 0.747624 resize: 1.00 convolve: 1 scaling: 2 basis: 7
AREA: 0.752974 resize: 1.00 convolve: 1 scaling: 2 basis: 8
AREA: 0.740661 resize: 1.00 convolve: 1 scaling: 2 basis: 9
AREA: 0.739026 resize: 1.00 convolve: 1 scaling: 2 basis: 10
AREA: 0.732781 resize: 1.00 convolve: 1 scaling: 2 basis: 11
AREA: 0.724569 resize: 1.00 convolve: 1 scaling: 2 basis: 12
AREA: 0.719200 resize: 1.00 convolve: 1 scaling: 2 basis: 13
AREA: 0.713733 resize: 1.00 convolve: 1 scaling: 2 basis: 14
AREA: 0.747411 resize: 1.00 convolve: 1 scaling: 2 basis: 15
AREA: 0.749819 resize: 1.00 convolve: 1 scaling: 2 basis: 16
AREA: 0.741024 resize: 1.00 convolve: 1 scaling: 2 basis: 17
AREA: 0.752860 resize: 1.00 convolve: 1 scaling: 2 basis: 18
AREA: 0.732482 resize: 1.00 convolve: 1 scaling: 2 basis: 19
AREA: 0.740274 resize: 1.00 convolve: 1 scaling: 2 basis: 20
AREA: 0.718788 resize: 1.00 convolve: 1 scaling: 2 basis: 21
AREA: 0.712555 resize: 1.00 convolve: 1 scaling: 2 basis: 22
AREA: 0.737737 resize: 1.00 convolve: 0 scaling: 1 basis: 1
AREA: 0.751525 resize: 1.00 convolve: 0 scaling: 1 basis: 2
AREA: 0.736418 resize: 1.00 convolve: 0 scaling: 1 basis: 3
AREA: 0.749927 resize: 1.00 convolve: 0 scaling: 1 basis: 4
AREA: 0.684259 resize: 1.00 convolve: 0 scaling: 1 basis: 5
AREA: 0.711860 resize: 1.00 convolve: 0 scaling: 1 basis: 6
AREA: 0.688779 resize: 1.00 convolve: 0 scaling: 1 basis: 7
AREA: 0.727700 resize: 1.00 convolve: 0 scaling: 1 basis: 8
AREA: 0.695392 resize: 1.00 convolve: 0 scaling: 1 basis: 9
AREA: 0.722738 resize: 1.00 convolve: 0 scaling: 1 basis: 10
AREA: 0.693837 resize: 1.00 convolve: 0 scaling: 1 basis: 11
AREA: 0.715082 resize: 1.00 convolve: 0 scaling: 1 basis: 12
AREA: 0.686888 resize: 1.00 convolve: 0 scaling: 1 basis: 13
AREA: 0.684657 resize: 1.00 convolve: 0 scaling: 1 basis: 14
AREA: 0.689389 resize: 1.00 convolve: 0 scaling: 1 basis: 15
AREA: 0.712016 resize: 1.00 convolve: 0 scaling: 1 basis: 16
AREA: 0.696188 resize: 1.00 convolve: 0 scaling: 1 basis: 17
AREA: 0.727771 resize: 1.00 convolve: 0 scaling: 1 basis: 18
AREA: 0.694133 resize: 1.00 convolve: 0 scaling: 1 basis: 19
AREA: 0.722430 resize: 1.00 convolve: 0 scaling: 1 basis: 20
AREA: 0.687758 resize: 1.00 convolve: 0 scaling: 1 basis: 21
AREA: 0.685376 resize: 1.00 convolve: 0 scaling: 1 basis: 22
AREA: 0.734786 resize: 1.00 convolve: 0 scaling: 2 basis: 1
AREA: 0.749562 resize: 1.00 convolve: 0 scaling: 2 basis: 2

AREA: 0.733983 resize: 1.00 convolve: 0 scaling: 2 basis: 3
AREA: 0.747595 resize: 1.00 convolve: 0 scaling: 2 basis: 4
AREA: 0.682295 resize: 1.00 convolve: 0 scaling: 2 basis: 5
AREA: 0.707229 resize: 1.00 convolve: 0 scaling: 2 basis: 6
AREA: 0.684367 resize: 1.00 convolve: 0 scaling: 2 basis: 7
AREA: 0.725146 resize: 1.00 convolve: 0 scaling: 2 basis: 8
AREA: 0.691156 resize: 1.00 convolve: 0 scaling: 2 basis: 9
AREA: 0.720091 resize: 1.00 convolve: 0 scaling: 2 basis: 10
AREA: 0.690956 resize: 1.00 convolve: 0 scaling: 2 basis: 11
AREA: 0.712476 resize: 1.00 convolve: 0 scaling: 2 basis: 12
AREA: 0.684451 resize: 1.00 convolve: 0 scaling: 2 basis: 13
AREA: 0.681919 resize: 1.00 convolve: 0 scaling: 2 basis: 14
AREA: 0.683811 resize: 1.00 convolve: 0 scaling: 2 basis: 15
AREA: 0.707771 resize: 1.00 convolve: 0 scaling: 2 basis: 16
AREA: 0.691657 resize: 1.00 convolve: 0 scaling: 2 basis: 17
AREA: 0.725210 resize: 1.00 convolve: 0 scaling: 2 basis: 18
AREA: 0.691145 resize: 1.00 convolve: 0 scaling: 2 basis: 19
AREA: 0.720126 resize: 1.00 convolve: 0 scaling: 2 basis: 20
AREA: 0.684926 resize: 1.00 convolve: 0 scaling: 2 basis: 21
AREA: 0.682738 resize: 1.00 convolve: 0 scaling: 2 basis: 22

Bibliography

[1] M. Abeles, E. Vaadia, and H. Bergman. Firing patterns of single units in the prefrontal cortex and neural network models. *Network*, 1:13–25, 1990.

[2] E.H. Adelson and J.R. Bergen. Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America A*, 2(2):284–299, 1985.

[3] Y. Adini and D. Sagi. Recurrent networks in human visual cortex: psychophysical evidence. *Journal of the Optical Society of America A*, 18(8): 2228–2236, 2001.

