UNIVERSITY OF CALIFORNIA, SAN DIEGO

A discriminant hypothesis for visual saliency: computational principles, biological plausibility and applications in computer vision

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy
in
Electrical Engineering (Signal and Image Processing)

by

Dashan Gao

Committee in charge:

Professor Nuno Vasconcelos, Chair
Professor Pamela Cosman
Professor Garrison W. Cottrell
Professor David J. Kriegman
Professor Truong Nguyen

2008

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

Chair

University of California, San Diego

2008

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to acknowledge many people for helping me during my doctoral work. Without their persistent help and support, I would not be able to complete this work.

First of all, I would like to express my deep and sincere gratitude to my supervisor, Professsor Nuno Vasconcelos, for his valuable time, personal guidance, and inspirational discussions. Many of his brilliant ideas have become the very foundation of the present thesis. His broad knowledge, his logic way of thinking, and deep insights to fundamental problems of vision science have been great value for me. His perpetual energy and enthusiasm in science had also motivated me. His understanding, encouraging and patient guidance have provided a solid basis for this work.

My research was supported in part by National Science Foundation, and Google Inc. I would like to thank our great sponsors for providing me such good opportunities to be able to work freely in this area.

I am also very grateful for having an exceptional doctoral committee and wish to thank Professor Pamela Cosman, Professor Garrison W. Cottrell, Professor David J. Kriegman, and Professor Truong Nguyen for their input, valuable discussions and accessibility.

I would like to thank all my colleagues and friends from SVCL lab at UCSD, Antoni B. Chan, Dr. Gustavo Carneiro, Sunhyoung Han, Vijay Mahadevan, Hamed Masnadi-Shirazi, and Nikhil Rasiwasia, for their continuous support, their friendship and the assistance in the past several years. I also owe a special note of gratitude to Sunhyoung Han for assisting me with the experiments, and to Antoni B. Chan and Vijay Mahadevan for proofreading this thesis.

During the last six years, I got to know so many friends in San Diego. I am particularly grateful to Dr. Junwen Wu, and her husband Dr. Junan Zhang, who have helped me since the very beginning of this long journey. Without them, my life in the U.S. could have been started miserably. I am also thankful to my friends

Ying Ji, Dan Liu, Dr. Fang Fang, Long Wang, Honghao Shan, Dr. Min Li, Dr. Yushi Shen, Dr. Deqiang Song, Dr. Lingyun Zhang, Wenyi Zhang, Dr. Haichang Sui, Dr. Zhou Lu, Yuzhe Jin, Ken Lin, and many others. Their friendship is the most precious gift during my PhD study.

I would like to express my great gratitude to my parents, who brought me to this world, who raised and taught me, who encouraged and supported me to pursuit my Ph.D. degree abroad in the first place, and who are always loving me and proud of me. I also owe my loving thanks to my wife, Zongjuan (Janet) Zhou, for giving up her career in China and coming to this country with me without hesitation, for standing beside and encouraging me whenever I feel frustrated, for being the breadwinner and taking care of my life for so many years, and for willing to make another sacrifice to move with me again. It is her endless and unwavering love that enabled me to finish this work. I therefore dedicate my dissertation to my wife, my father, mother, grandfather, grandmother, my sister, and I cannot leave out my cat, Milo, for being so supportive and loving all the time. Your love is the very foundation of my life!

The text of Chapter II, in part, is based on the materials as it appears in: D. Gao and N. Vasconcelos, Discriminant saliency for visual recognition from cluttered scenes. In *Proc. of Neural Information Processing Systems (NIPS)*, 2004. D. Gao and N. Vasconcelos. Decision-theoretic saliency: computational principles, biological plausibility, and implications for neurophysiology and psychophysics. Accepted for publication, *Neural Computation.* It, in part, has also been submitted for publication of the material as it may appear in D. Gao and N. Vasconcelos, Discriminant saliency for visual recognition. Submitted for publication, *IEEE Trans. on Pattern Analysis and Machine Intelligence.* The dissertation author was a primary researcher and an author of the cited materials.

The text of Chapter III, in part, is based on the material as it appears in: D. Gao and N. Vasconcelos. Decision-theoretic saliency: computational principles, biological plausibility, and implications for neurophysiology and psychophysics.

Accepted for publication, *Neural Computation.* The dissertation author was a primary researcher and an author of the cited material.

The text of Chapter IV, in part, is based on the materials as it appears in: D. Gao and N. Vasconcelos. Decision-theoretic saliency: computational principles, biological plausibility, and implications for neurophysiology and psychophysics. Accepted for publication, *Neural Computation.* D. Gao, V. Mahadevan and N. Vasconcelos On the plausibility of the discriminant center-surround hypothesis for visual saliency. Accepted for publication, *Journal of Vision.* It, in part, is also based on a co-authored work with N. Vasconcelos. The dissertation author was a primary researcher and an author of the cited materials.

The text of Chapter V, in part, is based on the materials as it appears in: D. Gao and N. Vasconcelos, Discriminant saliency for visual recognition from cluttered scenes. In *Proc. of Neural Information Processing Systems (NIPS)*, 2004. D. Gao and N. Vasconcelos, Discriminant Interest Points are Stable. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007. D. Gao and N. Vasconcelos, An experimental comparison of three guiding principles for the detection of salient image locations: stability, complexity, and discrimination. *The 3rd International Workshop on Attention and Performance in Computational Vision (WAPCV)*, 2005. It, in part, has also been submitted for publication of the material as it may appear in D. Gao and N. Vasconcelos, Discriminant saliency for visual recognition. Submitted for publication, *IEEE Trans. on Pattern Analysis and Machine Intelligence.* The dissertation author was a primary researcher and an author of the cited materials.

The text of Chapter VI, in part, is based on the material as it appears in: D. Gao, V. Mahadevan and N. Vasconcelos On the plausibility of the discriminant center-surround hypothesis for visual saliency. Accepted for publication, *Journal of Vision.* The dissertation author was a primary researcher and an author of the cited material.

The text of Chapter VII, in full, is based on a co-authored work with N.

Vasconcelos. The dissertation author was a primary researcher of this work.

VITA

1999	Bachelor of Engineering Automation, Tsinghua University, Beijing
2002	Master of Science Pattern Recognition and Artificial Intelligence, Tsinghua University, Beijing
2002–2008	Research Assistant Statistical and Visual Computing Laboratory Department of Electrical and Computer Engineering University of California, San Diego
2008	Doctor of Philosophy Electrical and Computer Engineering, University of California, San Diego

PUBLICATIONS

D. Gao and N. Vasconcelos. Decision-theoretic saliency: computational principles, biological plausibility, and implications for neurophysiology and psychophysics. Accepted for publication, *Neural Computation.*

D. Gao, V. Mahadevan and N. Vasconcelos On the plausibility of the discriminant center-surround hypothesis for visual saliency. *Journal of Vision*, 8(7), pp. 1-18, 2008

D. Gao and N. Vasconcelos, Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition. Submitted for publication, *IEEE Trans. on Pattern Analysis and Machine Intelligence.*

D. Gao, V. Mahadevan and N. Vasconcelos, The discriminant center-surround hypothesis for bottom-up saliency. In *Proc. Neural Information Processing Systems (NIPS)*, Vancouver, Canada, 2007.

D. Gao and N. Vasconcelos, Bottom-up saliency is a discriminant process. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, Rio de Janeiro, Brazil, 2007.

D. Gao and N. Vasconcelos, Discriminant Interest Points are Stable. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Minneapolis, 2007.

D. Gao and N. Vasconcelos, Integrated learning of saliency, complex features, and objection detectors from cluttered scenes. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, San Diego, 2005.

D. Gao and N. Vasconcelos, An experimental comparison of three guiding principles for the detection of salient image locations: stability, complexity, and discrimination. *The 3rd International Workshop on Attention and Performance in Computational Vision (WAPCV)*, San Diego, 2005.

D. Gao and N. Vasconcelos, Discriminant saliency for visual recognition from cluttered scenes. In *Proc. of Neural Information Processing Systems (NIPS)*, Vancouver, Canada, 2004.

ABSTRACT OF THE DISSERTATION

A discriminant hypothesis for visual saliency: computational principles,
biological plausibility and applications in computer vision

by

Dashan Gao
Doctor of Philosophy in Electrical Engineering
(Signal and Image Processing)
University of California, San Diego, 2008
Professor Nuno Vasconcelos, Chair

It has long been known that visual attention and saliency mechanisms play an important role in human visual perception. However, there have been no computational principles that could explain the fundamental properties of biological visual saliency. In this thesis, we propose, and study the plausibility of a novel principle for human visual saliency, which we denote as *discriminant saliency hypothesis*. The hypothesis states that all saliency decisions are optimal in a decision-theoretic sense. Under this formulation, optimality is defined in the minimum probability of error sense, under a constraint of computational parsimony. The discriminant saliency hypothesis naturally adapts to both stimulus-driven (bottom-up) and goal-driven (top-down) saliency problems, for which we derive the optimal discriminant saliency detectors, in an information-theoretic sense. Statistical properties of natural stimuli are also exploited in the derivation for the constraint of computational parsimony.

To study the biological plausibility of discriminant saliency, we show that, under the assumption that saliency is driven by linear filtering, the computations of discriminant saliency are completely consistent with the *standard neural architecture* in the primary visual cortex (V1). The discriminant saliency detectors are also applied to the set of classical displays, used in the study of human saliency behaviors, and shown to explain both qualitative and quantitative properties of human saliency. These results not only justify the biological plausibility of the discriminant hypothesis for saliency, but also offer explanations to the neural organization of perceptual systems. For example, we show that the basic neural structures in V1 are capable of computing the fundamental operations of statistical inference, e.g., assessment of probabilities, implementation of decision rules, and feature selection.

Finally, we evaluate the performance of the derived discriminant saliency detectors for computer vision problems. In particular, we apply the top-down saliency detector to the problem of weakly supervised learning for object recognition, and show that the detector outperforms the state-of-the-art saliency detectors in 1) capturing important information for object recognition tasks, 2) accurately localizing objects of interest from image clutter, 3) providing stable salient locations with respect to various geometric and photometric transformations, and 4) adapting to diverse visual attributes for saliency. We then evaluate the performance of the bottom-up discriminant saliency detector in the applications where no recognition is defined. In particular, we show that the bottom-up discriminant saliency implementation accurately predicts human eye fixation locations on natural scenes. In another application of discriminant saliency, we discuss a Bayesian framework to integrate top-down and bottom-up saliency outputs, where the top-down saliency is interpreted as a *focus-of-attention* mechanism. Experimental results show that this framework combines the selectivity of the top-down saliency with the localization ability of the bottom-up interest point detectors, and improves the object recognition performance.

Overall, the excellent performance of discriminant saliency in both biological and computer vision evaluations justifies the plausibility of discriminant hypothesis as an explanation for human visual saliency.

Chapter I

Introduction

I.A Human visual saliency

Biological vision systems, such as the human vision system, have a remarkable ability to automatically select and allocate attention to a few "relevant" locations in a scene [229, 149, 33, 86, 228]. This ability enables organisms to focus their limited perceptual and cognitive resources on the most pertinent subset of the available sensory data, facilitating learning and survival in everyday life. The deployment of visual attention is believed to be driven by *visual saliency* mechanisms, which is a fundamental, yet hard to define, property of vision systems, that had been known to exist for a number of elementary attributes of visual stimuli, including color, orientation, depth, and motion, among others [195, 222, 225, 22, 64, 133, 143].

In general, the saliency of a stimulus can be interpreted as its state or quality of standing out (relative to other stimuli) in a scene. As a result, a salient stimulus will often "pop-out" at the observer [195, 190, 196, 138], such as a red dot in a field of green dots, an oblique bar among a set of vertical bars, a flickering message indicator of an answering machine, or a fast moving object in a scene with mostly static or slow moving objects. Another direct effect of the saliency mechanism is that it helps the visual perceptual system to quickly organize visual information, such as texture segmentation [11, 12, 95, 97, 147], or grouping [10, 168]. For example, it was shown in [140] that upon the brief inspection of a pattern, such as that depicted in the leftmost display of Figure I.1, subjects report the global percept of a "triangle pointing to the left". This percept is quite robust to the amount of (random) variability of the distractor bars, and to the orientation of the bars that make up the vertices of the triangle. In fact, these bars do not even have to be oriented in the same direction: the triangle percept only requires that they have sufficient orientation contrast with their neighbors. Another example of this type of perceptual grouping, as well as some examples of texture segregation, are shown in Figure I.1. Below each display we present the saliency maps produced

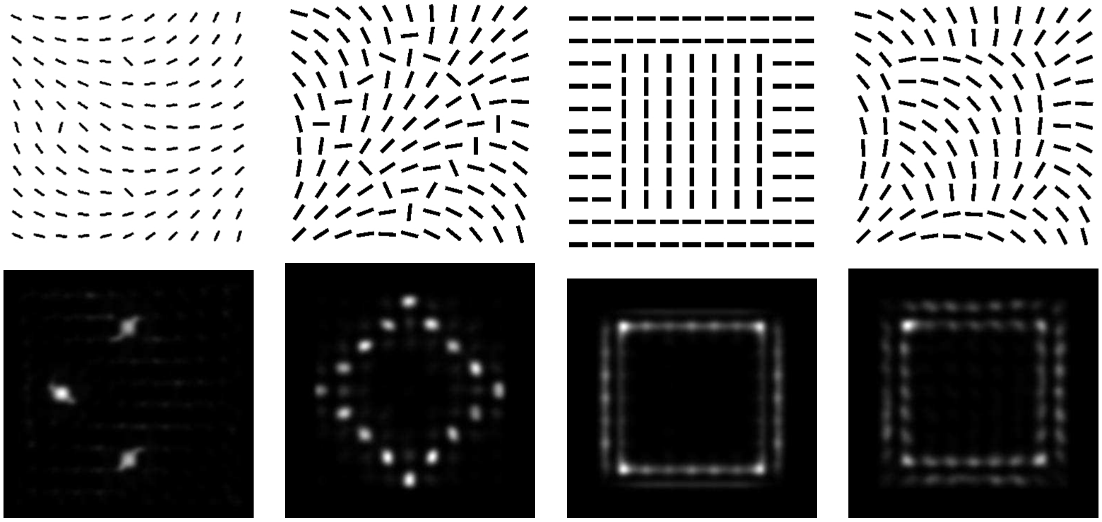


Figure I.1 Four displays (top row) and saliency maps produced by the algorithm proposed in this article (bottom row). These examples show that saliency analysis facilitates aspects of perceptual organization, such as grouping (left two displays), and texture segregation (right two displays).

by the saliency detector proposed in this work. Clearly, the saliency maps are informative of either the boundary regions or the elements to be grouped.

I.A.1 Two components of saliency

One common property of the above examples is that saliency is driven solely by the stimuli in each scene. However, psychological studies of visual attention have also shown that human saliency is not a single mechanism, but an interaction of two complementary mechanisms [90], *bottom-up* and *top-down* saliency. Bottom-up saliency is a fast, *stimulus-driven process*, which accounts for all of the aforementioned saliency examples. This mechanism is independent of any high-level visual tasks (such as recognition goals), and drives attention only by the properties of the stimuli in a visual scene. As another example, when you walk on a street, the traffic signs (or signals) always attract your attention, irrespective of whether you intend to look for them or not. Since it is purely stimulus-driven, bottom-up saliency is commonly believed to be a feed-forward visual processing in a nonconscious level, which is memory-free and reactive [106, 86, 115, 199]. Studies also indicate that the bottom-up saliency mechanism involves mostly localized processing: it typically arises from contrasts between a stimulus and its neighborhood. In fact, all of the pop-out examples mentioned above are accounted for by local stimulus contrast.

The other mechanism that guides the deployment of visual attention is a slower *memory-dependent* process, namely top-down saliency, which is determined by the (high-level) activities and visual tasks in which an organism is engaged. One important hallmark of top-down saliency is that, given the same scene (or the same pattern of visual stimuli), the most salient item(s) changes depending on the observer's tasks. For example, in a study of human eye movements [229], Yarbus recorded fixations and saccades that observers made while viewing natural objects and scenes. He showed that the patterns of saccades varied considerably for different questions that were asked to the observers prior to viewing the scene, for example, to estimate the economic level of the people in the scene, or to judge their ages. The studies of visual search experiments also indicate that for some types of displays, knowing the basic properties of a target (e.g. its color, shape, etc.) beforehand helps subjects to find a target much more efficiently than without the knowledge [135, 197, 29, 219, 222, 22].

Under the two-component saliency framework, both mechanisms can operate simultaneously and, for a given scene, the deployment of attention is believed to be determined by an interaction of the scene properties and the observer's set of attentional goals [228].

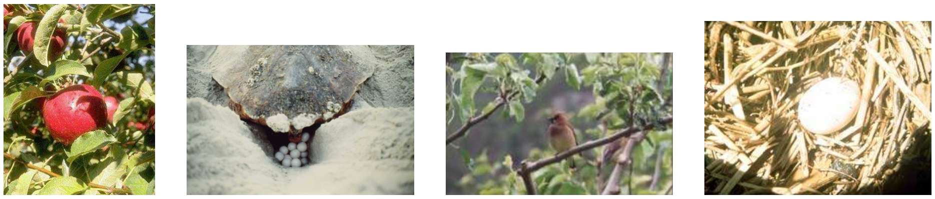
I.B Computational models for visual saliency

In recent years, there have been increasing efforts in introducing computational models for saliency mechanisms in both computer and biological literatures. In the computer vision community, although inspired by biological visual attention, little emphasis was given to replicating the psychophysical or physiological properties of human saliency. Instead, the majority of the research has been to develop saliency algorithms that are of direct interest to machine vision applications, such as object tracking and recognition. These studies are focused on extracting salient points (often called "interesting points") and applying them to build computer vision systems. On the other hand, in biological vision, most research addresses the understanding of how attentional mechanisms work, either through psychophysics experiments in psychology, or neural recordings in neurophysiology. Although a tremendous amount of knowledge about saliency has been amassed in this way, this literature is not rich in computational models. When such models are proposed, they tend to focus on high-level justifications for specific attention mechanisms, and do not necessarily translate into computer vision algorithms. In the following, we give an overview of the most popular saliency models/detectors in both literatures.

I.B.1 Saliency models in computer vision

The design of saliency detectors (often called *interest point detectors*) has been a significant subject of study in computer vision for several decades. Saliency detectors have been widely adopted in applications such as object tracking and recognition, and more recently, learning object detectors from weakly supervised (unsegmented) training examples [56, 184, 55, 111, 230, 32, 158]. In these applications, saliency is often justified as a pre-processing step that saves computation and improves robustness, facilitating the design of subsequent stages. As a result, most of the *existing saliency formulations proposed in this literature do not tie the optimality of saliency judgments to the specific goal of recognition, i.e. they are only bottom-up*, but focus on the extraction of image locations (*interest points*), that exhibit some *universally* desired, and mathematically well defined, properties such as stability under certain geometrical transformations.

Broadly speaking, saliency detectors in this literature can be divided into three major classes. The first, and most popular, class of saliency detectors treats the problem as the *detection of specific visual attributes*. Many detectors in this class emerged from research areas, such as structure-from-motion, or tracking [68, 60, 181, 200]. The most prevalent examples are edges and corners [68, 60, 181, 171, 200], but there have also been proposals for other low-level attributes, e.g. contours [177, 77, 4, 3, 150, 218], local symmetries [160, 72], and blobs [118]. These basic detectors can also be embedded in scale-space [116], to achieve detection invariance with respect to transformations such as scaling [126, 127], or affine mappings [127]. These bottom-up detectors have nice properties. For example, the salient image attributes can often be defined in mathematically explicit and optimal forms (e.g. [68]), which is desirable for the design of saliency detectors. The bottom-up detectors are also free of training, and mostly can be computed very efficiently. They, however, have significant limitations. Since the goals and constraints in object recognition are very different from those in the original domain where these detectors were proposed, the visual attributes deemed as salient may exist equally in a target and a background, and do not necessarily include any useful information for the recognition task at hand. Experimentally, a major drawback of these saliency detectors is that they do not generalize well for object recognition problems.



(a) (b) (c) (d)

Figure I.2 Challenging examples for existing saliency detectors. (a) apple among leaves; (b) turtle eggs; (c) a bird in a tree; (d) an egg in a nest.

For example, a corner detector will always produce a stronger response in a region that is strongly textured than in a smooth region, even though textured surfaces are not necessarily more salient than smooth ones. This is illustrated by the image of Figure I.2(a). While a corner detector would respond strongly to the highly textured regions of leaves and tree branches, it is not clear that these are more salient than the smooth apple. We would argue for the contrary. Similarly, in the image of Figure I.2(b), we present an example where contour-based saliency detection would likely fail. The image depicts a turtle laying eggs in the sand. While the eggs are arguably the most salient object in the scene, contour-based saliency would ignore them in favor of the large contours in the sand.

Some of these limitations are addressed by more recent, and generic, formulations of saliency. One idea that has recently gained some popularity is to define *saliency as image complexity*. Various complexity measures have been proposed in this context. For example, Yamada & Cottrell [226] defines saliency by the variance of Gabor filter responses over multiple orientations, while Sebe & Lew [174] equates saliency to the absolute value of the coefficients of a wavelet decomposition of the image, and Kadir & Brady [99] to the entropy of the distribution of local intensities. The major advantage of these *data-driven* definitions of saliency is a significant increase in flexibility, as they can detect any of the low-level attributes above (corners, contours, smooth edges, etc.), depending on the image under consideration. It is not clear, however, that saliency can always be equated with complexity. For example, Figure I.2 (c) and (d) show images containing complex regions, consisting of clustered leaves and straw that are not terribly salient. On the contrary, the much less complex image regions containing the bird or the egg appear to be significantly more salient. As with the first class, a key limitation of this class of detectors is that their salient points do not necessarily include any useful information for the recognition task at hand.

With respect object recognition applications, the third class of top-down saliency detectors is more interesting. The detectors of this class are normally trained for specific recognition problems under consideration. For example, authors in [66, 170, 214, 19] designed detectors based on the discriminant power of image regions (or features) for the classifications of an object class and a background class. In [136], top-down saliency is also measured by the signal-to-noise ratio (SNR) between target and background. Although top-down saliency detectors have been shown to have better performance for object recognition, especially in coping with image clutter, than bottom-up saliency detectors (see e.g., [66, 73, 65]), they are currently less popular in computer vision.

Finally a common limitation of all these saliency detectors in computer vision is that, although they are inspired by the saliency mechanisms of human vision, they seldom show any connection to biological vision, in terms of either the biological plausibility, or prediction of human saliency behaviors.

I.B.2 Saliency models in biological vision study

In the biological vision community, both the neurophysiological basis and psychophysical properties of visual saliency mechanisms have been extensively studied. Guided by these studies, most computational saliency models in this literature emphasize biological plausibility, and aim to replicate what is known about visual saliency and attention. With a few notable exceptions [219, 137], the overwhelming majority of these models have only considered bottom-up saliency mechanisms [106, 88, 163, 89, 115, 24, 67, 103, 85, 119], following the fact that the bottom-up visual pathway is better understood than its top-down counterpart, in terms of both the neural circuits involved and the resulting subject behavior.

Among the saliency models in this literature, three popular components are commonly adopted. The first component, which is also the first processing stage in most saliency models, is the extraction of early visual features. Inspired by the early visual pathway in biological vision, these features usually include low-level simple visual attributes, such as intensity contrast, color opponency, orientation, motion, and others (see e.g. [86, 88]). The second common component of many saliency models is the adoption of a "center-surround" formulation for bottom-up saliency (e.g. [88, 115, 219, 24, 67, 103, 85]). The "center-surround" formulation assumes that, in the absence of high-level (recognition) goals, saliency is determined by how distinct the stimulus at each location of the visual field is from its surrounding. This formulation is motivated by the ubiquity of "center-surround" mechanisms in the early stages of biological vision [108, 53, 81, 2, 98, 28, 104, 144], and has become dominant in this literature. The third common practice in the design of saliency detectors is the hypothesis of the existence of a *saliency map* [106], which can be generated through either the combination of intermediate feature-specific saliency maps [88, 219, 86], or the direct analysis of feature interactions [115]. The saliency map is a scalar, two-dimensional map whose activity topographically represents visual saliency, irrespective of the feature dimension that makes the location salient. On the basis of this scalar topographical representation, biasing attention to focus onto the most salient locations is reduced to drawing attention towards the locus of highest activity in the saliency map.

Given these commonly shared components, what differs between the computational saliency models is the strategy used to compute the saliency map. In what is perhaps the most popular model for bottom-up saliency [88], saliency is measured as the absolute difference between feature responses at a location and those in its neighborhood, in a center-surround fashion. This model has been shown to successfully replicate many observations from psychophysics [89, 151, 153], for both static and motion stimuli, and has been applied to the design of computer vision algorithms for robotics and video compression [84, 215, 182]. In [163], Rosenholtz measured the motion saliency of a target in a display as the number of standard deviations between the target velocity and the mean distractor velocity, and showed that it replicated a number of motion pop-out and asymmetry phenomena. On the other hand, in the famous Guided Search model [219], Wolfe emphasized the modulation of the bottom-up activation maps by top-down, goal-dependent, knowledge. In [115], Li argued that saliency maps are a direct product of the pre-attentive computations of primary visual cortex (V1), and implemented a saliency model inspired by the basic properties of the neural structures found in V1. This model has been shown to reproduce many psychophysical traits of human saliency, establishing a direct link between psychophysics and the physiology of V1. Lastly, in a recent proposal [103], Kienzle et al. relied on machine learning techniques to build a saliency model from recordings of human eye fixation on natural images, and showed that a center-surround receptive field emerged from the learned classifier.

While many of these saliency models are able to reproduce, to a certain extent, various known properties of biological vision, they lack a formal justification for their image processing steps in terms of a unifying computational principle for saliency. For example, it is not clear if these models are optimal in a well defined sense, whether that optimality is subject to any type of constraints (e.g., sparseness, computational parsimony, etc.), or whether they have any connection to the statistics of perceptual stimuli. With the absence of such a criterion it is difficult to evaluate, in an objective sense, the goodness of the proposed algorithms or to develop theories (and algorithms) for optimal saliency. Some more recent models have tried to address this problem, by deriving saliency mechanisms as optimal implementations of generic computational principles, such as the maximization of self-information [24], or "surprise" [85]. It is not yet clear how closely these models comply with classical psychophysics, since existing evaluations have been limited to the prediction of human eye fixation data. Finally, to the best of our knowledge, there has been few previous effort in either computer or biological vision literature to develop a unified formulation for both bottom-up and top-down saliency mechanisms (see, e.g., [231]).

I.C Contributions of the thesis

In this thesis we propose and investigate a new hypothesis for saliency mechanisms, which we refer to as *the discriminant saliency hypothesis*; saliency is, first and foremost, a discriminant process. Under this formulation, the saliency of a set of visual features is equated to the discriminant power of these features with respect to a classification problem, whose optimality is defined in a decision-theoretic sense under a constraint of computational parsimony. To justify the plausibility of the discriminant saliency hypothesis, one must address the following fundamental questions. First, what are the computational principles underlying the discriminant saliency hypothesis? Can it be applied to both bottom-up and top-down saliency mechanisms? How will it be implemented for each? Second, are the implementations biologically plausible, either physiologically or psychophysically, or both? For a valid formulation of saliency, this question is indispensable given the biological root of the saliency problem itself. Third, will the hypothesis lead to saliency detectors that significantly benefit problems in computer vision, especially recognition problems? This is very important for assessing the practical value of a saliency formulation. In this thesis, we answer each of the questions, by 1) providing a fully developed discriminant saliency formulation based on information theoretic principles, 2) investigating the biological plausibility of the hypothesis, and 3) studying the effectiveness of the derived saliency detectors in computer vision applications. The main contributions of this thesis are as follows.

I.C.1 Discriminant saliency hypothesis and its computational principles

As for the first contribution of the thesis, we propose the discriminant hypothesis for saliency: all saliency processes are optimal in a decision-theoretic sense with the constraint of computational parsimony. Under this formulation, the saliency of each location in the visual field is equated to the discriminant power of the image features with respect to a classification problem that opposes two classes of stimuli. The discriminant power of image features is measured in an information-theoretic sense, and the well known statistical properties of natural scenes are exploited to achieve computational parsimony. We then show that this hypothesis can be naturally implemented for both bottom-up and top-down saliency detectors.

I.C.2 Biological soundness of discriminant saliency

The second contribution lies in our efforts in collecting evidence to show the biological plausibility of the discriminant saliency hypothesis. In particular, we show that, by combining the discriminant saliency formulation with natural image statistics, the implementations of discriminant saliency are consistent with both neurophysiology and psychophysics of human saliency. With respect to neurophysiology, we show that under the assumption of natural image statistics, the computations of discriminant saliency can be implemented with a multi-layer neural network, which is consistent with the *standard neural architecture* in the primary visual cortex (V1), i.e., a combination of divisively normalized simple cell and complex cell [26, 71, 27, 1]. With respect to psychophysics, the ability of discriminant saliency to reproduce the classical behaviors of human saliency is evaluated. The experimental results show that discriminant saliency not only explains qualitative observations (such as pop-out for single feature search, disregard of feature conjunctions, and asymmetries between the existence and absence of a basic feature), but also makes surprisingly accurate quantitative predictions (such as the nonlinear aspects of human saliency perception, influence by the heterogeneity of the background, and the compliance of saliency asymmetries with Weber's law).

The significance of the consistency with neurophysiology and psychophysics is three-fold. First, these observations demonstrate, for the first time, a unifying computational principle of saliency that can be applied to explain both neurophysiological and psychophysical observations of early visual processing. Second, it provides a holistic functional justification for the standard architecture of V1; V1 has the capability to optimally detect salient locations in the visual field, when optimality is defined in a decision-theoretic sense and sensible simplifications are allowed for the sake of computational parsimony. Finally, the consistency implies that the basic neural structures in the early visual processing are capable of computing the fundamental operations of statistical inference (e.g., assessment of probabilities, implementation of decision rules, and feature selection) for visual signals that comply with the statistics of the natural world.

I.C.3 Applications of discriminant saliency in computer vision

In addition to comparisons with psychophysical and physiological properties of human saliency, we also evaluate the effectiveness of discriminant saliency detectors in solving various saliency problems of interest for computer vision. As object recognition is one of the most popular applications of saliency detectors, we first applied the top-down discriminant saliency detector to object recognition, particularly, the problem of learning from weakly supervised (unsegmented) training examples with cluttered background. Through extensive experiments, we show that the top-down discriminant saliency detector outperforms the state-of-the-art saliency principles with respect to a number of properties that are desirable for recognition: 1) the amount of information *relevant for the recognition task* which is captured by the salient points, 2) the ability to *localize* objects embedded in significant amounts of clutter, 3) the *robustness* of salient points to various image transformations and pose variability, and 4) the *richness* of the set of visual attributes that can be considered salient. We also compared the performance of the bottom-up discriminant saliency detector with other state-of-art saliency detectors, for the prediction of human eye movements on natural scenes in a free-viewing task. It is shown that the bottom-up discriminant saliency detector outperforms previously proposed methods.

I.C.4 Bayesian framework for integration of top-down and bottom-up saliency mechanisms

The final part of the thesis consists of an effort to study the connections between top-down and bottom-up saliency mechanisms. Since how the two mechanisms interact in biological vision, e.g., the underlying neural mechanisms, is not clearly understood, in this work, we only discuss applications in computer vision. In particular, we introduce 1) a probabilistic representation of salient locations, and 2) a Bayesian inference principle for the *integration of bottom-up and top-down saliency estimates*. The proposed Bayesian formulation is shown to have various interesting properties. First, it produces intuitive rules for the integration of the two saliency modes. Second, it supports the interpretation of top-down saliency as a focus-of-attention mechanism, which suppresses bottom-up salient points that are not relevant for the task of interest. Third, it provides evidence that bottom-up saliency has an important role when top-down routines are inaccurate (e.g. because they are learned from cluttered examples), but is not necessarily useful when the opposite holds. Fourth, it enables an explicit control of the relative weight of each saliency component in the final saliency estimates. Finally, it has a non-Bayesian interpretation as the simple multiplication of the two saliency maps, which enables a non-parametric extension of trivial computational complexity. The advantages of the Bayesian solution, over both top-down and bottom-up saliency in isolation, are illustrated in the context of recognition problems, both in terms of improving the ability to localize and segment objects from background clutter and preserving a great selectivity (recognition rate).

I.D Organization of the thesis

The rest of the thesis is organized as follows. In Chapter II, we introduce the discriminant saliency hypothesis and its computational principle in information-theoretic senses. We then investigate the statistical properties of natural scenes to achieve computationally efficient implementations of the discriminant saliency principles. The bottom-up and top-down implementations of the hypothesis are provided. In Chapter III, the question of the biological plausibility of the discriminant saliency, especially its neurophysiological plausibility, is investigated. The investigation reveals a functional justification of the basic neural structures in early visual processing. The study of the biological plausibility of discriminant saliency continues in Chapter IV, where we evaluate the ability of discriminant saliency to reproduce and explain the basic psychophysics of human saliency. Both the qualitative and quantitative properties of human saliency are considered in the context of classic visual search experiments. Chapter V and Chapter VI present results of various experiments designed to evaluate discriminant saliency as a solution for many saliency problems of significant interest in computer vision. In particular, we evaluate the performance of the top-down discriminant saliency detector on the problem of weakly supervised object detection from cluttered background in Chapter V, and the performance of the bottom-up discriminant saliency detector in other saliency problems, such as human fixation prediction, in Chapter VI. In Chapter VII, we discuss the Bayesian framework to integrate the output of the top-down and the bottom-up saliency detectors, and evaluate the resultant detector in the context of object recognition tasks. Conclusions are provided in Chapter VIII.

Chapter II

Discriminant saliency hypothesis and its computational principles

II.A Discriminant saliency hypothesis

The principle of discriminant saliency is rooted in a decision-theoretic interpretation of perception. Under this interpretation, perceptual systems evolve under the goal of producing decisions about the state of the surrounding environment that are optimal in a decision-theoretic sense, e.g. that have minimum probability of error. The evolutionary advantages of this type of optimality are evident: organisms that are less error-prone in identifying potential threats in the environment, are most likely to survive. This goal is complemented by the constraint, so called *computational parsimony*, that the brain has only limited computational power, and thus the perceptual mechanisms should be as efficient as possible. This constraint is essential to the sensory systems, and generally results in various representations, such as redundancy reduction [6], and sparse coding [146]. Saliency is one of the first steps of a visual system towards achieving the goal of understanding the surrounding environment, and is itself one representation of computational parsimony: it enables the organism to devote most computational resources to the locations of the visual field that are likely to provide most information of use to the decision-making process[1].

Compatible with the decision-theoretic interpretation of perception, we propose a hypothesis that *saliency is a discriminant process*. More specifically, saliency is defined with respect to two classes of stimuli: a *stimulus of interest* and a *null* hypothesis, composed of all the stimuli that are not salient. Given these two classes, the locations of the visual field that can be classified, with *lowest expected probability of error*, as containing the stimulus of interest are denoted as salient. In decision-theoretic terms, this is accomplished by 1) defining a binary classification problem that opposes the stimulus of interest to the null hypothesis, and 2) equating the saliency of each location in the visual field to the discriminant power (with respect to the classification problem) of the visual features extracted

[1] Note that, in this context, information does not necessarily correlate with *signal information* in the sense of Shannon [178].

from that location.

This discriminant formulation for saliency is clearly a departure from the other principles, and advances in at least two aspects. First, the definition is more generic and flexible. Because saliency is now defined with respect to two sets of visual stimuli, a set of salient visual features and the other set of the rest composing the null hypothesis, it is possible to apply this formulation to a broad class of saliency problems by assigning these two sets under different context. For example, by choosing appropriate instances of interest stimuli and null hypothesis, it is possible to specialize the discriminant saliency principle to either top-down or bottom-up saliency detection. If the saliency and null hypotheses are chosen respectively as the visual object class to be recognized and all other visual classes to be distinguished from the former in the visual recognition problem, the resulting saliency detector becomes top-down saliency detection. In this top-down context, saliency is contingent upon the existence of a collection of classes, and therefore for a given object, *different visual attributes will be salient in different recognition contexts.* For instance, while contours and shape will be most salient to distinguish a red apple from a red car, color and texture will be most salient when the same apple is compared to an orange. As illustrated by Figure II.1, this is consistent with perceptual saliency judgements. When a white fox is viewed against a forest, its color becomes very salient and recognition is easy. On the other hand, when the fox is presented against a background of white snow, color is no longer a salient feature and recognition becomes a lot more difficult. With respect to bottom-up applications where saliency is considered within one scene, the two sets of stimuli in discriminant saliency can be defined locally to oppose image attributes in one location from its surrounding regions, i.e. in a center-surround manner. In this set-up, saliency becomes contingent to the local context. For example, while a red dot looks very salient among a set of green dots, it is much less salient if embedded in a set of orange points. In this sense, discriminant saliency is flexible enough to detect any type of image features as salient for either top-down or bottom-up

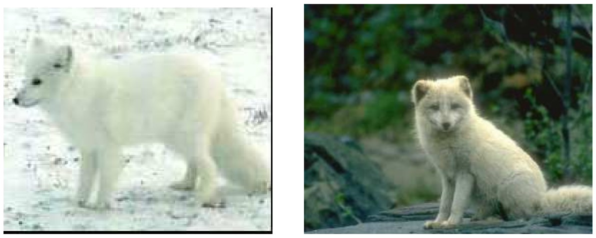


Figure II.1 The saliency of features like color depends on the viewing context.

implementations.

The second, and perhaps the most important, property of discriminant saliency is that it equates optimal saliency to the search for the most discriminant visual features for a binary classification problem. In particular, this is naturally formulated as an optimal feature selection problem: *optimal features for saliency are the most discriminant features for the binary classification problem that opposes the class of interest to the null classes.* From a computational standpoint, the search for discriminant features is a well-defined, and computationally tractable, problem that has been widely studied in the literature of decision-theory . We next consider efficient solutions to this problem.

II.B Computational principles for discriminant saliency

II.B.1 Minimum Bayes error

We start by recalling the well known result that, for the classification problem defined by 1) a feature space $\mathcal{X}$, and 2) a random variable Y that assigns $\mathbf{x} \in \mathcal{X}$ to one of M classes, $i \in \{1, \ldots, M\}$, the minimum probability of classification error is achieved by the *Bayes* classifier [47]

$$g^*(\mathbf{x}) = \arg\max_i P_{Y|\mathbf{X}}(i|\mathbf{x}). \tag{II.1}$$

This probability of error is denoted as the *Bayes error* (BE)

$$L^* = 1 - E_{\mathbf{x}}[\max_i P_{Y|\mathbf{X}}(i|\mathbf{x})], \tag{II.2}$$

where $E_{\mathbf{x}}$ means the expectation with respect to $P_{\mathbf{X}}(\mathbf{x})$. Since 1) the BE depends only on $\mathcal{X}$, not on the implementation of the classifier itself, 2) it lower bounds the probability of error of any classifier on $\mathcal{X}$, and 3) there is at least one classifier (the Bayes classifier) that achieves this lower bound, the minimization of BE is a natural optimality criteria for feature selection.

In practice, however, the feasibility of applying such criterion is constrained by its computational complexity. In particular, the implementation of discriminant saliency requires 1) the design of a large number of classifiers, for example, as many as the total number of object classes to recognize for a top-down context, or the number of image locations for a bottom-up context; and 2) classifier tuning whenever the visual concepts included in the two hypothetical sets changes, such as adding and deleting new classes to the recognition problem for top-down applications. It is therefore important to adopt criteria that are computationally efficient, preferably reusing computation from the design of one classifier to the next. It has, however, long been known that direct minimization of the BE is a difficult problem, due to the non-linearity associated with the $\max(\cdot)$ operator in (II.2). Consider, for example, the popular strategy of feature selection by sequential search, where at iteration n the previous best feature subset, $\mathbf{X}_{n-1}$, is augmented with the feature set $\mathbf{X}_a$ to obtain the best new solution $\mathbf{X}_n = (\mathbf{X}_a, \mathbf{X}_{n-1})$. When the goal is to minimize BE, such algorithms cannot be implemented efficiently because the $\max(\cdot)$ operator makes it impossible to express $E_{\mathbf{X}_n}[\max_i P_{Y|\mathbf{X}_n}(i|\mathbf{x}_n)]$ as a modular combination of $E_{\mathbf{X}_{n-1}}[\max_i P_{Y|\mathbf{X}_{n-1}}(i|\mathbf{x}_{n-1})]$ and a function of $\mathbf{X}_a$.

II.B.2 Infomax

An alternative optimality criteria is to select the features that are most informative about the class label [8, 18, 227, 156, 207, 212, 46]. This is frequently referred as the *infomax* criteria, due to its connections to the infomax principle for the organization of perceptual systems [117, 5, 7].

Definition 1. *Consider a M-class classification problem with observations drawn from a random variable $\mathbf{Z} \in \mathcal{Z}$, and a feature transformation $T : \mathcal{Z} \to \mathcal{X}$. $\mathcal{X}$ is an infomax feature space if and only if it maximizes the mutual information*

$$I(\mathbf{X}; Y) = \sum_i \int p_{\mathbf{X},Y}(\mathbf{x}, i) \log \frac{p_{\mathbf{X},Y}(\mathbf{x}, i)}{p_{\mathbf{X}}(\mathbf{x}) p_Y(i)} d\mathbf{x} \tag{II.3}$$

between class Y and feature vector $\mathbf{X}$.

The mutual information can also be written as

$$I(\mathbf{X}; Y) \;\; = \;\; \sum_i P_Y(i) KL\left[P_{\mathbf{X}|Y}(\mathbf{x}|i) || P_{\mathbf{X}}(\mathbf{x})\right] \tag{II.4}$$

where

$$KL[p\mathbf{x}||q\mathbf{x}] = \int p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})} d\mathbf{x} \tag{II.5}$$

is the Kullback-Leibler (K-L) divergence between the distributions $p(\mathbf{x})$ and $q(\mathbf{x})$. This is a measure of the average distance between each of the class conditional distributions $P_{\mathbf{X}|Y}(\mathbf{x}|i)$ and their average, $P_{\mathbf{X}}(\mathbf{x}) = \sum_i P_Y(i) P_{\mathbf{X}|Y}(\mathbf{x}|i)$ and gives an intuitive discriminant interpretation to the infomax solution: *the infomax space is that in which the distribution of each class is as different as possible from the mean distribution over all classes.* It therefore favors spaces where the classes are as separated as possible.