[4] S. Ahmed. Visit: An efficient computational model of human visual attention. *Technical Report 91-049, University of Illinois*, 1991.

[5] L. Alvarez and F. Morales. Affine morphological multiscale analysis of corners and multiple junctions. *International Journal of Computer Vision*, 2(25):95–107, 1997.

[6] R. A. Andersen, R. M. Bracewell, S. Barash, J. W. Gnadt, and L. Fogass. Eye position effects on visual, memory, and saccade-related activity in areas lip and 7a of macaque. *Journal of Neuroscience*, 10(4):1176–1196, 1990.

[7] C. Anderson and D. Van Essen. Shifter circuits: a computational strategy for dynamic aspects of visual processing. *Proceedings of the National Academy of Science*, 84:6297–6301, 1987.

[8] F. Attneave. Some informational aspects of visual perception. *Psychological Review*, 3:183–193, 1954.

[9] D. Attwell and S.B. Laughlin. An energy budget for signaling in the grey matter of the brain. *Journal of Cerebral Blood Flow and Metabolism*, 21: 1133–1145, 2001.

[10] R. Baddeley, L.F. Abbott, M.C. Booth, F. Sengpiel, T. Freeman, E.A. Wakeman, and E. Rolls. Responses of neurons in primary and inferior temporal visual cortices to natural scenes. *Proceedings of the Royal Society of London: B*, 264:1775–1783, 1997.

[11] H. Barlow. Single units and sensation: A neuron doctrine for perceptual psychology? *Perception*, 1:371–394, 1972.

[12] A. Baumberg. Reliable feature matching across widely separated views. *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, pages 774–781, 2000.

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

[14] I.N. Beloozerova, M.G. Sirota, and H.A. Swadlow. Activity of different classes of neurons of the motor cortex during locomotion. *Journal of Neuroscience*, 23:1087–1097, 2003.

[15] I. Biederman, J.C. Rabinowitz, A.L. Glass, and E.W. Stacy. On the information extracted from a glance at a scene. *Journal of Experimental Psychology*, 103:597–600, 103.

[16] E. Bisiach and G. Valler. Hemineglect in humans. *Handbook of Neuropsychology*, 222:1–195, 1988.

[17] J.W. Bisley and M.E. Goldberg. Neuronal activity in the lateral intraparietal area. *Science*, 299:81–86, 2003.

[18] R.T. Born and D.C. Bradley. Structure and function of visual area mt. *Annual Review of Neuroscience*, 28:157–189, 2005.

[19] M. Bravo and R. Blake. Preattentive vision and perceptual groups. *Perception*, 19:515–522, 1990.

[20] C. Breazeal and B. Scassellati. A context-dependent attention system for a social robot. *Proceedings of the sixteenth international joint conference on artificial intelligence*, pages 1146–1151, 1999.

[21] M. Brecht and B. Sakmann. Dynamic representation of whisker deflection by synaptic potentials in spiny stellate and pyramidal cells in the barrels and septa of layer 4 rat somatosensory cortex. *Journal of Physiology*, 23: 7940–7949, 2002.

[22] M. Brecht, M. Schneider, B. Sakmann, and T.W. Margrie. Whisker movements evoked by stimulation of single pyramidal cells in rat motor cortex. *Nature*, 427:704–710, 2004.

[23] N. Brenner, W. Bialek, and R. de Ruyter van Steveninck. Adaptive rescaling maximizes information transmission. *Neuron*, 26(3):695–702, 2000.

[24] D.E. Broadbent. *Perception and Communication*. London: Pergamon, 1958.

[25] N. D. B. Bruce and John K. Tsotsos. Saliency based on information maximization. *Advances in Neural Information Processing Systems*, 18:155–162, 2006.

[26] N.D.B. Bruce. Features that draw visual attention: An information theoretic perspective. *Neurocomputing*, 65-66:125–133, 2005.

[27] P. Burt. Attention mechanisms for vision in a dynamic world. *Proceedings Ninth International Conference on Pattern Recognition*, pages 977–987, 1988.

[28] M.C. Bushnell, M.E. Goldberg, and D.L. Robinson. Behavioral enhancement of visual responses in monkey cerebral cortex. i. modulation in posterior parietal cortex to selective visual attention. *Journal of Neurophysiology*, 46: 755–772, 1981.

[29] G.T. Buswell. How people look at pictures. *Chicago: University of Chicago Press. Hillsdale, NJ: Erlbaum*, 1935.

[30] M.W. Cannon and S.C. Fullencamp. A model for inhibitory lateral interaction effects in perceived contrast. *Vision Research*, 36(8):1115–1125, 1996.

[31] J.F. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8:679–698, 1986.

[32] J.F. Cardoso. High-order contrasts for independent component analysis. *Neural Computation*, 11(1):157–192, 1999.

[33] J. Cavanaugh and R.H. Wurtz. Subcortical modulation of attention counters change blindness. *Journal of Neuroscience*, 24:11236–11243, 2004.

[34] K.R. Cave. The feature gate model of visual selection. *Psychological Research*, 62:182–194, 1999.

[35] M.S. Caywood, B. Willmore, and D.J. Tolhurst. Independent components of color natural scenes resemble v1 neurons in their spatial and color tuning. *Journal of Neurophysiology*, 91:2859–2873, 2001.

[36] S.C. Chong and A. Treisman. Attentional spread in the statistical processing of visual displays. *Perception and Psychophysics*, 67:1–13, 2005.

[37] S. Christman. Hemispheric asymmetry in categorical versus coordinate processing of dynamic visual input. *Annual Meeting of the Psychonomic Society, Los Angeles, CA*, 1997.

[38] J.J. Clark and N.J. Ferrier. Modal control of an attentive vision system. *Proceedings of the Second International Conference on Computer Vision*, pages 514–523, 1988.

[39] C. Colby and M. Goldberg. Space and attention in parietal cortex. *Annual Review of Neuroscience*, 22:319–349, 1999.