With respect to computation, it has been shown in [206] that the infomax criterion enables efficient feature selection strategies. For example, consider for the strategy of feature selection by sequential search, where at iteration n the previous best feature subset, $\mathbf{X}_{1,n-1} = \{X_1, \ldots, X_{n-1}\}$, is augmented with the feature X_n to obtain the best new solution $\mathbf{X}_{1,n} = (\mathbf{X}_{1,n-1}, X_n)$. Authors in [206] showed that mutual information can be decomposed as following, by applying the chain rule of relative entropy [37] to (II.4),

$$\begin{aligned} I(\mathbf{X}_{1,n}; Y) &= \left\langle KL\left[P_{\mathbf{X}_{1,n}|Y}(\mathbf{x}_{1,n}|i) || P_{\mathbf{X}_{1,n}}(\mathbf{x}_{1,n})\right]\right\rangle_Y \\ &= \left\langle KL\left[P_{X_n|\mathbf{X}_{1,n-1},Y}(x_n|\mathbf{x}_{1,n-1}, i) || P_{X_n|\mathbf{X}_{1,n-1}}(x_n|\mathbf{x}_{1,n-1})\right]\right\rangle_Y \\ &\quad + \left\langle KL\left[P_{\mathbf{X}_{1,n-1}|Y}(\mathbf{x}_{1,n-1}|i) || P_{\mathbf{X}_{1,n-1}}(\mathbf{x}_{1,n-1})\right]\right\rangle_Y \\ &= I(X_n; Y|\mathbf{X}_{1,n-1}) + I(\mathbf{X}_{1,n-1}; Y). \end{aligned} \tag{II.6}$$

This allows an efficient implementation of the sequential search strategy, since the mutual information at iteration n can be computed as a sum of the same quantity at iteration $n-1$ and a term that depends on the additional feature X_n. This therefore makes the infomax principle more tractable than the minimization of BE. Finally, the two solutions are closely related and frequently similar [207, 206]. For all these reasons, we adopt the infomax principle as a criterion for salient features in this work.

II.C Computational parsimony and natural image statistics

While (II.6) enables the reuse of computation between consecutive feature selection iterations, the term $I(X_n; Y|\mathbf{X}_{1,n-1})$ can still be prohibitively expensive as the dimension of $\mathbf{X}_{1,n-1}$ increases since it requires high-dimensional density estimates. As we have previously mentioned, the constraint of computational parsimony suggests the search for approximations of (II.6) that enable efficient computations.

II.C.1 Natural image statistics for feature dependency

To achieve computational efficiency, we resort to the proposal of Attneave, Barlow, and others [5, 6, 7], that perception is tuned to the environment and is able to exploit the statistics of natural stimuli to reduce computational complexity. Of particular interest is a known statistical property of band-pass features, such as Gabor filters or wavelet coefficients, extracted from natural images: that such features exhibit strongly *consistent* patterns of dependence across a wide range of imagery [25, 79, 187]. One example of these regularities is illustrated by Figure II.2, which presents three images, the histograms of one coefficient of their wavelet decomposition, and the conditional histograms of that coefficient, given the state of the co-located coefficient of immediately coarser scale (known as its "parent"). Although the drastically different visual appearance of the images affects the scale (variance) of the marginal distributions, *their shape, or that of the conditional distributions between coefficients,* is quite stable. The observation that these distributions follow a canonical (bow-tie) pattern, which is simply rescaled to match the marginal statistics of each image, is remarkably consistent over the set of natural images. This "bow-tie" shaped distribution, in fact, has been widely observed for many natural image feature pairs [25], other than the "parent/child" feature pairs shown in Figure II.2. This consistency indicates that, even though the fine details of feature dependencies may vary from scene to scene, the coarse structure of such dependencies follows a universal statistical law that appears to hold for all natural scenes. This, in turn, suggests that feature dependencies are not greatly informative about the image class. The following theorem shows that, when this is the case, (II.3) can be drastically simplified.

Theorem 1. *Let* $\mathbf{X} = \{X_1, \ldots, X_d\}$ *be a collection of features, and* Y *the class label. If*

$$\frac{\sum_{i=1}^{d} [I(X_i; \mathbf{X}_{1,i-1}) - I(X_i; \mathbf{X}_{1,i-1}|Y)]}{\sum_{i=1}^{d} I(X_i; Y)} = 0 \tag{II.7}$$

where $\mathbf{X}_{1,i} = \{X_1, \ldots, X_i\}$, *then*

$$I(\mathbf{X}; Y) = \sum_{i=1}^{d} I(X_i; Y). \tag{II.8}$$

Proof. The proof of the theorem is given in [206]. ■

The left hand side of (II.7) measures the ratio between the information for discrimination contained in feature dependencies and that contained in the features themselves. While this ratio is usually non-zero, it is generally small for band-pass natural image features, and smallest in the locations where the features are most discriminant. Hence, the approximation of

$$I(\mathbf{X}; Y) \approx \sum_{k} I(X_k; Y) \tag{II.9}$$

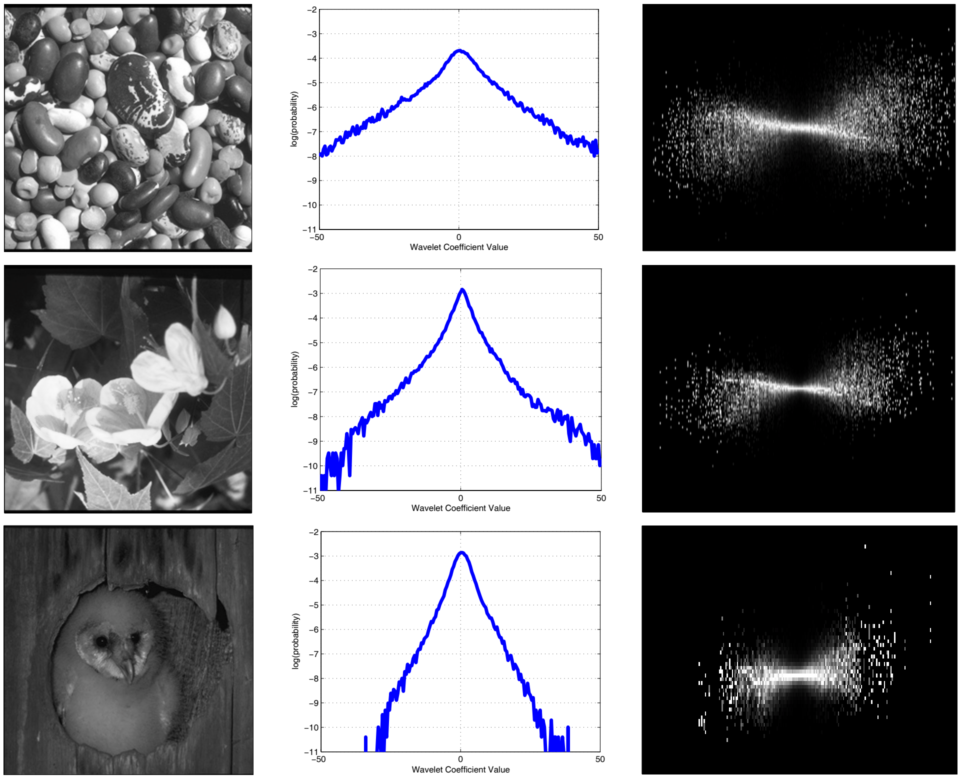


Figure II.2 Constancy of natural image statistics. Left: three images. Center: each plot presents the histogram of the same coefficient from a wavelet decomposition of the image on the left. Right: conditional histogram of the same coefficient, conditioned on the value of its parent. Note the constancy of the shape of both the marginal and conditional distributions across image classes.

is a sensible compromise between decision theoretic optimality and computational parsimony. Note that this approximation *does not* assume that the features are independently distributed, but simply that their dependencies are not informative about the class. This approximation has been widely tested in computer vision literature. For example, it has been shown that, for image classification problems, accounting for dependencies between feature pairs can be beneficial but their appears to be little gain in considering larger conjunctions [209, 206]. The gains from single feature to pairwise conjunctions are also not overwhelming. It has also been shown that large classes of texture can be synthesized from models that only enforce constraints on the marginal distributions of wavelet-like features [70, 232, 187]. In summary, the reduced infomax cost, in (II.9), enables a substantial computational simplification: because the mutual informations on the right hand side of (II.9) only require marginal density estimates, this computational cost can be drastically reduced.

II.C.2 The generalized Gaussian distribution

In the previous section, we showed that by exploiting the dependence properties of natural image features, the computation of the infomax principle can be drastically simplified. In fact, one important idea this work seeks is, in the spirit of Attneave, Barlow, and others [5, 6, 7], an interpretation of the optimal saliency detector as a mechanism that exploits the regularities of the visual world to implement the optimal solution to the saliency problem in a computationally efficient manner. In this section, we continue to apply other well known statistics of natural scenes to increase computational efficiency.

We start by noticing that the computation of (II.9) requires empirical estimates of the marginal mutual information $I(X_k; Y)$. These, in turn, require estimates of the marginal probability densities of features X_k, $P_{X_k}(x)$, and their class-conditional probability densities, $P_{X_k|Y}(x|i)$. Various studies in natural image processing showed that the probability densities of band-pass image features are well approximated by a generalized Gaussian distribution (GGD) [129, 54, 122, 34, 79, 16],

$$P_X(x; \alpha, \beta) = \frac{\beta}{2\alpha\Gamma(1/\beta)} \exp\left\{ -\left(\frac{|x|}{\alpha}\right)^{\beta} \right\}, \tag{II.10}$$

where $\Gamma(z) = \int_0^\infty e^{-t} t^{z-1} \mathrm{d}t$, $t > 0$, is the Gamma function, α a *scale* parameter, and β a *shape* parameter. The parameter β controls the decaying rate from the peak value, and defines a sub-family of the GGD (e.g. the Laplace family when $\beta = 1$ or the Gaussian family when $\beta = 2$).

The GGD has various interesting properties. First, various low-complexity methods exist for the estimation of the parameters (α, β), including the method

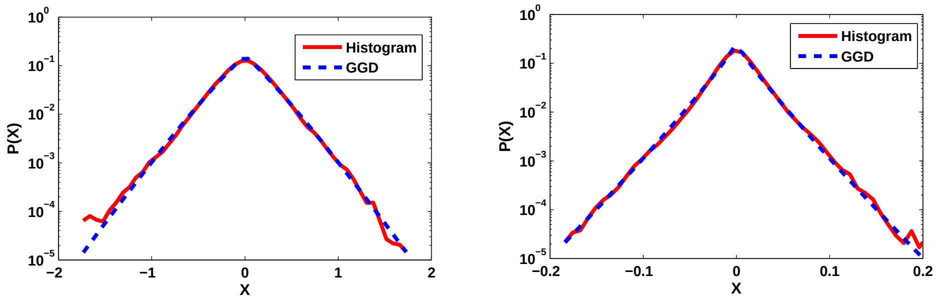


Figure II.3 Examples of GGD fits obtained with the method of moments.

of moments [179], maximum likelihood (ML) [43] and minimum mean-square-error [79]. In the implementation presented in this article, we have adopted the method of moments for all parameter estimation, because it is computationally more efficient. Under the method of moments, α and β are estimated through the relationships

$$\sigma^2 = \frac{\alpha^2\Gamma(\frac{3}{\beta})}{\Gamma(\frac{1}{\beta})} \quad \text{and} \quad \kappa = \frac{\Gamma(\frac{1}{\beta})\Gamma(\frac{5}{\beta})}{\Gamma^2(\frac{3}{\beta})}, \tag{II.11}$$

where σ^2 and κ are, respectively, the variance and kurtosis of X

$$\sigma^2 = E_X[(X - E_X[X])^2], \text{ and } \kappa = \frac{E_X[(X - E_X[X])^4]}{\sigma^4}.$$

This method has been shown to produce good fits to natural images [79]. Figure II.3 shows two examples of the fits that we obtained, with this method, for the responses of two Gabor filters.

Second, it leads to closed form solutions for various information theoretic quantities. For example, when both the class-conditional densities $P_{X|Y}(x|i)$ and the marginal density $P_X(x)$ are well approximated by the GGD, the mutual information $I(X;Y)$ has a closed form. This follows from (II.4)

$$I(X;Y) = \sum_i P_Y(i) KL\left[P_{X|Y}(x|i)||P_X(x)\right],$$

and [43],

$$KL[P_X(x;\alpha_1,\beta_1)||P_X(x;\alpha_2,\beta_2)] = \\ \log\left(\frac{\beta_1\alpha_2\Gamma(1/\beta_2)}{\beta_2\alpha_1\Gamma(1/\beta_1)}\right) + \left(\frac{\alpha_1}{\alpha_2}\right)^{\beta_2}\frac{\Gamma((\beta_2+1)/\beta_1)}{\Gamma(1/\beta_1)} - \frac{1}{\beta_1}. \tag{II.12}$$

It can also be shown that

$$H(X|Y=i) = \frac{1}{\beta_i} + \log \frac{2\alpha_i \Gamma(\frac{1}{\beta_i})}{\beta_i} \tag{II.13}$$

where $H(X|Y=i) = -\int P_{X|Y}(x|i) \log P_{X|Y}(x|i) dx$ is the entropy of feature X given its class label $Y = i$. These closed forms play an important role in the efficient implementation of discriminant saliency. In the following, we present top-down and bottom-up implementations of the discriminant saliency principle. These implementations are used to produce all saliency detection results presented in later chapters.

II.D Top-down discriminant saliency detector

We start with the implementation of a top-down discriminant saliency detector aiming at object recognition. As we have previously discussed, the discriminant saliency principle is intrinsically grounded on a classification problem, and thus can be naturally applied to top-down saliency detection. In the context of object recognition, the two classes of stimuli of discriminant saliency, the stimulus of interest and the null hypothesis, are simply the object class to be recognized and all other visual classes to be distinguished from the former in the visual recognition problem. Note that this assignment is applicable for either the single-class recognition problem which consists of an object class and a generic background class, or a multi-class recognition problem where more than one object classes are of interest. For the latter, a saliency detector is learned for each object class based on a *one-vs-all* classification problem, which opposes the object class under consideration to all other classes of interest. The design of a top-down discriminant saliency detector has two components: feature selection and saliency detection.

II.D.1 Discriminant feature selection

We have seen in Section II.C that, given a space $\mathcal{X}$ of band-pass features extracted from natural images, the best K-feature subset can be selected by computing the marginal mutual informations $M_k = I(Y; X_k)$, for all k, and selecting the K features of largest M_k. Note that such a simple feature selection strategy is possible also due to the fact that mutual information is always positive. The marginal mutual informations can be computed efficiently with (II.4) and (II.12). One final issue is that none of the feature selection costs considered so far is asymmetric: in general, discrimination does not differentiate between situations where 1) the feature is present (strong responses) in the object class of interest, but absent (weak response) in the null hypothesis, and 2) vice versa. Although both cases lead to low probability of error, feature absence is less interesting for saliency, which is an inherently asymmetric problem.

However, detecting if a feature is discriminant due to presence or absence in the class of interest is usually not difficult. For generalized Gaussian features, it suffices to note that feature absence produces a narrow GGD, close to a delta function, while feature presence increases the variance of the distribution (see Figure II.4 for an example). Since the former has lower entropy than the latter, discriminant features which are absent from the class of interest fail the test

$$H(X_k|Y=1) > H(X_k|Y=0), \tag{II.14}$$

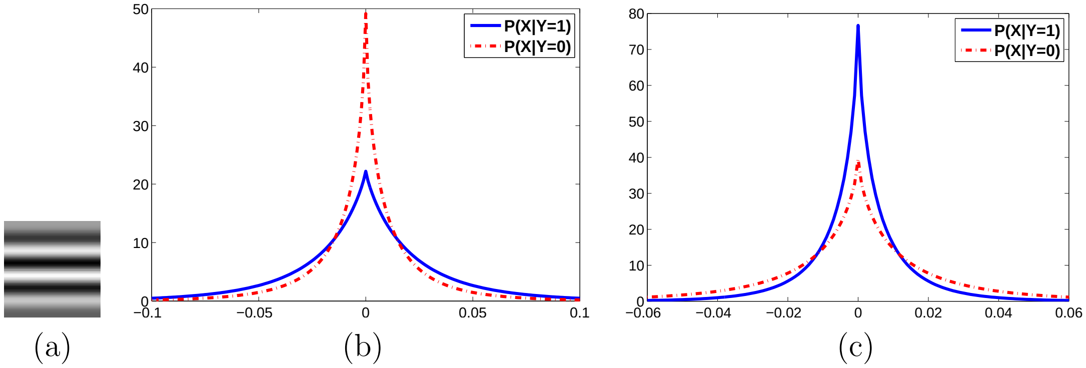
or, using (II.13),

$$\log\frac{\alpha_1}{\alpha_0} > \left(\frac{1}{\beta_0} - \frac{1}{\beta_1}\right) + \log\frac{\Gamma(\frac{1}{\beta_0})\beta_1}{\Gamma(\frac{1}{\beta_1})\beta_0}. \tag{II.15}$$

Such features should not be considered during feature selection.

II.D.2 Saliency detection

Given a set of selected salient features, in saliency detection, to be compatible with the biological plausibility and the central idea of discriminant saliency that takes band-pass features as basic elements, we adopt the classical proposal by Malik and Perona [119], which consists of a nonlinearity based on half-wave



(a) (b) (c)

Figure II.4 Illustrations of the conditional marginal distributions (GGDs) for the responses of a feature with horizontal bars (a), when (b) it is present (strong responses) in the object class ($Y = 1$) but absent (weak responses) in the null hypothesis ($Y = 0$), or (c) vice versa. Note that the absence of a feature always leads to narrower GGDs than the presence of the feature.

rectification, leading to the saliency map

$$S_D(l) = \sum_{j=1}^{2n} w_j x_j^2(l), \tag{II.16}$$

where l is an image location, $x_j(l), j = 1, \ldots, 2n$ a set of channels resulting from half-wave rectification of the outputs of n saliency filters $F_k, k = 1, \ldots, n$

$$\begin{aligned} x_{2k-1}(l) &= \max[(-I * F_k)(l), 0] \\ x_{2k}(l) &= \max[(I * F_k)(l), 0], \end{aligned} \tag{II.17}$$

I the input image, $*$ the convolution operator, and w_k weights which we set to the marginal mutual information. Salient locations are then located on the saliency map S_D by feeding it to a non-maximum suppression module, which has been shown to be prevalent in biological vision [154, 106, 149, 104]. In particular, the location of largest saliency is found, and its spatial scale set to the size of the region of support of the saliency filter with strongest response at that location. All neighbors within a circle of radius determined by this scale are then suppressed (set to zero) and the process is iterated. The overall procedure is illustrated in Figure II.5. We emphasize that, with respect to the Malik-Perona model, this saliency

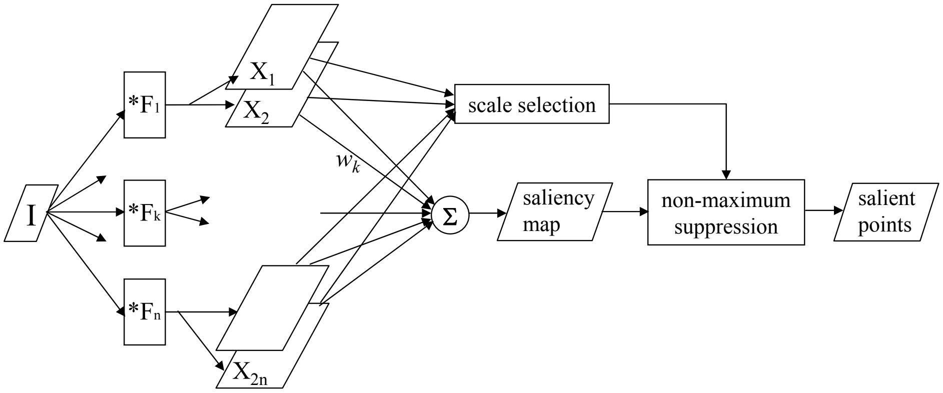


Figure II.5 Implementation of the top-down discriminant saliency detector.

detector contains a simple but very significant conceptual difference: both filters F_k and pooling weights w are chosen to maximize discrimination between the class of interest and the *all* class.

Determining the number of salient features

One parameter, required for the implementation of the top-down discriminant saliency detector, is the number of features, or filters, to include in the detector. To determine this parameter we start by noting that, if the output of a saliency detector is highly informative about the presence (or the absence) of the class of interest in its input images, it should be possible to classify the images (as belonging to the class of interest or not) by classifying the associated saliency maps. This suggests the use of a *saliency map classifier* as a means to determine the optimal number of features, using standard cross-validation procedures. We rely on a very simple saliency map classifier, based on a support vector machine (SVM) [205], which is applied to the histogram of the saliency map of (II.16) (from here on referred to as the *saliency histogram*) derived from each image. The accuracy of this classifier is also an objective measure of saliency detection performance that can be used to compare different detectors.

II.E Bottom-up implementation of discriminant saliency

As we have mentioned before, one nice property of discriminant saliency is that, by changing the definitions of the *stimulus of interest* and the null hypothesis, it can also be applied to bottom-up saliency detection. In this section, we consider the implementation of a bottom-up discriminant saliency detector.

II.E.1 Center-surround saliency

Recall that bottom-up saliency is a stimulus-driven mechanism which is memory free, and drives attention only by the properties of visual attributes in a scene. Biological vision studies have shown that bottom-up saliency is tightly connected to the the ubiquity of "center-surround" mechanisms in the early stages of visual processing. A significant body of psychophysical evidence suggests that an important role of this mechanism is to detect stimuli that are distinct from the surrounding background. For example, it has long been established that the simplest visual concepts, e.g. bars, can be highly salient when viewed against a background of similar visual concepts, e.g. other bars, that differ from them only in terms of low-level properties such as color or orientation. This center-surround property has been recognized as one of the fundamental guiding principles for the design of many psychophysical experiments, in the area of visual attention [195, 222, 225, 22, 64, 133, 143].

In addition to psychophysics, the same observations also emerged from neurophysiological studies of human vision [108, 53, 81, 2, 28, 104, 144]. For instance, anatomy studies of the primary visual cortex (V1) have shown that cells from this part of the brain are highly sensitive to oriented edges falling inside their receptive fields. In general, a cell in V1 fires vigorously when an edge of a certain orientation angle (so called *preferred orientation*) is inside its receptive field. However, the response of the same cell to the same orientation stimulus can be significantly inhibited, or excited, when some other orientation stimulus is

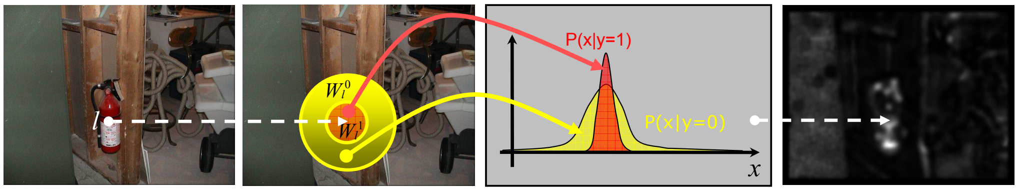


Figure II.6 Illustration of the discriminant center-surround saliency. Center and surround windows are analyzed at each location to infer the discriminant power of features at that location.

present immediately outside the receptive field [2, 104, 28, 102, 114].

Inspired by this evidence from biological vision, the center-surround formulation of bottom-up saliency has been widely exploited for the design of computational models for saliency (e.g., [88]). Interestingly, this center-surround formulation is also plausible under the discriminant saliency definition, where the background (surround) stimulus defines a *null* hypothesis, and salient visual features are those that best discriminate a foreground (center) stimulus from that null hypothesis. In particular, under the assumption that bottom-up saliency is driven by linear filtering, the visual stimulus is first linearly decomposed into a set of feature responses, and the saliency of each location is inferred from a sample of these responses. In this *discriminant center-surround saliency*, we hypothesize that the goal of the pre-attentive visual system is to optimally drive the deployment of attention and that, in the absence of high-level objectives, this reduces the saliency of each location to how distinct it is from the surround background. In decision-theoretic terms, it corresponds to 1) identifying the null hypothesis for the saliency of a location with the set of feature responses that surround it, and 2) defining bottom-up saliency as *optimal discrimination* between the responses at the location and its surround.

Mathematically, as illustrated in Figure II.6, discriminant saliency is measured by introducing two windows, $\mathcal{W}_l^0$ and $\mathcal{W}_l^1$, at each location l of the visual field. $\mathcal{W}_l^1$ is an inner window that accounts for a *center* neighborhood, and $\mathcal{W}_l^0$

an outer annulus that defines its *surround*. The responses of a pre-defined set of d features, henceforth referred to as *feature vectors*, are measured at all image locations within the two windows, and interpreted as observations drawn from a random process $\mathbf{X}(l) = (X_1(l), \ldots, X_d(l))$, of dimension d, conditioned on the state of a binary class label $Y(l) \in \{0, 1\}$. The feature vector observed at location j is denoted by $\mathbf{x}(j) = (x_1(j), \ldots, x_d(j))$, and feature vectors are independently drawn from the class-conditional probability densities $P_{\mathbf{X}(l)|Y(l)}(\mathbf{x}|i)$. Learning is supervised, in the sense that the assignment of feature vectors to classes is known: $\mathbf{x}(j)$ is drawn from class $Y(l) = 1$ when $j \in \mathcal{W}_l^1$ and from class $Y(l) = 0$ when $j \in \mathcal{W}_l^0$. For this reason, class $Y(l) = 1$ is denoted as the *center* class and class $Y(l) = 0$ as the *surround* class. Discriminant saliency defines the classification problem that assigns the observed feature vectors $\mathbf{x}(j), \forall j \in \mathcal{W}_l = \mathcal{W}_l^0 \cup \mathcal{W}_l^1$, into center and surround. The saliency judgement at an image location l is quantified by the sum of marginal information of (II.9), i.e.

$$\begin{aligned} S(l) &= \sum_{k=1}^{d} I_l(X_k; Y) \\ &= \sum_{k=1}^{d} \sum_{i=0}^{1} P_{Y(l)}(i) \cdot KL\left[P_{X_k(l)|Y(l)}(x_k(l)|i) || P_{X_k(l)}(x_k(l))\right] \end{aligned} \tag{II.18}$$

Note that the l subscript emphasizes the fact that the mutual information is defined locally, within $\mathcal{W}_l$. The function $S(l)$ is referred to as the *saliency map* and saliency detection consists of identifying the locations where (II.18) is maximal. These are the most informative locations with respect to the discrimination between center and surround. The overall implementation of the bottom-up saliency detector is summarized in Figure II.7, whose components are described in detail in the following sections.

II.E.2 Extraction of intensity and color features

As illustrated in Figure II.7, an input image is subject to a stage of feature decomposition. The choice of a specific set of features is not crucial for the

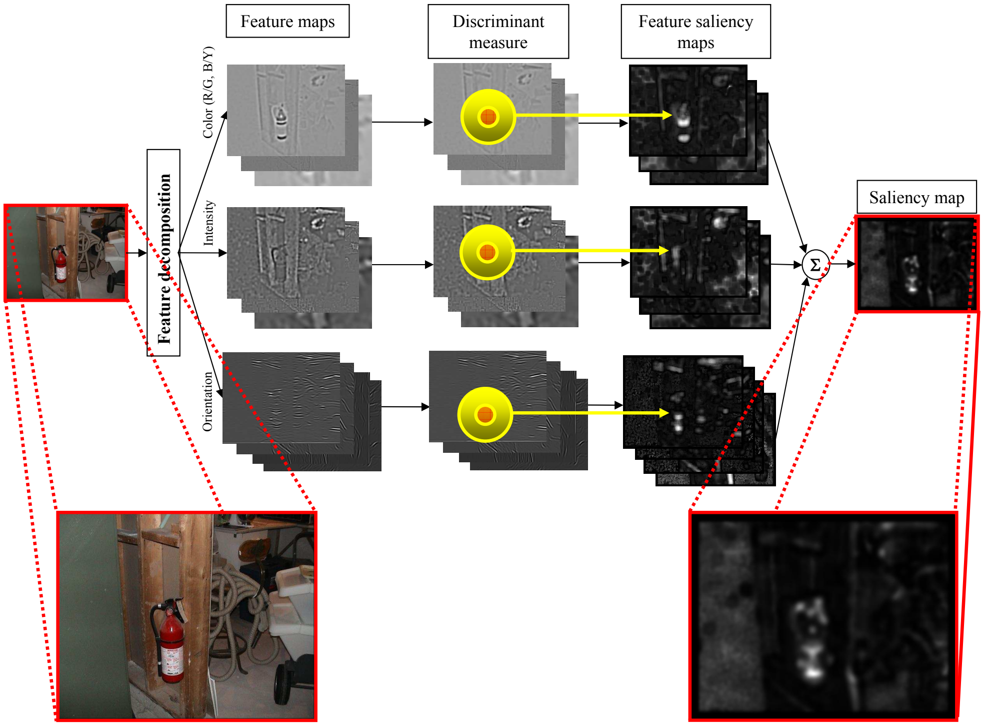


Figure II.7 Bottom-up discriminant saliency detector. The visual field is projected into feature maps that account for color, intensity, orientation, scale, etc. Center and surround windows are then analyzed at each location to infer the expected classification confidence power of each feature at that location. Overall saliency is defined as the sum of all feature saliency.

proposed saliency detector. We have obtained similar results with various types of wavelet or Gabor decompositions. In this work, we rely on a feature decomposition proposed in [88], which was loosely inspired by the earliest stages of biological visual processing. This establishes a common ground for comparison with the previous saliency literature. In this process, the input image is first decomposed into an intensity map (I), and four broadly-tuned color channels (R, G, B, and Y),

$$\begin{aligned} I &= (r+g+b)/3, \\ R &= \lfloor \tilde{r} - (\tilde{g}+\tilde{b})/2 \rfloor_+, \\ G &= \lfloor \tilde{g} - (\tilde{r}+\tilde{b})/2 \rfloor_+, \\ B &= \lfloor \tilde{b} - \tilde{(}r+\tilde{g})/2 \rfloor_+, \\ Y &= \lfloor (\tilde{r}+\tilde{g})/2 - |\tilde{r}-\tilde{g}|/2 \rfloor_+, \end{aligned}$$

where $\tilde{r} = r/I, \tilde{g} = g/I, \tilde{b} = b/I$, and $\lfloor x \rfloor_+ = \max(x, 0)$. The four color channels are in turn combined into two color opponent channels, $R-G$ for red/green and $B-Y$ for blue/yellow opponency. These and the intensity map are then convolved with three Laplacian of Gaussian (LoG; also known as Mexican hat wavelet) filters,

$$l(x,y) = -\frac{1}{\pi\sigma^4}\left(1 - \frac{x^2+y^2}{2\sigma^2}\right)\exp\left(-\frac{x^2+y^2}{2\sigma^2}\right),$$

with central frequencies ($\omega = \frac{1}{\sqrt{2}\pi\sigma}$) at 0.04, 0.08 and 0.16 cycles/pixel, to generate nine feature channels.

II.E.3 Gabor wavelets

The second set of features adopted in the implementation are orientation filters implemented by 2-D Gabor filters. A 2-D Gabor function is a sinusoid modulated by a Gaussian,

$$\begin{aligned} g(x,y) &= K\exp(-\pi(a^2(x-x_0)_r^2 + b^2(y-y_0)_r^2)) \\ &\quad \cdot \exp(j(2\pi F_0(x\cos\omega_0 + y\sin\omega_0) + P)), \end{aligned} \qquad \text{(II.19)}$$

with

$$\begin{aligned} (x-x_0)_r &= (x-x_0)\cos\theta + (y-y_0)\sin\theta \\ (y-y_0)_r &= -(x-x_0)\sin\theta + (y-y_0)\cos\theta. \end{aligned}$$

K, (a,b), θ, and (x_0, y_0) control the orientation and shape of the Gaussian envelope, and (F_0, ω_0) and P the spatial frequency and phase of the sinusoidal carrier. These parameters are usually defined so as to produce a tiling of the space/frequency volume.

It has been suggested that the linear components of simple cells in the primary visual cortex (V1) of higher vertebrates can be modeled by 2-D Gabor functions that satisfy certain neurophysiological constraints [38, 39, 124, 40, 109, 217, 41]. To produce an admissible wavelet basis, these Gabor functions are sometimes further constrained to have zero mean [112]. Gabor tilings have also been shown to be complete [39] and optimal for image representation, in the sense of minimizing joint uncertainty in space and frequency [112]. Finally, it has been shown that filters learned from natural images (including intensity, color and stereo), by sparse coding or independent components analysis (ICA), tend to be Gabor-like [146, 13, 78, 44, 203, 204].

In the context of discriminant saliency detection, our experience is that the precise choice of the Gabor function does not influence the overall saliency judgements in a significant manner. Rather than a particular wavelet, it appears to be more important to apply the wavelet decomposition across a wide range of scales, as these tend to produce different types of salient attributes. Figure II.8 shows an example of discriminant saliency for a texture from the Brodatz database. It can be seen from the figure that, while at the coarsest scale (4^{th} image from the left) the parallelism between the two horizontal lines and the symmetry between the two t-junctions on the left are deemed most salient, at the intermediate scale (3^{rd} image) the t-junction at the top-right of the image becomes more salient, and at the finer scale (2^{nd} image) the vertical bar located at the top-right of the image becomes dominant. By combining the various scales according to (II.9) all these attributes are deemed salient (see the rightmost saliency map in Figure II.8), even though the top right t-junction and the symmetry between the other two appear to dominate.

Therefore, in all experiments reported in this work, the Gabor decomposition was implemented with a dictionary of zero-mean Gabor filters at 3 spatial

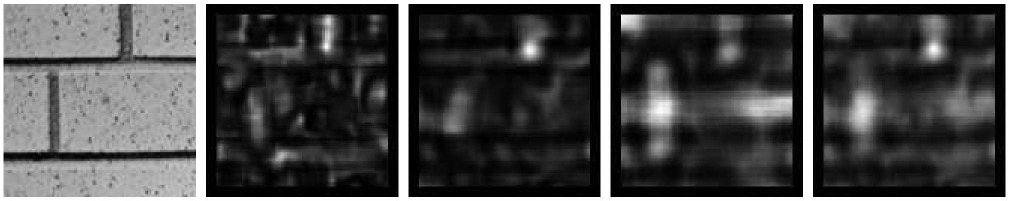


Figure II.8 Saliency maps for a texture (leftmost image) at 3 different scales (center images - fine to coarse scales from left to right), and the combined saliency map (rightmost). Note: the saliency maps are gamma corrected for best viewing on CRT displays.

scales (centered at frequencies of 0.08, 0.16, and 0.32 cycles/pixel) and 4 directions (evenly spread from 0 to π)[2]. Its algorithmic implementation follows the work of [123], and all Gabor channels are also subject to optimal least-squares denoising, implemented with soft-thresholding [31].

II.E.4 Other parameters

Given the feature decomposition of an input image, its saliency map is computed from (II.18), (II.4) and (II.12), with the parameters, α and β, of the GGD distribution estimated through the method of moments (II.11). It is worth to mention that the saliency detection performance does not depend critically on this parameter, e.g. our preliminary experiments showed that arbitrarily setting $\beta = 1$ produced qualitatively similar results.

The discriminant saliency detector has two free parameters: the size of the center and the surround windows. The choice of these two parameters is guided by available evidence from psychophysics and neurophysiology, where it is known that 1) human percepts of saliency depend on the density and size of the items in the display [144, 104], and 2) the strength of neural response is a function of

[2] Following the tradition of the image processing and computational modeling literatures, we measure all filter frequencies in units of "cycles/pixel (*cpp*)". For a given set of viewing conditions, these can be converted to the "cycle/degree of visual angle (*cpd*)" more commonly used in psychophysics. For example, in all psychophysics experiments discussed later, the viewing conditions dictate a conversion rate of 30 pixels/degree of visual angle. In this case, the frequencies of these Gabor filters are equivalent to 2.5, 5, and 10 *cpd*.

the stimulus that falls in the center and surround areas of the receptive field of a neuron [104, 2, 28, 114]. In particular, we mimic the common practice of making the size of the display items comparable to that of the classical receptive field (CRF) of V1 cells (see, e.g., [195, 81]), by setting the size of the center window to a value *comparable* to the size of the display items.

With respect to the surround, it is known that 1) pop-out only occurs when this area covers enough display items [144], and 2) there is a limit on the spatial extent of the underlying neural connections [104, 2, 28, 114]. Considering this biological evidence, the surround window was made 6 times larger than the center, at all image locations. Preliminary experimentation with these parameters has shown that the saliency results are not significantly affected by variations around the parameter values adopted.

Finally, to improve their intelligibility, the saliency maps shown in all figures were subject to smoothing, contrast enhancement (by squaring), and a normalization that maps the saliency value to the interval $[0, 1]$. This implies that absolute saliency values are not comparable across displays, but only within each saliency map.

II.F Acknowledgement

The text of Chapter II, in part, is based on the materials as it appears in: D. Gao and N. Vasconcelos, Discriminant saliency for visual recognition from cluttered scenes. In *Proc. of Neural Information Processing Systems (NIPS)*, 2004. D. Gao and N. Vasconcelos. Decision-theoretic saliency: computational principles, biological plausibility, and implications for neurophysiology and psychophysics. Accepted for publication, *Neural Computation.* It, in part, has also been submitted for publication of the material as it may appear in D. Gao and N. Vasconcelos, Discriminant saliency for visual recognition. Submitted for publication, *IEEE Trans. on Pattern Analysis and Machine Intelligence.* The dissertation author

was a primary researcher and an author of the cited materials.

Chapter III

Biological plausibility of discriminant saliency: Neurophysiology

We have, so far, considered efficient computer implementations of the discriminant saliency detectors. In a broad sense, the biological plausibility for the framework of discriminant saliency comes from the fact that its implementations (in Figure II.5 and Figure II.7) are compatible with most popular models for the early stages of biological vision, which consist of a multi-resolution image decomposition followed by some type of nonlinearity, and feature pooling [14, 119, 167, 106, 88, 219, 188, 59, 110]. For example, the central idea of discriminant saliency that the basic elements of saliency are features is fully consistent with the adoption of a multi-resolution decomposition as a front-end in these low-level vision models. The pooling of feature maps in the saliency measure of (II.16) and (II.18), can also be easily mapped into neural hardware by encoding them as firing rates of the pooled cells. Therefore, the remaining question is whether the saliency measures, i.e. the mutual information of (II.4), is biologically plausible. In the following sections, we will show that it is completely compatible with the widely accepted neural structures of early visual processing.

III.A Network representation of discriminant saliency

III.A.1 Maximum a posteriori (MAP) estimation for mutual information

In Chapter II, we saw that the bulk of the computations of discriminant saliency is based on the mutual information $I(X;Y)$ between a feature X and class label Y. We have also seen that, under the GGD assumption and parameter estimation with the method of moments, $I(X;Y)$ can be computed efficiently with (II.4) and (II.12). In this section, we consider the alternative of maximum a posteriori (MAP) estimation. We note that, for natural images, the shape parameter β is constrained to a range of values (in the vicinity of 1) that guarantees sparse distributions. The fact that, in the GGD, these values only change the exponent of $|x|$, indicates that a precise estimate of β is not critical. We have confirmed this with a number of preliminary experiments, which have shown that assuming $\beta = 1$ (Laplacian distribution) does not produce qualitatively significant differences from those achieved with the estimate of (II.11). Hence, in the following derivation, we assume that the shape parameter β of a GGD is known, and consider the computation of $I(X; Y)$ based on the estimate of the scale parameter α. When the sample size is small, accurate estimates frequently require some form of regularization, which can be implemented with recourse to Bayesian procedures. The parameter α is considered a random variable, and a distribution $P_\alpha(\alpha)$ introduced to account for prior beliefs in its configurations. Conjugate priors are a convenient choice, that produces simple estimators which enforce intuitive regularization. It turns out that, for the GGD, it is easier to work with the inverse scale than the scale itself.

Lemma 1. *Let $\theta = \frac{1}{\alpha^\beta}$ be the inverse scale parameter of the GGD. The conjugate prior for θ is a Gamma distribution*

$$P_\theta(\theta) = \text{Gamma}\left(\theta, 1 + \frac{\eta}{\beta}, \nu\right) = \frac{\nu^{1+\eta/\beta}}{\Gamma(1+\eta/\beta)} \theta^{\eta/\beta} e^{-\nu\theta}, \tag{III.1}$$

whose shape and scale are controlled by hyper-parameters η and ν, respectively. Under this prior, the maximum a posteriori (MAP) probability estimate of α, with respect to a sample $\mathcal{D} = \{x(1), \dots, x(n)\}$ of independent observations drawn from (II.10), is

$$\hat{\alpha}_{MAP} = \left[\frac{1}{\kappa}\left(\sum_{j=1}^{n} |x(j)|^\beta + \nu\right)\right]^{1/\beta}, \tag{III.2}$$

with $\kappa = \frac{n+\eta}{\beta}$.

Proof. The likelihood of the sample $\mathcal{D} = \{x(1), \dots, x(n)\}$ given θ is

$$P_{X|\theta}(\mathcal{D}|\theta) \;=\; \Pi_{j=1}^{n} P_{X|\theta}(x(j)|\theta) = \left(\frac{\beta\theta^{1/\beta}}{2\Gamma(1/\beta)}\right)^n \exp\left(-\theta \sum_{j=1}^{n} |x(j)|^\beta\right).$$

For the Gamma prior, application of Bayes rule leads to the posterior

$$\begin{aligned} P_{\theta|X}(\theta|\mathcal{D}) &= \frac{P_{X|\theta}(\mathcal{D}|\theta)P_\theta(\theta)}{\int_\theta P_{X|\theta}(\mathcal{D}|\theta)P_\theta(\theta)\mathrm{d}\theta} \\ &= \frac{1}{Z}\theta^{(n+\eta)/\beta}\exp\left(-(\sum_{j=1}^{n}|x(j)|^\beta+\nu)\theta\right), \end{aligned}$$

where Z is a normalization constant that does not depend on θ. Since this is a Gamma distribution, (III.1) is a conjugate prior for θ. Setting the derivative of $\log P_{\theta|X}(\theta|\mathcal{D})$ with respect to θ to zero[1], it follows that the MAP estimate is

$$\hat{\theta}_{MAP} = \frac{n+\eta}{\beta}\left(\sum_{j=1}^{n}|x(j)|^\beta+\nu\right)^{-1}.$$

Applying the change of variable from θ to α, leads to the MAP estimate of α,

$$\hat{\alpha}_{MAP} = \left[\frac{1}{\kappa}\left(\sum_{j=1}^{n}|x(j)|^\beta+\nu\right)\right]^{1/\beta}.$$

■

Given this estimate, for each of the classes, estimates of the posterior class probabilities $P_{Y|X}(c|x), c \in \{0,1\}$ can be computed as follows.

Lemma 2. *For a binary classification problem, with generalized Gaussian class-conditional distributions $P_{X|Y}(x|c)$ of parameters (α_c, β_c), $c \in \{0,1\}$, the posterior distribution for class $c = 0$ is*

$$P_{Y|X}(0|x) = s\left[\left(\frac{|x|}{\alpha_1}\right)^{\beta_1} - \left(\frac{|x|}{\alpha_0}\right)^{\beta_0} - K\right], \tag{III.3}$$

where

$$K = \log a + \log \pi + T, \tag{III.4}$$

$$a = \alpha_0/\alpha_1, \tag{III.5}$$

$$\pi = \frac{\pi_1}{\pi_0} \tag{III.6}$$

$T = \log\left(\frac{\beta_1\Gamma(\frac{1}{\beta_0})}{\beta_0\Gamma(\frac{1}{\beta_1})}\right)$, *$\pi_c = P_Y(c), c \in \{0,1\}$, are the prior probabilities for the two classes, and $s(x) = (1+e^{-x})^{-1}$ is a sigmoid.*

[1] It can also be shown that the second order derivative is non-negative, and strictly positive for $\theta > 0$.

Proof. Using Bayes rule and (II.10),

$$\begin{aligned} P_{Y|X}(0|x) &= \frac{P_{X|Y}(x|0)P_Y(0)}{P_{X|Y}(x|0)P_Y(0)+P_{X|Y}(x|1)P_Y(1)} \\ &= \frac{1}{1+\frac{P_{X|Y}(x|1)P_Y(1)}{P_{X|Y}(x|0)P_Y(0)}} \\ &= \frac{1}{1+\frac{\beta_1\pi_1\alpha_0\Gamma(\frac{1}{\beta_0})}{\beta_0\pi_0\alpha_1\Gamma(\frac{1}{\beta_1})}\frac{\exp\left\{-\left(\frac{|x|}{\alpha_1}\right)^{\beta_1}\right\}}{\exp\left\{-\left(\frac{|x|}{\alpha_0}\right)^{\beta_0}\right\}}} \\ &= \frac{1}{1+\exp\left(\left(\frac{|x|}{\alpha_0}\right)^{\beta_0}-\left(\frac{|x|}{\alpha_1}\right)^{\beta_1}+K\right)}, \qquad \text{(III.7)} \end{aligned}$$

where $K = \log a + \log \pi + T$, $a = \alpha_0/\alpha_1$, $\pi = \frac{\pi_1}{\pi_0}$, $T = \log\left(\frac{\beta_1\Gamma(\frac{1}{\beta_0})}{\beta_0\Gamma(\frac{1}{\beta_1})}\right)$. The lemma follows from the definition of the sigmoid, $s(x) = (1+e^{-x})^{-1}$. ∎

The combination of these two lemmas, and some information theoretic manipulation, lead to the desired estimates of the mutual information of $I(X;Y)$.

Theorem 2. *Consider a binary classification problem with generalized Gaussian class-conditional distributions $P_{X|Y}(x|i)$ of parameters (α_i, β_i), $i \in \{0,1\}$, where β_i is known and α_i is estimated, according to (III.2), from two samples, $\mathcal{D}_0$ for class $Y=0$ and $\mathcal{D}_1$ for class $Y=1$. The mutual information $I(X;Y)$ is,*

$$I(X;Y) = \frac{1}{|\mathcal{D}|}\sum_{x\in\mathcal{D}}\phi[g(x)], \qquad \text{(III.8)}$$

with $\mathcal{D} = \mathcal{D}_0 \cup \mathcal{D}_1$,

$$\phi(x) = s(x+\log\pi)\log\frac{s(x+\log\pi)}{\pi_1} + s(-x-\log\pi)\log\frac{s(-x-\log\pi)}{\pi_0}, \qquad \text{(III.9)}$$

$s(x) = (1+e^{-x})^{-1}$ is a sigmoid function, and

$$g(x) \;=\; \log\frac{P_{X|Y}(x|1)}{P_{X|Y}(x|0)} = \psi(x;\Phi_0) - \psi(x;\Phi_1) + \log a + T, \qquad \text{(III.10)}$$

where

$$\begin{aligned} \psi(x;\Phi_c) &= \frac{|x|^{\beta_c}}{\xi_c}, & \text{(III.11)} \\ \xi_c &= \frac{1}{\kappa_c}\left(\nu_c + \sum_{k\in\mathcal{D}_c}|x(k)|^{\beta_c}\right), & \text{(III.12)} \end{aligned}$$

$\Phi_c = (\kappa_c, \nu_c)^T$ is the vector of prior hyperparameters of class c, as defined in Lemma 1, and π, a and T are given in Lemma 2.

Proof. Let

$$\begin{aligned} g(x) &= \log \frac{P_{X|Y}(x|1)}{P_{X|Y}(x|0)} \qquad \text{(III.13)} \\ &= \left(\frac{|x|}{\alpha_0}\right)^{\beta_0} - \left(\frac{|x|}{\alpha_1}\right)^{\beta_1} + \log a + T, \end{aligned}$$

(III.10) follows from substituting α_0 and α_1 with their MAP estimates of (III.2). Combining with (III.3) of Lemma 2 leads to

$$P_{Y|X}(1|x) = s[g(x) + \log \pi],$$

and

$$P_{Y|X}(0|x) = s[-g(x) - \log \pi].$$

From the definition of mutual information,

$$I(X;Y) = E_X\left[\sum_i P_{Y|X}(i|x)\log\frac{P_{Y|X}(i|x)}{P_Y(i)}\right],$$

it follows that

$$I(X;Y) = E_X\left\{\phi[g(x)]\right\}, \qquad \text{(III.14)}$$

with $\phi(x)$ defined in (III.9). Given the set of feature responses $x \in \mathcal{D} = \mathcal{D}_0 \cup \mathcal{D}_1$, (III.14) can be estimated empirically by replacing expectation with sample means, i.e.

$$I(X;Y) = \frac{1}{|\mathcal{D}|}\sum_{x\in\mathcal{D}} \phi[g(x)], \qquad \text{(III.15)}$$

where the summation pools all feature response over the two sample sets. ■

III.B Neurophysiological plausiblity

To analyze the biological plausibility of the computations of mutual information, we note that for the values of a (defined in (III.5)) typical of natural

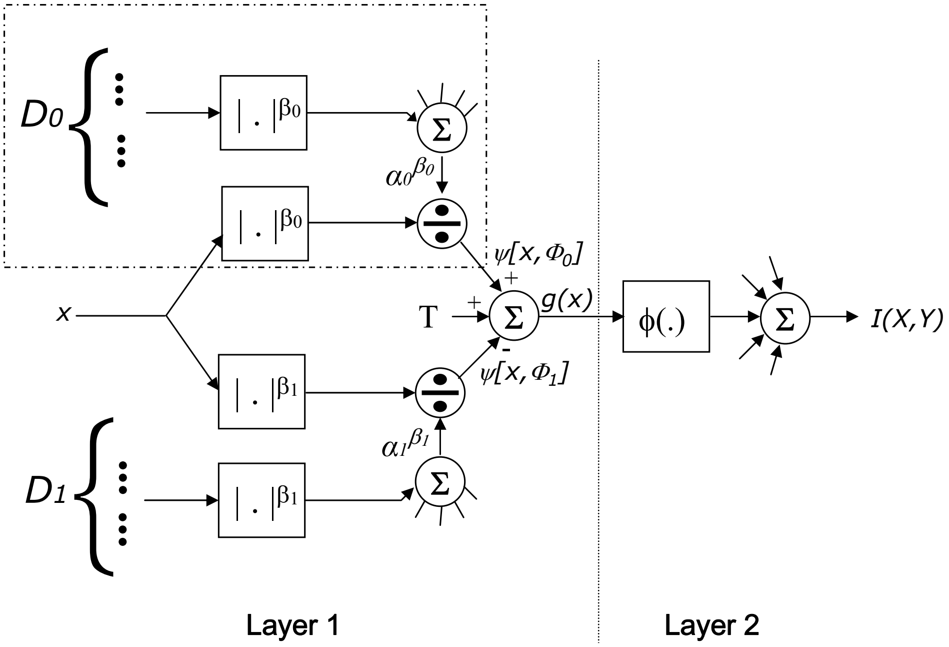


Figure III.1 A network representation of the computation of mutual information, $I(X, Y)$, between feature X and its class label Y.

image patches ($a \approx 1$), the computations of Theorem 2 can be implemented with the network of Figure III.1. In this section, we show that this is consistent with a number of well known properties of the neurophysiology of pre-attentive vision, particularly, the standard neural architecture of the primary visual cortex (V1).

III.B.1 Standard neural architecture of V1

The studies of biological vision have shown that early vision occurs mostly in V1, where cells are usually classified as simple and complex [80, 186, 26]. Classical studies focused on stimuli incident on the cell's receptive field, and simple cells were modeled as cascades of a linear filter and a rectifying non-linearity [132, 92] (as illustrated in Figure III.2 (a)). More recently, extensive physiological recordings have shown that simple cell responses can be strongly non-linear, including effects such as saturation [27], orientation masking [185], and cross-orientation suppression [17]. To explain these observations, an alternative view of simple

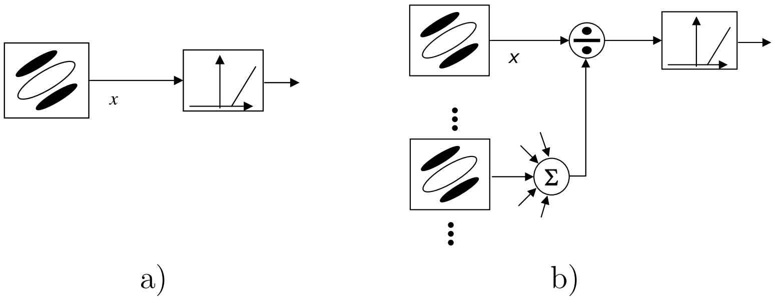


Figure III.2 Classical (a) and divisively normalized (b) models of simple cells in primary visual cortex.

cell response has emerged over the last two decades. Under this view, all of the above non-linearities are explained by the ability of V1 neurons to perform gain control [71]. Besides expanding their dynamic range, gain control enables simple cells to scale orientation tuning with contrast, i.e. to maintain a constant ratio between responses to stimulus of different orientations, independently of stimulus contrast [185, 27]. The implementation of this gain control requires an additional stage of *divisive normalization* of the cell response by that of others [71, 27, 173], as illustrated in Figure III.2 (b). The basic idea is to normalize the classical cell response through the division of its output by the pooled responses of a number of other cells, i.e.

$$x' = \frac{x}{\sigma + \sum_{j \in \mathcal{N}} w_j x(j)} \tag{III.16}$$

where x is the classical response, $\mathcal{N}$ the cell's pooling neighborhood, w_j a weight assigned to $x(j)$ within the neighborhood, and σ a regularization constant that controls the influence of $\mathcal{N}$ on the normalized firing rate x'. Note that (assuming $\sigma \ll \sum_{j \in \mathcal{N}} w_j x(j)$), as stimulus contrast increases, the same happens to the denominator of (III.16), and the cell response is divisively suppressed. Overall, the classical linear stage defines cell selectivity (e.g. orientation tuning), and the divisive stage guarantees that this selectivity holds over a large dynamic range of the cell's input. This enables the cell to significantly extend its input range, without a proportional increase in energy consumption.

In addition to simple cells, there is another type of cell in V1, named complex cell. Complex cells are orientation sensitive but location invariant, i.e. each complex cell responses to edges of a certain orientation within a large receptive field, regardless of the exact location. Complex cells are frequently modeled as units that pool squared and half-rectified outputs of simple cells with similar orientation, the energy model proposed by Adelson and Bergen [1]. We refer to the combination of complex and divisively normalized simple cells as the *standard V1 architecture* [26].

III.B.2 Neurophysiological plausibility of the MI network

It follows from Theorem 2 that the computations of mutual information, represented by the network of Figure III.1, are fully compatible with the standard architecture of V1. The theorem decomposes the computation of mutual information into three basic operations: (III.11) *divisively normalizes* each feature response by the responses of the feature in the sample $\mathcal{D}_c$, (III.10) computes the *differential* between the responses divisively normalized by the two samples, and (III.8) pools this differential response across the total sample $\mathcal{D}$, after application of the non-linearity $\phi(x)$ of (III.9). In general, the shape of $\phi(x)$ changes with the prior probabilities of the class label Y, $\pi_i, i \in \{0, 1\}$. In Bayes decision theory, different prior choices correspond to different cost structures. Although it would be interesting to consider an asymmetric setting when the classification problem is cost-sensitive, in this work, we consider a symmetric cost structure with equal prior probabilities, $\pi_0 = \pi_1 = 1/2$. Under this assumption, $\phi(x)$ of (III.9) can be simplified as

$$\phi(x; \pi_1 = 0.5) = s(x) \log s(x) + s(-x) \log s(-x) + \log(2) \tag{III.17}$$

As shown in Figure III.3, this non-linearity is very close to a hard-limited version of the quadratic function,

$$\tilde{\phi}(x) = 0.07x^2. \tag{III.18}$$

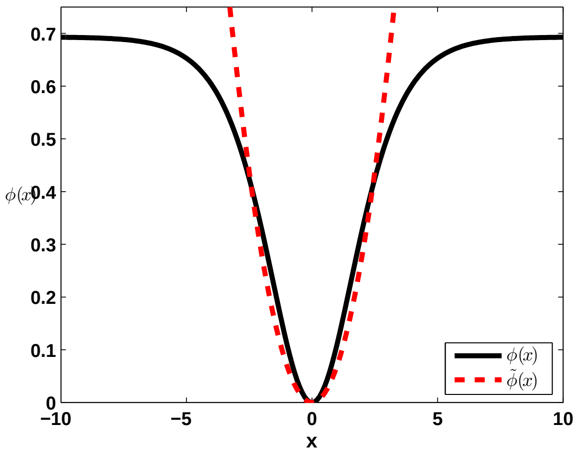


Figure III.3 Complex cell nonlinearity. $\phi(x; \pi_1 = 0.5)$ and its approximation by a quadratic function $\tilde{\phi}(x)$.

This quadratic form conforms to the quadratic non-linearity advocated by the energy model of complex cells [1].

If the step of (III.10) is omitted, these are really just the computations of the standard V1 architecture. This is probably best understood by momentarily disregarding the top branch (dashed box) in the first layer of the MI network, which accounts for the contribution of sample $\mathcal{D}_0$. The remaining network are exactly the standard V1 architecture: a stage of simple cells, divisively normalized by the outputs of their peers, subject to rectification by $\phi(\cdot)$ and pooled, in a manner akin to the classical energy model of complex cells. The implementation of the complete network simply requires the replacement of the divisively normalized simple cell by a cell which is *differentially* divisively normalized by the outputs of the cells belonging to $\mathcal{D}_0$ and $\mathcal{D}_1$.

III.C Statistical inference in V1

In addition to proving the biological plausibility of discriminant saliency, the consistency between the discriminant saliency computations and the basic neural architecture of V1 also offers a holistic functional justification for V1: that V1 has the capability to *optimally* detect salient locations in the visual field, when optimality is defined in a *decision-theoretic sense* and certain (sensible) approximations are allowed, for the sake of *computational parsimony*. Obviously, it is not likely that the whole of V1 would be uniquely devoted to saliency. This raises the question of whether the computational architecture discussed so far could be applied to the solution of generic inference problems. Answering this question, in the most general form, requires the derivation of a functional justification for the building blocks (cells) that compose V1. In what follows, we show that such a justification is indeed possible, but requires a minor extension of the current simple cell model. We show, however, that under this extension *the cells of the standard V1 architecture perform the fundamental operations of statistical inference*, for processes that conform to the statistics of natural images. We then discuss some interesting consequences of this finding.

III.C.1 Extended simple cell model

In the discussion above, the optimality of the standard V1 architecture for the maximization of (III.8) requires $a \approx 1$ in (III.5). While this approximation is acceptable for the saliency problem, it is possible to make the statistical interpretation of the saliency network of Figure III.1 *exact*. In fact, this only requires absorbing the two components of $\log a$ into $\psi(x; \Phi_0)$ and $\psi(x; \Phi_1)$, i.e. redefining these quantities as

$$\tilde{\psi}(x; \Phi_c) \quad = \quad \frac{|x|^{\beta_c}}{\xi_c} + \log \alpha_c. \tag{III.19}$$

Combining (II.10), (III.2), and (III.11) it is straightforward to show that, for generalized Gaussian stimuli, (III.19) is, up to a normalization constant, the estimate of

$$-\log P_{X|Y}(x|c) \tag{III.20}$$

resulting from the MAP estimation of the scale parameter α_c. Physiologically, the implementation of (III.19) requires a slight extension of the current standard simple

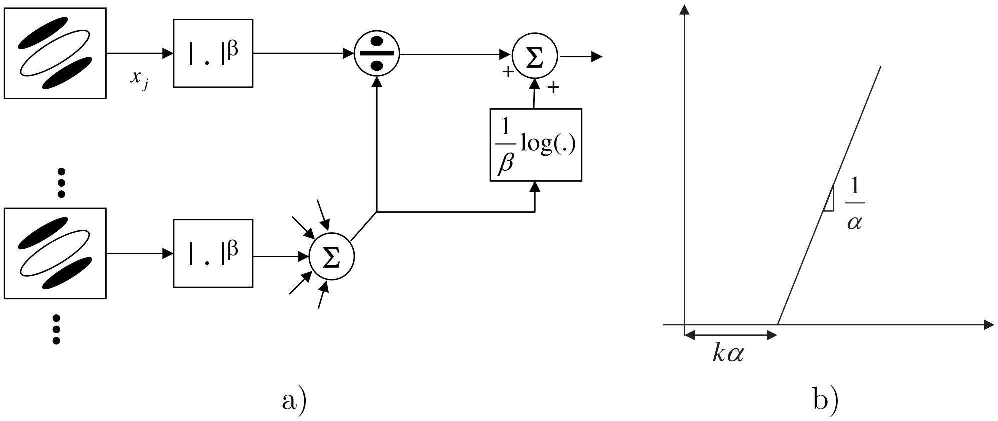


Figure III.4 Extension of the standard simple cell model that makes the probabilistic interpretation of the standard V1 architecture, summarized by Table III.1, exact. a) The log of the contrast α that (divisively) normalizes the cell response is added to it. b) The cell's curve of response has slope proportional to $1/\alpha$ and a shift to the right that is approximately linear in α.

cell model, which is depicted in Figure III.4. This extension consists of adding the log of the normalizing contrast α_c to the output of the cell, complementing the gain modulation of divisive normalization with a rightward shift of the response curve by $\alpha_c\,(\log 1/\alpha_c)^{1/\beta_c}$. For the (small) values of α_c typically found in natural scenes this shift is approximately linear in α_c. This extension is compatible with existing cell recording data [75, 30, 45] and there is even evidence that, when adaptation is considered, a shift occurs and is indeed proportional to the normalizing contrast (constant shifts of log contrast for multiplicative contrast increases) [145].

III.C.2 Fundamental operations of statistical inference

The existence of a one-to-one mapping between (III.19) and (III.20) is significant in the sense of showing that simple cells can be interpreted as probabilistic inference units, tailored to the statistics of natural stimuli. In fact, revisiting (III.8) after this modification, reveals that 1) *all components of the standard V1 architecture have a statistical interpretation*, and 2) this interpretation *covers the* *three fundamental operations of statistical inference: probability inference, decision rules, and feature selection.* The fundamental operation of statistical learning, *parameter estimation*, is also performed within the architecture, through the divisive normalization subjacent to all computations.

The statistical role of the different cell types is summarized in Table III.1, which suggests a clear functional distinction between simple and complex cells. While simple cells assess *probabilities*, differential simple cells implement *decision rules*, and complex cells are *feature detectors*. Physiologically, this is consistent with most aspects of the existing simple/complex cell dichotomy, e.g. the lack of location and polarity sensitivity of complex cells, but suggests a novel refinement of simple cells into two sub-classes: simple cells and differential simple cells. Simple cells conform to the currently accepted model, which is well known to explain most aspects of cell response within the classical receptive field (CRF) [173, 28]. Differential simple cells include additional divisive normalization from a region external to the CRF. They could explain the well documented observation that many cells are modulated by stimuli that fall outside this region [183, 175, 113, 28]. Note, in particular, that the subtraction of $\tilde{\psi}(x; \Phi_1)$ from $\tilde{\psi}(x; \Phi_0)$ can be either excitatory or inhibitory, depending on the stimulus contrasts inside and outside the CRF. The availability of two independent mechanisms to control the responses from the two regions appears necessary to explain the recordings from cells that exhibit this behavior. We intend to investigate this issue in detail, in future research.

Overall, the taxonomy of Table III.1 assigns much more credit to simple cells than simply performing signal processing operations, such as filtering and gain control. In fact, it suggests that the central operation for learning within V1 is the divisive normalization that takes place in these cells, either in the log-likelihood form of (III.19) or the log-likelihood ratio form of (III.10). The coincidence that divisive normalization also solves the signal processing challenge of gain control is an extremely fortunate one, arguably too fortunate for evolution to pass on by. At a more generic level, the taxonomy of Table III.1 also makes a compelling

Table III.1 V1 cells implement the atomic computations of statistical inference under the assumption of GGD statistics. All operations are based on empirical probability estimates derived from the regions used for divisive normalization. The computations are exact for the extended simple cell model of Figure III.4.

cell type	computation	function	description
simple	$\tilde{\psi}(x; \Phi_c)$	$-\log P_{X \mid Y}(x \mid c)$	negative log-likelihood
simple differential	$\tilde{\psi}(x; \Phi_0) - \tilde{\psi}(x; \Phi_1)$	$\log \frac{P_{X \mid Y}(x \mid 1)}{P_{X \mid Y}(x \mid 0)}$	log likelihood ratio
complex	$H(Y) - \langle \phi(g(x)) \rangle_{\mathcal{D}}$	$I(X; Y)$	mutual information

argument for the interpretation of brains as Bayesian inference engines, tuned to the statistics of the natural world. Note, in particular, that the exact shapes of the probability distributions of Table III.1 are determined by the MAP estimates of their parameters. These estimates are, in turn, defined by the two sample sets $\mathcal{D}_0$ and $\mathcal{D}_1$, specified by the lateral connections of divisive normalization. It follows that all probabilities could be computed with respect to distributions defined by arbitrary regions of the visual field, by simply relying on alternative topologies for these connections. Furthermore, since all computations are in the log domain, operations such as Bayes rule or the chain rule of probability can be implemented through simple pooling. Hence, in principle, *the architecture could implement optimal decisions for many other perceptual tasks*.

III.D Acknowledgement

The text of Chapter III, in part, is based on the material as it appears in: D. Gao and N. Vasconcelos. Decision-theoretic saliency: computational principles, biological plausibility, and implications for neurophysiology and psychophysics. Accepted for publication, *Neural Computation*. The dissertation author was a primary researcher and an author of the cited material.

Chapter IV

Prediction of psychophysics of human saliency

While physiological plausibility is important, an ultimate test for saliency models is whether it explains the psychophysics of human saliency. In this chapter, we address this question and demonstrate the ability of discriminant saliency to predict the well known psychophysical properties of human saliency. Due to the fact that there has been wider agreement on the fundamental properties of bottom-up saliency than its top-down counterpart in the literature, in this work, we only consider properties of human bottom-up attention. In particular, discriminant saliency is evaluated in the context of measuring stimulus similarity, which has been believed to play a critical role in guiding human saliency perception.

IV.A Stimulus similarity and saliency perception

We start with a brief review of the existing theories, in psychophysics, for visual saliency and its relation to the perception of stimulus similarity. The psychophysics of saliency and visual attention have been extensively studied in psychology literature. These studies have shown that the human perception of saliency in the visual field is mostly influenced by the interaction between the visual stimuli at a location and those surrounding it. For example, a significant body of psychophysical evidence indicates that the saliency mechanisms rely on measures of local contrast (dissimilarity) of elementary features, like intensity, color, or orientation, into which the visual stimulus is decomposed. Such contrast can produce perceptual phenomena such as texture segmentation [11, 12, 95, 97, 147], target pop-out [190, 196, 138], or even grouping [10, 168].

Motivated by these observations, many theories of visual saliency and attention mechanisms emphasize the importance of measuring stimulus similarities. For example, it is argued in [48] that the efficiency of a visual search task can be largely explained by measuring the similarity relationships both between the target item and the surrounding non-target items, and between different types of non-target items. The theory, however, did not dictate how the similarity could possibly be quantified, which had led to the historic debate on the correctness of the theory [192, 49, 193]. Part of the debate was focused on the question: how can stimulus similarity be measured, and precisely controlled, in the design of visual search experiments? Apparently, the answer to this question is not trivial. It requires a good understanding of each feature space, and is "likely to be reasonably complicated" [222, 219].

Since it is hard to define a good measure of stimulus similarity, a convenient compromise is to simply take *absolute difference* between feature responses to two different stimuli (e.g. [88, 192]). Since this *difference-based measure* is quite intuitive and likely to be biologically plausible, the models based on this measure [88, 86] have become quite popular, and have been applied to saliency detection in both static imagery and motion analysis, as well as to computer vision problems such as robotics, or video compression [215, 182, 84, 153].While it has been shown that the difference-based saliency model [88] can replicate some basic observations from psychophysics, it has significant limitations in four aspects. First, the difference-based saliency measure implies that visual perception relies on a linear measure of similarity. Such a measure does not account for the well known properties of higher level human judgements of similarity, which tend not to be symmetric or even compliant with Euclidean geometry [202, 162, 161]. Second, it does not provide functional explanations for the biological computations in visual processing. Third, the psychophysics of saliency offers strong evidence for the existence of both nonlinearities and asymmetries which are not easily reconciled with this measure. Fourth, even though the center-surround hypothesis intrinsically poses saliency as a classification problem that distinguishes center from surround, there exists little basis on which to justify difference-based measures as optimal in a classification sense. Although it is possible to overcome some of these limitations by adding nonlinear dynamics to the saliency models [89, 87] to mimic the known properties of pre-attentive vision, what is fundamentally missing in the difference-based models is a generic principle behind the neural organization of pre-attentive vision, or more general, a computational principle under the entire cognitive system.

In terms of general computational principles for perception systems, the discriminant saliency measure proposed in this work is very promising: it is not only decision-theoretically optimal and biologically plausible but also, more importantly, provides a functional justification for the neural organization of biological vision. In the following sections, we show that the proposed discriminant saliency consistently reproduces many human saliency behaviors. All the experiments are conducted in the context of visual search, where subjects are asked to detect a target object embedded in a distractor field on a display. It is shown that the center-surround discriminant saliency detector makes not only *qualitative*, but also *quantitative* predictions for the fundamental properties of human saliency in visual search experiments. It is our belief that quantitative predictions are essential to understand the biological plausibility of the discriminant saliency hypothesis. For example, we will see that the proposed discriminant saliency not only predicts, but also provides *analytical* explanations to each of the following properties:

1. while a target that is different from the distractors by a single feature "pops out" to an observer, the same does not happen when the difference is by a conjunction of two features.

2. the saliency perception of a target (among distractors) is nonlinear to the stimulus contrast, i.e. there exist threshold and saturation effects with the increase of the stimulus contrast between the target and the distractors.

3. saliency is affected by the similarity relationships between target and distractors, as well as between distractors. This influence is particularly interesting for heterogeneous distractors.

4. orientation categorization exists in visual search.

5. The saliency perception is asymmetric, and the saliency asymmetries exist

not only for the presence and absence of a feature, but also for the quantitative difference of a shared feature between target and distractor, and the asymmetries comply with Weber's law.

IV.B Single and conjunctive feature search

One classical observation from visual search experiments is that for basic features, such as color and orientation, the search for a target which differs from a set of distractors by a single feature is efficient, i.e. the target "pops-out". In such case, the response time is very short and independent to the number of distractors. However, the same does not occur when the difference is defined by a conjunction of two basic features. In this case, the response time is much longer, and also increases linearly to the number of items in the display[1]. Some examples of this behaviour are shown in the top row of Figure IV.1, where a target differs from a set of distractors in terms of (a) orientation, (b) color, and (c) a conjunction of orientation and color (green right-tilted bar among green left-tilted and red right-tilted bars). The saliency maps produced by discriminant saliency are shown below each display. Note that, like human subjects, the detector produces a very unambiguous judgement of saliency for single feature search ((a) and (b)), but is unable to assign a high saliency to the conjunctive target in (c) (bar in the 4^{th} line and 4^{th} column).

IV.B.1 Discussion

Various theories have been proposed in the literature to explain the difference between single and conjunctive searches [195, 48, 219, 210]. Among these explanations, the feature integration theory (FIT) [195, 197], is probably the most

[1] Note that although there were experimental evidences showing that, in certain cases, searching for conjunction of features could also be done efficiently [135, 191, 197, 221, 29], such efficient conjunctive feature search is unlikely to be driven by a purely bottom-up mechanism. It is likely to be a result of of top-down guidance, such as feature inhibition [197], activation [221, 219, 222], or both [137], which is beyond the scope of the current study.

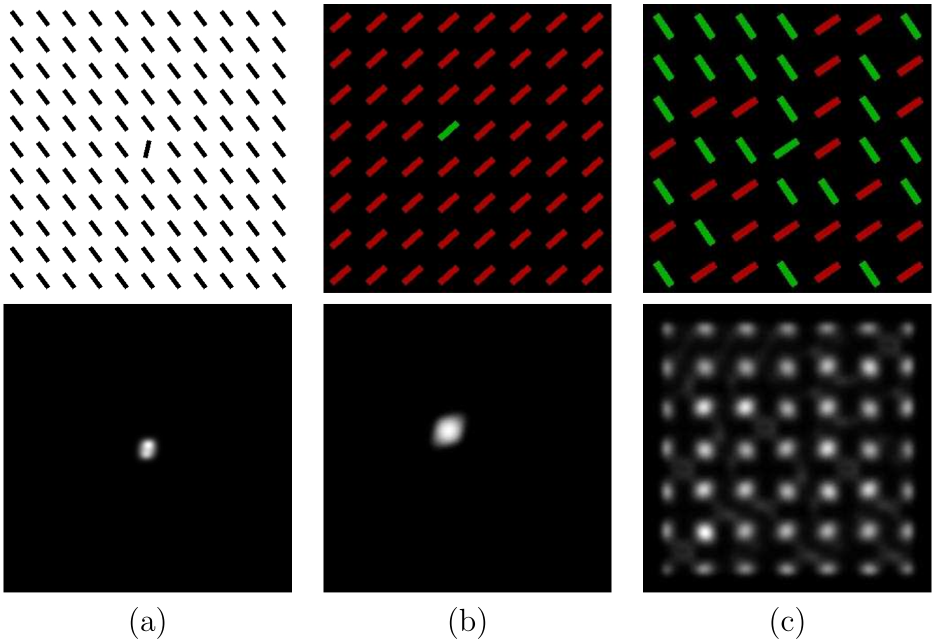


Figure IV.1 Saliency output for single basic features (orientation (a) and color (b)), and conjunctive features (c). Brightest regions are most salient.

influential one. The theory predicts that the visual stimulus is projected into feature maps that encode properties like color or orientation [222, 225]. Feature maps are then combined into a *master*, or *saliency* [106], map that drives attention, allowing top-down (recognition) processing to concentrate on a small region of the visual field. The saliency map is scalar and only registers the degree of relevance of each location to the search, not which features are responsible for it. Hence a target defined by a basic feature is highly salient and "pops-out", but a target defined by the conjunction of features does not.

While the theory explains why search for a conjunctive target is hard, it does not provide a computational explanation of why pre-attentive vision would choose to disregard feature conjunctions. However, discriminant saliency justifies this behavior, by explaining it as optimal, in a decision-theoretic sense, under sensible approximations that exploit the regularities of natural stimuli to achieve computational parsimony. Among these approximations, that of the mutual information by a sum of marginal mutual informations in (II.9) is the most significant one. It suggests that to the degree that (II.7) holds for natural scene statistics, i.e. that feature dependencies are not informative for discrimination of image classes, restricting search to the analysis of individual feature maps has no loss of optimality. The importance of feature dependencies to image classification has been tested in [209, 206], which showed that accounting for dependencies between feature pairs can be beneficial, but there appears to be little gain in considering larger conjunctions. While noticeable, the gains of pair-wise conjunctions over single features are not overwhelming, even for full-blown image classification. In the case of pre-attentive vision, by definition subject to tighter timing constraints, evolution could have simply deemed the gains of processing conjunctions unworthy of the inherent complexity.

IV.C Nonlinearity of saliency perception

Although the above judgements of pop-out are interesting, they are purely qualitative, and therefore anecdotal. Given the simplicity of the displays, it is not hard to conceive of other center-surround operations that could produce similar results. For example, it has been shown that a difference-based saliency detector [88, 89] can easily replicate the above observation on single and conjunctive feature search. To address this problem, we introduce an alternative evaluation strategy, in this section, based on the comparison of *quantitative predictions*, made by the saliency detector and available human data. It is our belief that quantitative predictions are essential for an *objective* comparison of different saliency principles, as well as for an analytical explanation of the saliency mechanisms.

We start the quantitative study with a well known observation that human saliency perception is nonlinear to local feature contrast between target and distractors [15, 152, 48, 134, 192, 219, 61, 211, 143, 148]. Among various visual stimulus modalities in the early visual processing, we consider orientation stimuli in this experiment simply because they are most frequently studied in the psychological literature [80, 82, 92, 132, 42, 11, 12, 95, 97, 147, 190, 196, 138, 10, 168]. We also notice that although it has been shown that the human perception of saliency is nonlinear with respect to local orientation contrast (the orientation differences between a target and the distractors) [139, 61, 224, 219, 110, 22, 121], most of the early studies pursued only the threshold at which these events occur. Examples include the threshold at which a (previously non-salient) target pops-out [61, 139], two formerly indistinguishable textures segregate [110, 96], a "serial" visual search becomes "parallel", or vice versa [195, 224, 130]. In the context of objective evaluation, these studies are less interesting than a posterior set, which also measured the saliency of pop-out targets above the detection threshold [141, 131, 159].

A direct quantitative measure of human saliency perception is, however, not trivial. For this, Nothdurft [141] designed experiments where he compared pop-out from local orientation differences with pop-out from luminance differences. In particular, each display contained both a luminance and an orientation target (shown against background fields of distractors). Subjects were asked to report which of the two targets were perceived faster (more salient) in each display. The experiment was repeated with different luminance and orientation contrasts, and the luminance scaling was carefully calibrated to ensure linear increments at all levels. The luminance of the target which produced an equal preference rating for the two targets was taken as a measure of saliency for orientation difference. Nothdurft showed that the saliency of a target increases with orientation contrast, but in a non-linear manner, exhibiting both threshold and saturation effects: 1) there exists a threshold below which the effect of pop-out vanishes, and 2) above this threshold saliency increases rapidly with orientation contrast, saturating after certain point. The overall relationship has a *sigmoidal* shape, with lower (upper) threshold t_l (t_u).

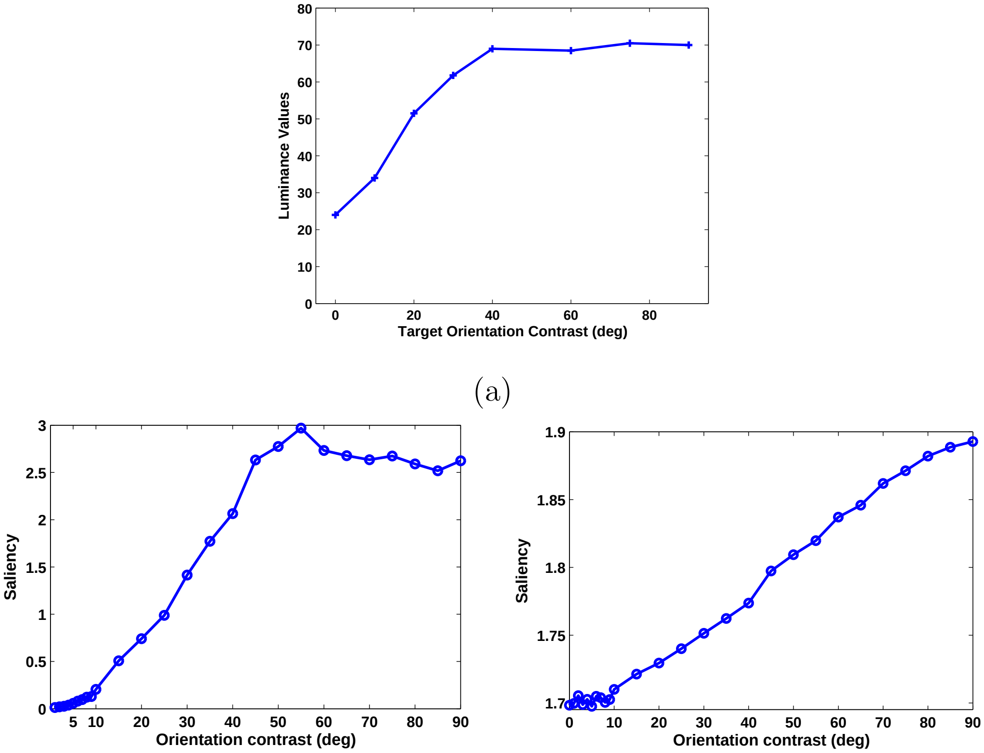
Figure IV.2 (a) presents the results of this experiment, which are reproduced from [141], where the saliency perception of orientation is measured for a set of displays with a homogeneous distractor field. We repeated this experiment by applying the discriminant saliency detector to the similar set of displays with only orientation targets. In particular, each display contains a distractor field of identical bars with a random orientation, and a target which is defined by orientation contrast (one example display is illustrated in Figure IV.1 (a)). The discriminant saliency is measured at the target, and averaged across all displays with the same orientation contrast. The result is presented in Figure IV.2 (b), where the average discriminant saliency of the target is plotted as a function of the orientation contrast. Interestingly, like the human saliency curve shown in Figure IV.2 (a), the discriminant saliency curve increases slowly when the orientation contrast is below a lower threshold $t_l \approx 10°$, rising rapidly afterwards, and then is saturated after the upper threshold $t_u \approx 40°$. This strong nonlinear behavior matches human saliency perception, and suggests that, up to certain normalization factor[2], the discriminant saliency provides a good quantitative prediction of human visual saliency. The same experiment was repeated for a popular difference-based saliency model [89][3] which, as illustrated by Figure IV.2 (c), exhibited no quantitative compliance with human performance.

IV.C.1 Discussion

There have been various explanations for the threshold and saturation effect, but most of them are highly hypothetical. For example, Nothdurft [141] speculated that it is due to some mechanism nonlinearly related to target contrast, particularly, which reflects the nonlinear properties of orientation tuning profiles of cortical cells. The authors in [131], on the other hand, explained the saturation effect as the consequence of the fact that the orientation contrast leads to the perception surface boundaries (in texture segmentation), whose strength, once perceived, is almost independent of the changes in the magnitude of orientation contract. To the best of our knowledge, there has been no previous attempt for an

[2] Note that an exactly numerical comparison of the two plots is not meaningful since saliency was measured under two different units.

[3] Results obtained with the MATLAB implementation by [215].



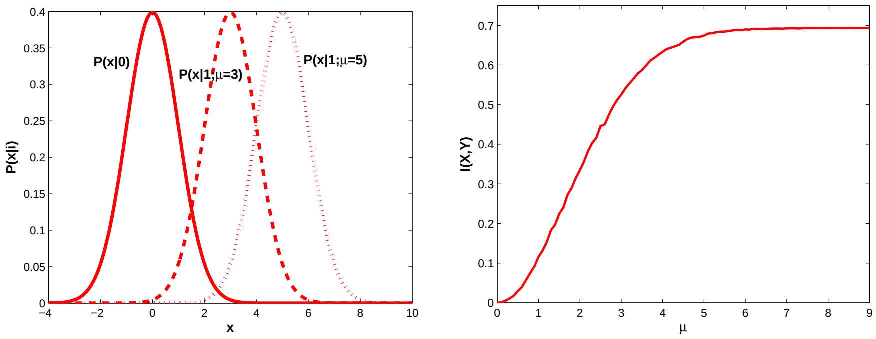
(a)

(b)

(c)

Figure IV.2 The nonlinearity of human saliency responses to orientation contrast (reproduced from Figure 9 of Nothdurft (1993)) (a) is replicated by discriminant saliency (b), but not by the model of Itti & Koch (2000) (c).

analytical explanation of the nonlinear behavior of saliency. Discriminant saliency, however, offers such an explanation: the nonlinearity originates naturally from the adoption of mutual information as a measure of stimulus contrast. This is intuitive from the fact that given any pair of class-conditional feature distributions for a binary classification problem, the mutual information between the feature and the class label is alway bounded between 0 and $\log 2$. Figure IV.3 illustrates a simple example for the mutual information measured for the case of 1-D Gaussian conditional densities. Suppose the two class-conditional probability density functions both follow Gaussian distribution with unit variance, i.e. $P_{X|Y}(x|0) = \mathcal{N}(x, 0, 1)$ and $P_{X|Y}(x|1) = \mathcal{N}(x, \mu, 1)$. The mean of the class $Y = 0$ is fixed at 0, and that of the class $Y = 1$ is a free parameter μ (as illustrated in Figure IV.3(a)). The



(a) (b)

Figure IV.3 Illustration of the nonlinear nature of mutual information. (a) Two class-conditional probability densities, each is a Gaussian with unit variance. The Gaussian of class $Y = 0$, $P_{X|Y}(x|0)$, has a fixed mean at 0, while that of class $Y = 1$, $P_{X|Y}(x|1)$, takes various mean values, determined by μ. (b) The mutual information between feature and class label, $I(X;Y)$, for (a) is plotted as a function of μ.

mutual information between random variables X and Y is plotted, as a function of parameter μ, in Figure IV.3(b). We can see that mutual information exhibits a strongly nonlinear behavior, which resembles the shape of the human saliency perception curve.