[40] C.L. Colby, J.R. Duhamel, and M.E. Goldberg. Visual, presaccadic, and cognitive activation of single neurons in monkey lateral intraparietal area. *Journal of Neurophysiology*, 76:2841–2852, 1996.

[41] C.E. Connor, J.L. Gallant, D.C. Preddie, and D.C. Van Essen. Responses in V4 depend on the spatial relationship between stimulus and attention. *Journal of Neurophysiology*, 75:1306–1308, 1996.

[42] M. Corbetta, F.M. Miezin, S. Dobmeyer, G.L. Shulman, and S.E. Petersen. Attentional modulation of neural processing of shape color and velocity in humans. *Science*, 278:1556–1559, 1990.

[43] C. Cortes and M. Mohri. Confidence intervals for the area under the roc curve. *Advances in Neural Information Processing Systems*, 17:305–312, 2005.

[44] S. Culhane. *Implementation of an Attentional Prototype for Early Vision.* University of Toronto, M.Sc. Thesis, 1992.

[45] E.B. Cuttrell and R.T. Marrocco. Electrical microstimulation of primate posterior parietal cortex initiates orienting and alerting components of covert attention. *Experimental Brain Research*, 144:103–113, 2002.

[46] A.R. Damasio, H. Damasio, and C.H. Chang. Neglect following damage to frontal lobe or basal ganglia. *Neuropsychologia*, 18:123–132, 1980.

[47] D. Van Dantzig. On the consistency and power of Wilcoxon's two sample test. *Koninklijke Nederlandse Akademie van Weterschappen Series A*, 54: 1–8, 1915.

[48] M.P. Davey and J.M. Zanker. Detection the orientation of short lines in the periphery. *Australian New Zealand Journal of Opthalmology*, 26:s104–s107, 1998.

[49] G. Deco and J. Zihl. Top-down selective visual attention: A neurodynamical approach. *Visual Cognition*, 8(1):119–140, 1997.

[50] R. Desimone. Visual attention mediated by biased competition in extrastriate visual cortex. *Philosophical Transactions of the Royal Society of London*, 353: 1245–1255, 1998.

[51] R. Desimone and J. Duncan. Neural mechanisms of selective visual attention. *Annual Reviews of Neuroscience*, 18:193–222, 1995.

[52] J.A. Deutsch and D. Deutsch. Attention: Some theoretical considerations. *Psychological Review*, 87:272–300, 1963.

[53] M. DeWeese, M. Wehr, and A. Zador. Binary spiking in auditory cortex. *Journal of Neuroscience*, 23:7940–7949, 2003.

[54] B. Draper and A. Lionelle. Evaluation of selective attention under similarity transforms. *Workshop on Attention and Performance in Computer Vision*, pages 31–38, 2003.

[55] J. Driver. Object segmentation and visual neglect. *Behavioral Brain Research*, 71:135–146, 1995.

[56] J. Driver and J.B. Mattingly. Parietal neglect and visual awareness. *Nature Neuroscience*, 1:17–22, 1998.

[57] J. Driver, G.C. Baylis, and R.D. Rafal. Preserved figure-ground segmentation and symmetry perception in visual neglect. *Nature*, 360:73–75, 1992.

[58] R.O Duda, P.E. Hart, and D.G. Stork. *Pattern Classification: Second Edition.* New York: John Wiley, 2001.

[59] J. Duncan and G.W. Humphreys. Visual search and stimulus similarity. *Psychological Review*, 96(3):433–458, 1989.

[60] S. Eifuku and R.H. Wurtz. Response to motion in extrastriate area msti: Centre-surround interactions. *Journal of Neurophysiology*, 80(11):282–296, 1998.

[61] L. Elazary and L. Itti. Interesting objects are visually salient. *Journal of Vision*, 8(3):1–15, 2008.

[62] R. Engbert and R. Kliegl. Microsaccades uncover the orientation of covert attention. *Vision Research*, 43(9):1035–1045, 2003.

[63] J.T. Enns and R.A. Rensink. Sensitivity to three-dimensional orientation in visual search. *Psychological Science*, 1:323–326, 1990.

[64] E.A. Essock. The oblique effect of stimulus identification considered with respect to two classes of oblique effects. *Perception*, 9:37–46, 1980.

[65] K.K. Evans and A. Treisman. Perception of objects in natural scenes: is it really attention free? *Journal of Experimental Psychology: Human Perception and Performance*, 31(6):1476–1492, 2005.

[66] J.H. Fecteau and D.P. Munoz. Salience, relevance, and firing: apriority map for target selection. *Trends in Cognitive Sciences*, 10(8):382–389, 2006.

[67] D.J. Felleman and D.C. Van Essen. Distributed hierarchical processing in primate cerebral cortex. *Cerebral cortex*, 1:1–47, 1991.

[68] V.P. Fererra, T.A. Nealey, and J.H.R. Maunsell. Responses in macaque area V4 following inactivation of the parvocellular and magnocellular LGN pathways. *Journal of Neuroscience*, 14:2080–2088, 1994.

[69] D.J. Field and B.A. Olshausen. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

[70] G.R. Fink, R.J. Dolan, P.W. Halligan, J.C. Marshall, and C.D. Frith. Space-based and object-based visual attention: shared and specific neural domains. *Brain*, 120:2013–2028, 1997.

[71] P. Foldiak. Sparse coding in the primate cortex. *The Handbook of Brain Theory and Neural Networks*, 2nd Edition:1064–1068, 2002.

[72] W. Fries. Cortical projections to the superior colliculus in the macaque monkey, a retrograde study using horseradish peroxidase. *Journal of Computational Neurology*, 230:55–76, 1984.

[73] F. Fritz, C. Seifert, L. Paletta, and H. Bischof. Attentive object detection using an information theoretic saliency measure. *Proceedings of the Second Workshop on Attention and Performance in Computer Vision*, pages 29–41, 2004.

[74] G. Galfano, E. Betta, and M. Turatto. Inhibition of return in microsaccades. *Experimental Brain Research*, 159:400–404, 2004.

[75] J.I. Gold and M.N. Shadlen. Neural computations that underlie decisions about sensory stimuli. *Trends in Cognitive Science*, 5:10–16, 2001.

[76] M. Grabowecky, L.C. Robertson, and A. Treisman. Preattentive processes guide visual search: evidence from patients with unilateral visual neglect. *Journal of Cognitive Neuroscience*, 5:288–302, 1993.

[77] M.R. Greene and A. Oliva. Natural scene categorization from conjunctions of ecological global properties. *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, In Press, 2006.

[78] S. Grossberg. Adaptive pattern classification and universal recoding. *Biological Cybernetics*, 23:187–202, 1976.

[79] P.E. Haenny, J.H.R. Maunsell, and P.H. Schiller. State dependent activity in monkey visual cortex. *Experimental Brain Research*, 69:245–259, 1988.

[80] Z.M. Hafed and J.J. Clark. Microsaccades as an overt measure of covert attention shifts. *Vision Research*, 42:2533–2545, 2002.

[81] B.C. Hansen, E.A. Essock, Y. Zheng, and J.K. DeFord. Perceptual anisotropies in visual processing and their relation to natural image statistics. *Network: Computation in Neural Systems*, 14:501–526, 2003.

[82] C. Harris and M. Stephens. A combined corner and edge detector. *Alvey Vision Conference*, pages 147–151, 1988.

[83] P.Y. He and E. Kowler. The role of location probability in the programming of saccades: implications for "center-of-gravity" tendencies. *Vision Research*, 29(9):1165–81, 1989.

[84] Z.J. He and K. Nakayama. Surfaces versus features in visual search. *Nature*, 359:231–233, 1992.

[85] G. Heidemann. Focus-of-attention from local color symmetries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(7):817–830, 2004.

[86] K.M. Heilman and E. Valenstein. Frontal lobe neglect in man. *Neurology*, 22: 660–664, 1992.

[87] K.M. Heilman, R.T. Watson, and E. Valenstein. *Neglect and Related disorders*. Oxford, UK: Oxford University Press, 1993.

[88] S.A. Hillyard and L.A. Vento. Event-related brain potentials in the study of visual selective attention. *Proceedings of the National Academy of Science, USA*, 95:781–787, 1998.

[89] J.M. Hopf, S.J. Luck, K. Boelmans, M.A. Schoenfeld, C.N. Boehler, J. Rieger, and H.J. Heinze. The neural site of attention matches the spatial scale of perception. *Journal of Neuroscience*, 26(13):3532–3540, 2006.

[90] T. S. Horowitz, E.M. Fine, D.E. Fencsik, S. Yurgenson, and J. M. Wolfe. Fixational eye movements are not an index of covert attention. *Psychological Science*, 18(4):356–363, 2006.

[91] P. O. Hoyer and A. Hyvrinen. A multi-layer sparse coding network learns contour coding from natural images. *Vision Research*, 42(12):1593–1605, 2002.

[92] X. Huand, T.D. Albright, and G.R. Stoner. Adaptive surround modulation in cortical area mt. *Neuron*, 53(5):761–770, 2007.

[93] J. Hullman, W. T. Winkel, and F. Boselie. Concavities as a basic feature in visual search: evidence from search asymmetries. *Perception and Psychophysics*, 62:162–174, 2000.

[94] A. Hyvrinen, M. Gutmann, and P.O. Hoyer. Statistical model of natural stimuli predicts edge-like pooling of spatial frequency channels in v2. *BMC Neuroscience*, 6(12), 2005.

[95] Aapo Hyvrinen and Patrik O. Hoyer. A two-layer sparse coding model learns simple and complex cell receptive fields and topography from natural images. *Vision Research*, 41(18):2413–2423, 2001.

[96] L. Itti and P. Baldi. Bayesian surprise attracts human attention. *Advances in Neural Information Processing Systems*, 19:1–8, 2006.

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

[98] 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.

[99] P. Perona J. Harel, C. Koch. Graph-based visual saliency. *Advances in Neural Information Processing Systems 19*, pages 545–552, 2007.

[100] W. James. *The Principles of Psychology*. Holt, New York, 1890.

[101] K.W. Janer and J.V. Pardo. Deficits in selective attention following bilateral anterior cingulotomy. *Journal of Cognitive Neuroscience*, 3:231–241, 1991.

[102] M. Jazayeri and J.A. Movshon. Optimal representation of sensory information by neural populations. *Nature Neuroscience*, 9:690–696, 2006.

[103] J.C. Johnston and J.L. McLelland. Perception of letters in words: seek not and ye shall find. *Science*, 184(142):1192–1194, 1974.

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

[105] S. Kastner and M.A. Pinsk. Visual attention as a multilevel selection process. *Cognitive, affective & behavioral neuroscience*, 4(4):483–500, 2004.

[106] S. Kastner and L.G. Ungerleider. Mechanisms of visual attention in the human cortex. *Annual Review of Neuroscience*, 23:315–341, 2000.

[107] S. Kastner, P. De Weerd, J.M. Maisog, and R. Desimone anc L.G. Ungerleider. Sensory interactions in the human visual system: A functional MRI study. *Society of Neuroscience Abstracts*, 23:1396, 1997.

[108] S. Kastner, P. De Weerd, R. Desimone, and L.G. Ungerleider. Mechanisms of directed attention in the human extrastriate cortex as revealed by functional MRI. *Science*, 282:108–111, 1998.

[109] S. Kastner, M.A. Pinsk, P. De Weerd, R. Desimone, and L.G. Ungerleider. Increased activity in human visual cortex during directed attention in the absence of visual stimulation. *Neuron*, 22:751–761, 1999.

[110] C. Koch and S. Ullman. Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology*, 4:219–227, 1985.

[111] J.J. Koenderink and A.J. van Doorn. Representation of local geometry in the visual system. *Biological Cybernetics*, 55:367–375, 1987.

[112] H. Koesling, E. Carbone, and H. Ritter. Tracking of eye movements and visual attention. *University of Bielefeld Technical Report*, 2002.

[113] H. Kondo and H. Komatsu. Suppression on neuronal responses by a metacontrast masking stimulus. *Neuroscience Research*, 36(1):27–33, 2000.