We can also analyze this property more rigorously, by studying the computations of the discriminant saliency. In fact, one can show that the nonlinearity is a result of combining mutual information with the generalized Gaussian marginal distributions. Recall in Chapter III, we have shown, through Theorem 2, that the computations of mutual information can be implemented by the saliency network of Figure III.1. We redraw this network in Figure IV.4, and present, in each box in the figure, the outputs at the intermediate stages of the network, for the above experiment on orientation contrast. The computation, at each stage, corresponds respectively to (up to some constant) the computations of negative log-likelihood, $-log(P_{X|Y}(x|1)) \sim \psi(x) = \left(\frac{|x|}{\alpha_1}\right)^{\beta_1}$, absolute log-likelihood ratio between the center and the surround classes, $\left|log\left(\frac{P_{X|Y}(x|1)}{P_{X|Y}(x|0)}\right)\right| = |g(x)|$, conditional mutual informa-

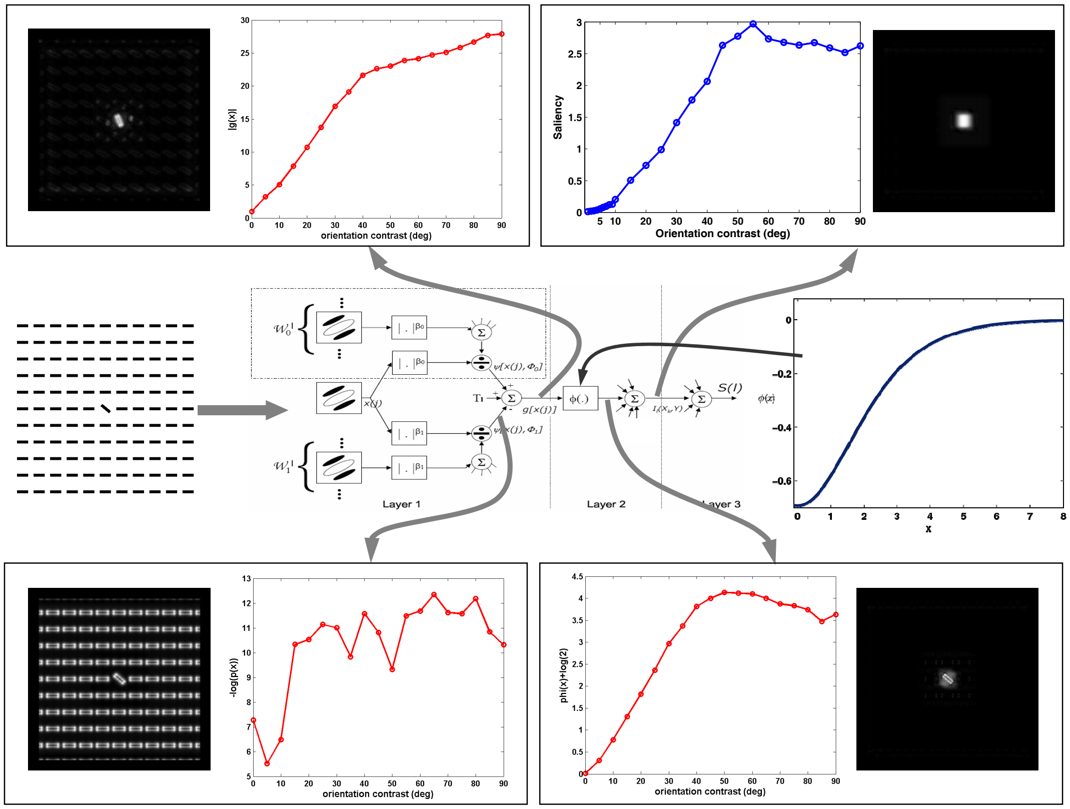


Figure IV.4 Illustration of the output at each stage of the discriminant saliency network for the orientation contrast experiment.

tion, $I(Y; X = x) = \phi[g(x)]$, and the discriminant saliency, $S(X) = I(X; Y)$. We present, within each box, the average output of the corresponding stage at the target as a function of orientation contrast, as well as the entire output for one example display shown to the left of the network.

At least two interesting observations can be drawn by comparing these outputs. First, the nonlinear behavior exists to a certain extent throughout the network, however, it is most strongly exhibited after $\phi(x)$ whose functional shape is shown (only for $x > 0$) to the right of the network. Second, among all these outputs, the saliency output (in the upper-right box) is the one that resembles the human saliency perception the most. This observation not only supports the plausibility of the discriminant saliency hypothesis, but also rules out the possibility of some other principles as driving principles for saliency. For example, the output of $\psi(x)$ represents a previous proposal that defines saliency as the negative log-likelihood of feature responses (also referred to as self-information) (e.g., [163, 24]). Intuitively the proposal is quite plausible, but, as we can see from the figure that $\psi(x)$ responded strongly to both the target and the distractors in the example display, and did not make the target stand out as it should. This is because log-likelihood considers only individual feature response, but not the discrimination between target and distractors, which in turn does not suppress distractors in the display. The plot of $\psi(x)$ also appeared to be quite noisy and unstable, which does not replicate human saliency perception. Another possible principle, the absolute log-likelihood-ratio ($|g(x)|$), considers the discrimination between target and distractors, so it is more robust at eliminating distractors in the background and responds only to the target. However, its response curve does not show a strong nonlinearity. In this aspect, the transformation of $\phi(x)$ significantly increases the nonlinearity of the responses, especially the saturation effect. The final pooling stage smoothed the previous output, and produced the saliency measure which resembles human saliency perception. These comparisons indicate that although each component of the saliency network contributes to the saliency detection, none of them alone is a biologically plausible solution for saliency.

IV.D Distractor heterogeneity and search surface

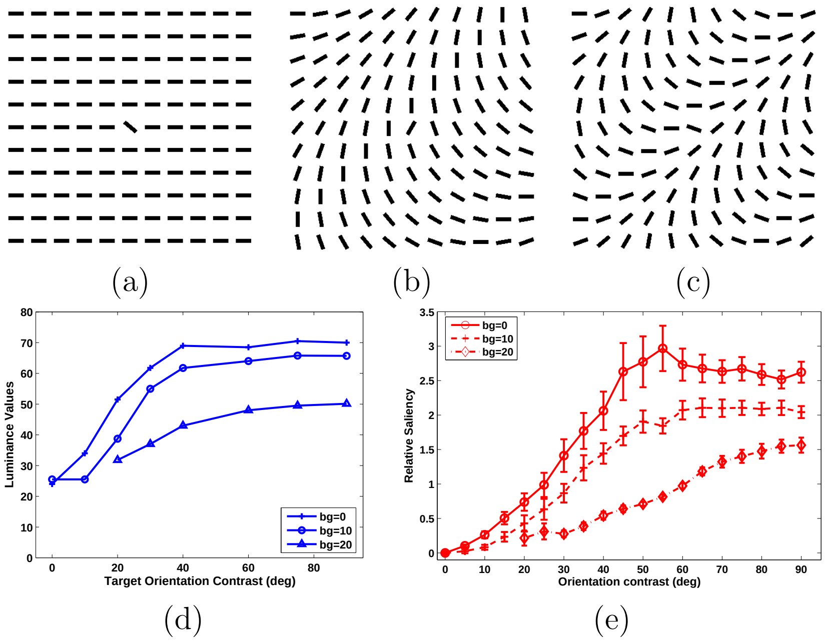
Besides similarity relationships between target and distractor, human saliency perception is also affected by similarity between distractors, i.e. the homogeneity of the distractor field. For example, it is shown in [191] that search for a blue target bar among a set of distractors with randomly mixed colors (red, green and white), or for a horizontal target bar among a set of vertical, left diagonal and right diagonal bars, is significantly slower than the controlled case, where the distractors contain only one type of stimulus, i.e. they are homogeneous. It is also reported in a letter search experiment [48] that, when the target is an upright

"L" and distractors are "L"s rotated 90° clockwise or counterclockwise from the target position, the slope of the response time (RT), i.e., the average search time for each item in a display, is much steeper than that with all distractors rotated in the same direction. Similar observations have also been seen by various research groups (e.g., [130, 192, 139, 140, 141, 224, 222, 131, 210, 164]).

Using the same protocol as in the previous orientation contrast experiment, Nothdurft [141] quantitatively measured the influence of heterogeneous distractors to human saliency percepts. In particular, he showed subjects the displays with the target defined by the orientation contrast, as described before, but with respect to a heterogeneous distractor field. The homogeneity of the distractors, i.e. the orientation directions of the background bars, was varied by adding a constant angle value (bg) when going from element to element along rows or columns in the raster. The "target-distractor orientation contrast" was defined as the difference in orientation between the actual target and a virtual background element at the target's position. Three examples of such displays are shown in Figure IV.5 (a)-(c), for $bg = 0°, 10°, 20°$ with a target-distractor orientation contrast of 40°. The human saliency perception curves resulting from this experiment are presented in Figure IV.5 (d), and the discriminant saliency predictions on the same set of displays are plotted in Figure IV.5 (e). It is clear that both the human saliency and the discriminant saliency drop continuously and exhibit weaker threshold and saturation effects, when the the distractor field becomes heterogeneous (plots marked with $bg = 10$ and $bg = 20$). For $bg = 20°$, both curves show only slight threshold and saturation effects. Comparing with the two plots, we can see that the discriminant saliency provides quantitatively similar behavior as human.

IV.D.1 Heterogeneity in an irrelevant dimension

Although it is in general true that saliency is significantly reduced when the distractor field becomes heterogeneous, various visual search experiments (e.g., [191, 48]) have shown that search is not affected if the heterogeneity of the distrac-


(a) (b) (c)

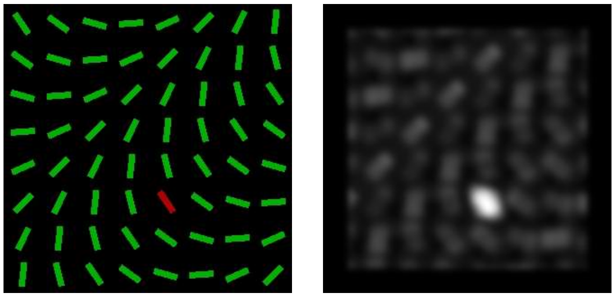
(d) (e)

Figure IV.5 Example displays of different orientation variations of distractor bars ((a) $bg = 0°$, (b) $bg = 10°$, and (c) $bg = 20°$), and the corresponding saliency judgements from (d) human subjects (Northdurft, 1993a), and (e) discriminant saliency, plotted as a function of orientation contrast.

tors exists only in an irrelevant dimension (the feature dimension that does not differentiate the target and the distractors). For instance, in the display shown in Figure IV.6 (a), since the target is different from the distractors in color (the relevant dimention), the variation of the distractor field in orientation (irrelevant dimension) does not affect human performance of the search for the target. The experiment on the discriminant saliency also produces the same observation. This is illustrated in Figure IV.6 (b), where the target pops out in the discriminant saliency map.

IV.D.2 Discussion

To explain the influence of the distractor heterogeneity on the efficiency of visual search, Treisman et al. [191] resorted to the Feature Integration Theory [195]. They argued that for search with heterogeneous distractors, a distractor

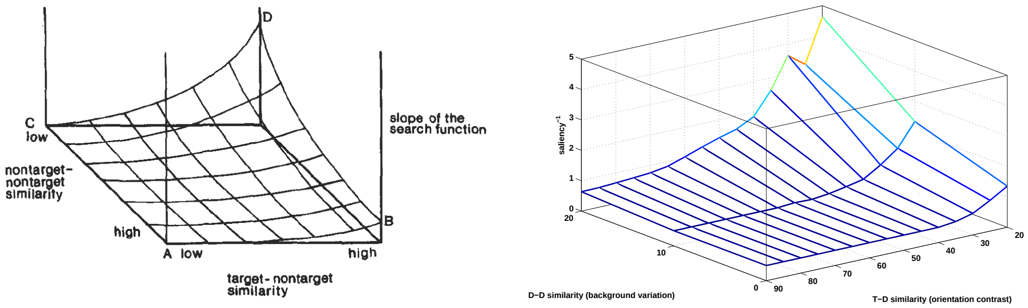


(a) (b)

Figure IV.6 A display with background heterogeneity in an irrelevant dimension (a) does not affect the discriminant saliency measure at the target (b).

contrasts not only with the target, but also with other distractors within the relevant dimension, it is therefore necessary to locate the specific map for the target. On the other hand, the more different maps are activated due to the heterogeneity of the distractors, the more similar to the target the nearer distractor value is likely to be, which makes the localization of the specific map for the target even harder. The two factors together produce a slow search for target with heterogeneous distractors.

In another influential attentional engagement theory (AET) for visual search, Duncan and Humphreys [48] from a more interesting point of view, explained the effects of heterogeneous distractors in a unifying framework based on two types of stimulus similarities, target-nontarget (T-N) similarity and nontarget-nontarget (N-N) similarity. They proposed that the two types of similarities affect not only the complexity of the target template in a top-down processing, but also the local perceptual grouping of the items in the bottom-up processing. Both a highly complex target template and less grouped nontargets increase the search time significantly. Hence, by manipulating T-N and N-N similarity, it is possible to make a search task arbitrarily easy or arbitrarily difficult. In particular, they hypothesized that the influence of T-N similarity and N-N similarity to the slope of RT in a visual search task can be described as a continuous search surface in a three-dimensional parameter space, which is illustrated in Figure IV.7 (a). The



(a)

(b)

Figure IV.7 The search surface for stimulus similarities hypothesized by Duncan & Humpreys (1989) (a) is reproduced by discriminant saliency (b).

search surface has the following four basic properties: 1) when the T-N similarity is low, the saliency prediction is high, and the search is always highly efficient, which is irrespective of N-N similarity (curve AC in the figure); 2) when N-N similarity is maximal (i.e., the distractors are identical, or homogeneous), T-N similarity has a relatively small effect (curve AB); 3) when when N-N similarity is reduced (i.e., the distractor field becomes heterogeneous), T-N similarity becomes more important (curve CD); and 4) when T-N similarity is high, N-N similarity has a very substantial effect (curve BD). Overall the worst performance happens when T-N similarity is high and N-N similarity is low (point D).

Although describing the efficiency of a search task by the similarity relationships between stimuli is in general unquestionable, quantifying these similarities is not trivial at all. Unfortunately, the AET theory [48] did not propose a solution. However, without an objective measure of stimulus similarity, precisely controlling the similarity between the target and the distractors, in the design of visual search experiments, becomes hard and often controversial (see, [192, 49, 193]). As pointed out by Wolfe [221, 224], “the lack of a proper similarity measure also raises practical difficulties for models of visual search and attention”.

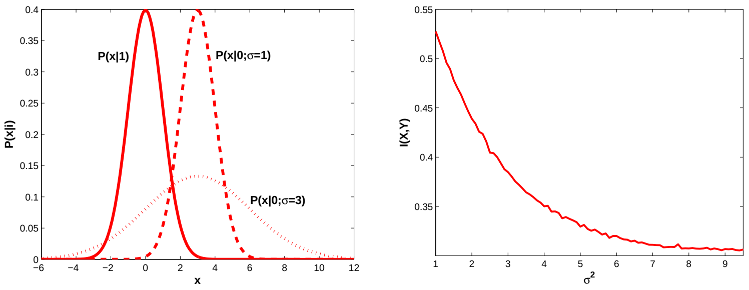
This controversy, nevertheless, can be resolved by introducing mutual information as a measure of stimulus similarity. As we have shown in the last experiments that the discriminant saliency quantitatively predicted human saliency perceptions of orientation contrast for both homogeneous and heterogeneous distractors. In fact, under a simple assumption between saliency and search time, the discriminant saliency prediction on the orientation contrasts (Figure IV.5 (e)) can be shown to consistently replicate the search surface depicted in [48]. Considering the close relationship between saliency judgement and search time [219, 88, 89, 157], we assume that the slope of RT is *qualitatively* inversely proportional to saliency magnitude[4], and draw the curves of RT slope in the space spanned by the T-N similarity (orientation contrast) and N-N similarity (variation of the distractor homogeneity) as in [48]. Noting the fact that RT slope saturated at certain saliency level after targets "pop-out" [61], we also upper bound the saliency by a proper threshold, before converting it to RT slope. The surface presented by the discriminant saliency on orientation stimulus is illustrated in Figure IV.7 (b). The surface suggests that when the orientation contrast between the target and the distractors is large, i.e. low T-N similarity, the RT slope is small and hardly affected by the background variation. The latter, however, affects the quality of search significantly when the orientation contrast is small, i.e. high T-N similarity. It is clear that orientation contrast plays a more significant role when the background variation is large (e.g. $bg = 20$) than when the nontarget is homogeneous ($bg = 0$). All these observations match the proposal in [48] and the search surface illustrated in Figure IV.7 (a), suggesting that the mutual information measure adopted in the discriminant saliency provides a competent similarity measure for pre-attentive visual features.

The influence of distractor heterogeneity on mutual information can be in fact intuitively explained. To show this, we consider the case where the target does not change when the distractors become heterogeneous, and assume that at each image location, the center window covers only one item (either a target or a distractor). At the location of the target, since the center window contains only the target, the distribution of the feature responses within this region remains unchanged

[4] Note that this approximation is only for illustration purpose. No claim is made about the quantitative relationship between the saliency judgement and RT slope, which is beyond the scope of this work.

when the distractor becomes heterogeneous. However, the heterogeneous distractors contained in the surround window generate less consistent feature responses, which in turn increases the variance of feature distribution in the surround. The two distributions, therefore, have larger overlap compared with the homogeneous case. From the decision-theoretic standpoint, this decreases the discrimination between the two classes, and leads to a smaller mutual information, i.e. a less salient target. The whole process can, again, be illustrated by a simple example, where the mutual information, $I(X;Y)$, is computed for a binary classification problem with Gaussian conditional distributions. As illustrated in Figure IV.8 (a), changing the distractor heterogeneity is equivalent to changing the variance σ^2 of the distribution $P(x|0)$, while keeping its mean and the distribution $P(x|1)$ fixed. The plot of Figure IV.8 (b) shows that $I(X;Y)$ decreases significantly as σ^2 increases.

On the other hand, a similar analysis can be applied to infer the saliency of distractors. When both the center and the surround windows cover only distractors, the distributions of feature responses in the two windows, which used to be identical in the homogeneous case, become different. This difference increases the saliency (or mutual information) at each distractor. The distractor saliency increase is interesting for visual search experiments, especially when the search of a target is guided by bottom-up saliency cues. In such a task, if a target is the only item whose saliency value is significantly greater than those of the distractors in the display, the subject's attention will be immediately directed to the target location, which leads to a fast search. If, however, the saliency value of the distractors increases so that it is comparable to that of the target, the subject's attention is likely to be directed first to distractors before reaching the target, resulting in a slow search. This suggests that, in visual search, the relative saliency value between the target and the distractors is more important than their absolute values. In fact, in some cases, the heterogeneous distractors may increase the saliency of the target, but it also increases the saliency value of distractors, which altogether reduces the search efficiency. The next experiment illustrates such an example.



(a) (b)

Figure IV.8 Illustration of the effect of distractor heterogeneity on the mutual information. (a) Two class-conditional probability densities, each is a Gaussian with mean values at $x = 0$ and $x = 3$, respectively. The Gaussian of class $Y = 1$, $P_{X|Y}(x|1)$, has a unit variance, while that of class $Y = 0$, $P_{X|Y}(x|0)$, takes various variance values, determined by σ. (b) The mutual information between feature and class label, $I(X;Y)$, for (a) is plotted as a function of σ^2.

The displays used in this experiment are illustrated in Figure IV.9 (a)-(c), with the target in the center of each display. The displays represent three different target-distractor orientation configurations:

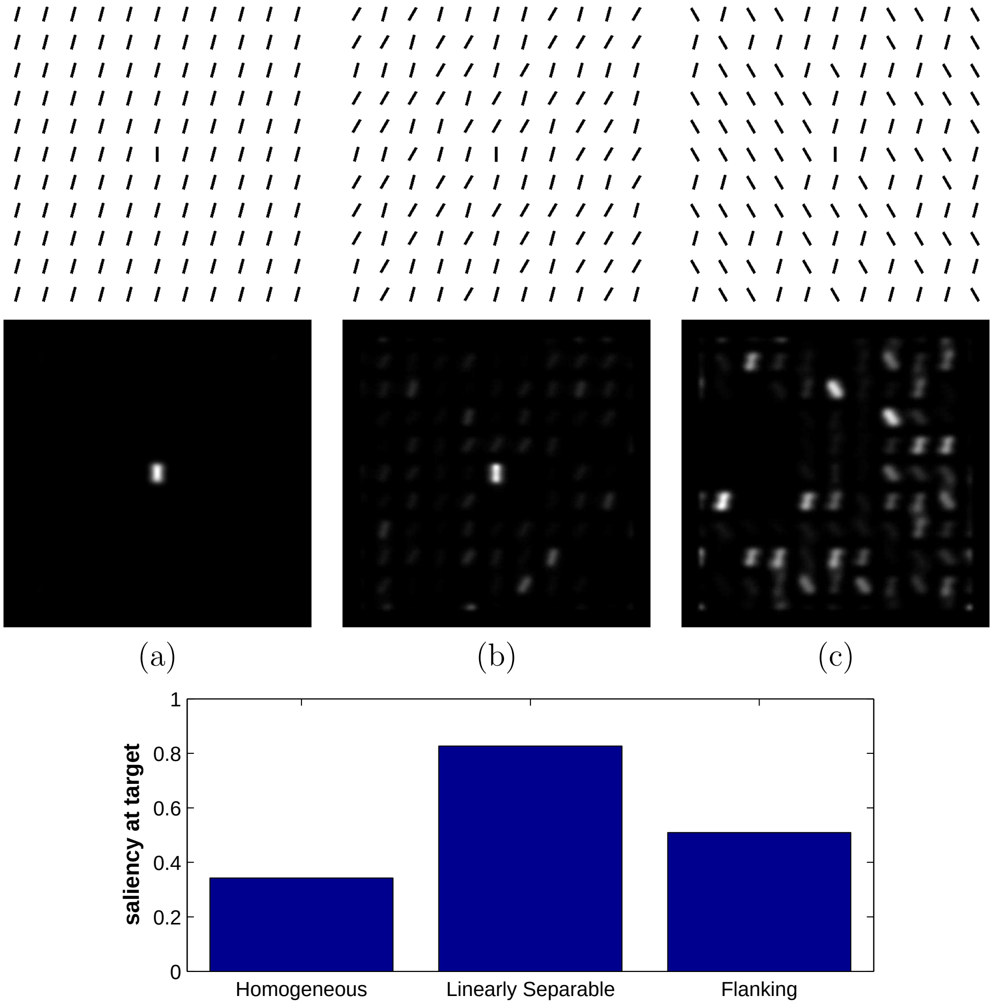
Homogeneous : Target: 0°; distractors: 15°. Distractors are homogeneous.

Tilted right Target: 0°; distractors: 15°, 30°. Distractors are heterogeneous, but all distractors are tilted to the right of the target orientation, i.e. in orientation dimension, target orientation is *linearly separable* from those of the distractors.

Flanking Target: 0°; distractors: 15°, −30°. Distractors are heterogeneous, and the target orientation is flanked by those of the distractors: half of the distractors are tilted to the left of the target orientation, and the other half to the right.

Note that for the two heterogeneous configurations, half of the distractors have 30° difference in orientation from the target, which is larger than the 15° orientation contrast in the *homogeneous* case. As suggested by Figure IV.2, increasing orientation contrast should increase the target saliency. This is confirmed by the plot of Figure IV.9 (d), which shows that the discriminant saliency of the target for the homogeneous case is significantly weaker than those for the heterogeneous cases. However, the heterogeneity of distractors also increases the saliency of the distractors, and therefore reduces the efficiency of the search. This can be seen from the saliency maps shown under each display of Figure IV.9. For the homogeneous case (display (a)), the target stands out against a clear background, while for the heterogeneous cases (displays (b) and (c)), the targets are embedded in more noisy distractor fields, which thus are less evident than the former. This example shows that although heterogeneous distractors may sometimes increase the saliency of the target, they always increase the difficulty of visual search, which is consistent with human experimental data [164].

Another interesting property of saliency can also be observed from this experiment by comparing the saliency maps for display (b) and (c) of Figure IV.9. Although both displays have heterogeneous distractors, the target in display (b) shows a much stronger saliency peak than those of the distractors, representing an easier search task. The target in display (c), however, has much weaker saliency value than the distractors, indicating a difficult search task. Such an observation has been widely reported in human experiments, and is frequently explained as the *linear separability* of the target and the distractors in the relevant feature dimension [196, 50, 51, 9, 224, 220, 164]. From the discriminant saliency point of view, however, we can explain this property by measuring the heterogeneity of the distractors. For the two displays, although the orientation differences between the target and the distractors within each display are both 15° and 30°, the orientation difference between the two types of distractors in each display are very different: 15° for *tilted right*, but 45° for *flanking*. It is almost obvious that the distractors in the *flanking* display produce higher saliency values than those in the *tilted right* display, suggesting a much harder search task.



(a) (b) (c)

(d)

Figure IV.9 Orientation flanking and linear separability.

IV.E Orientation categorization and coarse feature coding

Although orientation is undoubtedly one of the few basic features that are coded in the early visual processing, and neurons are tuned to all orientation angles [92, 93, 94], it seems that not all orientations are equally coded in preattentive vision. For example, it was shown in [61] that given a fixed orientation difference between target and homogeneous distractors, different configurations of the orientations of the target and the distractos lead to different discriminability. In [192], it was discovered that while the search of targets that are defined by conjunctions of “standard” features (such as vertical and horizontal in orientations, or red and blue in color) is very efficient, search of “non-standard” conjunction target gave a much steeper RT and more illusory conjunctions. Similar behavior was also observed in [224] that although, in general, the efficiency of searching for an orientation target declines when the orientation of the distractors becomes heterogeneous, the search can be significantly facilitated if the orientations of the target and the distractors fall into some special patterns, for example, if the orientations of the distractors could be grouped into categories which are different from that of the target orientation. In particular, the authors in [224] suggested that there are at least four orientation categories, namely “steep”, “shallow”, “tilted-left”, and “tilted-right”, are coded in pre-attentive vision.

Figure IV.10 presents the three displays used in [224] to justify the “steep” as an orientation category. In each display, the orientation differences between a target (the central bar) and the set of heterogeneous distractors (with two different orientations) are constants, namely 40° and 60°. The displays are different, therefore, only in terms of the orientation configurations listed below,

Steep Target: $-10°$; distractors: $-50°$, $50°$. Target is the only “steep” item.

Steepest Target: $10°$; distractors: $-30°$, $70°$. Target is “steepest” but not the only steep item

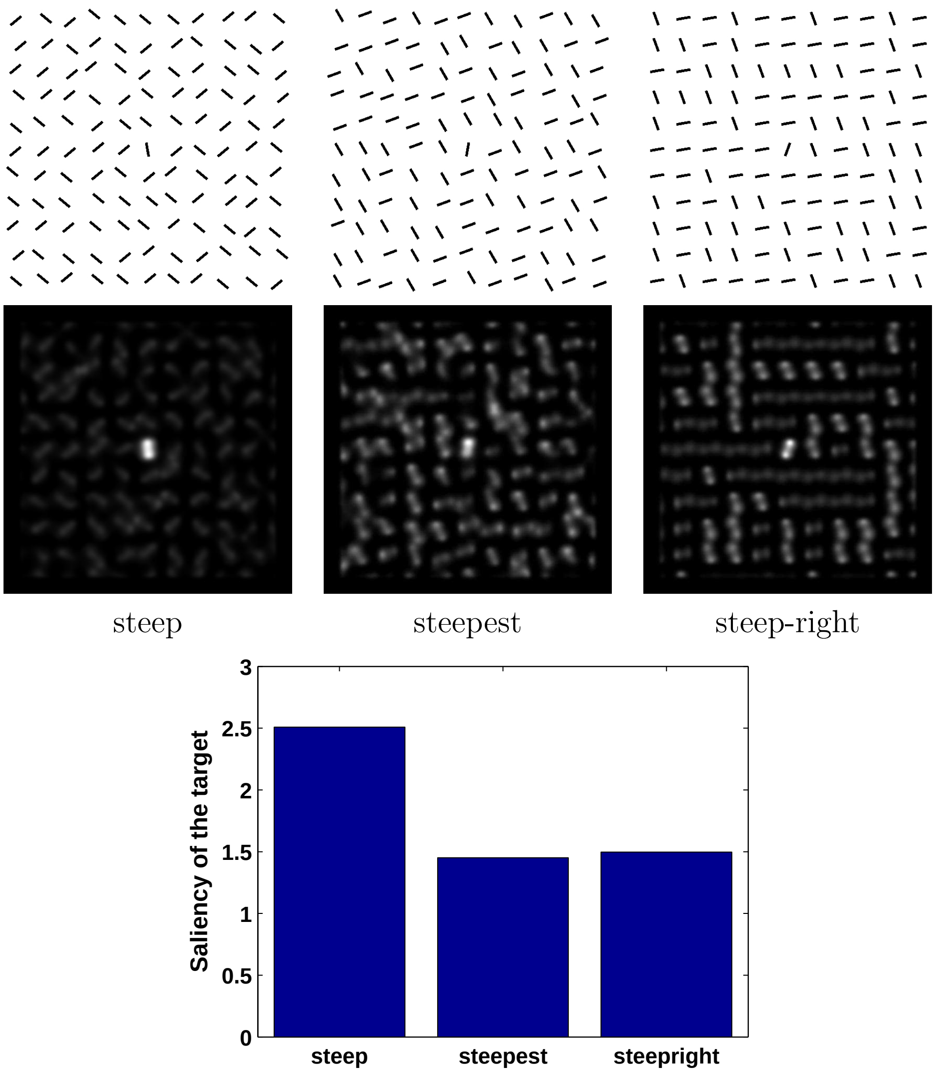
Steep-right Target: $20°$; distractors: $-20°$, $80°$. Target is defined conjunctively by “steep” and “tilted to the right”.

It was found in [224] that while most of the subjects had shallow target trial slopes (less than 3.0 ms/item) for the “steep” condition, few of them could perform so efficiently for the “steep-right” and “steepest” conditions. In other words, when the target is the only steep item, the search is significantly more efficient than it is in other geometrically equivalent conditions. To examine how discriminant saliency predicts this property, we applied the saliency detector to these displays. The resulting saliency maps are presented, under each display, in Figure IV.10.

In the figure, we also present a bar plot of the saliency magnitude at each target for the three displays. Consistent with human behavior, the saliency map for the "steep" display shows a single dominant saliency peak at the target, while the other two maps show saliency peaks at both the targets and the distractors, where the saliency values of the targets are much less dominant than it is in the "steep" case. This indicates that the search for the "steep" target is efficient, but that for the others are not. The fact that the "steep" target has a significantly higher saliency value than the other targets also conforms to human data.

IV.E.1 Discussion

One popular explanation for these observations is the *coarse coding hypothesis* which states that only a few broadly tuned "standard" feature detectors are available in the pre-attentive level. This hypothesis was first illustrated by Treisman in her original work of feature integration theory as a drawing of a few orientation feature maps [195], and was later formalized as a hypothesis [194]. The hypothesis suggests that coarse coding be "a general property of vision in conditions that preclude focused attention, such as search tasks under time pressure and discrimination judgments with brief exposures" [192]. In [61, 62], the authors argued that only two broadly-tuned orientation channels, one vertical and one horizontal, are required to explain some properties of simple orientation tasks, although they later discovered that more orientations seem to be necessary for some other pre-attentive orientation processing [63]. On the other hand, in [224], the channels tuning to the orientation categories were assumed to be coded additionally to the continuous orientationally tuned channels in early vision. Although this assumption makes it easier to explain the efficient search of a unique orientation category, it raises practical difficulties for developing saliency models that can be simulated on real images [219]. In the implementation of discriminant saliency, we have followed Treisman's proposal and decompose the features into four broadly tuned color channels (red, green, blue and yellow), and four Gabor channels with



steep steepest steep-right

discriminant saliency at the target

Figure IV.10 Orientation categories.

different preferred orientations (vertical, horizontal, left and right diagonal orientations). Details of this implementation were introduced in Section II.E. The fact that this discriminant saliency implementation reproduces the orientation categorization experiment indicates that the assumption of the additional orientation category channels in [224] is not necessary. What is more critical, in our opinion, is the choice of a proper measure of stimulus similarity that explains basic properties of human pre-attentive vision which, in this case, turns out to be the combination of decision-theoretic formulation of feature similarity with the proposal of coarse coding.

One remaining issue is the relationship between the specific orientations adopted in the current implementation and those available to the pre-attentive visual system. The fact that the discriminant saliency detector performs well in the above experiment of orientation categorization, however, does not necessarily mean that the four orientations adopted in the detector coincide with the ones deployed in human pre-attentive processing. Nonetheless, we believe that given the connections between discriminant saliency network and the neural structures in V1, it would be interesting to use the discriminant saliency detector as a tool to study these underlying feature channels. This will require more evidence on the ability of the detector to predict psychophysical and physiological observations of human pre-attentive vision, and is worth future investigation.

IV.F Visual search asymmetries

One other classic hallmark of human saliency perception is its asymmetries in visual search tasks: while a target with some stimulus A "pops-out" in a distractor field of another stimulus B, the saliency of the target vanishes when the two stimuli are exchanged for the target and the distractors. This phenomenon was first thoroughly documented by Treisman and her colleagues through a series of visual search experiments [198, 196]. They found that while, in general, the presence in the target of a feature absent from the distractors produces pop-out, the reverse (pop-out due to the absence, in the target, of a distractor feature) does not hold. For example, for the pair of examples illustrated in the first row of Figure IV.11, they showed that the search for the target "Q" on the left display, which differs from the distractor "O"s by the presence of an additional feature (a vertical bar), produces only a flat RT slope, but the search for the target "O" among "Q"s, on the right display, is difficult and gives a steep RT slope. Other examples of search asymmetries, such as single bar versus pair of bars and vertical bar versus tilted bar, are also illustrated in Figure IV.11. The study of search asymmetries has been made into an important diagnostic tool in studying the pre-attentive features in visual attention [198, 196, 223], such as orientation [196, 61, 224], color [198], motion [166, 165], curvature [107], 3-D depth [52, 155, 107], and others (see, [222]).

It worth to mention that there are possibly other sources of search asymmetries, besides the presence versus absence of a basic feature. For example, it was shown in [196] that it is easier to find deviations among canonical (or standard) stimuli than vice versa, which could explain the observation of the search asymmetry between a tilted item among vertical items and a vertical item among tilted items. The authors of [58] also showed that subjects were faster to reject familiar, normal letters than to reject unfamiliar, mirror-reversed letters. Hence, they were faster to find the unfamiliar item among familiar items than vice versa. Similarly, Nothdurft [142] presented evidence that it is easier to find the unfamiliar inverted face among up-right faces than vice versa. Such search asymmetries have been generalized to the argument that "novelty" should be regarded as a basic feature (see, e.g., [91, 69, 216, 180, 120]). However, it is likely that these search asymmetries involve higher level stages of visual processing, such as top-down learning, and are beyond the scope of the current study, where we only consider comparisons to search asymmetries caused by the presence and absence of the basic features.

As in previous experiments, we applied discriminant saliency to the set of classic displays used in [196, 198], and present the resulting saliency maps under each display in Figure IV.11. Interestingly, discriminant saliency exhibits strong asymmetric behaviors. As can be seen from the saliency maps, there is always a unique conspicuous saliency peak at the target location on the left displays, indicating a "pop-out" effect. No such effect, however, is observed on the saliency maps for the right displays.

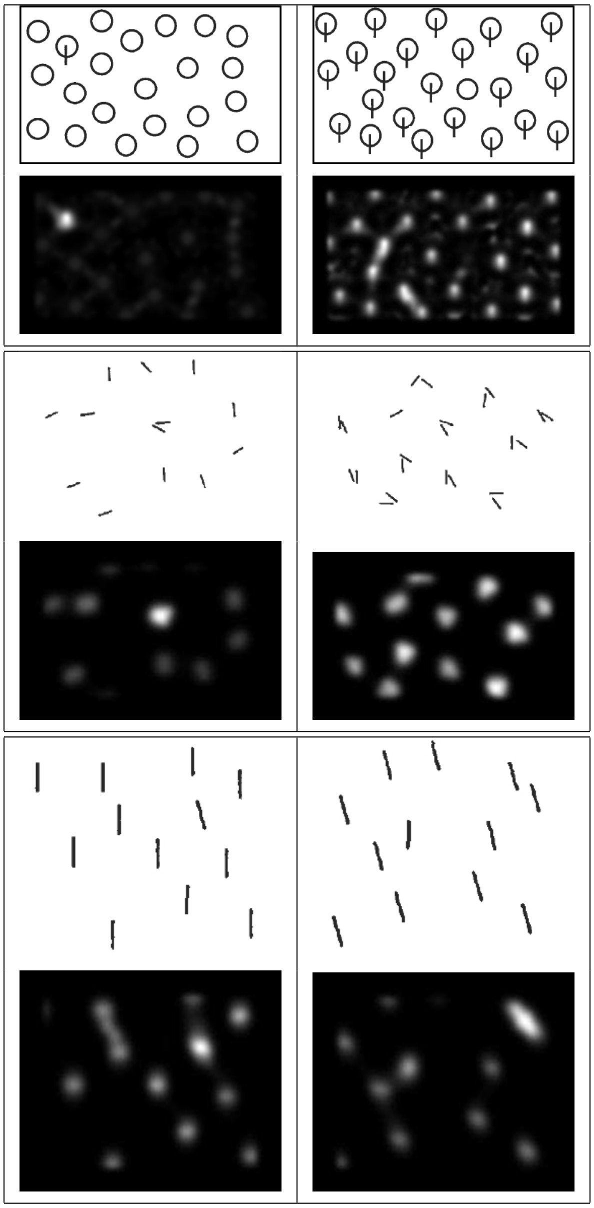


Figure IV.11 Examples of pop-out asymmetries for discriminant saliency. Left: a target that differs from distractors by presence of a feature is very salient. Right: a target that differs from distractors by absence of the same feature is much less salient.

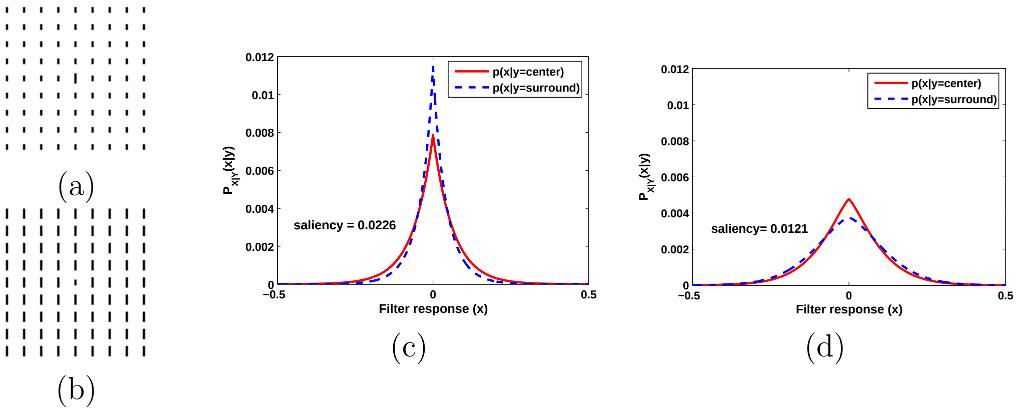
IV.F.1 Discussion

Consistent with the Feature Integration Theory, Treisman et al. [198] argued that all these asymmetries can be explained by the presence and absence of a basic feature in the pre-attentive processing. In particular, when a target is defined by the presence of an additional feature (absent in the distractors) that is positively coded in pre-attentive vision, it generates unique activity on that feature map, and hence can be detected without focused attention. On the other hand, when the target is defined by the absence of a feature, the target feature must be localized, therefore focused and serial scanning is required. While there are other possible accounts of the search asymmetries(e.g., [61, 48, 210]), the explanation of presence and absence of a feature has obtained a wide agreement [223].

Among all of the examples shown in Figure IV.11, the pair of examples in the bottom row is particularly interesting. This search asymmetry was first observed in [196], which showed that while a tilted bar is easy to find among a set of vertical bars, a vertical bar among tilted bars is not. Such behavior of human perception of orientation differences was observed by many others through visual search experiments (e.g., [61, 224]). It is explained in [196] that the tilted orientation represents a deviating value from a standard or reference value represented by the vertical orientation. The deviating stimulus produces substantial activity in the standard channels, but is distinguished from the standards by the additional activity it generates in detectors for a positively coded dimension of deviation from the standards. Therefore, the asymmetry is due to the presence and absence of the deviating stimuli. On the other hand, it is argued in [224] that this is due to the fact that, preattentively, orientations were categorized as “steep”, “shallow”, “tilted-left” or “tilted-right”. While both the vertical and the tilted item share the category label “steep”, the vertical target is defined by its absence of the “tilted” category.

Although varied in the assumptions of specific feature channels, all these explanations seem to support the proposal of “coarse coding”. Implemented with the “coarse coding” assumption, the discriminant saliency also gives a similar explanation: the search asymmetry between tilted bar among vertical bars and vertical bar among tilted bars comes from the fact that the tilted bar produces activity on the *horizontal* orientation filter, while the vertical bar does not. In other words, it is the presence and absence of a horizontal feature which accounts for the asymmetries. The fact that the asymmetries of discriminant saliency are consistent with the asymmetries of visual search is quite interesting because the discriminant saliency measures the similarity of the stimuli between the center and the surround windows. This consistency not only indicates that the search of primitive visual features is significantly affected by the similarities of the stimuli at the target and the distractors, but also provides important evidences for the connections between the asymmetry of similarity judgment [162, 202] and asymmetry of visual search. The authors in [196] also discussed the possible connections between the two types of asymmetries. However, due to the lack of a meaningful similarity measure, their discussion was only hypothetical. In this work, the adoption of mutual information as a measure of stimulus similarity bridges the two seemingly disjoint properties of human perception.

To investigate, in more detail, the asymmetries of discriminant saliency, we notice that it originates from the asymmetric changes of the distributions of feature responses in the center and the surround window. The adoption of mutual information for saliency makes it possible to capture these asymmetric changes. This can be demonstrated by an experiment using the following two displays, each of which contains two types of line stimuli, long and short vertical line segments. The long line segment is assigned to the target and the short ones to the distractors in the display shown in Figure IV.12 (a), and vice versa for the display of Figure IV.12 (b). The size of the center window of the discriminant saliency detector was chosen such that, when placed at the target location, the center window covers both the target and some of the distractors. To make the demonstration more intuitive, only one vertical Gabor filter was used to compute saliency. In Figure IV.12 (c), we plot the two conditional distributions of the filter responses, which were estimated from the center and the surround windows at the target location for the display of Figure IV.12 (a). Similarly, the two conditional



(a)

(b)

(c)

(d)

Figure IV.12 Asymmetry of saliency measure for a target of a longer line segment (a) and a shorter line segment (b) from background of line segments of the same length. Plots (c) & (d) illustrate the estimated distributions of the responses of a vertical Gabor filter at the target and the background for display (a) and (b) respectively.

distributions at the target location for the display of Figure IV.12 (b) are plotted in Figure IV.12 (d). Comparing the two plots, we can see that exchanging the stimuli of the target and the distractor did not simply lead to a swap of the two distributions of feature responses. Instead, it caused significant shape changes of the distributions, indicating two distinct classification problems at the target locations of the two displays. Intuitively, the two conditional distributions of (c) are more different than those of (d), which indicates an easier classification problem, or a higher saliency value at the target for display (a) than (b). This asymmetry is very well quantified by the discriminant saliency (0.0226 for (a) and 0.0121 for (b)), another prediction consistent with human saliency behavior [196].