[114] S. Kullback. The kullback-leibler distance. *The American Statistician*, 41: 340–341, 1987.

[115] S. Kullback and R.A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.

[116] M. Kusunoki, J. Gottlieb, and M.E. Goldberg. The lateral intraparietal area as a salience map: The representation of abrupt onset, stimulus motion, and task relevance. *Vision Research*, 40(10-12):1459–1468, 2000.

[117] J. Laubrock, R. Engbert, and R. Kliegl. Microsaccade dynamics during covert attention. *Vision Research*, 45:721–730, 2005.

[118] J. Laubrock, R. Engbert, M. Rolfs, and R. Kliegl. Microsaccades are an index of covert attention: Commentary on horowitz, fine, fencsik, yurgenson, and wolfe. *Psychological Science*, 18:364–366, 2007.

[119] T.S. Lee and S. Yu. An information theoretic framework for understanding saccadic eye movements. *Advances in Neural Information Processing Systems*, 12:834–840, 2000.

[120] T.W. Lee, M. Girolami, and T.J. Sejnowski. Independent component analysis using an extended infomax alorithm for mixed subgaussian and supergaussian sources. *Neural Computation*, 11:417–441, 1999.

[121] T.W. Lee, T. Wachtler, and T.J. Sejnowski. Color opponency is an efficient repreentation of spectral properties in natural scenes. *Vision Research*, 17: 2095–2103, 2002.

[122] P. Lennie. The cost of cortical computation. *Current Biology*, 13:493–497, 2003.

[123] D.S. Levine. *Introduction to Neural and Cognitive Modeling*. Mahwah, New Jersey: Lawrence Erlbaum Associates, 2000.

[124] Z. Li. A saliency map in primary visual cortex. *Trends in Cognitive Science*, 6:9–16, 2002.

[125] T. Lindeberg. Scale-space theory for discrete signals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(1):234–254, 1990.

[126] T. Lindeberg. Direct estimation of affine image deformation using visual front-end operations with automatic scale selection. *Proceedings of the 5th International Conference on Computer Vision*, pages 134–141, 1995.

[127] T. Lindeberg and J. Garding. Shape-adapted smoothing in estimation of 3-d shape cues from affine deformations of local 2-d brightness structure. *Image and Vision Computing*, 15(6):415–434, 1997.

[128] R. Linsker. From basic network principles to neural architecture: emergence of spatial-opponent cells. *Proceedings of the National Academy of Science USA*, 83(19):7508–7512, 1986.

[129] R. Linsker. From basic network principles to neural architecture: emergence of orientation-selective cells. *Proceedings of the National Academy of Science USA*, 83(21):8390–8394, 1986.

[130] R. Linsker. From basic network principles to neural architecture: emergence of orientation columns. *Proceedings of the National Academy of Science USA*, 83(22):8779–8783, 1986.

[131] R. Linsker. Self-organization in a perceptual network. *IEEE Computer*, 21(3): 105–117, 1988.

[132] S.J. Luck, L. Chelazzi, S.A. Hillyard, and R. Desimone. Neural mechanisms of spatial selective attention in areas V1, V2, and V4 of macaque visual cortex. *Journal of Neurophysiology*, 77:24–42, 1997.

[133] G. Maehara, M. Okubo, and C. Michimata. Effects of background color on detecting spot stimuli in the upper and lower visual fields. *Brain and Cognition*, 55:558–563, 2004.

[134] J.C. Marshall and P.W. Halligan. The yin and the yang of visuo-spatial neglect: a case study. *Neuropsychologia*, 32:1037–1057, 1994.

[135] A. Martinez, L. Anllo-Vento, M.I. Sereno, L.R. Frank, R.B. Buxton, D.J. Dubowitz, E.C. Wong, H. Hinrichs, H.J. Heinze, and S.A. Hillyard. Involvement of striate and extrastriate visual cortical areas in spatial attention. *Nature Neuroscience*, 2:364–369, 1999.

[136] J.H.R. Maunsell, G. Sclar, T.A. Nealey, and D.D. DePriest. Extraretinal representations in area V4 in the macaque monkey. *Visual Neuroscience*, 7: 561–573, 1991.

[137] J.H.R. Maunsell, G.M. Ghose, J.A. Assad, C.J. McAdams, C.E. Boudreau, and B.D. Noerager. Visual response latencies of magnocellular and parvocellular LGN neurons in macaque monkeys. *Visual Neuroscience*, 16:1–14, 1999.

[138] J.A. Mazer and J.L. Gallant. Goal-related activity in v4 during free viewing visual search: Evidence for a ventral stream visual salience map. *Neuron*, 40: 1241–1250, 2003.

[139] M.M. Mesulam. A cortical network for directed attention and unilateral neglect. *Annals of Neurology*, 10:309–325, 1981.

[140] M.M. Mesulam. Spatial attention and neglect, parietal, frontal and cingulate contributions to the mental representation and attentional targeting of salient extrapersonal events. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 354:1325–1346, 1999.

[141] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *International Journal of Computer Vision*, 65(1/2):43–72, 2005.

[142] M.M. Mller M.M. and R. Hbner. Can the spotlight of attention be shaped like a doughnut? *Psychological Science*, 13(2):119–124(6), 2002.

[143] T.A. Mondor and M.P. Bryden. On the relation between visual spatial attention and visual field asymmetries. *Quarterly Journal of Experimental Psychology Section A - Human Experimental Psychology*, 44(3):529–555, 1992.

[144] T. Moore and K.M. Armstrong. Selective gating of visual signals by microstimulation of frontal cortex. *Nature*, 421:370–373, 2003.

[145] T. Moore and M. Fallah. Microstimulation of the frontal eye field and its effects on covert spatial attention. *Journal of Neurophysiology*, 91:152–162, 2004.

[146] J. Moran and R. Desimone. Selective attention gates visual processing in the extrastriate cortex. *Science*, 229:782–784, 1985.

[147] H. Moravec. Visual mapping by a robot rover. *1979*, pages 598–600, 1979.

[148] B.C. Motter. Focal attention produces spatially selective processing in visual cortical areas V1, V2, and V4 in the presence of competing stimuli. *Journal of Neurophysiology*, 70:909–919, 1993.

[149] B.C. Motter. Neural correlates of attentive selection for color or luminance in extrastriate area V4. *Journal of Neuroscience*, 14:2178–2189, 1994.

[150] M.C. Mozer. *The Perception of Multiple Objects*. MIT Press, 1991.

[151] A.L. Nagy and R.R. Sanchez. Critical color differences determined with a visual search task. *Journal of the Optical Society of America*, 7(10):1209–1217, 1990.

[152] A.L. Nagy and G. Thomas. Distractor heterogeneity, attention and color in visual search. *Vision Research*, 43:1541–1552, 2003.

[153] J. Najemnik and W.S. Geisler. Optimal eye movement strategies in visual search. *Nature*, 434:387–391, 2005.

[154] K. Nakayama and G.H. Silverman. Serial and parallel processing of visual feature conjunctions. *Nature*, 320:264–265, 1991.

[155] D. Navon. Forest before trees: Precedence of global features in visual perception. *Cognitive Psychology*, 9:353–383, 1977.

[156] E. Niebur and C. Koch. A model for the neuronal implementation of selective visual attention based on temporal correlation among neurons. *Journal of Computational Neuroscience*, 1(1):141–158, 1994.

[157] E. Niebur and C. Koch. Control of selective visual attention: Modeling the 'where' pathway. *Advances in Neural Information Processing Systems*, 8:802–808, 1996.

[158] E. Niebur, C. Koch, and C. Rosin. An oscillation based model for the neural basis of attention. *Vision Research*, 33:2789–2802, 1993.

[159] A.C. Nobre, G.N. Sobestyen, D.R. Gitelman, M.M. Mesulam, and R.S.J. Frackowiak. Functional localization of the system for visuospatial attention using positron emission tomography. *Brain*, 120:515–533, 1997.

[160] A.C. Nobre, T. Allison, and G. McCarthy. Modulation of human extrastriate visual processing by selective attention to colours and words. *Brain*, 121: 1357–1368, 1998.

[161] H.C. Nothdurft. The role of features in preattentive vision: Comparison of orientation, motion and color cues. *Vision Research*, 33:1937–1958, 1993.

[162] D. Obradovic and G. Deco. Information maximization and independent component analysis: Is there a difference? *Neural Computation*, 10(8):2085–2101, 1998.

[163] F.H. O'Connor, M.M. Fukui, and M.A. Pinsk ans S. Kastner. Attention modulates responses in the human lateral geniculate nucleus. *Nature Neuroscience*, 5(11):1203–1209, 2002.

[164] A. Ohman, A. Flykt, and F. Esteves. Emotion drives attention: Detecting the snake in the grass. *Journal of Experimental Psychology: General*, 130(3): 466–478, 2001.

[165] K. Okajima. Binocular disparity encoding cells generated through an infomax based learning algorithm. *Neural Networks*, 17(7):953–962, 2004.

[166] B.A. Olhausen, C.H. Anderson, and D.C. van Essen. A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *The Journal of Neuroscience*, 13(11):4700–4719, 1993.

[167] A. Oliva. Gist of the scene. *in the Encyclopedia of Neurobiology of Attention. L. Itti, G. Rees, and J.K. Tsotsos (Eds.), Elsevier, San Diego, CA*, pages 251–256, 2005.

[168] A. Oliva and A. Torralba. Scene-centered description from spatial envelope properties. *Proceedings of the 2nd International Workshop on Biologically Motivated Computer Vision, Eds: H. Bulthoff, S.W. Lee, T. Poggio and C. Wallraven*, pages 263–272, 2002.

[169] A. Oliva and A. Torralba. Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research*, In Press, 2006.

[170] B.A. Olshausen and D.J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

[171] L.A. Olzak and P.I. Laurinen. Contextual effects in fine spatial discriminations. *Nature*, 381(6583):607–609, 2005.

[172] N. Ouerhani and H. Hugli. Computing visual attention from scene depth. *Proceedings of the 15th International Conference on Pattern Recognition*, pages 375–378, 2000.

[173] G. Palm. Evidence, information and surprise. *Biological Cybernetics*, 42: 57–68, 1981.

[174] D.H. Palmer. *Vision Science: Photons to Phenomenology*. Cambridge Massachusetts: MIT Press, 1999.

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

[176] H. Pashler. Detecting conjunctions of color and form. *Perception and Psychophysics*, 41:191–201, 1987.

[177] J. Perez-Orive, O. Mazor, G.C. Turner, S. Cassanaer, R.I. Wilson, and G. Laurent. Oscillations and sparsening of odor representations in the mushroom body. *Science*, 297:359–365, 2002.

[178] Y. Petrov and S.P. McKee. The effect of spatial configuration on surround suppression of contrast sensitivity. *Journal of Vision*, 6(3):224–238, 2006.

[179] M. Pomplun. Saccadic selectivity in complex visual search displays. *Vision Research*, 46:1886–1900, 2006.

[180] M.J. Posner and S.E. Petersen. The attention system of the human brain. *Annual Review of Neuroscience*, 13:25–42, 1990.

[181] E.O. Postma, H.J. van der Herik, and P.T.W. Hudson. SCAN: A scalable model of attentional selection. *Neural Networks*, 10(6):993–1015, 1997.

[182] C.M. Privitera and L.W. Stark. Algorithms for defining visual regions-of-interest: Comparison with eye fixations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:9:970–982, 2000.

[183] R.D. Rafal. Neglect. *Current Opinion in Neurobiology*, 4:231–236, 1994.

[184] V.S. Ramachandran. Perception of shape from shading. *Nature*, 331:163–166, 1988.

[185] K. Rapantzikos and N. Tsapatsoulis. Enhancing the robustness of skin-based face detection schemes through a visual attention architecture. *Proceedings of the IEEE International Conference on Image Processing*, pages 1298–1301, 2005.

[186] K. Rapantzikos, Y. Avrithis, and S. Kollias. Handling uncertainty in video analysis with spatiotemporal visual attention. *Proceedings of the 14th IEEE International Conference on Fuzzy Systems*, pages 213–217, 2005.