Compliance with Weber's law

To explain the asymmetries between feature presence and absence, as well as those between more and less of the quantitative change of a feature, Treisman et al. proposed a pooled response and group-scanning hypothesis [190, 198, 196]. The hypothesis assumes that subjects check a pooled response to the relevant feature over a group of items, thus they are able to find the target if the pooled response over a group containing the target becomes sufficiently larger than that over a group containing only distractors. They also suggested that Weber's law determines the discriminability of groups of a given size when they do and do not contain a target. This law states that the size of the *just noticeable difference* is a constant proportion of the background activation level. According to Weber's law, with certain level of discriminability, subjects can compare groups of large numbers of items when distractors produce a low level activity, but they can only compare groups of smaller numbers of items, when the distractors produce a high level of activity, in order to keep the same discriminability level. Scanning groups of fewer items over the entire display requires more time than scanning groups of larger numbers of items, which leads to the search asymmetries between more and less of a feature as well as between its presence and absence. To show the evidence that search asymmetries obey Weber's law, Treisman et al. designed a set of experiments (Experiment 1a in [196]) in which the subjects were presented with displays, such as the one shown in Figure IV.13 (a), where the target (a vertical bar) differed from the distractors (a set of identical vertical bars) only in terms of its length. They showed that while interchanging the target and the distractors led to asymmetry, the search time is approximately the same for a target either longer or shorter than the distractors but of the same amount, when the length of the distractor is fixed, i.e. obeying Weber's law.

What is interesting is that the computations of discriminant saliency also comply with Weber's law. As we have seen in Chapter III, one important computation of discriminant saliency is *divisive normalization*, $\frac{|x(s)|^\beta}{\frac{1}{|\mathcal{W}_s|}\sum_{j\in\mathcal{W}_s}|x(j)|^\beta}$, which normalizes the response of a filter location s by the responses, of the same feature, at neighboring locations. Rewriting this term as $\frac{|x(s)|^\beta-\frac{1}{|\mathcal{W}_s|}\sum_{j\in\mathcal{W}_s}|x(j)|^\beta}{\frac{1}{|\mathcal{W}_s|}\sum_{j\in\mathcal{W}_s}|x(j)|^\beta}$, we can see it has exactly the form of Weber's law. We repeated the Experiment 1a of [196] described above with discriminant saliency, and confirmed the compliance with Weber's law. In Figure IV.13 (b), we present a scatter plot of the discriminant

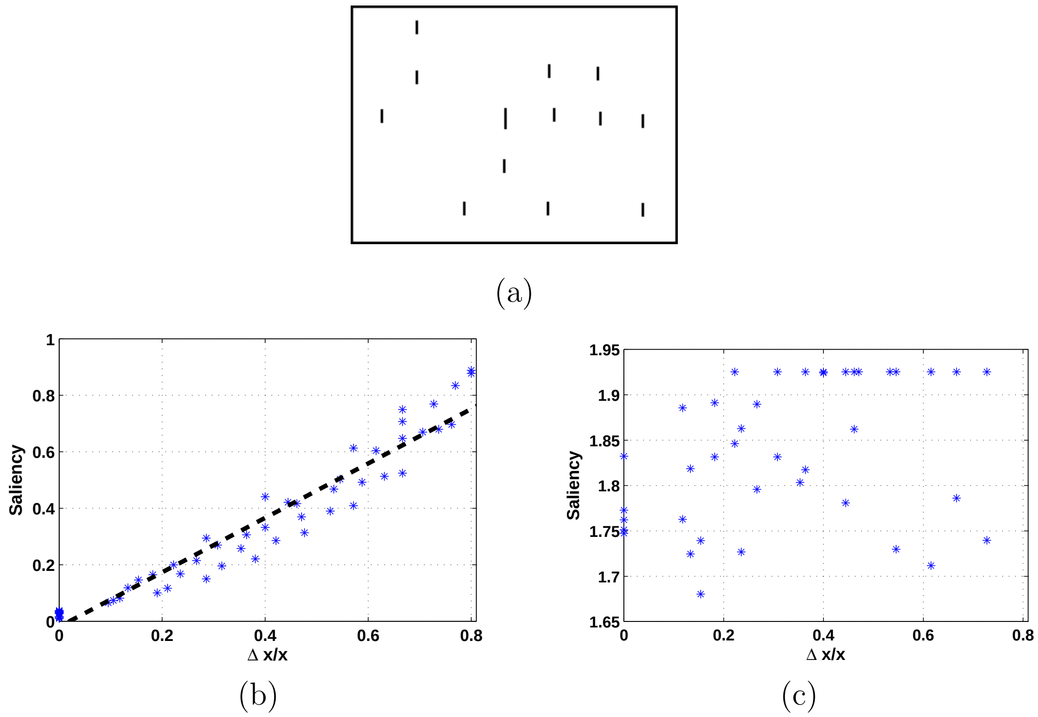


Figure IV.13 An example display (a) and performance of saliency detectors (discriminant saliency (b) and the model of Itti & Koch (2000) (c)) on Treisman's Weber's law experiment (Experiment 1a in [196]).

saliency measurements across the set of displays, as a function of the ratio between the difference of target/distractor length and distractor length. Each point in the plot corresponds to the target saliency in one display, and the dashed line shows that, like human perception, discriminant saliency follows Weber's law: target saliency is approximately proportional to the difference of target/distractor length, but subject to the normalization of the distractor length. For comparison, Figure IV.13 (c) presents the corresponding scatter plot for the model of [89], which does not replicate human performance.

Following what we have been doing in the previous experiments, we analyze in the following how discriminant saliency changes as a function of the changes of the target and distractor lengths. To simplify the computations without qualitatively changing the discriminant saliency measure, we used the following assumptions and approximations. First, in the GGD representation of feature responses, we assume $\beta = 1$ for all GGD's. Second, in the computation of discriminant saliency, we used the fact that the nonlinear operation $\phi(x)$ of (III.9) can be (qualitatively) well approximated by a linear soft threshold operation, $\phi\prime(x)$,

$$\phi\prime(x) = s_{0.35}(0.14 * x + 0.35) + s_{0.35}(-0.14 * x + 0.35). \tag{IV.1}$$

This approximation is illustrated in Figure IV.14. Last, for simplicity, only one Gabor feature (of vertical orientation) is assumed in the following derivation.

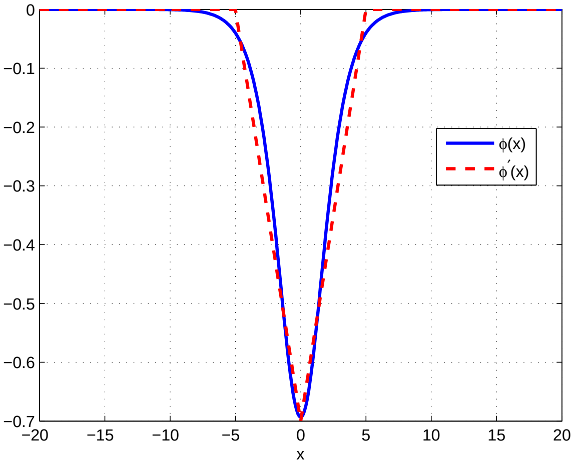


Figure IV.14 The nonlinear operation $\phi(x)$ can be well approximated by a linear soft threshold operation $\phi\prime(x)$.

As shown in Figure IV.14, when the change of the line length is small, $g(x) \propto \frac{|x|}{\alpha_0} - \frac{|x|}{\alpha 1}$ of (III.10) falls mostly into the linear part of $\phi\prime(x)$, hence the computation of saliency can be further approximated as

$$S(x) \approx \hat{S}(x) =< |g(x)| >_{\mathcal{W}} = \left| < |x| >_{\mathcal{W}} (\frac{1}{\alpha_0} - \frac{1}{\alpha_1}) \right| \tag{IV.2}$$

where $\mathcal{W} = \mathcal{W}_0 \cap \mathcal{W}_1$, $< \cdot >_{\mathcal{W}}$ means averaging over the neighborhood $\mathcal{W}$, and α_0 and α_1 are estimated over the center and the surround by ML estimates,

$$\alpha_0 =< |x| >_{\mathcal{W}_0}, \alpha_1 =< |x| >_{\mathcal{W}_1} . \tag{IV.3}$$

Noting that $< |x| >_{\mathcal{W}}$ can be written as a linear combination of α_0 and α_1,

$$< |x| >_{\mathcal{W}} = \tau \cdot \alpha_0 + (1 - \tau) \cdot \alpha_1, \tag{IV.4}$$

with $0 < \tau = \frac{size(surroundwindow)}{size(centerwindow)+size(surroundwindow)} < 1$, we rewrite $\hat{S}(x)$ as

$$\begin{aligned}\hat{S}(x) &= |[\tau\alpha_0 + (1-\tau)\alpha_1](1/\alpha_0 - 1/\alpha_1)| \\ &= |2\tau - 1 - \tau\alpha_0/\alpha_1 + (1-\tau)\alpha_1/\alpha_0| \,. \qquad \text{(IV.5)}\end{aligned}$$

Assume that, in the above experiment, both the target and the distractor have initial length L, which changes to $L + \Delta L_0$ for distractor and $L + \Delta L_1$ for target, in each display. We also assume that, at the target location, the center window covers not only the target, but also n neighboring distractors. Given the fact that the Gabor feature is a linear filter, the following approximations can be used,

$$\begin{aligned}\alpha_0 &\approx K \cdot (L + \Delta L_0), \qquad \text{(IV.6)} \\ \alpha_1 &\approx K \cdot \frac{L + \Delta L_1 + n(L + \Delta L_0)}{n+1}, \qquad \text{(IV.7)}\end{aligned}$$

where K is a constant. The saliency approximation $\hat{S}(x)$ at the target, as a function of the lengths of the distractor and the target, can be then written as

$$\begin{aligned}&\hat{S}(L + \Delta L_0, L + \Delta L_1) \\ &= \left| 2\tau - 1 - \tau \frac{K \cdot (L + \Delta L_0)(n+1)}{K \cdot [(n+1)L + \Delta L_1 + n\Delta L_0]} \right. \\ &\quad \left. + (1-\tau) \frac{K \cdot [(n+1)L + \Delta L_1 + n\Delta L_0]}{K \cdot (L + \Delta L_0)(n+1)} \right| \\ &= \left| 2\tau - 1 - \tau \frac{(1 + \frac{\Delta L_0}{L})(n+1)}{n + 1 + \frac{\Delta L_1}{L} + \frac{n\Delta L_0}{L}} + (1-\tau) \frac{n + 1 + \frac{\Delta L_1}{L} + \frac{n\Delta L_0}{L}}{(1 + \frac{\Delta L_0}{L})(n+1)} \right|\end{aligned}$$

let $y_0 = \frac{\Delta L_0}{L}$ and $y_1 = \frac{\Delta L_1}{L}$, representing the relative length changes of the target and the distractor, then

$$\begin{aligned}\hat{S}(y_0, y_1) &= |2\tau - 1 - \tau \frac{(n+1)(1+y_0)}{n+1+y_1+ny_0} + (1-\tau)\frac{n+1+y_1+ny_0}{(n+1)(1+y_0)}| \\ &= \left| \frac{y_1 - y_0}{(n+1)(y_0+1)} - \frac{\tau(y_1-y_0)^2}{(n+1)(y_0+1)(ny_0+y_1+n+1)} \right| . \qquad \text{(IV.8)}\end{aligned}$$

The saliency representation in (IV.8) has some interesting properties. First, when the distractor length is fixed, i.e. $y_0 = 0$, it becomes

$$\hat{S}(y_1) = \left| \frac{y_1}{n+1} - \frac{\tau y_1^2}{(n+1)(y_1+n+1)} \right| , \qquad \text{(IV.9)}$$

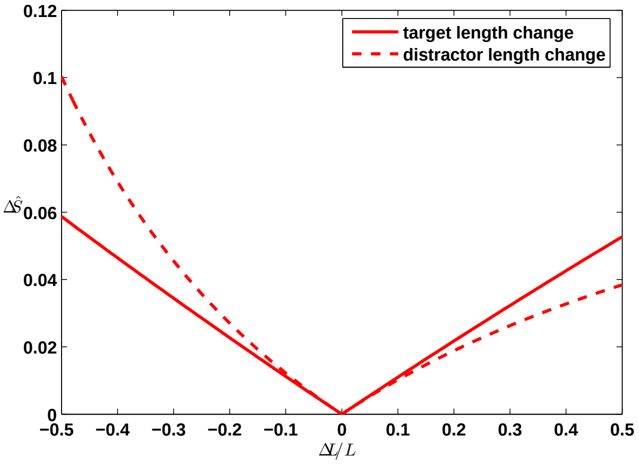


Figure IV.15 The target saliency $\hat{S}(y_0)$ and $\hat{S}(y_1)$.

which is plotted in Figure IV.15 (the solid line) for the parameters used in the simulation ($\tau = 35/36$ and $n = 8$). We can see that $S(y_1)$ is linear to the relative change of the target length, and is also symmetric with respect to $y_1 = 0$. This is exactly compliant with the Weber's law hypothesis in [196]. Second, if on the other hand, we fix the target length, i.e. let $y_1 = 0$, and only change the distractor length, (IV.8) becomes

$$\hat{S}(y_0) = \left| \frac{y_0}{(n+1)(y_0+1)} + \frac{\tau y_0^2}{(n+1)(y_0+1)(ny_0+n+1)} \right|. \qquad \text{(IV.10)}$$

Drawing $\hat{S}(y_0)$ in the same plot as $\hat{S}(y_1)$ in Figure IV.15 (the dashed line), we can clearly see the asymmetries of saliency. Considering two displays, where the target and the distractor are exchanged from one to the other, they are equivalent to increasing the lenth of the target by ΔL in one display, while increasing that of the distractors by ΔL in the other diaplsy. The target saliency of the two displays therefore corresponds to the values of $\hat{S}(y_0)$ and $\hat{S}(y_1)$ in Figure IV.15, with the same $\Delta L/L$ on the x-axis. The plot indicates that the target saliency is always higher when the target is longer than the distractor than the reverse, i.e. an asymmetric behavior of discriminant saliency.

Finally, it is worth noting that the compliance of discriminant saliency with human search asymmetries provides a unified justification for the seemly disjoint observations from both neurophysiology and psychophysics, namely divisive normalization and saliency asymmetries. These are, in some sense, the central components of the neurophysiology of V1 and the psychophysics of visual search. Divisive normalization explains a rich set of neural behaviors that cannot be accommodated by the classic model of "linear filtering plus non-linearity", search asymmetries are one of the most heavily studied properties of visual search. Discriminant saliency provides a unified *functional* justification to these observations: optimal decision making, that exploits the statistical structure of natural images to achieve computational efficiency, and is possible with biological hardware.

Group scanning theory

The derivation in the previous section provides another interesting property of discriminant saliency that both (IV.9) and (IV.10) increases as the number of distractors n, covered by the center window, decreases. In other words, reducing the size of the center window will always increase the saliency of the target so that the search of the target becomes easier. Figure IV.16 plots $\hat{S}(y_1)$ and $\hat{S}(y_1)$ for the following parameter settings: $y_1 = \Delta L/L = -0.3$, $\tau = 35/36$ for $\hat{S}(y_1)$, and $y_0 = \Delta L/L = -0.3$, $\tau = 35/36$ for $\hat{S}(y_0)$. We can see that a short target among long distractors can be as salient as a long target among short distractors, if only the center window is small enough (e.g., $n = 1$). This property suggests a strategy for search: start with a large center window (e.g., the whole display as a group) to compute saliency, and then gradually reduce the size of the center window until the saliency of the target "pops-out". This strategy is, in fact, the "group scanning" hypothesis suggested in [196]. The coincidence that discriminant saliency supports not only the Weber's law explanation of visual search, but also the grouping scanning strategy confirms, once again, that discriminant saliency is a biologically plausible measure of human saliency.

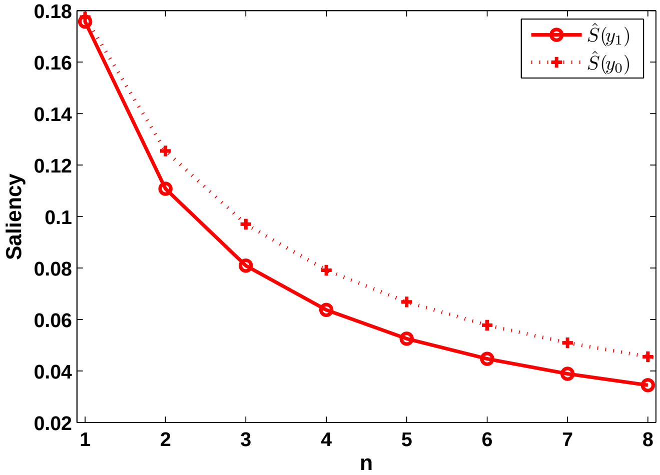


Figure IV.16 Change of discriminant saliency as a function of the number of distractors (n) covered by the center window.

IV.G Acknowledgement

The text of Chapter IV, in part, is based on the materials as it appears in: D. Gao and N. Vasconcelos. Decision-theoretic saliency: computational principles, biological plausibility, and implications for neurophysiology and psychophysics. Accepted for publication, *Neural Computation.* D. Gao, V. Mahadevan and N. Vasconcelos On the plausibility of the discriminant center-surround hypothesis for visual saliency. Accepted for publication, *Journal of Vision.* It, in part, is also based on a co-authored work with N. Vasconcelos. The dissertation author was a primary researcher and an author of the cited materials.

Chapter V

Object recognition with top-down discriminant saliency

We have seen, so far, that the discriminant saliency is 1) physiologically plausible and 2) able to make accurate predictions of the psychophysical behaviors of human saliency. This encourages us to examine its performance as a solution for computer vision problems. In fact, in computer vision literature, it has recently become quite popular to adopt saliency detectors as a front-end of object recognition systems [56, 46, 73, 184]. In these applications, the use of saliency detectors eliminates image regions that are not interesting for recognition, and often significantly reduces the computational complexity of the recognition system.

Although it seems natural to adopt, in object recognition, top-down saliency detectors which are expected to provide informative image regions for the specific object to recognize, this has been rarely the case in computer vision. On the contrary, the majority of the recognition systems in this literature use bottom-up saliency detectors (e.g. [56, 46, 184, 111, 20, 32, 230, 158]). The frequently used bottom-up detectors are, for example, Harris detector [68, 60, 127], scale saliency detector [100], and MSER detector [125] (see Chapter I for an overview). Since these detectors do not tie the optimality of saliency judgements to the specific goal of recognition, the detected locations may not necessarily be informative or discriminant for the objects to recognize.

In this chapter we report results of various experiments designed to characterize the performance of the top-down discriminant saliency detector (described in Section II.D), and compare it to alternative saliency principles adopted in computer vision systems. In particular, the discriminant saliency detector (DSD) is compared with some popular representatives from the literature, which we refer to as *classic* saliency detectors: the scale saliency detector (SSD) [99], the Harris-Laplace (HarrLap) [127], the Hessian-Laplace (HesLap) [127], and the maximally stable extremal region (MSER) detector [125]. The results presented for SSD, HarrLap, HesLap and MSER were produced with the binaries available from `http://www.robots.ox.ac.uk/~timork/salscale.html` and `http://www.robots.ox.ac.uk/~vgg/research/affine/detectors.html`. The default parameter settings, suggested by the authors, were used in all experiments.

V.A Detection of object categories

We start with a comparison on the problem of detecting object categories in cluttered imagery.

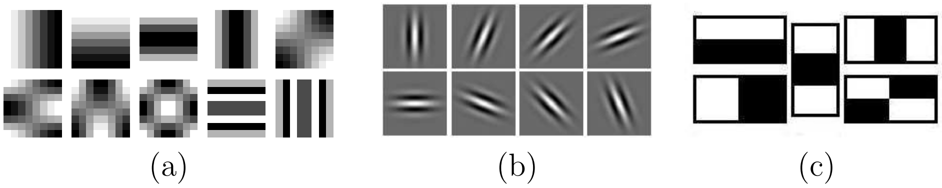
V.A.1 Experimental set-up

This comparison is based on the popular Caltech image database[1], using the set-up proposed in [56]. In particular, six image classes, faces (435 images), motorbikes (800 images), airplanes (800 images), rear view of cars (800 images), spotted-cats (200 images), and side view of cars (550 training and 170 test images) were used as the class of interest ($Y = 1$). The Caltech class of "background" images was used, in all cases, as the *all* class ($Y = 0$). Except for the class of car side views, where explicit training and test assignments are provided, the images in each class were randomly divided into training and testing sets, each containing half of the image set. All saliency detectors were applied to the test images, producing a saliency map per image and detector[2]. These saliency maps were histogrammed and classified by an SVM, as described in section II.D.2. For each saliency detector, the SVM was trained on the histograms of saliency responses of the training set. Detection performance was evaluated with the 1 minus the receiver-operating characteristic equal-error-rate (EER) measure, i.e., 1 minus the rate at which the probabilities of false positives and misses are equal ($1 - EER$).

DSD was evaluated with three feature sets commonly used in the vision literature. The first was a *multi-scale version of the discrete cosine transform* (DCT). Each image was decomposed into a four-level Gaussian pyramid, and the DCT features obtained by projecting each level onto the 8×8 DCT basis functions.

[1] Available from `http://www.vision.caltech.edu/archive.html`.

[2] For detectors that do not produce a saliency map (e.g. HarrLap), the latter was approximated by a weighted sum of Gaussians, centered at the salient locations, with covariance determined by shape of the salient region associated with each location, and weight determined by its saliency.



(a) (b) (c)

Figure V.1 Some of the basis functions in the (a) DCT, (b) Gabor, and (c) Harr feature sets.

The so-called DC coefficient (average of the image patch) was discarded for all scales, in order to guarantee lighting invariance. As shown in Figure V.1 (a), many of the DCT basis functions can be interpreted as detectors of perceptually relevant image attributes, including edges, corners, t-junctions, and spots. The second was a *Gabor wavelet* based on a filter dictionary of 4 scales and 8 directions (evenly spread from 0 to π, as shown in Figure V.1 (b)). It was also made scale-adaptive by application to a four-level Gaussian pyramid. The third set was the *Haar wavelet* based on the five basic features shown in Figure V.1 (c). By varying the size and ratio of the width and height of each rectangle, we generated a set with a total of 330 features. Haar wavelets have recently become very popular in the vision literature, due to their extreme computational efficiency [213], which makes them highly appealing for real-time processing. This can be important for certain applications of discriminant saliency.

Overall, seven *SVM-based saliency map classifiers* were compared: three based on implementations of DSD with the three feature sets, and four based on the classic detectors. As additional benchmarks, we have also tested two classification methods. The first is an SVM identical to that used for saliency-map classification, but applied directly to the images (after stacking all pixels in a column) rather than to saliency histograms. It is referred to as the *pixel-based classifier*. The second is the *constellation classifier* of [56]. While the former is an example of the simplest possible solution to the problem of detecting object categories in clutter, the latter is a representative of the state-of-the-art in this area.

Table V.1 Saliency detection accuracy in the presence of clutter.

	DSD			SSD	Harr-	Hes-	MSER	pixel	constel-
	DCT	Gabor	Harr		Lap	Lap			lation
Faces	**97.2**	95.4	93.1	77.3	56.2	64.5	55.3	93.1	96.4
Bikes	**96.3**	96.0	93.5	81.3	86.0	88.5	81.5	87.8	92.5
Planes	93.0	93.5	**94.8**	78.7	75.3	81.5	87.8	90.3	90.2
Cars(rear)	**100.00**	98.1	99.9	90.9	89.0	86.3	75.5	99.5	90.3
Spotted-cats	**95.0**	92.8	94.3	79.0	56.0	52.0	65.0	81.0	90.0
Cars(side)	**94.1**	93.4	93.8	55.9	54.7	62.4	71.2	59.4	88.5
Average	**96.2**	94.9	94.9	77.2	69.5	72.5	72.7	85.2	91.3

V.A.2 Detection accuracy

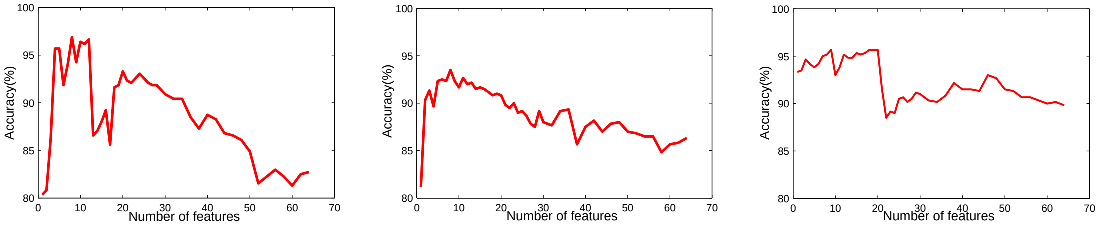
Table V.1 presents the classification accuracy achieved by the seven classifiers. With respect to saliency principles, all classifiers based on classic detectors (SSD, HarrLap, HesLap, and MSER) perform poorly. While the average rate of DSD varies between 94.9 and 96.2% (depending on the feature set) the performance of the classic detectors is between 69.5 (HarrLap) and 72.7% (MSER). This shows that saliency maps produced by the latter are not always informative about the presence/absence of the class of interest in the images to classify. Somewhat surprisingly, given that the Caltech images contain substantial clutter (e.g., see Figure V.3), the performance of the simple pixel-based classifier is very reasonable (average rate of 85.2%). It is, nevertheless, inferior to that of the constellation classifier (average rate of 91.3%), for all but the "Car rear views" and "Planes" classes. The final observation is that even the latter is clearly inferior to all DSD-based classifiers, which achieve the overall best accuracies. While we do not claim that the DSD-based saliency classifier is the ultimate solution to the problem of detecting object classes in clutter, these results support the claims that discriminant saliency 1) produces saliency maps that are informative about the class of interest, and 2) is more effective in doing so than techniques, such as SSD or HarrLap, commonly used in the recognition literature [56, 46, 184].

It should also be noted that the comparison above is somewhat unfair to the constellation classifier, which tries to solve a more difficult problem than that considered in this experiment. While the question of interest here is "is class x present in the image or not?" the constellation classifier can actually localize the object from the class of interest (e.g. a face) in the image. The reasonable performance of the pixel-based classifier in this experiment indicates that it is probably not necessary to solve the localization problem to achieve good detection rates on Caltech. In fact, the best detection rates published on this database are, to the best of our knowledge, achieved by classifiers that do not even attempt to solve the localization problem [176]. The question of whether discriminant saliency can be used to localize the regions associated with the class of interest is analyzed in the following section.

V.A.3 Features

Regarding the relative performance of the different feature sets, while the DCT appears to be the clear winner, all feature sets achieved high accuracy. This implies that discriminant saliency is not overly dependent on a unique set of features. However, a close inspection of Table V.1 also suggests that further performance improvements should be possible by designing specific features for each class. Note, for example, that the Haar features achieved the top performance in the "Airplane" class, where the elongated airplane bodies are very salient. This is mostly due to the fact that one of the Haar basis functions (bottom left of Figure V.1 (c)) is close to a matched filter for this feature. An interesting question is, therefore, how to augment the discriminant saliency principle with feature extraction, i.e. the ability to learn the set of features which are most discriminant for the class of interest (rather than just selecting a subset from a previously defined feature collection). This is discussed in [65]. In all subsequent experiments, the DSD is based on the DCT feature set.

A final question is the sensitivity to the number of features declared salient during feature selection. Figure V.2 presents the variation, with this number, of



(a) (b) (c)

Figure V.2 Classification accuracy vs number of features used by the DSD for (a) faces, (b) motorbikes and (c) airplanes.

the detection accuracy of the DCT set, for three of the classes (the curves for the others are similar and were omitted for brevity, and similar results were observed with the Gabor and Haar sets). In general, accuracy is approximately constant over a range of feature cardinalities (as shown in (a) and (c)), but there are also cases where it decays monotonically with cardinality (as in (b)). The rate of decay is, however, slow and, in all cases, there is a significant range of cardinalities where performance is close to the optimal, suggesting that discriminant saliency is robust to variations of this parameter. Visual inspection of saliency maps obtained with different numbers of features has also shown no substantial differences with respect to the saliency maps obtained with the optimal cardinality.

V.B Object localization

Although the detection accuracies of the previous section are a good sign for discriminant saliency, the ultimate performance measure for a saliency detector is its ability to localize the image regions associated with the class of interest. To evaluate the performance of the DSD under this criterion we conducted two sets of experiments.

V.B.1 Subjective evaluation

We started by visually inspecting all saliency maps. As exemplified by Figure V.3, this revealed that DSD is superior to the classic detectors in its ability to localize instances of the class of interest. The figure presents examples of saliency maps generated by DSD, and the locations of highest saliency, according to the five detectors. While DSD is able to disregard background clutter, focusing on instances of the target class, many of the locations detected by the other methods are uninformative about the latter.

V.B.2 Objective evaluation

A second set of experiments targeted an objective evaluation of the localization ability of the various saliency detectors. It was based on a protocol proposed in [100], which exploits the fact that, although there is a fair amount of intra-class variation on Caltech (e.g., faces of different people appear with different expressions and under variable lighting conditions), there is enough commonality of pose (e.g., all faces shown in frontal view) to allow the affine mapping of the images of each class into a common coordinate frame. The frame associated with each class was estimated, by Kadir et al. [100], by manually clicking on corresponding points in each of the images of the class. The stability of the salient locations when mapped to the common coordinate frame is a measure of the localization ability of the saliency detector. In particular, a mapped salient location, R_a, is considered to match the reference image if there exists a salient location, R_f, in the latter such that the *overlap error* is sufficiently small, i.e.

$$1 - \frac{R_a \cap R_f}{R_a \cup R_f} < \epsilon, \tag{V.1}$$

where $\cap$ represents intersection, and $\cup$ union. To avoid favoring matches between larger salient points, the reference region was normalized to a radius of 30 pixels before matching, as suggested by [128]. The matching threshold ϵ was set to 0.4. The localization score Q is defined as

$$Q = \frac{Total\ number\ of\ matches}{Total\ number\ of\ locations}. \tag{V.2}$$

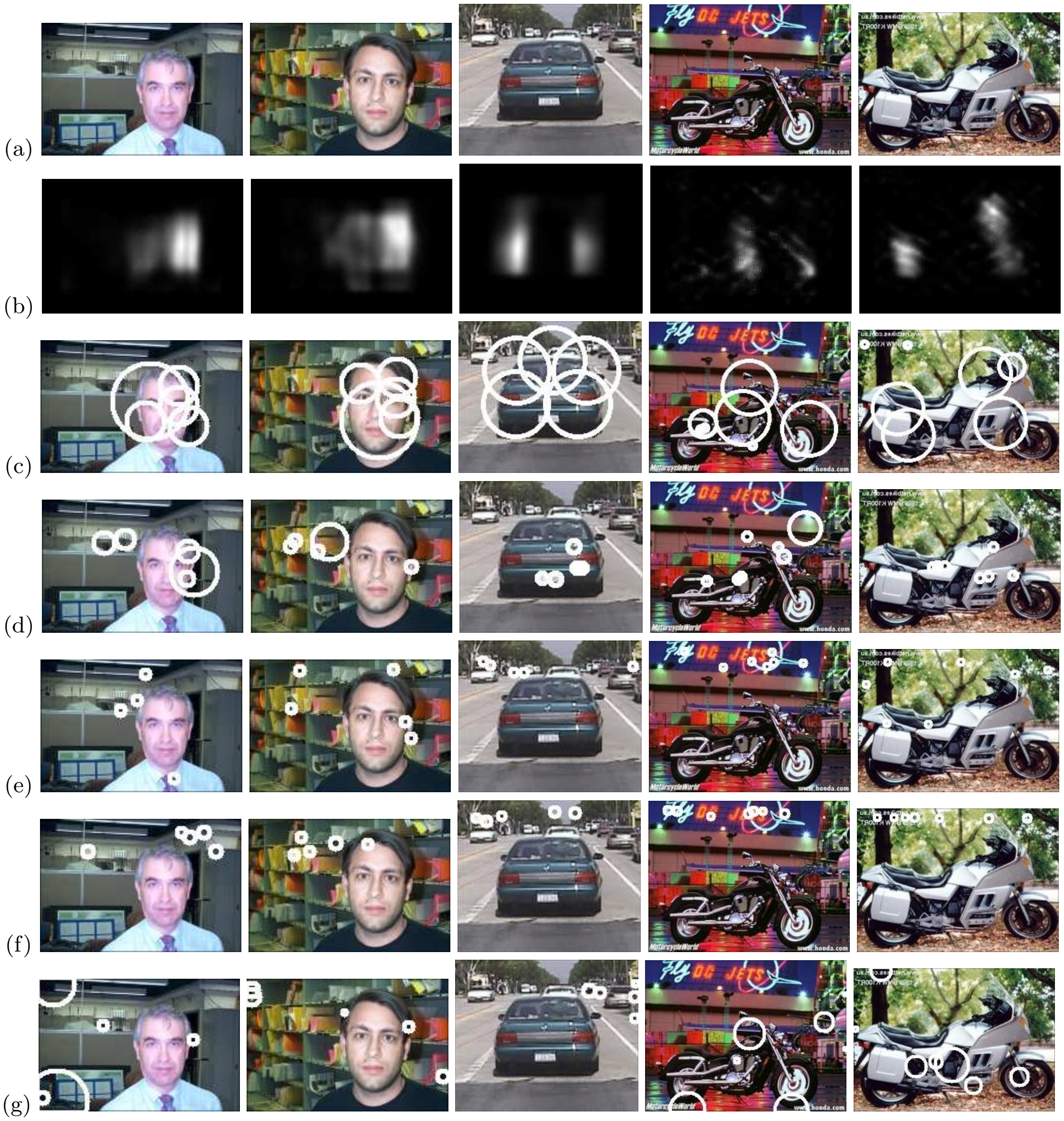
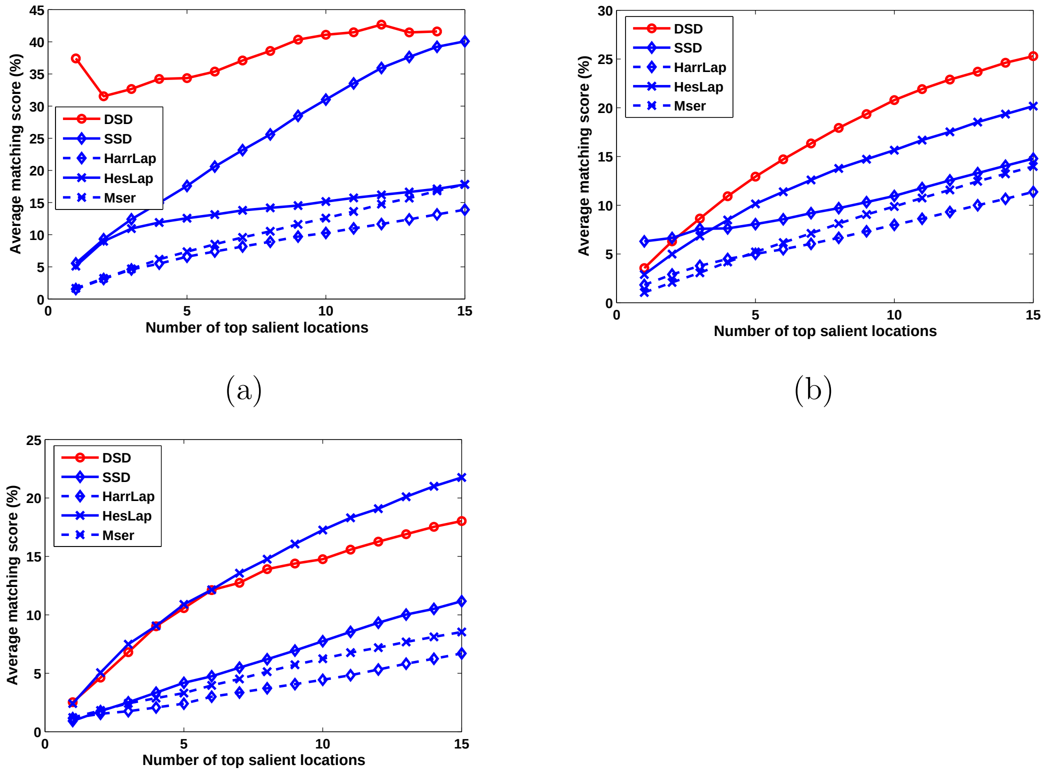


Figure V.3 Original images (a) , saliency maps generated by DSD (b) and a comparison of salient locations detected by: (c) DSD, (d) SSD, (e) HarrLap, (f) HesLap, and (g) MSER. Salient locations are the centers of the white circles, the circle radii representing scale. Only the first (5 for faces and cars, 7 for motorbikes) locations identified by the detector as most salient are marked.

If N locations are detected for each of the M images in the database, the score Q_i of reference image i is

$$Q_i = \frac{N_M^i}{N(M-1)}, \tag{V.3}$$



(a)

(b)

(c)

Figure V.4 Localization accuracy of various saliency detectors for (a) face, (b) motorbike, and (c) car.

where N_M^i is the total number of matches between that image and all other $M-1$ images in the database. The overall score Q is the average of Q_i over the entire database, and is evaluated as a function of the number of detected regions per image.

The localization ability of the five detectors was compared on the three Caltech object classes (face, motorbike, and rear views of cars) for which alignment ground truth is available [100]. As shown in Figure V.4, discriminant saliency performed better than most other methods, for all classes. On faces, only SSD produced competitive results, and only for a relatively large number of salient points. On motorbike SSD performed best with a single salient point (SSD is particularly good at finding the circular bike wheels) but its performance degraded

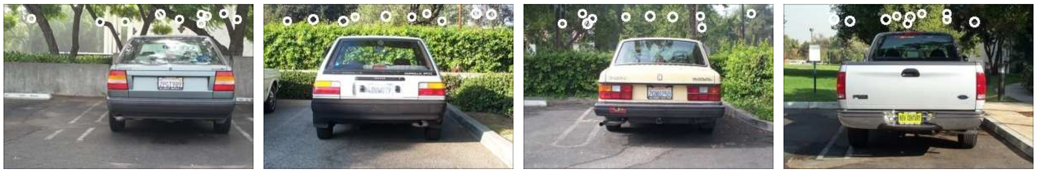


Figure V.5 Examples of salient locations detected by HesLap on images of car rear views.

quickly. On this class, only HesLap achieved a score consistently higher than half of that obtained by DSD, which once again produced the best overall results. On car rear views, DSD outperformed all methods but HesLap. It should be emphasized that these results must be considered in conjunction with Table V.1. The fact that a saliency detector produces highly localized salient points is not very useful if these are not co-located with the target objects. This is illustrated in Figure V.5, where it can be seen that, for car rear views, HesLap frequently produces salient points which are stable but irrelevant for recognition. On the other hand, DSD tends to produce salient points that are not only stable, but also localized within the visual class of interest. This is illustrated by Figure V.6, which presents examples of salient locations for all Caltech classes, illustrating the robustness of DSD-based object localization to substantial variability in appearance and significant amounts of clutter. Typically, high localization accuracy is achieved with a few salient locations.

V.C Repeatability of salient locations

We have shown, so far, that the top-down discriminant saliency produces salient locations which are more informative about the objects to recognize than other saliency mechanisms. In what follows we evaluate its stability under various generic image transformations. This is the task for which many bottom-up saliency detectors are proposed to be optimal, or close to optimal.

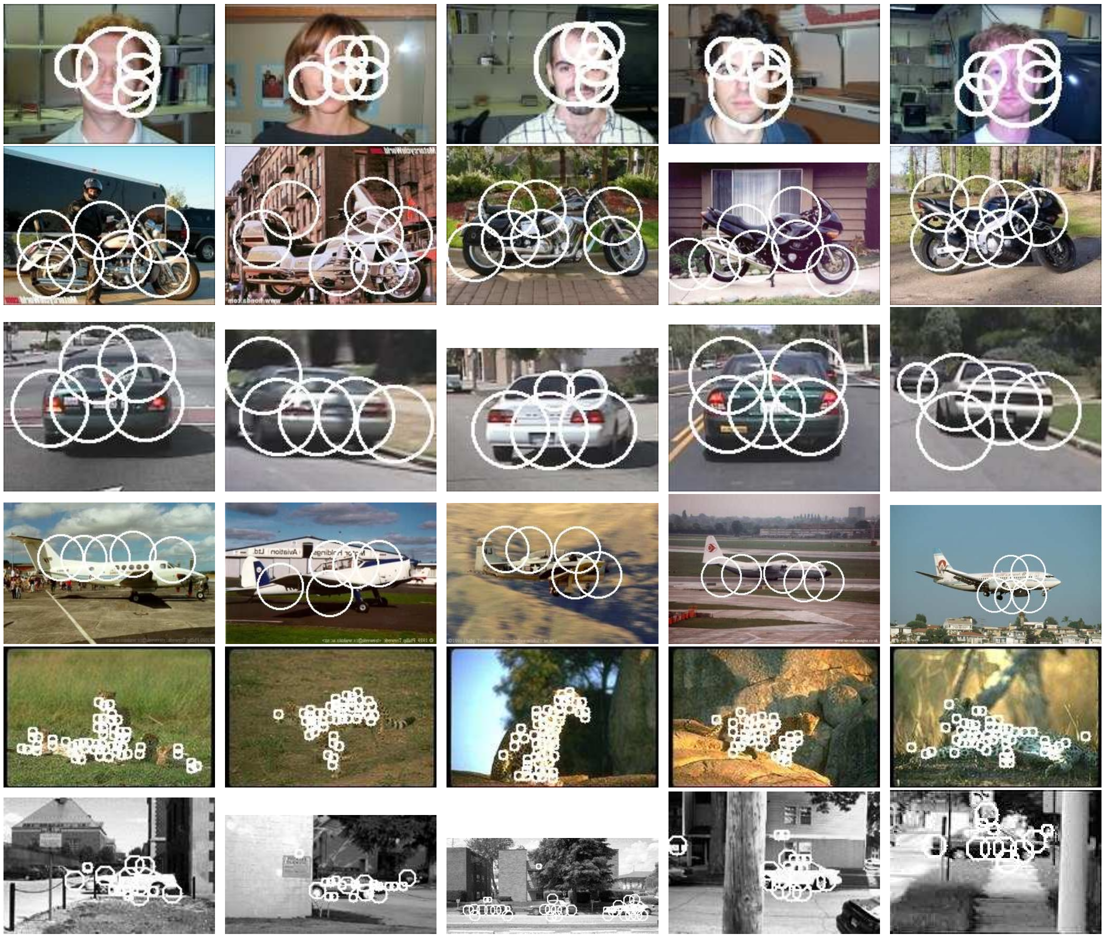


Figure V.6 Examples of discriminant saliency detection on Caltech image classes.

V.C.1 Experimental protocol

Ideally, the salient locations extracted from a scene should be unaffected by variations of the (scene-independent) parameters that control the imaging process, e.g. lighting, geometric transformations such as rotation and scaling, and so forth. Mikolajczyk et al. [128] have devised an experimental protocol for evaluating the repeatability of salient points under various such transformations. The protocol includes 8 classes of transformations, each class consisting of 6 images produced by applying a set of transformations, from the same family, to a common scene. The transformations include joint scaling and rotation, changes of viewpoint angle (homographies), blurring, JPEG artifacts, and lighting. Scale + rotation, view point changes, and blurring are applied to sets of two scenes which can be roughly characterized as textured (e.g. images of tree bark or of a brickwall) or structured (e.g. an outdoors scene depicting a boat or a wall covered with graffiti).

To measure the repeatability of a saliency detector, the protocol uses the first image of each class as a reference image, and maps the rest of the five images to the coordinate frame of the reference. The salient points detected on each of the five images are then matched with those detected on the reference image for correspondences. Salient points falling out of the common frame of each pair of images are eliminated before matching. Corresponding points between a pair of images are mapped using the criterion of (V.1). Again, the reference region was normalized to a radius of 30 pixels before matching, as suggested by [128]. The matching threshold, ϵ, was set to 0.4. The repeatability score for a given pair of images is computed as the ratio between the number of correspondences and the smaller of the number of regions in the pair.

Extending the protocol for learning

Since the protocol of [128] does not define training and test images, we propose an extension applicable to learning-based methods. This extended protocol is based on various rounds of experiments. At the k^{th} round, the first k images of a given class are treated as a training set for that class, and the repeatability scores of the learned saliency detector are measured on the remaining $6-k$ images. This is accomplished by matching the interest points detected on these images to the reference image, which is the k^{th} image. When $k = 1$, i.e. train on the first image and test on all remaining images, this reduces to the protocol of [128], but larger values of k enable a quantification of the improvement of stability with the richness of the training set. The new protocol is illustrated in Figure V.7 for $k = 1$ and 2. In the experiment reported below, the repeatability score of DSD is measured for $k = \{1, 2, 3\}$, and compared to the bottom-up detectors (SSD, HarrLap, HesLap, and MSER), operating under the same test protocol (i.e., using image

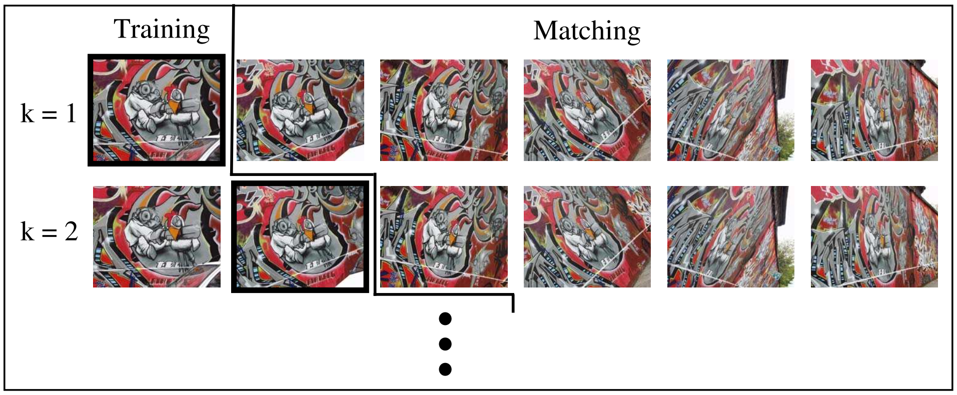
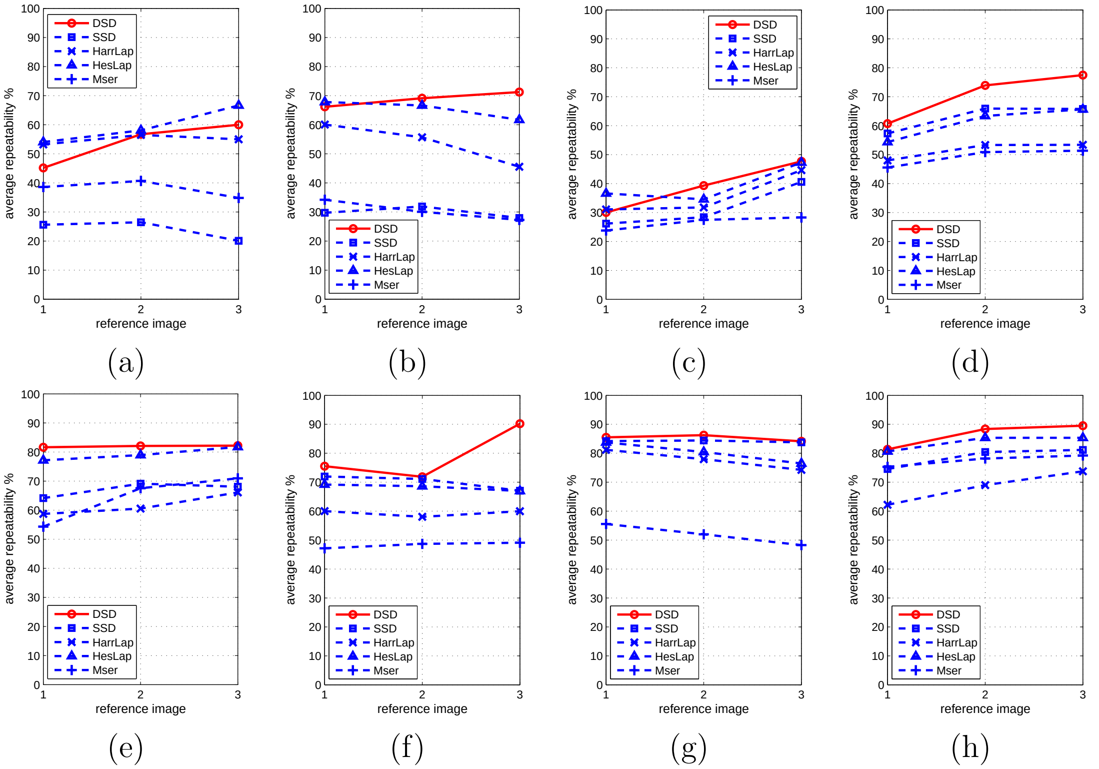


Figure V.7 Extended protocol for the evaluation of the repeatability of learned interest points. At the k^{th} round, the detector is trained on the first k images, and the repeatability score measured by matching the remaining images to the reference, which is set to the last training image, and shown with thick boundaries.

k as a reference). To deal with the extreme variations of scale of this dataset, we implemented a simple multi-resolution extension of DSD: discriminant salient points were first detected at each layer of a Gaussian pyramid decomposition of the image and, at each salient point location, the layer of largest saliency was selected. This type of processing is already included in all other detectors.

V.C.2 Results

The average repeatability scores obtained (across the set of test images) by each saliency detector are shown, as a function of the reference image number k, in Figure V.8. A more detailed characterization, presenting the repeatability score of each test image and each of the values of k, is shown in Figure V.9-Figure V.12. The plots on the left columns of Figure V.9-Figure V.12 are equivalent to those of [128], the ones on the center and right columns correspond to $k = 2$ and $k = 3$, respectively. In all plots, the extent of the transformation between the reference (image k) and the test image (whose number is shown) increases with the latter. Note that Figure V.8 presents the average of the repeatabilities in Figure V.9-Figure V.12. For example, Figure V.8 (a), presents the average of each curve in



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure V.8 Repeatability of salient locations under different conditions: *scale + rotation* ((a) for structure & (b) for texture); *viewpoint angle* ((c) for structure & (d) for texture); *blur* ((e) for structure & (f) for texture); *JPEG compression* (g); and *lighting* (h).

the top row of Figure V.9, as a function of k.

The following conclusions can be reached from the figures. First, a richer training set improves the performance of DSD for all transformations. This improvement occurs not only in absolute terms, but also comparatively to the other methods. This shows that the principle of discriminant learning is a good idea from a repeatability point of view. It enables the design of detectors which can be made more invariant by simply increasing the richness of the transformations covered by their training sets. Second, DSD is competitive with the other techniques even when the set of positive training examples is a single image. In this case, DSD achieves the top repeatability scores for five of the eight classes ((d)-(h)), is very close to the best for another (b), and is always better than at least two of the

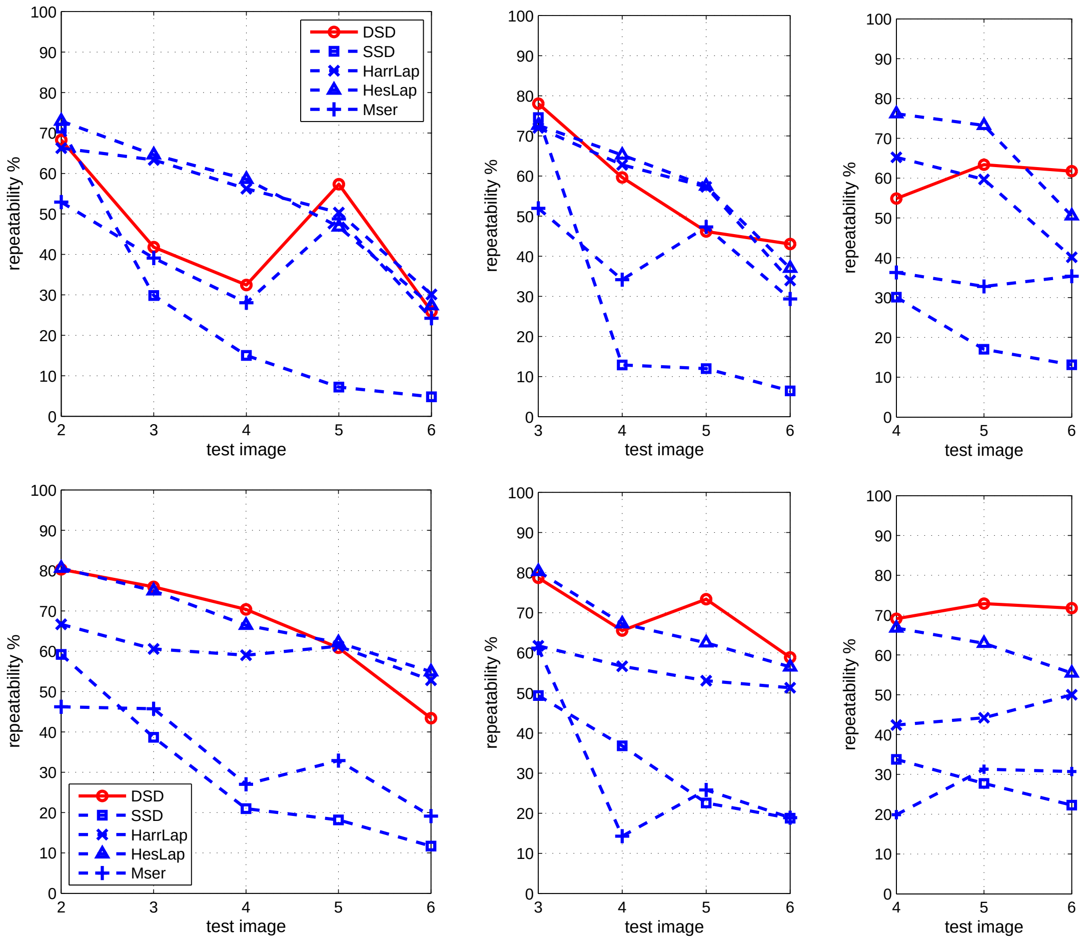


Figure V.9 Repeatability of salient locations under *scale + rotation* changes ((top) structure & (bottom) texture) with different number of training images for DSD: $k = 1$ (left), 2 (middle), and 3 (right).

classical algorithms. Finally, when the most diverse training sets are used ($k = 3$) DSD has the top scores for all but one class.

It is also interesting to analyze these results by transformation and image class. With respect to transformations, DSD is the most robust method in the presence of blurring, JPEG artifacts and lighting transformations (Figure V.8 (e-h)) independently of the degree of training. It also achieves the best performance for changes of viewpoint angle, but this can require more than one example (c). Its worst performance occurs under combinations of scale and rotation, where it is always inferior to HesLap for small amounts of training data, and sometimes infe-

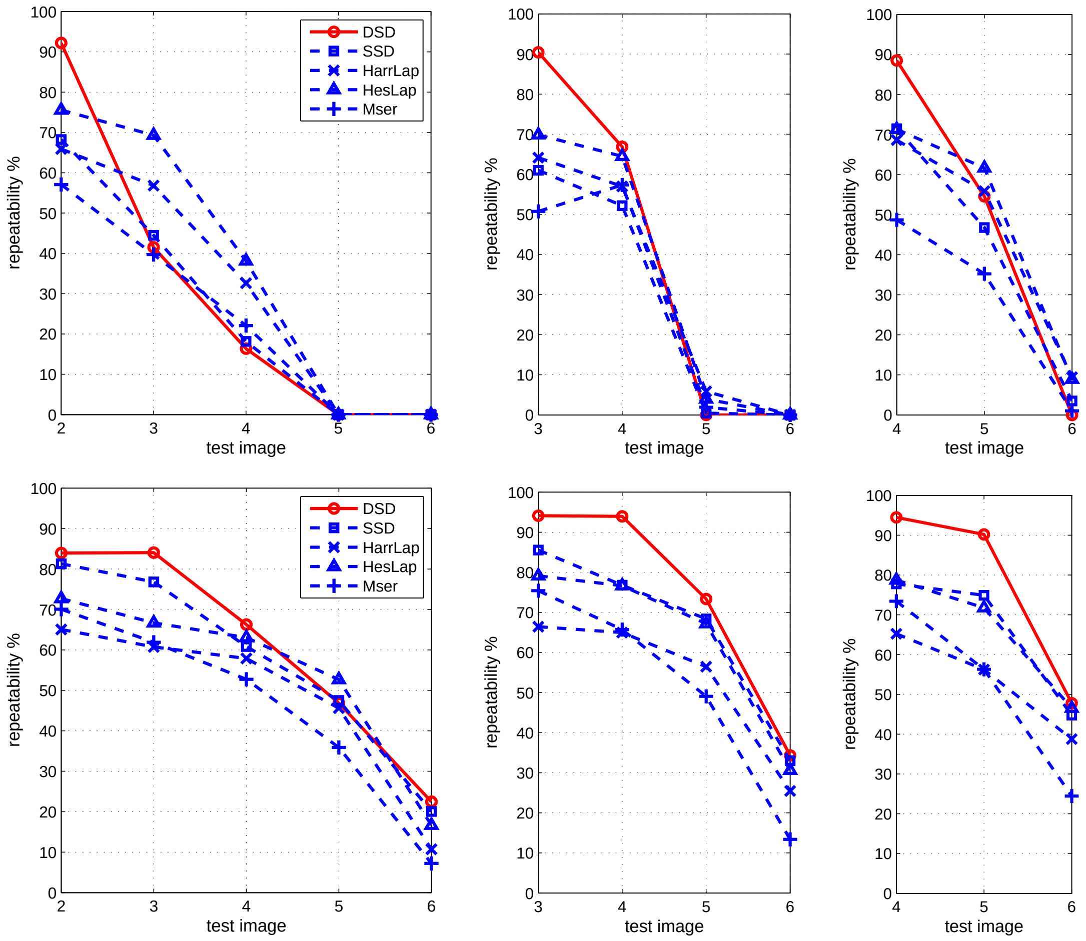


Figure V.10 Repeatability of salient locations under *viewpoint angle* changes ((top) structure & (bottom) texture) with different number of training images for DSD: $k = 1$ (left), 2 (middle), and 3 (right).

rior even for the largest training sets. With respect to image class, it is interesting to note that the robustness of DSD to geometric transformations is better for texture ((b) & (d)) than for structured scenes ((a) & (c)). While, for the former, DSD achieves the best, or close to the best, performance at all training levels, for structured scenes DSD is less invariant than at least one of the classic detectors in all training regimes.

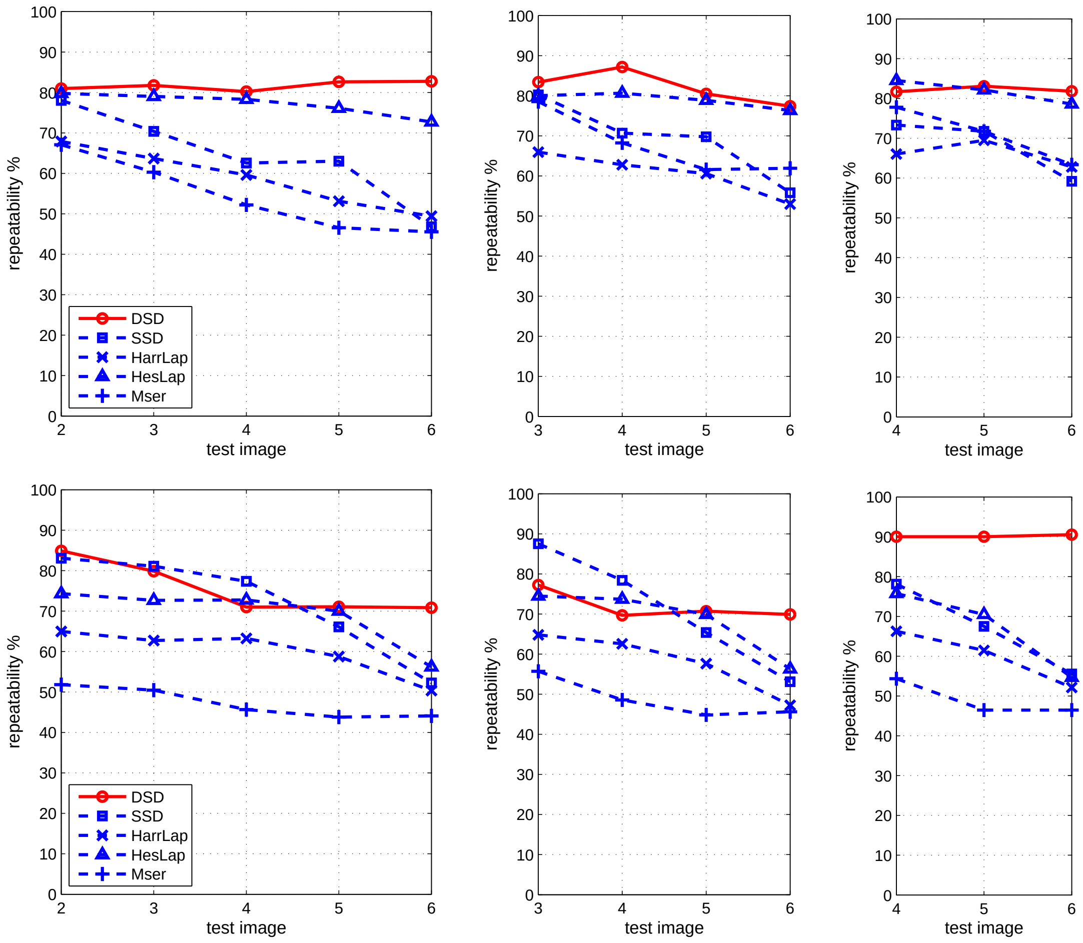


Figure V.11 Repeatability of salient locations under *blurring* ((top) structure & (bottom) texture) with different number of training images for DSD: $k = 1$ (left), 2 (middle), and 3 (right).

Invariance to 3D rotation

To evaluate invariance to more general transformations, such as 3D rotation, we measured the repeatability of the salient points produced by all methods on the Columbia Object Image Library (COIL-100) [74]. This is a library of images of 100 objects, containing 72 images from each object, obtained by rotating the object in 3D by 5^o between consecutive views. The appearance changes due to 3D rotation make COIL more challenging than the database of [128], for methods that explicitly encode invariance. To avoid saliency ambiguities due to large

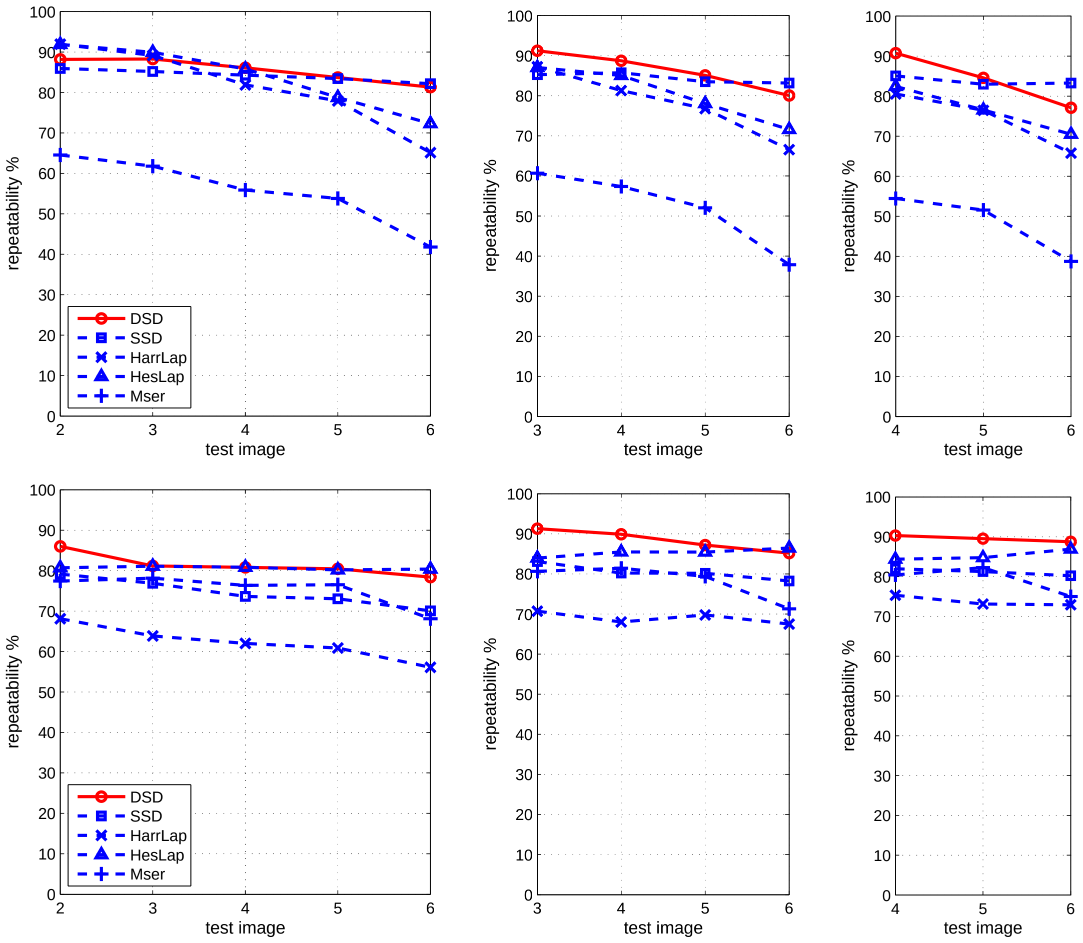


Figure V.12 Repeatability of salient locations under *JPEG compression* (top) and *lighting* (bottom) changes with different number of training images for DSD: $k = 1$ (left), 2 (middle), and 3 (right).

view-angle change (e.g. the front of an object is not visible from the rear) we used six consecutive views of each object for training and the next three adjacent views (subsampled from the next six adjacent original views so as to produce a separation of 10^o of rotation between views) for testing. For each image, the ten most salient locations were computed, and each salient location was considered stable if it appeared in all three test images. The overall stability score was measured with (V.2).

Table V.2 lists the stability score achieved by the five saliency detectors,

Table V.2 Stability results on COIL-100.

	DSD	SSD	HarrLap	HesLap	MSER
Stability(%)	**74.7**	52.2	46.5	47.0	57.3

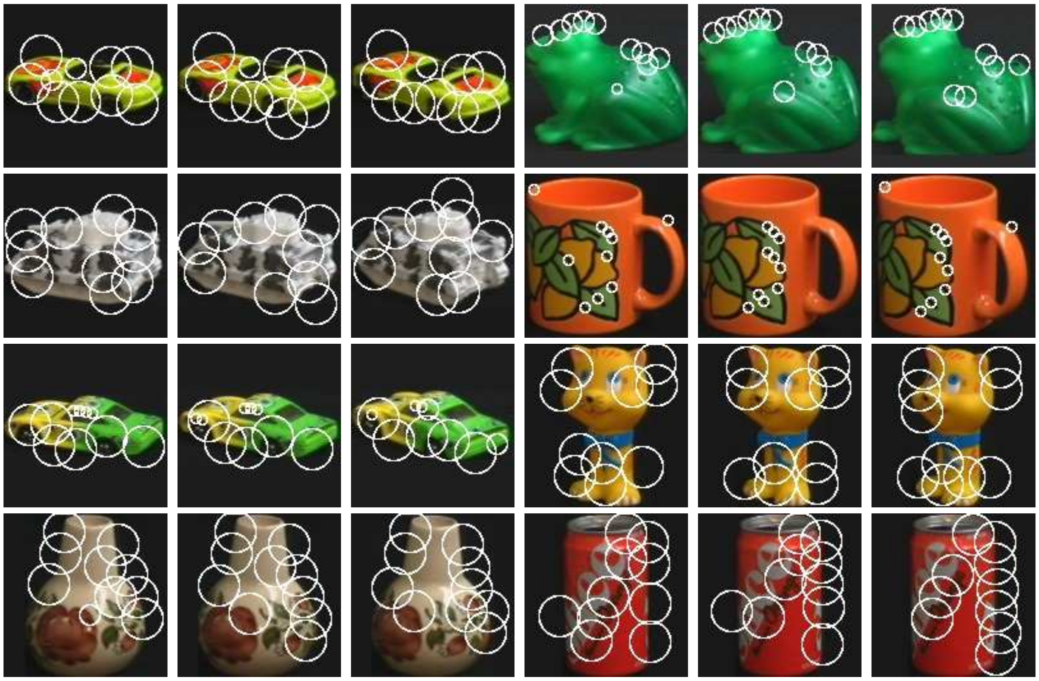


Figure V.13 Examples of salient locations detected by DSD for COIL.

showing that all classic detectors produce less stable salient points than those of DSD. Figure V.13 shows that the locations detected by the latter maintain a consistent appearance as the object changes pose. This implies that discriminant saliency selects features which are “consistently salient” for the whole set of object views in the image class. These are features that exhibit small variability of response within the class of interest, while discriminating between this class and all others. On the other hand, the classical (bottom-up) definitions of saliency are only optimally stable for specific classes of spatial transformations (e.g., affine), which do not approximate well enough the transformations found in a database like COIL-100.

V.C.3 Discussion

Overall, the results of the repeatability test illustrate some of the trade-offs associated with learning based (top-down) saliency detectors, such as DSD. On one hand, the ability to select specific features for the class under consideration increases not only the discriminant power but also the stability of saliency detection. It appears that the principle of discriminant learning is a good idea even from a repeatability point of view. It enables the design of detectors which can be made more invariant by simply increasing the richness of the transformations covered by their training sets. This is a property that bottom-up routines lack, sometimes leading to dramatic variability of repeatability scores across classes (see the curves of SSD on Figure V.8 for an example), or even a clear inability to deal with some types of transformations (as is the case on COIL-100). On the other hand, the generalization ability of a top-down detector depends on the quality of its training data and the complexity of the mappings that must be learned. In Figure V.8, this can be seen by the consistent loss of performance for smaller training sets, and the greater difficulties posed by structured scenes, when compared to texture. When little training data is available, or the mappings have great complexity, explicit encoding of certain types of invariance (as done by the classic bottom-up detectors) can be more effective. In this sense, the combination of top-down and bottom-up saliency detectors, to optimally balance the trade-off between learning and pre-specification of invariance, could be beneficial. We will investigate this point in detail in Chapter VII.

V.D The diversity of discriminant saliency attributes

We finalize with a qualitative experiment designed to illustrate the richness of the set of visual attributes that can be declared salient under the discriminant saliency principle. This experiment was based on the Brodatz texture database [23] which, in addition to a great variety of salient attributes - e.g. corners, contours, regular geometric figures (circles, squares, etc.), texture gradients, crisp and soft edges, etc - places two significant challenges to existing saliency detectors: 1) the need to perform saliency judgments in highly textured regions, and 2) a great diversity of shapes for the salient regions associated with different texture classes. The Brodatz database was divided into a training and test set, using a set-up commonly adopted for texture retrieval (described in detail in [208]). The salient features of each class were computed from the training set, and the test images used to produce all saliency maps. The process was repeated for all texture classes, on a one-vs-all setting (class of interest against all others) with each class sequentially considered as the "one" class.

As illustrated by Figure V.14, none of the challenges posed by Brodatz seems very problematic for discriminant saliency. Note, in particular, that the latter does not appear to have any difficulty in 1) ignoring highly textured background areas in favor of a more salient foreground object (two leftmost images in the top row), which could itself be another texture, 2) detecting as salient a wide variety of shapes, contours of different crispness and scale, or 3) even assigning strong saliency to texture gradients (rightmost image in the bottom row). This robustness is a consequence of the fact that salient features are selected according to both the class of interest and the set of images in the *all* class.

V.E Acknowledgement

The text of Chapter V, in part, is based on the materials as it appears in: D. Gao and N. Vasconcelos, Discriminant saliency for visual recognition from cluttered scenes. In *Proc. of Neural Information Processing Systems (NIPS)*, 2004. D. Gao and N. Vasconcelos, Discriminant Interest Points are Stable. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007. D. Gao and N. Vasconcelos, An experimental comparison of three guiding principles for the detection of salient image locations: stability, complexity, and discrimination.

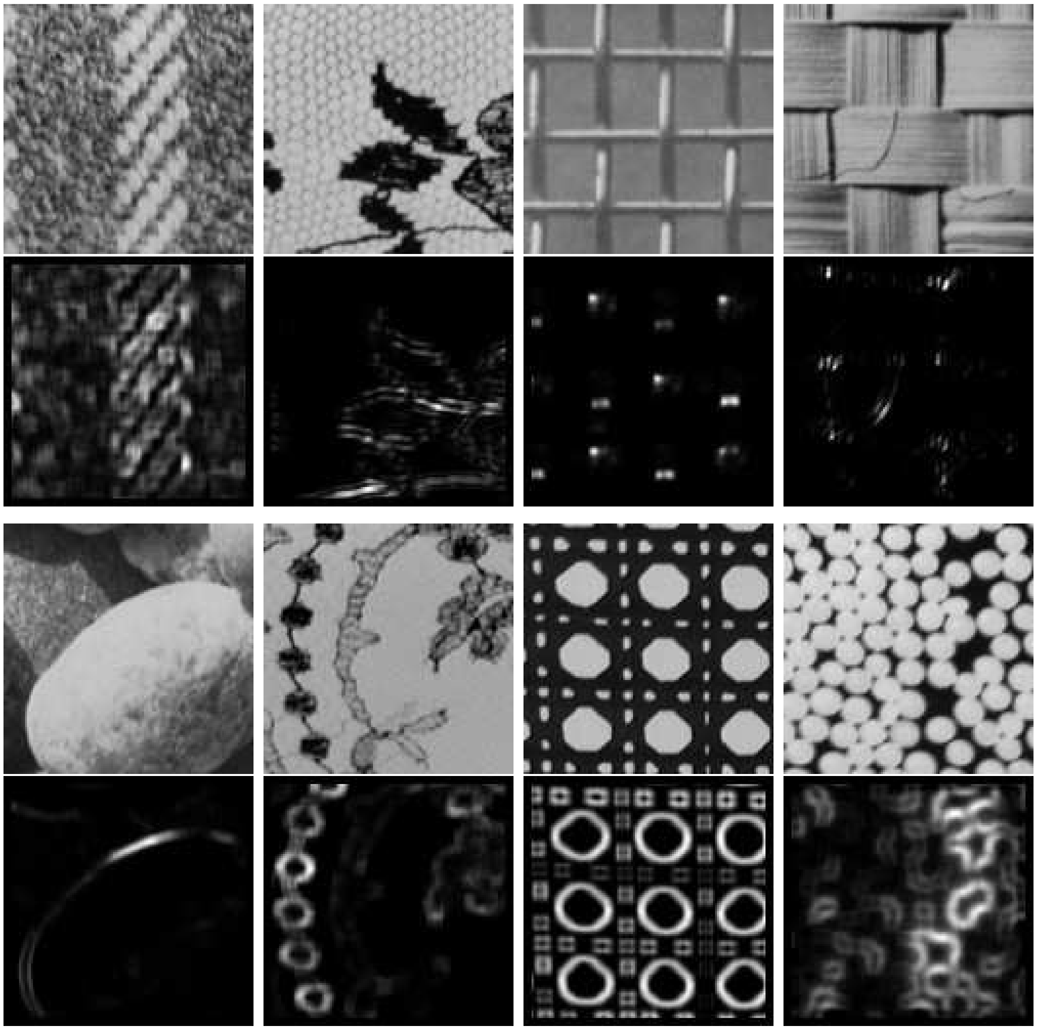


Figure V.14 Saliency maps obtained on various textures from Brodatz. Bright pixels flag salient locations.

The 3rd International Workshop on Attention and Performance in Computational Vision (WAPCV), 2005. It, in part, has also been submitted for publication of the material as it may appear in D. Gao and N. Vasconcelos, Discriminant saliency for visual recognition. Submitted for publication, *IEEE Trans. on Pattern Analysis and Machine Intelligence*. The dissertation author was a primary researcher and an author of the cited materials.

Chapter VI

Prediction of human eye movements by bottom-up discriminant saliency

In the previous chapter we have shown that the top-down discriminant saliency leads to better localization and classification accuracy for object recognition problems, than the existing saliency detectors. However, for applications where no recognition problems is defined, the use of bottom-up saliency detectors is more appropriate. In this chapter we present, for such circumstances, the application of the bottom-up discriminant saliency detector described in Section II.E. In particular, we consider the problems of predicting human eye fixations. The output of the bottom-up discriminant saliency detector is compared to both human performance, and state-of-the-art results.

VI.A Predicting human eye movements

To evaluate the ability of the bottom-up discriminant saliency detector to predict human eye fixation locations, we compared the discriminant saliency maps obtained from a collection of natural images to the eye fixation locations recorded from human subjects, in a free-viewing task.

VI.A.1 Eye movement data and performance metric

The eye-fixation data were collected by Bruce and Tsotsos [24], from 20 subjects and 120 different natural color images, depicting urban scenes (both indoor and outdoor). The images were presented in 1024×768 pixel format on a 21-in. CRT color monitor. The monitor was positioned at viewing distance of 75 cm; consequently, the image presented subtended 32° horizontally and 24° vertically, i.e. approximately 30 pixels per degree of visual angle. All images were presented in random order, to each subject for 4 seconds, with a mask inserted between consecutive presentations. Subjects were given no instructions, and there were no predefined initial fixations. A standard non-head-mounted gaze tracking device (Eye-gaze Response Interface Computer Aid (ERICA) workstation) was applied to record the eye movements. All participants had normal or correct-to-normal vision.

The comparison between saliency predictions and human eye movements was based on a metric proposed in [189]. The basic idea is that, by defining a threshold, a saliency map can be quantized into a binary mask that classifies each image location as either a fixation or non-fixation. Using the measured human eye fixations as ground truth, a receiver operator characteristic (ROC) curve is produced by varying the quantization threshold. In this context, labeling a human fixation as a non-fixation is a false negative and labeling a human non-fixation as a fixation is a false positive. Overall, this procedure quantifies the goodness of the saliency detector at predicting human performance. Perfect prediction corresponds to an ROC area (area under the ROC curve) of 1, while chance performance reduces it to 0.5. Since the metric makes use of all saliency information in both the human fixations and the saliency detector output, it has been adopted in various recent studies [24, 67, 103]. The predictions of discriminant saliency were compared to those of the methods of [89] and [24]. As an absolute benchmark, we also computed the "inter-subject" ROC area [67], which measures fixation consistency between human subjects. For each subject, a "human saliency map" was derived from the fixations of all other subjects, by convolving these fixations with a circular 2-D Gaussian kernel. The standard deviation (σ) of this kernel was set to 1° of visual angle ($\approx$ 30 pixels), which is approximately the radius of the fovea. The "inter-subject" ROC area was then measured by comparing subject fixations to this saliency map, and averaging across subjects and images.

VI.A.2 Results

Table VI.1 presents average ROC areas for all detectors, across the entire image set[1], as well as the "inter-subject" ROC area. It is clear that discriminant saliency achieves the best performance among the three saliency detectors.

[1] It should be noted that the results of [89, 24], for both this table and all subsequent figures, were optimized for this particular image set, by tuning of model parameters. This was not done for discriminant saliency, whose results were produced with the parameter settings of the previous section.

Saliency model	Discriminant	Itti et al. [89]	Bruce et al. [24]	Inter-subject
ROC area	0.7694	0.7287	0.7547	0.8766

Table VI.1 ROC areas for different saliency models with respect to all human fixations.

Nevertheless, because there is still a non-negligible gap to human performance, we studied in greater detail the relationship between the output of saliency algorithms and the subjects' fixations. In [189], Tatler et al. observed that early human fixation locations are more consistent than later ones. As shown in Figure VI.1, this observation holds for the fixation data used in these experiments. In particular, the figure shows that the inter-subject ROC area decreases dramatically as the number of fixated locations increases. The first two locations have significantly higher ROC area than all others. This indicates that, while the first few eye movements are most likely to be driven by bottom-up processing, top-down influences dominate the viewing process after that. Given no specific task, the subjects' attention is likely to be dominated by the interpretation of the objects in the scene, or other forms of top-down guidance. It is, therefore, questionable that any fixations beyond the first or second should be used to evaluate bottom-up detectors.

The ROC area curves in Figure VI.1 also reveal that all bottom-up detectors achieve the best performance at the second fixation. This is unlike the inter-subject performance, which is more consistent for the first fixation. The discrepancy is most likely due to a "central fixation bias" [189]: subjects tend to be biased towards the image center even when there is no initial central fixation point. This bias is illustrated in Figure VI.2, which shows the average inter-subject saliency map for the first and second fixations (average taken across subjects and images). It is clear that the first fixation is very likely to be near the image center, while the second exhibits significantly more diversity.

Taking these observations into account, we compared the performance of the three saliency detectors, using only the first two fixations, and as a function of the inter-subject ROC area. The results are shown in Figure VI.3, where the

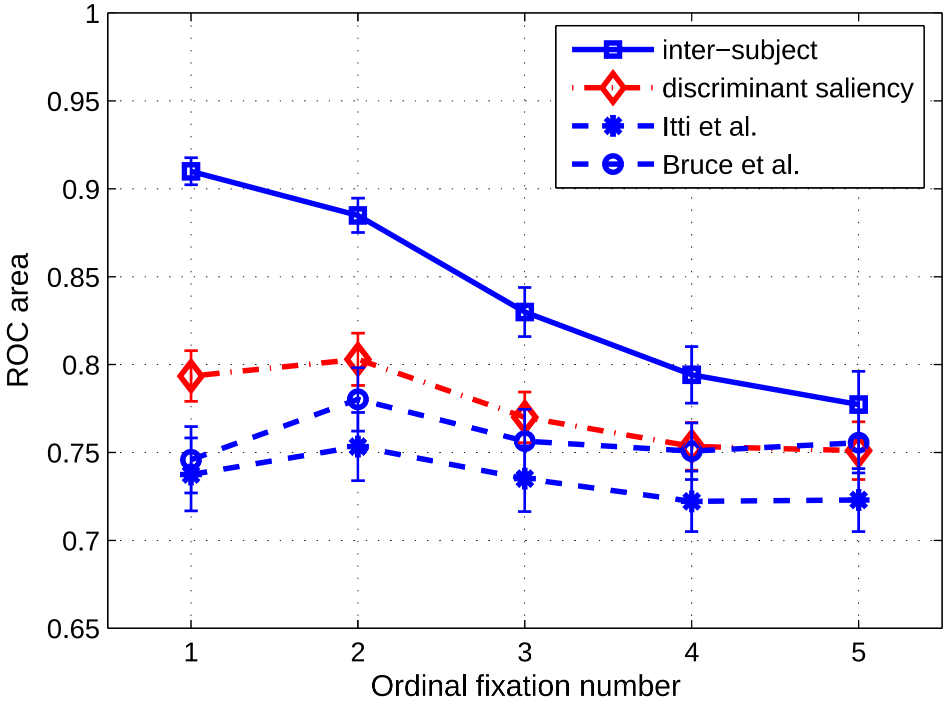


Figure VI.1 ROC area for ordinal eye fixation locations.

thin dotted line represents perfect correlation with human performance. Note that, for all detectors, best performance occurs when inter-subject consistency is highest. Since saliency judgements driven uniquely by bottom-up, stimulus-driven, processing are likely to be constant across subjects, this is the region where it makes most sense to evaluate saliency detection with eye fixation data. In this region, the performance of discriminant saliency (0.85) is close to 90% of that of humans (0.95), while the other two detectors achieve close to 85% (0.81).

Overall, the bottom-up discriminant saliency detector performed best at predicting human fixations among all compared saliency models, both for the entire set of fixations, and for the first two. It also exhibited greater correlation with human performance at all levels of inter-subject consistency, but especially when the latter is large. This is the regime in which saliency is most likely to be due uniquely to bottom-up, stimulus-driven, cues.

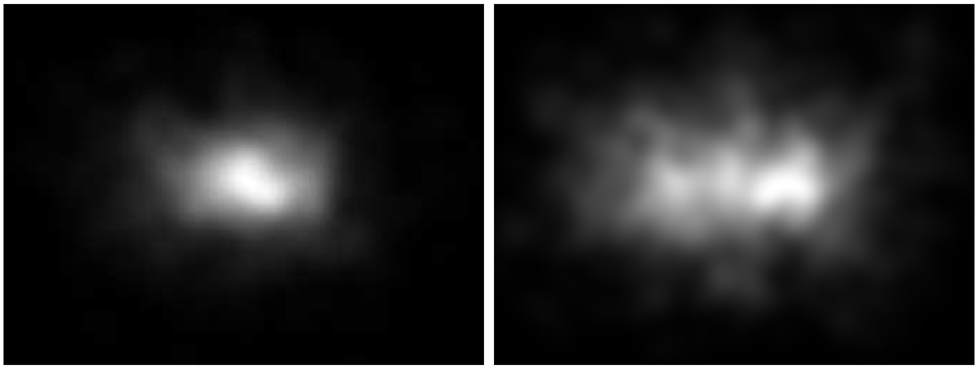


Figure VI.2 Inter-subject saliency maps for the first (left) and the second (right) fixation locations.

VI.B Acknowledgement

The text of Chapter VI, in part, is based on the material as it appears in: D. Gao, V. Mahadevan and N. Vasconcelos On the plausibility of the discriminant center-surround hypothesis for visual saliency. Accepted for publication, *Journal of Vision*. The dissertation author was a primary researcher and an author of the cited material.