[187] G. Rees, R.S.J. Frackowiak, and C.D. Frith. Two modulatory effects of attention that mediate object categorization in human cortex. *Science*, 275: 835–838, 1997.

[188] L.W. Renninger, J. Coughlan, P. Verghese, and J. Malik. An information maximization model of eye movements. *Advances in Neural Information Processing Systems*, 17:1121–1128, 2004.

[189] L.W. Renninger, P. Verghese, and J. Coughlan. Where to look next? eye movements reduce local uncertainty. *Journal of Vision*, 7(3):6:1–17, 2007.

[190] R.A. Rensink, J.K. O'Regan, and J.J. Clark. To see or not to see: the need for attention to perceive changes in scenes. *Psychological Science*, 8:368–373, 1997.

[191] J.H. Reynolds, L. Chelazzi, and R. Desimone. Attention and contrast have similar effects on competitive interactions in macaque area V4. *Society of Neuroscience Abstracts*, 22:1197, 1999.

[192] D.L. Rhodes and L.C. Robertson. Visual field asymmetries and allocation of attention in visual scenes. *Brain and Cognition*, 50(1):95–115, 2002.

[193] G. Rizzolatti, L. Riggio, I. Dascola, and C. Umilt. Reorienting attention across the horizontal and vertical meridians: evidence in favor of a premotor theory of attention. *Neuropsychologia.*, 25(1A):31–40, 1987.

[194] L. Roberts. Machine perception of 3d solids. *Optical and Electro-optical Information Processing*, MIT Press, 1965.

[195] P.R. Roelfsema, V.A. Lamme, and H. Spekreijse. Object-based attention in the primary visual cortex of the macaque monkey. *Nature*, 395:376–381, 1998.

[196] M. Rolfs, J. Laubrock, and R. Kliegl. Shortening and prolongation of saccade latencies following microsaccades. *Experimental Brain Research*, 169:369–376, 2004.

[197] R. Rosenholtz. A simple saliency model predicts a number of motion pop-out phenomena. *Vision Research*, 39:3157–3163, 1999.

[198] R. Rosenholtz. Visual search for orientation among heterogeneous distractors: Experimental results and implications for signal detection theory models of search. *Journal of Experimental Psychology*, 27(4):985–999, 2001.

[199] R. Rosenholtz. Search asymmetries? what search asymmetries? *Perception & Psychophysics*, 63(3):476–489, 2001.

[200] R. Rosenholtz, A.L. Nagy, and N.R. Bell. The effect of background color on asymmetries in color search. *Journal of Vision*, 4(3):224–240, 2004.

[201] P.A. Sandon. Simulating visual attention. *Journal of Cognitive Neuroscience*, 2(3):213–231, 1989.

[202] J.D. Schall and K.G. Thompson. Neural selection and control of visually guided eye movements. *Extrastriate cortex of Primates, Cerebral Cortex (Rockland K.S. et al, eds.)*, pages 527–538, 1997.

[203] S.J. Schein and R. Desimone. Spectral properties of v4 neurons in the macaque. *Journal of Neuroscience*, 10(10):3369–3389, 1990.

[204] W.X. Schneider. Vam: Neuro-cognitive model for visual attention, control of segmentation, object recognition, and space-based motor action. *Visual Cognition*, 2:331–375, 1995.

[205] C.E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 1948.

[206] B.M. Sheliga, L. Craighero, L. Riggio, and G. Rizzolatti. Effects of spatial attention on directional manual and ocular responses. *Experimental Brain Research*, 114:339–351, 1997.

[207] Z.M. Shen, W.F. Xu, and C.Y. Li. Cue-invariant detection of centre surround discontinuity by v1 neurons in awake macaque monkey. *Journal of Physiology*, 583:581–592, 2007.

[208] S.G. Solomon, J.W. Pierce, and P. Lennie. The impact of suppressive surrounds on chromatic properties of cortical neurons. *Journal of Neuroscience*, 24(1):148–160, 2004.

[209] D.C. Somers, A.M. Dale, A.E. Seiffert, and R.B. Tootell. Functional mri reveals spatially specific attentional modulation in human visual cortex. *Proceedings of the National Academy of Science*, 96:1663–1668, 1999.

[210] H. Spitzer, R. Desimone, and J. Moran. Increased attention enhances both behavioral and neuronal performance. *Science*, 240:338–340, 1988.

[211] D. Tadin and J.S. Lappin. Optimal size for perceiving motion decreases with contrast. *Vision Research*, 45:2059–2064, 2005.

[212] D. Tailor, L. Finkel, and G. Buchsbaum. Color opponent receptive fields derived from independent component analysis of natural images. *Vision Research*, 40:2671–2676, 2000.

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

[214] L.T. Thompson and P.J. Best. Place cells and silent cells in the hippocampus of freely behaving rats. *Journal of Neuroscience*, 9:2382–2390, 1989.

[215] S.P. Tipper and M. Behrmann. Object-centred not scene-based visual neglect. *Journal of Experimental Psychology: Human Perception and Performance*, 22(5):1261–1278, 1996.

[216] N. Tishby, F. Pereira, and W. Bialek. Shifts in selective visual attention: Towards the underlying neural circuitry. *Proceedings of the 37th annual Allerton Conference on Communication, Control and Computing*, 1999.

[217] T.N. Topper. Selection mechanisms in human and machine vision. *University of Waterloo Ph.D. thesis*, 1991.

[218] A. Torralba and A. Oliva. Statistics of natural image categories. *Network: Computation in Neural Systems*, 14:391–412, 2006.

[219] J.T. Townsend. A note on the identification of serial and parallel processes. *Perception and Psychophysics*, 10:161–163, 1971.

[220] J.T. Townsend. Serial and within-stage independent parallel model equivalence on the minimum completion time. *Journal of Mathematical Psychology*, 14:219–239, 1976.

[221] J.T. Townsend. Serial and parallel processing: sometimes they look like tweedledum and tweedledee but they can and should be distinguished. *Psychological Science*, 1:46–54, 1990.