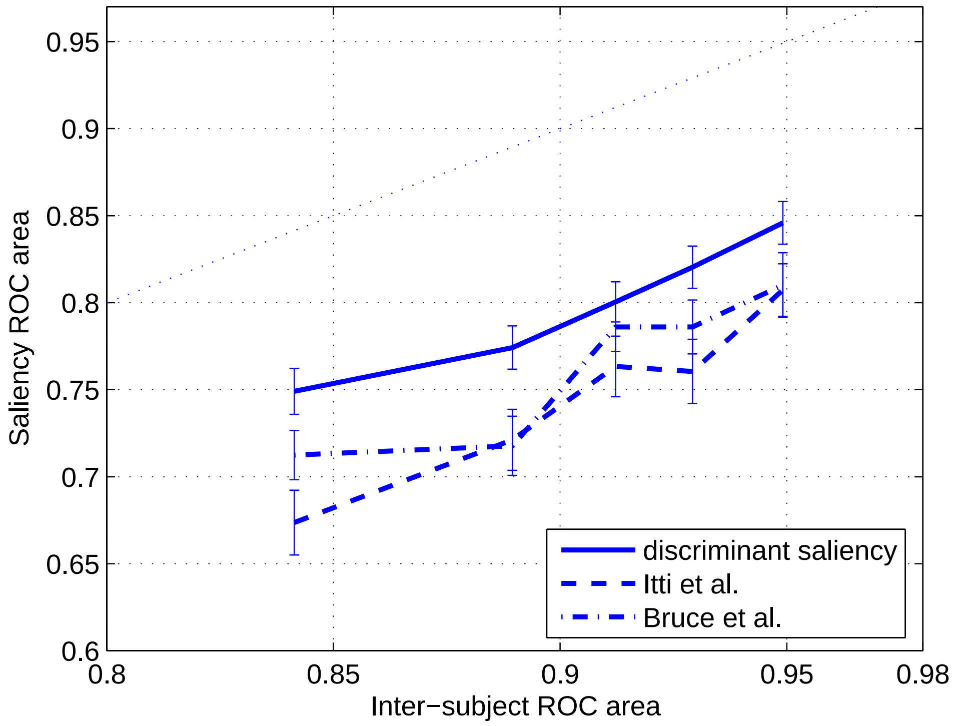


Figure VI.3 Average ROC area, as a function of inter-subject ROC area, for the saliency algorithms discussed in the text.

Chapter VII

Bayesian integration of top-down and bottom-up saliency mechanisms

In Chapter V, we briefly discussed the trade-off between top-down and bottom-up saliency detection. In this chapter, we investigate this issue in more detail. We note that, in the study of biological vision, although there has been psychophysical evidence that the bottom-up (BU) and top-down (TD) attention mechanisms can operate simultaneously, and, for a given scene, the deployment of attention is determined by an interaction of the two modes, its underlying neural mechanisms are not yet clear [21, 228, 33, 76, 36, 219, 201]. For this reason, in the following, we focus our discussions only on computer vision applications, and particularly, object recognition.

As we have mentioned before, for computer vision, both the BU and the TD strategies have their advantages and limitations. BU routines can be made mathematically optimal with respect to universally desirable properties for saliency detection. For example, the popular Harris [68] and Förstner [60] interest point detectors are optimal saliency detectors under a generic cost functional that equates saliency with *repeatability, or invariance to geometric image transformations, of salient points* [172]. BU saliency also tends to be free from computationally intensive training requirements and can usually be implemented with very low complexity. On the other hand, due to the absence of a task-driven focus, BU routines *can only* be optimal in very generic senses, and the resulting salient points are rarely the best for specific applications, such as object recognition. While this illustrates the importance of task-specificity, there could also be clear inconvenience in the adoption of purely TD principles. In particular, because the implementation of these principles usually requires some form of learning from examples, their performance can be sensitive to factors such as insufficient amounts of training data, or training set noise. The latter is a major liability for applications involving cluttered imagery, where one of the main attentional goals is exactly to separate the signal (e.g. objects of interest) from the noise (e.g. background clutter). When the noise level is significant, it may be simply impossible to obtain accurate saliency estimates, and TD mechanisms can, at best, behave as coarse *focus of attention* mechanisms. Combining these with stimulus-driven (BU) saliency (e.g. the detection of corners or contours) could lead to more localized, and therefore accurate, saliency judgements.

There is, nevertheless, a poor understanding on how to combine these TD models with those used for BU saliency in computer vision. The prevalent solution is to either ignore the latter [213, 66] or simply use it as a *pre-filter* of image locations to be processed by TD routines (e.g., [56, 46, 170], see also Chapter V). Both of these strategies are somewhat problematic. Ignoring BU saliency assumes that it is possible to accurately design all saliency stages under task-specific goals. While the recent success in areas such as face detection shows that this is possible when certain conditions are met, e.g. availability of clean training sets and tolerance to large training complexity, there is little evidence that it can be done when such conditions do not hold. Reducing BU saliency to a pre-filter for TD saliency can be a solution to the problem of computational complexity, but could be otherwise problematic. In general, the optimality criteria that guide the design of BU mechanisms are completely unrelated to the task-dependent definitions of TD saliency and it is, therefore, not uncommon for BU pre-processors to summarily eliminate image information highly relevant for TD saliency [66]. Intuitively, the importance of BU saliency should be larger when TD estimates are not accurate, than when they are. This advises the adoption of strategies that integrate saliency information derived from the two saliency modes, rather than hard decisions based on BU saliency. Ideally, it should even be possible to control the relative contribution of the two components.

This is the problem that we address in this chapter, where we 1) introduce a probabilistic formulation of saliency, and 2) argue for the adoption of Bayesian inference principles for the *integration of BU and TD saliency estimates*. The proposed Bayesian formulation is shown to have various interesting properties. First, it produces intuitive rules for the integration of the two saliency modes. Second, it supports the interpretation of TD saliency as a focus-of-attention mechanism which suppresses BU salient points that are not relevant for the task of interest. Third, it provides evidence that BU saliency has an important role when TD routines are inaccurate (e.g. because they are learned from cluttered examples), but is not necessarily useful when the opposite holds. Fourth, it enables explicit control of the relative weight of each saliency component in the final saliency estimates. Finally, it has a non-Bayesian interpretation as the simple multiplication of the two saliency maps, that enables a non-parametric extension of trivial computational complexity. The advantages of the Bayesian solution, over both TD and BU saliency in isolation, are illustrated in the context of recognition problems, both in terms of improved recognition rates and the ability to localize and segment objects from background clutter.

VII.A Bayesian integration

We start from the view of perception as a problem of Bayesian inference [105], under which saliency is naturally formulated as a problem where an observer tries to infer the location of salient scene features, from potentially noisy visual observations. For this, the observer relies on mid-level vision routines that combine information from low-level stages of the visual system (BU mechanisms) with feedback from the higher-level areas (TD mechanisms). BU saliency detectors produce *task-independent* estimates of saliency location which are *well localized* (reduced uncertainty) but *not necessarily relevant for achieving particular goals*. For example, a contour-based detector, that localizes with equally great accuracy the outline of a face, a boulder, or a soccer ball. While, in the absence of high-level feedback, the visual system will respond equally to all these stimuli, when goals become available (e.g. the observer decides to look for faces but not boulders), TD mechanisms are activated to modulate these responses. They produce *goal-driven* saliency estimates which have *greater selectivity* for the image regions that are *relevant for the task at hand* than those produced by bottom-up mechanisms.

If, in addition to *selective*, TD mechanisms were also *accurate* (e.g. capable of localizing the outline of faces with great accuracy while being completely non-responsive to boulders or soccer-balls), there would probably be no need for BU mechanisms. In practice, however, a number of reasons may make this impossible: there may be a limited amount of time or computation available for training TD routines (in order to guarantee a plastic visual system), or the training data may not be clean enough to enable highly accurate estimates (e.g. training is based on cluttered examples). In such situations, it would seem logical for TD learning to maximize *selectivity* (impossible to achieve with BU mechanisms), e.g. by producing routines capable of coarsely identifying image regions containing faces but not accurate enough to precisely outline their contours. The resulting saliency estimates could then be combined with those produced by BU mechanisms to achieve the desired combination of *selectivity and accuracy*.

This process is illustrated in Figure VII.1. The figure depicts (a) an image from the Caltech database [56], and the associated saliency maps produced by two saliency detectors, a BU Harris-Laplace detector [127], and a TD discriminant saliency detector (see Chapter II and Chapter V)[1]. Note how the BU saliency map is very accurate (highly localized responses) but not selective for the face (responds strongly to a large number of corners in the background), while the TD saliency maps are very selective for the face but less accurate. Note also how the TD detector trained with carefully cropped examples is significantly more accurate than that trained with cluttered images. While the former is not likely to benefit greatly from the combination with the BU saliency map (it is accurate enough by itself), this combination tremendously improves the accuracy of the latter, as can be seen from (e). In this case, TD saliency becomes more of a *focus of attention* mechanism that suppresses the spurious responses of BU saliency while emphasizing the responses which fall inside the object of interest (in the example of

[1] Note that we adopt these two detectors for this, and all following, experiments in this chapter. The choice of the detectors is mainly due to their simplicity and the fact that software for their implementations are publicly available. We do not claim that this is necessarily the best combination, and the Bayesian formulation proposed in this work is in no way restricted to them.

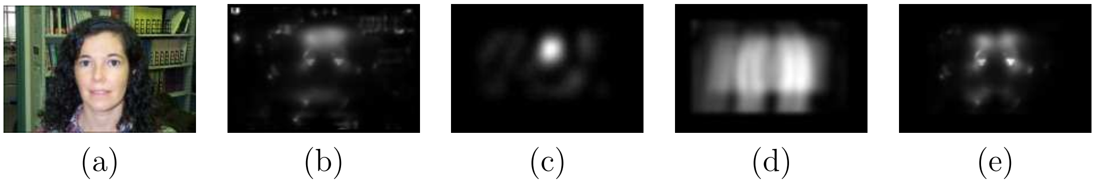


Figure VII.1 Illustration of non-parametric Bayesian saliency. (a) input image, and saliency maps produced by (b) Harris-Laplace [127], (c) the TD discriminant saliency detector when trained with cropped faces, (d) the TD discriminant saliency detector when trained with cluttered images of faces (images such as (a)), and (e) the combination of (b) and (d) with the method of section VII.B.5.

the figure, the net effect is to declare the eyes as the most salient image locations).

Given that there is always a degree of uncertainty associated with saliency estimates, it seems natural to rely on a probabilistic formalism for the combination of BU and TD saliency. Under this formalism, instead of *salient locations*, saliency routines produce *probability distributions of saliency location over the image plane*. The greater accuracy of BU mechanisms translates into distributions that decay more quickly from their peaks (e.g. a mixture of a large number of components of very small variance), while the greater selectivity of TD routines originates a greater concentration of the probability mass (a mixture of a few components of sizeable variance). Faced with a static scene[2], e.g. a picture containing several people in front of a rocky formation, the visual system starts by resorting to BU mechanisms to produce a *prior distribution* for salient locations, e.g. one that assigns high probability to the contours of both faces in the foreground and boulders in the background. As the observer establishes goals for saliency, e.g. looking for faces, TD mechanisms produce a saliency distribution which is combined with the BU prior through the principles of Bayesian inference. The resulting posterior distribution combines the accuracy of the prior with the selectivity of the TD estimates, e.g. by assigning high probability to contours of faces but not those of boulders. If the observer refines the goals, e.g. looking for a particular person, TD

[2] While the formalism could be extended to moving scenes, we only address the static case in this work.

mechanisms react by producing a distribution of smaller entropy, e.g. concentrated around that person's face. This distribution is then combined with the current posterior as is usual in sequential Bayesian inference, e.g. methods commonly used for visual tracking [83, 101, 35], to produce a new posterior distribution that assigns a high probability to the outline of the face of interest and a low probability to the rest of the image.

VII.B Bayesian saliency model

In this section, we introduce a concrete model for the implementation of the Bayesian formulation discussed above. We start by outlining the main features of the model, and then discuss the derivation of the posterior solution. The case where both TD and BU saliency maps have a single salient point is considered first, followed by the more general situation of multiple BU and a single TD point, and finally the full-generality case where both maps have multiple salient points.

VII.B.1 Model outline

Location uncertainty is encoded by associating a Gaussian distribution (defined over image coordinates) with each salient point. This lends itself to mathematically tractable inference (saliency maps, containing salient locations and their relative saliency strength, are represented as Gauss mixtures) and conforms to the time honored psychophysical metaphor of visual attention as a spotlight (that raises the observer's awareness to portions of the visual field) [169, 154]. Given the mixture distributions associated with the BU and TD components, the posterior distribution for the true, but unknown, salient locations is also a Gauss mixture. An analytical solution is derived for its parameters, which are expressed as closed-form functions of the parameters of the component mixtures. A hyper-parameter is introduced in the prior distribution to control the relative importance of the contributions of BU and TD saliency to the posterior estimates. This enables adaptation of the prior's influence according to the accuracy of the TD estimates. For example, when training is based on cluttered examples, the TD estimates should be considered less accurate and a larger weight given to BU saliency. On the other hand, when training is clutter-free, the prior distribution should be made closer to uniform, making its contribution to the posterior solution much less significant. It is shown that this ability to control the balance between BU and TD saliency estimates enables performance superior to that achievable in the absence of such balance.

VII.B.2 Single salient point

A salient point $\mathbf{s}$ is characterized by three parameters: its saliency strength α, image location $\mathbf{x}$, and scale σ. In this work, it is assumed that both the strength and scale are known[3]. When the application, to the image, of a TD saliency detector results in a salient point $\mathbf{x}^{td}$, of scale σ^{td}, this point is modeled as an observation from a Gaussian random variable $\mathbf{X} = (x, y)$ of covariance $\mathbf{\Sigma} = (\sigma^{td})^2\mathbf{I}$ and centered on the true, but unknown, salient location μ,

$$P_{\mathbf{X}|\mu}(\mathbf{x}^{td}|\mu) = \mathcal{G}(\mathbf{x}^{td}, \mu, (\sigma^{td})^2\mathbf{I}).$$

As is usual in Bayesian inference, the uncertainty about the true location μ is formalized by considering this parameter a random variable and introducing a prior $P_\mu(\mu)$, derived from a BU saliency principle. Assuming that a BU saliency detector produced a salient point $\mathbf{s}^{bu} = (\alpha^{bu}, \mu^{bu}, \sigma^{bu})$, this location prior is also assumed Gaussian

$$P_\mu(\mu) = \mathcal{G}(\mu, \mu^{bu}, (\sigma^{bu})^2\mathbf{I}).$$

The posterior distribution for the true salient location is then

$$P_{\mu|\mathbf{X}}(\mu|\mathbf{x}^{td}) = \mathcal{G}(\mu, \mu^s, (\sigma^s)^2\mathbf{I}), \tag{VII.1}$$

[3] While, in practice, this is not strictly true, there is usually a fair amount of tolerance to errors in these parameters. For example, it is common to simply classify points as salient or non-salient, in which case a measure of saliency strength is not even required. With respect to the scale parameter, it is common practice to consider only a finite set of possible scales. Since the selection of the best among these with small error is usually feasible, the assumption of known scale is a reasonable one.

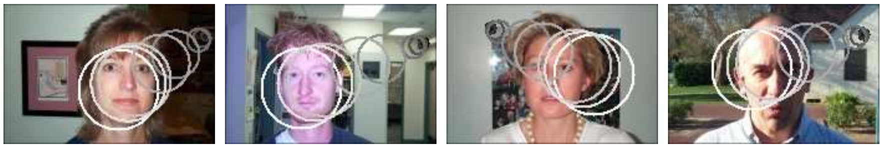


Figure VII.2 The posterior distribution (circle) of the most salient location as a function of the hyper-parameter σ. Brighter circles indicate larger values of σ: in all images the black (white) circle represents the most salient point detected by the BU (TD) detector.

with

$$\mu^s = \frac{(\sigma^{bu})^2}{(\sigma^{bu})^2 + (\sigma^{td})^2}\mathbf{x}^{td} + \frac{(\sigma^{td})^2}{(\sigma^{bu})^2 + (\sigma^{td})^2}\mu^{bu}, \quad (\sigma^s)^2 = \frac{(\sigma^{bu})^2(\sigma^{td})^2}{(\sigma^{bu})^2 + (\sigma^{td})^2}. \quad \text{(VII.2)}$$

The relative importance of the TD and BU saliency maps, can be controlled by multiplying the prior variance by a hyper-parameter σ, i.e. by replacing σ^{bu} with $\sigma \cdot \sigma^{bu}$ in the equations above. Note that, as $\sigma \rightarrow \infty$, $\mu^s = \mathbf{x}^{td}$ and $\sigma^s \rightarrow \sigma^{td}$, making the posterior distribution equal to the Gaussian associated with the TD salient point $\mathbf{s}^{td}$. On the other hand, when $\sigma \rightarrow 0$, $\mu^s = \mu^{bu}$ and $\sigma^s \rightarrow 0$, making the posterior distribution equal to the delta function centered in the location of the BU salient point μ^{bu}. This is illustrated by Figure VII.2 where the most salient point produced by a (BU) Harris-Laplace detector [127] is combined with the most salient point produced by the (TD) discriminant saliency detector of [66]. While, when $\sigma \approx 0$, the posterior is highly localized around the BU point, as σ increases it converges to the distribution resulting from TD saliency.

VII.B.3 Multiple bottom-up salient points

When there are various BU salient points $\{\mathbf{s}^{bu}_1, \ldots, \mathbf{s}^{bu}_n\}$, any of them could be responsible for the observed salient location $\mathbf{x}^{td}$ produced by the TD saliency detector. To account for this we introduce a hidden variable Y, such that $Y = k$ when $\mathbf{s}^{bu}_k$ is the responsible BU salient point, and the following generative model:

1. the k^{th} BU salient point is chosen with probability $P_Y(k) = \alpha^{bu}_k / \sum_j \alpha^{bu}_j$.

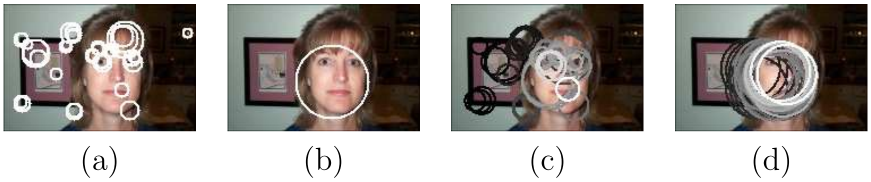


Figure VII.3 Modulation of the focus of attention mechanism, associated with TD saliency, by σ. Images show salient locations detected by (a) Harris-Laplace, (b) discriminant, (c) Bayesian ($\sigma^2 = 6$), and (d) Bayesian ($\sigma^2 = 200$) detectors. Brighter circles indicate stronger saliency.

2. the prior density for location becomes $P_{\mu|Y}(\mu|k) = \mathcal{G}(\mu, \mu_k^{bu}, (\sigma_k^{bu})^2\mathbf{I})$.

3. the observed salient location $\mathbf{x}^{td}$ is sampled from the distribution $P_{\mathbf{X}|\mu}(\mathbf{x}|\mu)$.

Given $\mathbf{x}^{td}$, the posterior for the unknown salient location can be shown to be

$$P_{\mu|\mathbf{X}}(\mu|\mathbf{x}^{td}) = \sum_k \mathcal{G}(\mu, \mu_k^s, (\sigma_k^s)^2\mathbf{I})\pi(\mathbf{x}^{td}, \mathbf{s}_k^{bu}) \tag{VII.3}$$

with

$$\pi(\mathbf{x}^{td}, \mathbf{s}_k^{bu}) = \frac{\mathcal{G}(\mu_k^{bu}, \mathbf{x}^{td}, [(\sigma^{td})^2 + (\sigma_k^{bu})^2]\mathbf{I})\alpha_k^{bu}}{\sum_j \mathcal{G}(\mu_j^{bu}, \mathbf{x}^{td}, [(\sigma^{td})^2 + (\sigma_j^{bu})^2]\mathbf{I})\alpha_j^{bu}},$$

and μ_k^s and σ_k^s as given in (VII.2) with μ^{bu} and σ^{bu} replaced by μ_k^{bu} and σ_k^{bu} respectively.

It is interesting to compare this distribution to that of the case of a single BU salient point: the posterior is now a mixture of Gaussians of the form of (VII.1), each weighted according to the link function $\pi(\mathbf{x}^{td}, \cdot)$. Up to a constant, this is a Gaussian centered on the observed salient location $\mathbf{x}^{td}$ produced by the TD detector, and penalizes the contributions of BU salient points which are located far from this observation. It enables the interpretation of the TD saliency detector as a *focus of attention* operator that suppresses BU salient points which are not discriminant for the object of interest.

As before, the relative importance of the BU and TD saliency maps can be controlled by multiplying all prior variances by a hyper-parameter σ. This can be exploited to modulate the focus of attention mechanism as illustrated in Figure VII.3, where we present the top TD and the 40 top BU salient points for one image, and the posterior distribution for the salient location obtained with two values of σ. Note that, as σ increases, attention is more narrowly focused on the salient points located inside the object of interest, in this case a face.

VII.B.4 Multiple TD and BU salient points

We have, so far, shown that a TD salient point can be interpreted as a focus-of-attention operator that produces a Bayesian estimate of the true, but unknown, salient location $P_{\mu|\mathbf{X}}(\mu|\mathbf{x}^{td})$ of the form of (VII.3). The TD salient point $\mathbf{s}^{td} = (\alpha^{td}, \mathbf{x}^{td}, \sigma^{td})$ associated with $\mathbf{x}^{td}$ can, therefore, be viewed as an *attentional hypothesis* about which image area is most likely to contain discriminant information for the object of interest.

Under this interpretation, a collection of TD salient points $\{\mathbf{s}^{td}_1, \ldots, \mathbf{s}^{td}_m\}$ is nothing more than a set of attentional hypotheses regarding the location of the target visual concept. This suggests the introduction of a second hidden variable Y', such that $Y' = l$ when the l^{th} attentional hypothesis holds, and the following generative model for salient locations:

1. the l^{th} attentional hypothesis is chosen with probability $P_{Y'}(l) = \frac{\alpha^{td}_l}{\sum_j \alpha^{td}_j}$.
2. a salient observation $\mathbf{x}^{td}_l$ is then sampled according to the generative model in the previous section, conditioning all probabilities on the value of Y', i.e.,

$$P_{\mathbf{X}|\mu,Y'}(\mathbf{x}|\mu, l) = \mathcal{G}(\mathbf{x}, \mu, (\sigma^{td}_l)^2\mathbf{I}).$$

Using the fact that BU saliency is independent of the attentional hypothesis, i.e. $P_{\mu|Y,Y'}(\mu|k, l) = P_{\mu|Y}(\mu|k)$ and $P_{Y|Y'}(k|l) = P_Y(k)$, it follows that the posterior for salient location, under the l^{th} attentional hypothesis, is

$$P_{\mu|Y',\mathbf{X}}(\mu|l, \mathbf{x}^{td}_l) = \sum_k \mathcal{G}(\mu, \mu^s_{k,l}, (\sigma^s_{k,l})^2\mathbf{I})\pi_l(\mathbf{x}^{td}_l, \mathbf{s}^{bu}_k)$$

with $\mu^s_{k,l}$ and $\sigma^s_{k,l}$ as given by (VII.2) with μ^{bu}, $\mathbf{x}^{td}$, σ^{bu} and σ^{td} replaced by μ^{bu}_k, $\mathbf{x}^{td}_l$, σ^{bu}_k, and σ^{td}_l respectively, and $\pi_l(\mathbf{x}, \mathbf{s}^{bu}_k) = \frac{\mathcal{G}(\mu^{bu}_k, \mathbf{x}, [(\sigma^{td}_l)^2 + (\sigma^{bu}_k)^2]\mathbf{I})\alpha^{bu}_k}{\sum_j \mathcal{G}(\mu^{bu}_j, \mathbf{x}, [(\sigma^{td}_l)^2 + (\sigma^{bu}_j)^2]\mathbf{I})\alpha^{bu}_j}$. The overall posterior distribution is then

$$P_{\mu|\mathbf{X}}(\mu|\{\mathbf{x}_1^{td},\ldots,\mathbf{x}_m^{td}\}) \;=\; \sum_{k,l}\mathcal{G}(\mu,\mu_{k,l}^s,(\sigma_{k,l}^s)^2\mathbf{I})\beta(\mathbf{x}_l^{td},\mathbf{s}_k^{bu}) \qquad \text{(VII.4)}$$

with

$$\beta(\mathbf{x}_l^{td},\mathbf{s}_k^{bu}) = \frac{\mathcal{G}(\mu_k^{bu},\mathbf{x}_l^{td},[(\sigma_l^{td})^2+(\sigma_k^{bu})^2]\mathbf{I})\alpha_k^{bu}\alpha_l^{td}}{\sum_{i,j}\mathcal{G}(\mu_i^{bu},\mathbf{x}_j^{td},[(\sigma_j^{td})^2+(\sigma_i^{bu})^2]\mathbf{I})\alpha_i^{bu}\alpha_j^{td}}.$$

Note that this is a mixture of posterior distributions of the form of (VII.3), i.e. a mixture of the $n \times m$ Gaussians associated with all pairs of BU and TD salient points. As before, the link function $\beta(\mathbf{x}^{td},\cdot)$ is, up to constants, a Gaussian centered on the observed salient location $\mathbf{x}^{td}$ produced by the TD detector, and penalizes the contributions of BU salient points located far from it. The relative importance of the TD and BU saliency maps can still be controlled by multiplying all prior variances by a hyper-parameter σ.

VII.B.5 Non-parametric interpretation

An interesting low-level interpretation of the posterior distribution (VII.4), that does not require Bayesian inference, can be obtained by noting that, up to constants,

$$\mathcal{G}(\mathbf{x},\mu_k^{bu},(\sigma_k^{bu})^2\mathbf{I})\mathcal{G}(\mathbf{x},\mathbf{x}_l^{td},(\sigma_l^{td})^2\mathbf{I}) = \mathcal{G}(\mathbf{x},\mu_{k,l}^s,(\sigma_{k,l}^s)^2\mathbf{I})\mathcal{G}(\mu_k^{bu},\mathbf{x}_l^{td},[(\sigma_l^{td})^2+(\sigma_k^{bu})^2]\mathbf{I})$$

with $\mu_{k,l}^s$ and $(\sigma_{k,l}^s)^2$ as given above. It follows that the posterior distribution of (VII.4) is the product of the mixtures,

$$\sum_k \frac{\alpha_k^{bu}}{\sum_i \alpha_i^{bu}}\mathcal{G}(\mathbf{x},\mu_k^{bu},(\sigma_k^{bu})^2\mathbf{I}), \quad \text{and} \quad \sum_l \frac{\alpha_l^{td}}{\sum_i \alpha_i^{td}}\mathcal{G}(\mathbf{x},\mathbf{x}_l^{td},(\sigma_l^{td})^2\mathbf{I}),$$

associated with the two saliency detectors, when the true salient locations are μ_k^{bu} and $\mathbf{x}_l^{td}$. Noting that the mixture representation is a probabilistic approximation to the observed saliency maps, this enables a completely non-parametric representation of the posterior for the true salient location as the simple element-wise multiplication of the two saliency maps (plus normalization). This was, in fact, the procedure used to create the Bayesian saliency map of Figure VII.1 (e). What is lost, under this non-parametric interpretation, is the ability to introduce the hyper-parameter σ that modulates the strength of the focus-of-attention mechanism associated with TD saliency.

VII.C Experimental results

To evaluate the performance of Bayesian saliency, we relied on the Caltech database [56]. Four image classes (faces, motorbikes, airplanes, and rear-cars) were used as the classes of interest, and a set of background images was used as the negative class, as proposed in [56]. Two representative saliency detectors, a (BU) Harris-Laplace (HarrLap) detector [127], and the TD discriminant saliency (DiscSal) detector, were selected to implement the Bayesian saliency detector. The sets of salient points produced by the two detectors were first fused into a Bayesian saliency (BayesSal) map according to (VII.4), and the centers of the resulting Gauss mixture were then selected as salient points.

VII.C.1 Salient locations

We start by examining the salient locations detected for different object classes. Figure VII.4 presents some examples of the salient locations produced by the three detectors (locations with saliency strength lower than 50% of the largest are omitted). Note how Bayesian saliency combines the two (BU and TD) saliency components in an intuitive manner: while DiscSal forces HarrLap to focus in the area of the object of interest, the addition of HarrLap improves the accuracy of the TD location estimates by reducing the variance of the Gaussian components. As a side benefit, it also helps “clean up” some of the unstable locations detected by DiscSal (see columns 5-7).

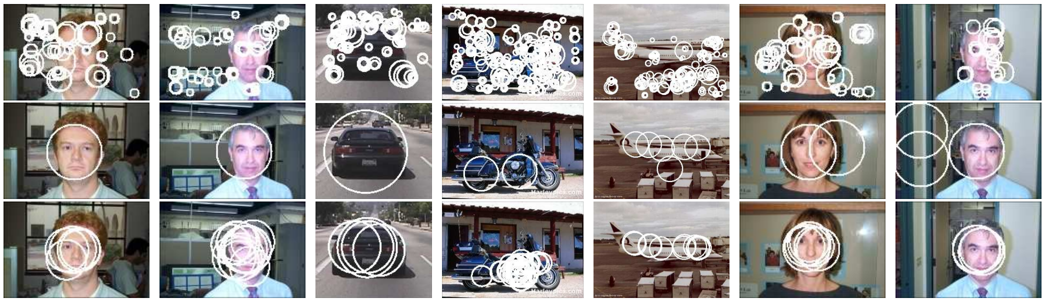


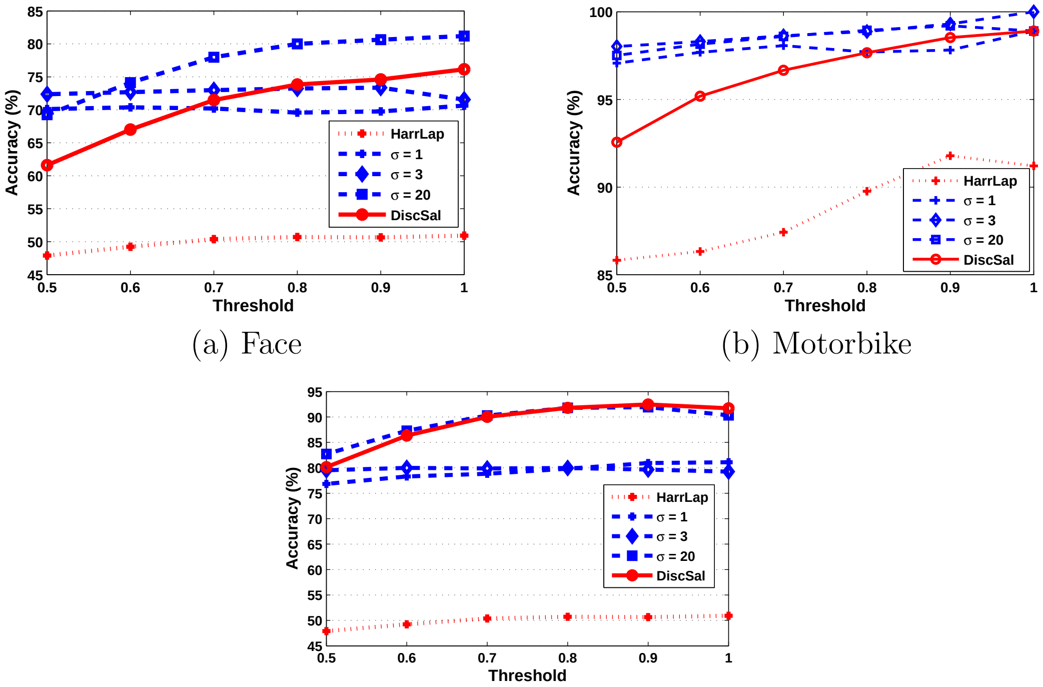
Figure VII.4 Examples of Bayesian saliency. (top) HarrLap, (middle) DiscSal and (bottom) BayesSal.

VII.C.2 Accuracy

To obtain an objective characterization of the accuracy improvements achievable with BayesSal, we designed two experiments. The first measured how well the salient points produced by the three detectors were localized inside the image region covered by the object of interest. The second measured how accurately a segmentation algorithm based on the salient points could identify that image region.

Salient point localization

The set of saliency map locations with saliency strength greater than a threshold (set to $Th_{sal}*$(maximum saliency strength), with $Th_{sal} \in \{0, 0.1, \ldots, 1\}$) was first selected. The number of locations inside the ground truth (a manually produced bounding box of the object) was counted, and *accuracy* was measured by the ratio between the number of locations inside the ground truth and the total number of locations. This measure was then averaged over all images in the test set. Figure VII.5 shows the accuracy achieved, as a function of the threshold Th_{sal}, for faces and motorbikes (results on the other two classes were similar, and are omitted for brevity), with the three saliency detectors (and various values of σ for BayesSal). In (a) and (b) DiscSal was learned from cluttered examples, while cropped faces were used in (c).



(a) Face

(b) Motorbike

(c) Face without clutter

Figure VII.5 Accuracy of salient locations produced by the BayesSal (with various values of σ), DiscSal and HarrLap saliency detectors.

Several interesting observations can be made from the figure. First, HarrLap performed quite poorly, confirming the expectation that BU detectors do not provide much information about the object of interest. Second, in the cases where TD saliency was learned from cluttered examples, (a) and (b), BayesSal achieved the highest accuracy for a large range of values of σ. Note, in particular, the significant improvements (up to 9% absolute points) over DiscSal. Third, BayesSal did not improve over DiscSal when the latter was trained without clutter (c), in fact exhibiting lower accuracy for most values of σ. These two observations support the conclusion that *BU saliency can play an important role in visual saliency, by increasing the accuracy of TD saliency estimates when these cannot be reliably learned, but can also be detrimental, when this is not the case.*

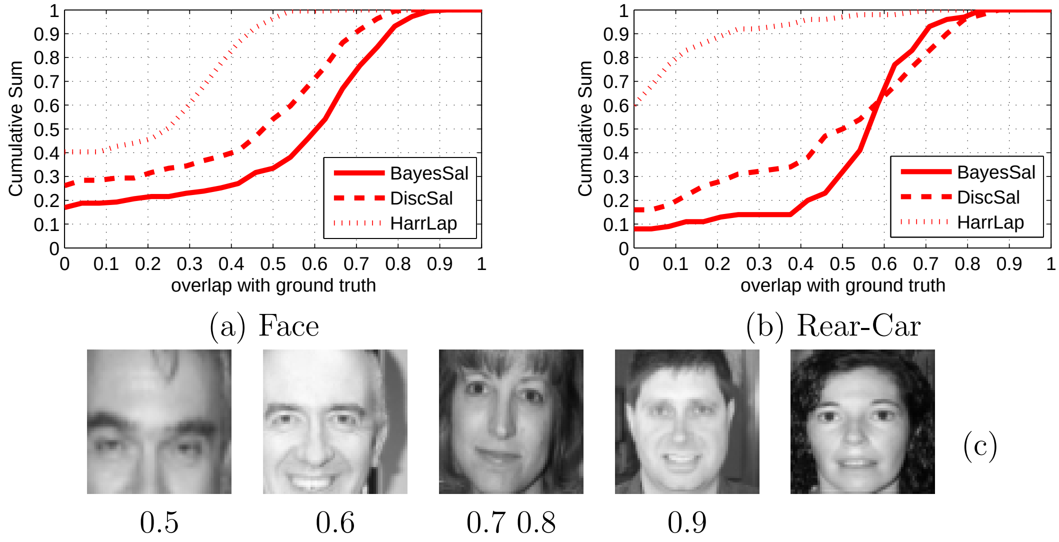
VII.C.3 Segmentation of samples

After showing that the BayesSal achieved better localization performance than the TD saliency alone, we tested its performance in object segmentation experiments. In particular, a variation of the RANSAC algorithm [57] was implemented to align and segment the object of interest from the test images. Image locations were first sampled according to the distribution defined by each saliency map. Locations from pairs of images were then matched, and an affine transformation between them estimated, using RANSAC. All images were then mapped into a common coordinate frame to create an object template. Finally, the matched image locations, which overlapped with the region of support of the template, were segmented from each image. The algorithm was applied to two classes, face and car-rear, and the quality of the segmented examples evaluated by comparing them with manual ground truth. The relative overlap between the segmented example and the ground truth was measured by

$$\text{overlap}(A, B) = \frac{|A \cap B|}{|A \cup B|}, \tag{VII.5}$$

where A, B are two bounding boxes and $|A|$ the area of A. The accuracy of the different saliency detectors was measured by the cumulative distribution function of the relative overlap of the segmented examples produced by them. Ideally, all the examples would have 100% overlap, i.e. the cumulative distribution would be a delta function located at 100%. Figure VII.6 shows cumulative distributions achieved by the three detectors for the two image classes (in this experiment, DiscSal is trained with cluttered images).

It is worth mentioning that, because RANSAC has some ability to reject poor matches, many of the HarrLap salient points that land outside of the object of interest are rejected. The poor performance of HarrLap is, in this case, due to another problem: because the salient points it produces tend to be highly localized, the resulting saliency maps tend to have holes, leaving a significant percentage of the area of the target uncovered. While this is undesirable for the segmentation



(a) Face

(b) Rear-Car

(c)

Figure VII.6 (a, b) Cumulative distribution of overlap between segmented examples and ground truth; (c) Illustrative examples of segmented faces with overlap measures ranging from 0.5 to 0.9.

task considered in these experiments, it could be beneficial for other tasks. In any case, it shows that HarrLap points tend to be highly localized. DiscSal, on the other hand, suffers from the opposite problem. Because training was based on cluttered examples, its saliency estimates are not very accurate and the saliency maps tend to "bleed" beyond the boundaries of the objects of interest. Overall, although the resulting segmentations are not perfect, they are better than those produced by HarrLap. The best results are, however, achieved with BayesSal, which further improves on the DiscSal performance. This is due to the ability of BayesSal to use the highly localized HarrLap estimates as a regularizer for the less accurate DiscSal estimates. In result, BayesSal estimates tend to exhibit less "bleeding" beyond object borders, and produce better segmentations. To provide a sense for the quality of the segmented patches, examples of faces segmented with various values of overlap are also shown in Figure VII.6(c). Figure VII.7 shows segmented faces produced with DiscSal and BayesSal. Note how the faces automatically extracted with the latter tend to cover a much larger region of the

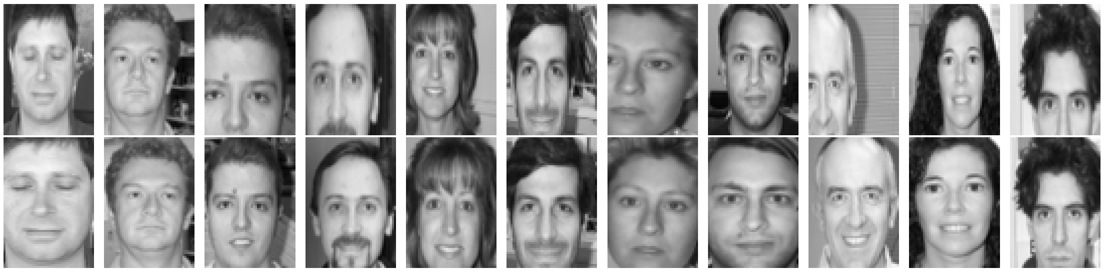


Figure VII.7 Face templates automatically extracted from saliency estimates produced by DiscSal (top) and BayesSal (bottom).

segmented template than those produced by the former.

VII.C.4 Selectivity

While the previous experiments have already shown that BU saliency maps are much less selective than those achievable with TD saliency, we designed a final experiment to exclusively measure selectivity. This experiment consisted of comparing the performance of the different saliency detectors on an object detection task. In particular, we used the simple *SVM-based saliency maps classifier* proposed in Section II.D.2, which consists of feeding a histogram of saliency map intensities to a support vector machine (SVM), and measuring the probability of classification error. The experiment quantifies how relevant the extracted saliency information is for recognition purposes, a measure of how selective the saliency estimates are of the object of interest. The performance of each classifier was measured by 1 minus the receiver-operating characteristic (ROC) equal-error rate (EER), i.e., 1 minus the rate at which the probabilities of false positives and misses are equal. As presented in Table VII.1, BayesSal produced better classification results than the two individual saliency detectors, DiscSal and HarrLap. Note, in particular, how BayesSal explores the selectivity of DiscSal to significantly improve on the prior saliency maps produced by HarrLap. On the other hand the improvements of BayesSal over DiscSal are not stellar. This was expected, since BU salient points have very little selectivity and are only rarely helpful from this

Dataset	BayesSal	DiscSal	HarrLap	constellation [56]
Faces	**98.5**	97.2	61.9	96.4
Motorbikes	**96.5**	96.3	74.8	92.5
Airplanes	**93.9**	93.0	80.2	90.2
Car Rear	**100.0**	100.0	92.7	90.3

Table VII.1 SVM classification accuracy based on different detectors.

point of view. Finally, for completeness, the table also presents the results, on this database, of a state-of-the-art method for recognition from cluttered scenes (the constellation-based classifier of [56]). Despite its simplicity, the saliency-based classifier achieves better recognition rates.

VII.D Acknowledgement

The text of Chapter VII, in full, is based on a co-authored work with N. Vasconcelos. The dissertation author was a primary researcher of this work.

Chapter VIII

Conclusions

The ability of human and other organisms to allocate their limited perceptual and cognitive resources to a few most pertinent subset of sensory data, significantly facilitates learning and survival. While it has long been known that visual attention and saliency mechanisms play a fundamental role in this process, the studies of saliency have been mostly restricted to collecting experimental observations or building heuristic models to replicate the former. There has not been a definition of saliency that could explain the fundamental properties of biological visual saliency. In this thesis, we proposed and studied a novel formulation of saliency, which we denoted as the *discriminant saliency hypothesis*, that all saliency mechanisms are discriminant processes. Our study provided answers to three sets of questions: 1) How does the hypothesis translate into a computational formulation of saliency? What is the optimality of the formulation? how can computational efficiency be achieved? And is the solution applicable to both bottom-up and top-down saliency? 2) Is the discriminant saliency hypothesis biologically plausible? Can it be implemented by the known neural structures in biological visual processing? Can it replicate, both qualitatively and quantitatively, psychophysics of human visual saliency? If so, does it provide any insights or explanations to the neural computations in early visual processing? 3) Does the discriminant saliency hypothesis lead to saliency detectors that benefit problems of interest in computer vision? How do they compare to state-of-the-art saliency detectors?

With respect to the first set of questions, we showed that the hypothesis naturally defines saliency as discriminant feature selection for a classification problem. The optimal solution of this problem is provided by the Bayes decision theory which can be approximated, efficiently and effectively, by the information-theoretic solution, *the maximization of mutual information.* The mutual information solution is consistent with the previous proposals for the organization of perceptual systems, i.e. the *infomax* principle. Resorting to the hypothesis that perception is tuned to the statistical properties of the natural environment, we showed that the discriminant saliency can be implemented in an extremely computationally efficient manner. Besides computational efficiency, the discriminant saliency hypothesis is also suitable for different application domains. In this work, we derived discriminant saliency detectors for both bottom-up and top-down applications by relying on, respectively, *center-surround* and *one-vs-all* assignments of the opposing stimuli in the classification problem.