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

[223] A. Treisman and S. Gormican. Feature analysis in early vision: Evidence from search asymmetries. *Psychological Review*, 95:15–48, 1988.

[224] S. Treue and J.C. Martinez. Feature-based attention influences motion processing gain in macaque visual cortex. *Nature*, 399:575–579, 1999.

[225] S. Treue and J.H.R. Maunsell. Attentional modulation of visual processing in cortical areas MT and MST. *Nature*, 382:539–541, 1996.

[226] J.K. Tsotsos. A complexity level analysis of immediate vision. *International Journal of Computer Vision*, 2:303–320, 1988.

[227] J.K. Tsotsos. *Visual Attention Mechanisms; The Selective Tuning Model.* Kluwer Academic, 2003.

[228] J.K. Tsotsos, S. Culhane, W. Wai, Y. Lai, N. Davis, and N. Nuflo. Modeling visual attention via selective tuning. *Artificial Intelligence*, 1-2:507–547, 1995.

[229] T. Tuytelaars and L. Van Gool. Content-based image retrieval based on local addinely invariant regions. *International Conference on Visual Information Systems*, pages 493–500, 1999.

[230] L. Uhr. Layered recognition cone networks that preprocess, classify, and describe. *IEEE Transactions on Computing*, 21:758–768, 1972.

[231] G. Vallar. *The anatomical basis of spatial neglect in humans. Appearing in Unliateral Neglect: Clinical and Experimental Studies.* Hillsdale, NJ: Erlbaum, 1993.

[232] G. Vallar and D. Porani. The anatomy of unilateral neglect after right-hemisphere stroke lesions: a clinical/CT-scan correlation study in man. *Neuropsychologia*, 24:609–622, 1987.

[233] J.H. van Hateren and A. van der Schaaf. Independent component filters of natural images compared with simple cells in the primary visual cortex. *Proceedings of the Royal Society of London B*, 265:359–366, 1998.

[234] R. Vandenberghe, J. Duncan, P. Dupont, R. Ward, and J. Poline. Attention to one of two features in left or right visual field: a positron emission tomography study. *Journal of Neuroscience*, 17:3739–3750, 1997.

[235] L.A. Vento, S.J. Luck, and S.A. Hillyard. Spatiotemporal dynamics of attention to color: evidence from human electrophysiology. *Human Brain Mapping*, 6:216–238, 1998.

[236] P. Verghese. Visual search and attention: A signal detection theory approach. *Neuron*, 31:523–535, 2001.

[237] W.E. Vinje and J.L. Gallant. Sparse coding and decorrelation in primary visul cortex during natural vision. *Science*, 287:1273–1276, 2000.

[238] W.E. Vinje and J.L. Gallant. Natural stimulation of the nonclassical receptive field increases information transmission efficiency in v1. *Journal of Neuroscience*, 22:2904–2915, 2002.

[239] C. von der Malsberg. The correlation theory of brain function. *Internal Report 81-2, Department of Neurobiology, Max-Planck Institute for Biophysical Chemistry, Gottigen Germany*, 1981.

[240] H. von Helmholtz. Handbuch der physiologischen optik. *Leipzig: Leopold Voss*, Section 28, 1867.

[241] T. Wachtler, T.W. Lee, and T.J. Sejnowski. Chromatic structure of natural scenes. *Journal of the Optical Society of America A*, 18(1):65–77, 2001.

[242] R.T. Watson and K.M. Heilman. Thalamic neglect. *Neurology*, 29:1003–1007, 1979.

[243] G. Westheimer. The distribution of preferred orientations in the peripheral visual field. *Vision Research*, 43(1):53–37, 2003.

[244] G. Westheimer. Meridional anisotropy in visual processing, implications for the neural site of the oblique effect. *Vision Research*, 43(22):2281–2289, 2003.

[245] R. P. Wildes. A measure of motion salience for surveillance applications. *Proceedings of the IEEE International Conference on Image Processing*, pages 183–187, 1998.

[246] R. P. Wildes and J.R. Bergen. Qualitative spatiotemporal analysis using an oriented energy representation. *Proceedings of the European Conference on Computer Vision*, pages 768–784, 2000.

[247] A.L. Williams, K.D. Singh, and A.T. Smith. Surround modulation measured with fmri in the visual cortex. *Journal of Neurophysiology*, 89(1):525–533, 2003.

[248] J.M. Wolfe. Guided search 2.0: A revised model of visual search. *Psychonomic Bulletin and Review*, 1:202–238, 1994.

[249] J.M. Wolfe. What can 1,000,000 trials tell us about visual search? *Psychological Science*, 9(1):33–39, 1998.

[250] J.M. Wolfe. Visual search. *Attention. H. Pashler. UK: University College London Press*, 1998.

[251] J.M. Wolfe and R.K. Cave. Guided search: An alternative to feature integration theory for visual search. *Journal of Experimental Psychology: Human Perception and Performance*, 15:419–443, 1989.

[252] J. Xing and D.J. Heeger. Centre-surround interactions in foveal and peripheral vision. *Vision Research*, 40:3065–3072, 2000.

[253] J. Xing and D.J. Heeger. Measurement and modeling of centre-surround suppression and enhancement. *Vision Research*, 41:571–583, 2001.

[254] A.L. Yarbus. Eye movements and vision. *Plenum: New York*, 1967.

[255] C. Yu and D.M. Levi. Surround modulation in human vision unmasked by masking experiments. *Nature*, 3(7):724–728, 2000.

[256] C. Yu, A.K. Yu, and D.M. Levi. Surround modulation of perceived contrast and the role of brightness induction. *Journal of Vision*, 1:18–31, 2001.

[257] C. Yu, A.K. Klein, and D.M. Levi. Cross-and iso-oriented surrounds modulate the contrast response function: The effect of surround contrast. *Journal of Vision*, 3:527–540, 2003.

[258] B. Zhang, J. Zheng, I. Watanabe, I. Maruko, H. Bi, E.L. Smith, and Y. Chino. Delayed maturation of receptive field centre/surround mechanisms in v2. *Proceedings of the National Academy of Sciences*, 102(16):5862–5867, 2005.