Regarding the biological plausibility of discriminant saliency, we showed that under the assumptions of natural image statistics, the computation of discriminant saliency is completely consistent with the *standard neural architecture* in the primary visual cortex (V1), i.e. a combination of divisively normalized simple cells and complex cells. We have also applied discriminant saliency to a set of classical displays used in the studies of human saliency behaviors, and showed that discriminant saliency not only explains the qualitative observations (such as pop-out for single feature search, disregard of feature conjunctions, and asymmetries between the existence and absence of a basic feature), but also makes surprisingly accurate quantitative predictions. These include the nonlinear aspects of human saliency perception, the influences of background heterogeneity on percepts of saliency, and the compliance of saliency asymmetries with Weber's law. Such consistency between discriminant saliency and biological saliency not only demonstrates the biological plausibility of the former, but also offers explanations to the latter. For example, it provides a holistic functional justification for the standard architecture of V1: that V1 has the capability to optimally detect salient locations in the visual field, when optimality is defined in a decision-theoretic sense and sensible simplifications are allowed for the sake of computational parsimony. Furthermore, we showed that under a minor extension of the currently prevalent simple cell model, the basic neural structures in V1 are capable of computing the fundamental operations of statistical inference: assessment of probabilities, implementation of decision rules, and feature selection.

Finally, with respect to computer vision applications, we first applied the top-down implementation of discriminant saliency to the problem of weakly supervised learning for object recognition. The detector was shown to outperform the state-of-the-art saliency detectors in computer vision in terms of 1) capturing important information for object recognition tasks, 2) accurately localizing objects of interest in clutter, 3) providing stable salient locations with respect to various geometric and photometric transformations, and 4) adapting to diverse visual attributes for saliency. In the applications where no object recognition is defined, we also showed that the bottom-up discriminant saliency detector accurately predicts human eye fixation locations on natural scenes during a free-viewing process. In another application of discriminant saliency, we introduced a Bayesian framework for the integration of top-down and bottom-up saliency, where the top-down saliency is interpreted as a *focus-of-attention* mechanism. Experimental results showed that this framework combines the selectivity of the top-down saliency with the localization ability of the bottom-up interest point detectors, and improves object recognition performance.

Bibliography

[1] E. Adelson and J. Bergen, "Spatiotemporal energy models for the perception of motion," *Journal of the Optical Society of America A*, vol. 2, no. 2, pp. 284–299, 1985.

[2] J. Allman, F. Miezin, and E. McGuinness, "Stimulus specific responses from beyond the classical receptive field: neurophysiological mechanisms for local-global comparisons in visual neurons." *Annual Review Neuroscience*, vol. 8, pp. 407–430, 1985.

[3] T. Alter and R. Basri, "Extracting salient curves from images: An analysis of the saliency network," *Int'l J. Comp. Vis.*, vol. 27, no. 1, pp. 51–69, 1998.

[4] H. Asada and M. Brady, "The curvature primal sketch," *IEEE Trans. PAMI*, vol. 8, no. 1, pp. 2–14, 1986.

[5] F. Attneave, "Some Informational Aspects of Visual Perception," *Psychological Review*, vol. 61, pp. 183–193, 1954.

[6] H. B. Barlow, "Possible principles underlying the transformation of sensory messages," in *Sensory Communication*, W. A. Rosenblith, Ed. Cambridge, MA: MIT Press, 1961, pp. 217–234.

[7] ——, "Redundancy Reduction Revisited," *Network: Computation in Neural Systems*, vol. 12, pp. 241–253, 2001.

[8] R. Battiti, "Using Mutual Information for Selecting Features in Supervised Neural Net Learning," *IEEE Trans. Neural Networks*, vol. 5, no. 4, pp. 537–550, July 1994.

[9] B. Bauer, P. Jolicoeur, and W. B. Cowan, "Visual search for color targets that are or are not linearly separable fromdistractors," *Visual Research*, vol. 36, pp. 1439–1465, 1996.

[10] J. Beck, "Effect of orientation and of shape similarity on perceptual grouping," *Perception & Psychophysics*, vol. 1, pp. 300–302, 1966.

[11] ——, "Perceptual grouping produced by changes in orientation and shape," *Science*, vol. 154, pp. 538–540, 1966.

[12] ——, “Similarity grouping and peripheral discriminability under uncertainty,” *American Journal of Psychology*, vol. 85, pp. 1–19, 1972.

[13] A. J. Bell and T. J. Sejnowski, “The ‘independent components’ of natural scenes are edge filters,” *Vision Research*, vol. 37, no. 23, pp. 3327–3338, 1997.

[14] J. R. Bergen and E. H. Adelson, “Early vision and texture perception,” *Nature (London)*, vol. 333, pp. 363–364, 1988.

[15] J. R. Bergen and B. Julesz, “Rapid discrimination of visual patterns,” *IEEE Transactions on Systems, Man and Cybernetics*, vol. 19, pp. 857–863, 1983.

[16] K. A. Birney and T. R. Fisher, “On the modeling of dct and subband image data for compression,” *IEEE Transactions on Image Processing*, vol. 4, pp. 186–193, 1995.

[17] A. Bonds, “Role of inhibition in the specification of orientation selectivity of cells in the cat striate cortex,” *Visual Neuroscience*, vol. 2, pp. 41–55, 1989.

[18] B. Bonnlander and A. Weigand, “Selecting input variables using mutual information and nonparametric density estimation,” in *Proc. IEEE International ICSC Symposium on Artificial Neural Networks*, 1994.

[19] E. Borenstein and S. Ullman, “Learn to segment,” in *Proc. European Conference on Computer Vision.* Springer, 2004, pp. 315–328.

[20] C. Bouveyron, J. Kannala, C. Schmid, and S. Girard, “Object localization by subspace clustering of local descriptors,” in *ICVGIP*, 2006.

[21] J. Braun, “Visual search among items of different salience: removal of visual attention mimics a lesion in extrastriate area v4,” *J. Neurosci.*, vol. 14, pp. 554–567, 1994.

[22] M. Bravo and K. Nakayama, “The role of attention in different visual search tasks,” *Perception and Psychophysics*, vol. 51, pp. 465–472, 1992.

[23] P. Brodatz, *Textures: A Photographic Album for Artists and Designers.* Dover, NewYork, 1966.

[24] N. Bruce and J. Tsotsos, “Saliency based on information maximization,” in *Advances in Neural Information Processing Systems 18*, Y. Weiss, B. Schölkopf, and J. Platt, Eds. Cambridge, MA: MIT Press, 2006, pp. 155–162.

[25] R. Buccigrossi and E. Simoncelli, “Image compression via joint statistical characterization in the wavelet domain,” *IEEE Transactions on Image Processing*, vol. 8, pp. 1688–1701, 1999.

[26] M. Carandini, J. Demb, V. Mante, D. Tolhurst, Y. Dan, B. Olshausen, J. Gallant, and N. Rust, "Do we know what the early visual system does?" *Journal of Neuroscience*, vol. 25, pp. 10 577–10 597, 2005.

[27] M. Carandini, D. Heeger, and A. Movshon, "Linearity and normalization in simple cells of the macaque primary visual cortex," *Journal of Neuroscience*, vol. 17, pp. 8621–8644, 1997.

[28] J. Cavanaugh, W. Bair, and J. Movshon, "Nature and interaction of signals from the receptive field center and surround in macaque V1 neurons," *Journal of Neurophysiology*, vol. 88, pp. 2530–2546, 2002.

[29] K. Cave and J. Wolfe, "Modeling the role of parallel processing in visual search," *Cognitive Psychology*, vol. 22, pp. 225–271, 1990.

[30] F. Chance, L. Abbott, and A. Reyes, "Gain modulation from background synaptic input," *Neuron*, vol. 35, pp. 773–782, 2002.

[31] S. G. Chang, B. Yu, and M. Vetterli, "Adaptive wavelet thresholding for image denoising and compression," *IEEE Transactions on Image Processing*, vol. 9, no. 9, pp. 1532–1546, 2000.

[32] O. Chum and A. Zisserman, "An exemplar model for learning object classes," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, 2007, pp. 1–8.

[33] M. M. Chun and J. M. Wolfe, "Visual attention," in *Blackwell Handbook of Perception*, B. Goldstein, Ed. Oxford, UK: Blackwell Publishers Ltd., 2001, pp. 272–310.

[34] R. Clarke, *Transform Coding of Images.* Academic Press, 1985.

[35] D. Comaniciu, V. Ramesh, and P. Meer, "Real-time tracking of non-rigid objects using mean shift," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2000, pp. 142–149.

[36] M. Corbetta, J. M. Kincade, J. M. Ollinger, M. P. McAvoy, and G. L. Shulman, "Voluntary orienting is dissociated from target detection in human posterior cortex," *Nature Neurosci.*, vol. 3, pp. 292–297, 2000.

[37] T. Cover and J. Thomas, *Elements of Information Theory.* New York: John Wiley & Sons Inc., 1991.

[38] J. Daugman, "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters," *Journal of the Optical Society of America A*, vol. 2, no. 7, pp. 1362–1373, 1985.

[39] ——, “Complete discrete 2-d gabor transform by neural networks for image analysis and compression,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, no. 7, pp. 1169–1179, 1988.

[40] R. De Valois, D. Albrecht, and L. Thorell, “Spacial frequency selectivity of cells in macaque visual cortex,” *Vision Research*, vol. 22, pp. 545–559, 1982.

[41] R. De Valois and K. De Valois, *Spacial vision.* New York: Oxford University Press, 1988.

[42] R. L. De Valois, E. W. Yund, and N. Hepler, “The orientation and direction selectivity of cells in macaque visual cortex,” *Vision Research*, vol. 22, pp. 531–544, 1982.

[43] M. N. Do and M. Vetterli, “Wavelet-based texture retrieval using generalized gaussian density and kullback-leibler distance,” *IEEE Transactions on Image Processing*, vol. 11, no. 2, pp. 146–158, 2002.

[44] E. Doi, T. Inui, T.-W. Lee, T. Wachtler, and T. J. Sejnowski, “Spatiochromatic receptive field properties derived from information-theoretic analyses of cone mosaic responses to natural scenes,” *Neural Computation*, vol. 15, no. 2, pp. 397–417, 2003.

[45] B. Doiron, A. Longtin, N. Berman, and L. Maler, “Subtractive and divisive inhibition: Effect of voltage-dependent inhibitory conductances and noise,” *Neural Computation*, vol. 13, pp. 227–248, 2000.

[46] G. Dorko and C. Schmid, “Selection of scale-invariant parts for object class recognition,” in *Proc. IEEE ICCV*, 2003, pp. 634–640.

[47] R. Duda, P. Hart, and D. Stork, *Pattern Classification.* John Wiley & Sons, 2001.

[48] J. Duncan and G. Humphreys, “Visual search and stimulus similarity,” *Psychological Review*, vol. 96, pp. 433–458, 1989.

[49] ——, “Beyond the search surface: visual search and attentional engagement,” *Journal of Experimental Psychology: Human Perception and Performance*, vol. 18, no. 2, pp. 578–588, 1992.

[50] M. D’zmura, “color in visual search,” *vision research*, vol. 31, no. 6, pp. 951–966, 1991.

[51] M. D’Zmura and P. Lennie, “Attentional selection of chromatic mechanisms,” *Investigative Ophthalmology and visual science*, vol. 29, p. 162, 1988.

[52] J. T. Enns and R. A. Rensik, “Preattentive recovery of three-dimensional orientation from line drawings,” *Psychological Review*, vol. 98, no. 3, pp. 335–351, 1991.

[53] C. Enroth-Cugell and J. G. Robson, “The contrast sensitivity of retinal ganglion cells of the cat,” *Journal of Physiology*, vol. 187, pp. 517–522, 1966.

[54] N. Farvardin and J. W. Modestino, “Optimum quantizer performance for a class of non-gaussian memoryless sources,” *IEEE Trans. Information Theory*, vol. 30, no. 3, pp. 485–497, 1984.

[55] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman, “Learning object categories from google”s image search,” in *Proc. IEEE International Conference on Computer Vision (ICCV).* Washington, DC, USA: IEEE Computer Society, 2005, pp. 1816–1823.

[56] R. Fergus, P. Perona, and A. Zisserman, “Object class recognition by unsupervised scale-invariant learning,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2. IEEE Computer Society, 2003, pp. 264–271.

[57] M. A. Fischler and R. C. Bolles, “Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography,” *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[58] J. H. Flowers and D. J. Lohr, “How does familiarity affect visual search for letter strings,” *Perception & Psychophysics*, vol. 37, pp. 557–567, 1985.

[59] I. Fogel and D. Sagi, “Gabor filters as texture discriminator,” *Biol. Cybern.*, vol. 61, pp. 103–113, 1989.

[60] W. Förstner, “A framework for low level feature extraction,” in *Proc. European Conference on Computer Vision.* Springer, 1994, pp. 383–394.

[61] D. H. Foster and P. A. Ward, “Asymmetries in oriented-line detection indicate two orthogonal filters in early vision,” in *Proceedings: Biological Sciences*, vol. 243, 1991, pp. 75–81.

[62] ——, “Horizontal-vertical filters in early vision predict anomalous line-orientation frequencies.” in *Proceedings of the Royal Society London*, ser. B, vol. 243, 1991, pp. 75–81.

[63] ——, “Orientation contrast vs orientation in line-target detection,” *Vision research*, vol. 35, no. 6, pp. 733–738, 1995.

[64] A. Found and H. J. Müller, “Searching for unknown feature targets on more than one dimension:further evidence for a ‘dimension weighting’ account,” *Perception and Psychophysics*, vol. 58, no. 1, pp. 88–101, 1995.

[65] D. Gao and N. Vasconcelos, “Integrated learning of saliency, complex features, and object detectors from cluttered scenes,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition.* IEEE Computer Society, 2005, pp. 282–287.

[66] ——, "Discriminant saliency for visual recognition from cluttered scenes," in *Advances in Neural Information Processing Systems 17*, L. K. Saul, Y. Weiss, and L. Bottou, Eds. Cambridge, MA: MIT Press, 2005, pp. 481–488.

[67] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Advances in Neural Information Processing Systems 19*, B. Schölkopf, J. Platt, and T. Hoffman, Eds. Cambridge, MA: MIT Press, 2007, pp. 545–552.

[68] C. Harris and M. Stephens, "A combined corner and edge detector," in *Alvey Vision Conference*. University of Manchester, Manchester, UK, 1988, pp. 147–151.

[69] K. J. Hawley, W. A. Johnston, and J. M. Farnham, "Novel popout with nonsense string: Effects of object length and spatial predictability," *Perception and Psychophysics*, vol. 55, pp. 261–268, 1994.

[70] D. Heeger and J. Bergen, "Pyramid-based Texture Analysis/Synthesis," in *Proc. ACM SIGGRAPH*, 1995, pp. 229–238.

[71] D. Heeger, "Normalization of cell responses in cat striate cortex," *Visual Neuroscience*, vol. 9, pp. 181–197, 1992.

[72] G. Heidemann, "Focus-of-attention from local color symmetries," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 7, pp. 817–830, 2004.

[73] A. B. Hillel, D. Weinshall, and T. Hertz, "Efficient learning of relational object class models," in *Proc. IEEE International Conference on Computer Vision*. IEEE Computer Society, 2005, pp. 1762–1769.

[74] H.Murase and S. Nayar, "Visual learning and recognition of 3-d objects from appearance," *Int'l J. Comp. Vis.*, vol. 14, pp. 5–24, 1995.

[75] G. Holt and C. Koch, "Shunting inhibition does not have a divisive effect on firing rates," *Neural Computation*, vol. 9, pp. 1001–1013, 1997.

[76] J. B. Hopfinger, M. H. Buonocore, and G. R. Mangun, "The neural mechanisms of top-down attentional control," *Nature Neurosci.*, vol. 3, pp. 284–291, 2000.

[77] R. Horaud, F. Veillon, and T. Skordas, "Finding geometric and relational structures in an image," in *Proc. ECCV*, 1990, pp. 274–384.

[78] P. O. Hoyer and A. Hyvärinen, "Independent component analysis applied to feature extraction from colour and stereo images," *Network: Computation in Neural Systems*, vol. 11, pp. 191–210, 2000.

[79] J. Huang and D. Mumford, “Statistics of Natural Images and Models,” in *Proceedings IEEE Conference on Computer Vision and Pattern Recognition.* IEEE Computer Society, 1999, pp. 541–547.

[80] D. H. Hubel and T. N. Wiesel, “Receptive field, binocular interaction, and functional architecture of in the cat’s visual cortex,” *Journal of Physiology*, vol. 160, pp. 106–154, 1962.

[81] ——, “Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat,” *Journal of Neurophysiology*, vol. 28, pp. 229–289, 1965.

[82] ——, “Receptive field and functional architecture of monkey striate cortex,” *Journal of Physiology*, vol. 195, pp. 215–243, 1968.

[83] M. Isard and A. Blake, “Condensation - conditional density propagation for visual tracking,” *Int’l J. Comp. Vis.*, vol. 29, no. 1, pp. 5–28, 1998.

[84] L. Itti, “Automatic foveation for video compression using a neurobiological model of visual attention,” *IEEE Transactions on Image Processing*, vol. 13, no. 10, pp. 1304–1318, 2004.

[85] L. Itti and P. Baldi, “A principled approach to detecting surprising events in video,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, San Siego, CA, Jun 2005, bu ; cv ; eye ; su, pp. 631–637.

[86] L. Itti and C. Koch, “Computational modeling of visual attention,,” *Nature Rev. Neurosci.*, vol. 2, no. 3, pp. 194–203, March 2001.

[87] ——, “Feature combination strategies for saliency-based visual attention systems,” *Journal of Electronic Imaging*, vol. 10, no. 1, pp. 161–169, 2001.

[88] 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*, vol. 20, no. 11, pp. 1254–1259, 1998.

[89] L. Itti and C. Koch, “A saliency-based search mechanism for overt and covert shifts of visual attention,” *Vision Research*, vol. 40, pp. 1489–1506, 2000.

[90] W. James, *The Principles of Psychology.* Cambridge, MA: Harvard Univ. Press, 1981, Originally published in 1890.

[91] W. A. Johnston, K. J. Hawley, and J. M. Farnham, “Novel popout: Empirical boundaries and tentative theory,” *Journal of Experimental Psychology: Human Perception and Performance*, vol. 19, pp. 140–153, 1993.

[92] J. P. Jones and L. A. Palmer, “An evaluation of the two-dimensional gabor filter model of simple receptive fields in cat striate cortex,” *Journal of Neurophysiology*, vol. 58, pp. 1233–1258, 1987.

[93] ——, "The two-dimensional spatial structure of simple receptive fields in cat striate cortex," *Journal of Neurophysiology*, vol. 58, no. 6, pp. 1187–1211, 1987.

[94] ——, "The two-dimensional spectral structure of simple receptive fields in cat striate cortex," *Journal of Neurophysiology*, vol. 58, no. 6, pp. 1212–1232, 1987.

[95] B. Julesz, "Experiments in the visual perception of texture," *Scientific American*, vol. 232, no. 4, pp. 34–43, 1975.

[96] ——, "A theory of preattentive texture discrimination based on first order statistics of textons," *Biology and Cybernetics*, vol. 41, pp. 131–138, 1981.

[97] ——, "A brief outline of the texton theory of human vision," *Trends in Neuroscience*, vol. 7, pp. 41–45, 1984.

[98] ——, "Texton gradients: the texton theory revisited," *Biological Cybernetics*, vol. 54, pp. 245–251, 1986.

[99] T. Kadir and M. Brady, "Scale, saliency and image description," *International Journal of Computer Vision*, vol. 45, pp. 83–105, 2001.

[100] T. Kadir, A. Zisserman, and M. Brady, "An affine invariant saliency region detector," in *Proc. ECCV*, 2004, pp. 228–241.

[101] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Trans. ASME, J. Basic Eng.*, vol. 82D, pp. 35–45, 1960.

[102] M. K. Kapadia, M. Ito, C. D. Gilbert, and G. Westheimer, "Improvement in visual sensitivity by changes in local context: parallel studies in human observers and in v1 of alert monkeys," *Neuron*, vol. 15, no. 4, pp. 843–856, 1995.

[103] W. Kienzle, F. A. Wichmann, B. Schölkopf, and M. O. Franz, "A nonparametric approach to bottom-up visual saliency," in *Advances in Neural Information Processing Systems 19*, B. Schölkopf, J. Platt, and T. Hoffman, Eds. Cambridge, MA: MIT Press, 2007, pp. 689–696.

[104] J. J. Knierim and D. C. Van Essen, "Neuronal responses to static texture patterns in area V1 of the alert macaque monkey," *Journal of Neurophysiology*, vol. 67, no. 4, pp. 961–980, 1992.

[105] D. C. Knill and W. Richards, *Perception as Bayesian Inference.* OX: Cambridge University Press, 1996.

[106] C. Koch and S. Ullman, "Shift in selective visual attention: towards the underlying neural circuitry," *Human Neurobiology*, vol. 4, pp. 219–227, 1985.

[107] A. Kristjánsson and P. U. Tse, "Curvature discontinuities are cues for rapid shape analysis," *Perception and Psychophysics*, vol. 63, no. 3, pp. 390–403, 2001.

[108] S. W. Kuffler, "Discharge patterns and functional organization of mamalian retina," *Journal of Neurophysiology*, vol. 16, pp. 37–68, 1953.

[109] J. Kulikowski and P. Bishop, "Fourier analysis and spatial representation in the visual cortex," *Experientia*, vol. 37, pp. 160–163, 1981.

[110] M. S. Landy and J. R. Bergen, "Texture segregation and orientation gradient," *Vision Research*, vol. 31, pp. 679–691, 1991.

[111] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2006.

[112] T. S. Lee, "Image representation using 2d gabor wavelets," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 10, pp. 959–971, 1996.

[113] J. Levitt and J. Lund, "Contrast dependence of contextual effects in primate visual cortex," *Nature*, vol. 387, pp. 73–76, 1997.

[114] C. Li and W. Li, "Extensive integration field beyond the classical receptive field of cat'sstriate cortical neurons-classification and tuning properties," *Vision Research*, vol. 34, no. 18, pp. 2337–2355, 1994.

[115] Z. Li, "A saliency map in primary visual cortex," *Trends in Cognitive Sciences*, vol. 6, no. 1, pp. 9–16, 2002.

[116] T. Lindeberg, "Scale-space theory: A basic tool for analyzing structures at different scales," *J. Applied Statistics*, vol. 21, no. 2, pp. 224–270, 1994.

[117] R. Linsker, "Self-organization in a perceptual network," *IEEE Computer*, vol. 21, no. 3, pp. 105–117, 1988.

[118] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. IEEE International Conference on Computer Vision.* IEEE Computer Society, 1999, pp. 1150–1157.

[119] J. Malik and P. Perona, "Preattentive texture discrimination with early vision mechanisms," *Journal of the Optical Society of America A*, vol. 7, no. 5, pp. 923–932, May 1990.

[120] P. Malinowski and R. Hübner, "The effect of familiarity on visual-search performance: Evidence for learned basic features," *Perception and Psychophysics*, vol. 63, no. 3, pp. 458–463, 2001.

[121] V. Maljkovic and K. Nakayama, “Priming of popout: I. role of features,” *Memory & Cognition*, vol. 22, no. 6, pp. 657–672, 1994.

[122] S. G. Mallat, “A theory for multiresolution signal decomposition: The wavelet representation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 674–693, 1989.

[123] B. S. Manjunath and W. Y. Ma, “Texture feature for browsing and retrieval of image data,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 8, pp. 837–842, 1996.

[124] S. Marcelja, “Mathematical description of the responses of simple cortical cells,” *Journal of the Optical Society of America*, vol. 70, pp. 1297–1300, 1980.

[125] J. Matas, O. Chum, M. Urban, and T. Pajdla, “Robust wide-baseline stereo from maximally stable extremal regions,” *Image and Vision Computing*, vol. 22, no. 10, pp. 761–767, September 2004.

[126] K. Mikolajczyk and C. Schmid, “Indexing based on scale invariant interest points,” in *Proc. ICCV*, 2001, pp. 525–531.

[127] ——, “Scale and affine invariant interest point detectors,” *Int’l J. Comp. Vis.*, vol. 60, no. 1, pp. 63–86, 2004.

[128] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. V. Gool, “A comparison of affine region detectors,” *Int’l J. Comp. Vis.*, vol. 65, pp. 43–72, 2005.

[129] J. W. Modestino, “Adaptive nonparametric detection techniques,” in *Nonparametric Methods in Communications*, P. Papantoni-Kazakos and D. Kazakos, Eds. New York: Marcel Dekker, 1977, pp. 29–65.

[130] G. Moraglia, “Display organization and the detection of horizontal line segments,” *Perception and Psychophysics*, vol. 45, no. 3, pp. 265–272, 1989.

[131] I. Motoyoshi and S. Nishida, “Visual response saturation to orientation contrast in the perception of texture boundary,” *Journal of the Optical Society of America A*, vol. 18, no. 9, pp. 2209–2219, 2001.

[132] J. A. Movshon, I. D. Thompson, and D. J. Tolhurst, “Spatial summation in the receptive fields of simple cells in the cat’s striate cortex,” *Journal of Physiology*, vol. 283, pp. 53–77, 1978.

[133] H. J. Müller, D. Heller, and J. Ziegler, “Visual search for singleton feature targets within and across feature dimensions,” *Perception and Psychophysics*, vol. 57, no. 1, pp. 1–17, 1995.

[134] A. L. Nagy and R. R. Sanchez, "Critical color differences determined with a visual search task," *Journal of the Optical Society of America A*, vol. 7, pp. 1209–1217, 1990.

[135] K. Nakayama and G. H. Silverman, "Serial and parallel processing of visual feature conjunctions," *Nature*, vol. 320, pp. 264–265, 1986.

[136] V. Navalpakkam and L. Itti, "An integrated model of top-down and bottom-up attention for optimal object detection," in *Proc. IEEE CVPR*, 2006, pp. 2049–2056.

[137] ——, "Search goal tunes visual features optimally," *Neuron*, vol. 53, no. 4, pp. 605–617, 2007.

[138] H. C. Nothdurft, "The role of local contrast in pop-out of orientation, motion and color," *Investigative Ophthalmology and Visual Science*, vol. 32, no. 4, p. 714, 1991.

[139] ——, "Texture segmentation and pop-out from orientation contrast," *Vision Research*, vol. 31, no. 6, pp. 1073–1078, 1991.

[140] ——, "Feature analysis and the role of similarity in preattentive vision," *Perception and Psychophysics*, vol. 52, no. 4, pp. 355–375, 1992.

[141] ——, "The conspicuousness of orientation and motion contrast," *Spatial Vision*, vol. 7, pp. 341–363, 1993.

[142] ——, "Faces and facial expression do not pop-out," *Perception*, vol. 22, pp. 1287–1298, 1993.

[143] ——, "The role of features in preattentive vision: Comparison of orientation, motion and color cues," *Vision Research*, vol. 33, no. 14, pp. 1937–1958, 1993.

[144] ——, "Salience from feature contrast: variations with texture density," *Vision Research*, vol. 40, pp. 3181–3200, 2000.

[145] I. Ohzawa, G. Sclar, and R. Freeman, "Contrast gain control in the cat's visual system," *Journal of Neurophysiology*, vol. 54, pp. 651–667, 1985.

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

[147] R. K. Olson and F. Attneave, "What variables produce similarity grouping?" *American Journal of Psychology*, vol. 83, pp. 1–21, 1970.

[148] J. Palmer, C. T. Ames, and D. T. Lindsey, "Measuring the effect of attention on simple visual search," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 19, pp. 108–130, 1993.

[149] S. E. Palmer, *Vision Science: Photons to Phenomenology.* The MIT Press, 1999.

[150] P. Parent and S. W. Zucker, "Trace inference, curvature consistency, and curve detection," *IEEE Trans. PAMI*, vol. 11, no. 8, pp. 823–839, 1989.

[151] D. J. Parkhurst, K. Law, and E. Niebur, "Modeling the role of salience in the allocation of overt visual attention," *Vision Research*, vol. 42, no. 1, pp. 107–123, 2002.

[152] H. Pashler, "Target-distractor discriminability in visual search," *Perception & Psychophysics*, vol. 41, pp. 385–392, 1987.

[153] R. Peters, A. Iyer, L. Itti, and C. Koch, "Components of bottom-up gaze allocation in natural images," *Vision Research*, vol. 45, no. 18, pp. 2397–2416, 2005.

[154] M. I. Posner, "Orientation of attention," *Quart. J. Experimental Psychology*, vol. 32, pp. 3–25, 1980.

[155] F. H. Previc and J. L. Blume, "Visual search asymmetries in three-dimensional space," *Vision Research*, vol. 33, no. 18, pp. 2697–704, 1993.

[156] J. Principe, D. Xu, and J. Fisher, "Information-Theoretic Learning," in *Unsupervised Adaptive Filtering, Volume 1: Blind-Source Separation*, S. Haykin, Ed. Wiley, 2000.

[157] C. Privitera and L. Stark, "Algorithms for defining visual regions-of-interest: comparison with eye fixations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 970–982, 2000.

[158] P. Quelhas, F. Monay, J.-M. Odobez, D. Gatica-Perez, and T. Tuytelaars, "A thousand words in a scene," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 9, pp. 1575–1589, 2007.

[159] D. Regan, "Orientation discrimination for bars defined by orientation texture," *Perception*, vol. 24, pp. 1131–1138, 1995.

[160] D. Reisfeld, H. Wolfson, and Y. Yeshurun, "Context-free attentional operators: The generalized symmetry transform," *Intl J. Comp. Vis.*, vol. 14, pp. 119–130, 1995.

[161] I. Rock, "The perception of disoriented figures," *Scientific American*, vol. 230, no. 1, pp. 78–85, 1974.

[162] E. Rosch, "Cognitive reference points," *Cognitive Psychology*, vol. 7, no. 4, pp. 532–547, 1975.

[163] R. Rosenholtz, "A simple saliency model predicts a number of motion popout phenomena," *Vision Research*, vol. 39, pp. 3157–3163, 1999.

[164] ——, "Visual search for orientation among heterogeneous distractors:experimental results and implications for signal detection theory models of search," *J. Experimental Psychology*, vol. 27, no. 4, pp. 985–999, 2001.

[165] ——, "Search asymmetries? what search asymmetries?" *Perception and Psychophysics*, vol. 63, no. 3, pp. 476–489, 2001.

[166] C. S. Royden, J. M. Wolfe, and N. Klempen, "Visual search asymmetries in motion and optic flow fields," *Perception and Psychophysics*, vol. 63, no. 3, pp. 436–444, 2001.

[167] D. Sagi, "The psychophysics of texture segmentation," in *Early Vision and Beyond*, T. Papathomas, Ed. MIT Press, 1996.

[168] D. Sagi and B. Julesz, ""where" and "what" in vision," *Science*, vol. 228, pp. 1217–1219, 1985.

[169] ——, "Enhanced detection in the aperture of focal attention during simple shape discrimination tasks," *Nature*, vol. 321, pp. 693–695, 1986.

[170] B. Schiele and J. Crowley, "Where to look next and what to look for," in *Intelligent Robots and Systems (IROS).* World Scientific, 1996, pp. 1249–1255.

[171] C. Schmid and R. Mohr, "Local grayvalue invariants for image retrieval," *IEEE Trans. PAMI*, vol. 19, no. 5, pp. 530–534, 1997.

[172] C. Schmid, R. Mohr, and C. Bauckhage, "Comparing and evaluating interest points," in *Proc. ICCV.* IEEE Computer Society Press, January 1998. [Online]. Available: http://perception.inrialpes.fr/Publications/1998/SMB98

[173] O. Schwartz and E. Simoncelli, "Natural signal statistics and sensory gain control," *Nature Neuroscience*, vol. 4, pp. 819–825, 2001.

[174] N. Sebe and M. S. Lew, "Comparing salient point detectors," *Pattern Recognition Letters*, vol. 24, no. 1-3, pp. 89–96, Jan. 2003.

[175] F. Sengpiel, A. Sen, and C. Blakemore, "Characteristics of surround inhibition in cat area 17," *Experimental Brain Research*, vol. 116, pp. 216–228, 1997.

[176] T. Serre, L. Wolf, and T. Poggio, "Object recognition with feature inspired by visual context," in *Proc. IEEE Conf. CVPR*, 2005, pp. 994–1000.

[177] A. Sha'ashua and S. Ullman, "Structural saliency: the detection of globally salient structures using a locally connected network," in *Proc. IEEE International Conference on Computer Vision*, 1988, pp. 321–327.

[178] C. E. Shannon, “A mathematical theory of communication,” *The Bell system technical journal*, vol. 27, pp. ”379–423, 623–656”, 1948.

[179] K. Sharifi and A. Leon-Garcia, “Estimation of shape parameter for generalized gaussian distributions in subband decompositions of video,” *IEEE Transactions on Circuits Syst. Video Technol.*, vol. 5, no. 1, pp. 52–56, 1995.

[180] J. Shen and E. M. Reingold, “Visual search asymmetry: The influence of stimulus familiarity and low-level features,” *Perception and Psychophysics*, vol. 63, no. 3, pp. 464–475, 2001.

[181] J. Shi and C. Tomasi, “Good features to track,” in *Proc. IEEE Conf. CVPR*, 1994, pp. 593–600.

[182] F. Shic and B. Scassellati, “A behavioral analysis of computational models of visual attention,” *Journal International Journal of Computer Vision*, vol. 73, pp. 159–177, 2007.

[183] A. M. Sillito, K. L. Grieve, H. E. Jones, J. Cudeiro, and J. Davis, “Visual cortical mechanisms detecting focal orientation discontinuities,” *Nature*, vol. 378, pp. 492–496, 1995.

[184] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman, “Discovering objects and their localization in images,” in *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2005, pp. 370–377.

[185] B. C. Skottun, A. Bradley, G. Sclar, I. Ohzawa, and R. S. Freeman, “The effects of contrast on visual orientation and spacial frequency discrimination: A comparison of single cell and behavior,” *Journal of Neurophysiology*, vol. 57, no. 3, pp. 773–786, 1987.

[186] B. Skottun, R. D. Valois, D. Grosof, J. Movshon, D. Albrecht, and A. Bonds, “Classifying simple and complex cells on the basis of response modulation,” *Vision Research*, vol. 31, pp. 1079–1086, 1991.

[187] A. Srivastava, A. Lee, E. Simoncelli, and S. Zhu, “On advances in statistical modeling of natural images,” *Journal of Mathematical Imaging and Vision*, vol. 18, pp. 17–33, 2003.

[188] A. Sutter, J. Beck, and N. Graham, “Contrast and spatial variables in texture segregation: Testing a simple spatial-frequency channels model,” *Percept. Psychophys*, vol. 46, pp. 312–332, 1989.

[189] B. W. Tatler, R. J. Baddeley, and I. D. Gilchrist, “Visual correlates of fixation selection: effects of scale and time,” *Vision Research*, vol. 45, pp. 643–659, 2005.

[190] A. Treisman, “Preattentive processing in vision,” *Computer vision, Graphics, & Image Processing*, vol. 31, pp. 156–177, 1985.

[191] ——, "Features and objects: The fourteenth bartlett memorial lecture," *Quarterly Journal of Experimental Psychology*, vol. 40A, no. 2, pp. 201–237, 1988.

[192] ——, "Search, similarity, and integration of features between and within dimensions," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 17, no. 3, pp. 652–676, 1991.

[193] ——, "Spreading suppression or feature integration? a reply to duncan and humphreys (1992)," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 18, no. 2, pp. 589–593, 1992.

[194] ——, "The perception of features and objects," in *Attention: Selection, awareness, and control*, A. Baddeley and L. Weiskrantz, Eds. Oxford: Clarendon Press, 1993, pp. 5–35.

[195] A. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive Psychology*, vol. 12, no. 1, pp. 97–136, 1980.

[196] A. Treisman and S. Gormican, "Feature analysis in early vision: Evidence from search asymmetries," *Psychological Review*, vol. 95, pp. 15–48, 1988.

[197] A. Treisman and S. Sato, "Conjunction search revisited," *Journal of Experimental Perception and Performance*, vol. 16, pp. 459–478, 1990.

[198] A. Treisman and J. Souther, "Search asymmetry: A diagnostic for preattentive processing of separable features," *Journal of Experimental Psychology: General*, vol. 114, pp. 285–310, 1985.

[199] S. Treue, "Visual attention: the where, what, how and why of saliency," *Current Opinion in Neurobiology*, vol. 13, pp. 428–432, 2003.

[200] B. Triggs, "Detecting keypoints with stable position, orientation, and scale under illumination changes." in *Proc. ECCV*, 2004, pp. 100–113.

[201] J. K. Tsotsos, S. M. Culhane, W. Y. K. Winky, Y. Lai, N. Davis, and F. Nuflo, "Modeling visual attention via selective tuning," *Artif. Intell.*, vol. 78, no. 1-2, pp. 507–545, 1995.

[202] A. Tversky, "Features of similarity," *Psychological Review*, vol. 84, p. 327C352, 1977.

[203] J. H. van Hateren and D. L. Ruderman, "Independent component analysis of natural image sequences yields spatiotemporal filters similar to simple cells in primary visual cortex," in *Proc. Royal Society ser. B*, vol. 265, 1998, pp. 2315–2320.

[204] J. H. van Hateren and A. van der Schaaf, "Independent component filters of natural images compared with simple cells in primary visual cortex," in *Proc. Royal Society ser. B*, vol. 265, 1998, pp. 359–366.

[205] V. N. Vapnik, *The Nature of Statistical Learning Theory.* NY: Springer-Verlag, 1995.

[206] M. Vasconcelos and N. Vasconcelos, "Natural image statistics and low complexity feature selection," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, In press.

[207] N. Vasconcelos, "Feature selection by maximum marginal diversity," in *Advances in Neural Information Processing Systems 15*, S. T. S. Becker and K. Obermayer, Eds. Cambridge, MA: MIT Press, 2003, pp. 1351–1358.

[208] N. Vasconcelos and G. Carneiro, "What is the role of independence for visual regognition?" in *Proc. ECCV*, Copenhagen, Denmark, 2002.

[209] N. Vasconcelos and M. Vasconcelos, "Scalable discriminant feature selection for image retrieval and recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, 2004, pp. 770–775.

[210] P. Verghese, "Visual search and attention: A signal detection theory approach," *Neuron*, vol. 31, pp. 523–535, 2001.

[211] P. Verghese and K. Nakayama, "Stimulus discriminability in visual search," *Vision Research*, vol. 34, no. 18, pp. 2453–2467, 1994.

[212] M. Vidal-Naquet and S. Ullman, "Object recognition with informative features and linear classification," in *Proc. ICCV*, Nice, France, 2003.

[213] P. Viola and M. Jones, "Robust real-time object detection," in 2^{nd} *Int. Workshop on Statistical and Computational Theories of Vision Modeling, Learning, Computing and Sampling*, July 2001.

[214] K. Walker, T. Cootes, and C. Taylor, "Locating salient object features," in *Proc. British Machine Vision Conference.* British Machine Vision Association, 1998, pp. 557–566.

[215] D. Walther and C. Koch, "Modeling attention to salient proto-objects," *Neural Networks*, vol. 19, pp. 1395–1407, 2006.

[216] Q. Wang, P. Cavanagh, and M. Green, "Familiarity and pop-out in visual search," *Perception and Psychophysics*, vol. 56, no. 5, pp. 495–500, 1994.

[217] M. Webster and R. De Valois, "Relationships between spatial frequency and orientation tuning of striate cortex cells," *Journal of the Optical Society of America A.*, vol. 2, no. 7, pp. 1124–1132, 1985.

[218] L. Williams and D. Jacobs, "Stochastic completion fields: a neural model of illusory contour shape and salience," in *Proc. IEEE ICCV*, 1995, pp. 408–415.

[219] J. M. Wolfe, "Guided search 2.0: A revised model of visual search," *Psychonomic Bulletin & Review*, vol. 1, no. 2, pp. 202–238, 1994.

[220] ——, "Guided search 4.0: Current progress with a model of visual search," in *Integrated models of cognitive systems*, W. D. Gray, Ed. New York: Oxford University Press, 2007, pp. 99–119.

[221] J. M. Wolfe, K. R. Cave, and S. L. Franzel, "Guided search: An alternative to the feature integration model for visual search," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 15, pp. 419–433, 1989.

[222] J. M. Wolfe, "Visual search," in *Attention*, H. Pashler, Ed. UK: Psychology Press, 1998, pp. 13–74.

[223] ——, "Asymmetries in visual search: An introduction," *Perception & Psychophysics*, vol. 63, no. 3, pp. 381–389, 2001.

[224] J. M. Wolfe, S. R. Friedman-Hill, M. I. Stewart, and K. M. O'Connell, "The role of categorization in visual search for orientation," *Journal of Experimental Psychology: Human Perception & Performance*, vol. 18, pp. 34–49, 1992.

[225] J. M. Wolfe and T. Horowitz, "What attributes guide the deployment of visual attention and how do they do it?" *Nature Reviews Neuroscience*, vol. 5, pp. 495–501, 2004.

[226] K. Yamada and G. W. Cottrell, "A model of scan paths applied to face recognition," in *Proceedings of the Seventeenth Annual Cognitive Science Conference.* Pittsburgh, PA: Mahwah: Lawrence Erlbaum, 1995, pp. 55–60.

[227] H. Yang and J. Moody, "Data Visualization and Feature Selection: New Algorithms for Nongaussian Data," in *Proc. NIPS*, Denver, USA, 2000.

[228] S. Yantis, "Control of visual attention," in *Attention*, H. Pashler, Ed. East Sussex, UK: Psychology Press, 1998, pp. 223–256.

[229] A. Yarbus, *Eye movements and vision.* New York: Plenum, 1967.

[230] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture andobject categories: A comprehensive study," *International Journal of Computer Vision*, vol. 73, no. 2, pp. 213–238, 2007.

[231] L. Zhang, M. H. Tong, and G. W. Cottrell, “Information attracts attention: A probabilistic account of the cross-race advantage in visual search,” in *Proceedings of the 29th Annual Cognitive Science Conference*. Nashville, Tennessee.: Mahwah: Lawrence Erlbaum, 2007, pp. 749–754.

[232] S. Zhu, Y. Wu, and D. Mumford, “Filters, Random field And Maximum Entropy (FRAME): Towards a Unified Theory for Texture Modeling,” *International Journal of Computer Vision*, vol. 27, no. 2, pp. 107–126, 1998.