Computational Modeling and Psychophysics in Low- and Mid-Level Vision

Thesis by

Xiaodi Hou

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

2014

(Defended May 7, 2014)

To my family.

Acknowledgements

I would like to express the deepest gratitude to my *Doctor-Father*, Professor Christof Koch. In my 6-year marathon in the PhD program, Christof has spent great effort guiding me to become a true scientist. With his trust, patience, encouragement, and support, I was able to explore interesting topics as free as Siegfried, from Baltimore harbor to Santa Monica beach, from border-ownership cells to compositional models. Outside the lab, I am greatly indebted to Christof for the sincere help that he offered during the hard times of my life. I also fondly remember our never-ending discussions about tastes of shoe colors, some particular IT company, and the tarantula in Titus Canyon.

Since 2011, I am fortunate enough to be co-supervised by Professor Alan Yuille from UCLA. I enjoyed every single meeting with this British gentleman, to get exposed to his encyclopedic knowledge. More fantastically, Alan has always kept his office door open to me. I would like to thank Alan for his mini math lectures that have greatly helped my research; for his generous financial support during the last quarter of my PhD program; and for every meal he had skipped due to our prolonged discussions.

I am also grateful to other members of my thesis committee, Professor Shinsuke Shimojo and Professor Pietro Perona, who have constantly been providing insightful suggestions for my thesis research.

It is my privilege to have Jonathan Harel, Yin Li, Liwei Wang and Bolei Zhou as my collaborators. I would like to thank them for accepting the invitations to my crazy projects, in which we have shared both ecstacies and disappointments. The days we fought back-to-back are invaluable memories to me.

I would also like to thank my friends: Jonathan Harel, Lingsen Meng, Virgil Griffith, Piercesare Grimaldi, Ueli Rutishauser, Jiannis Taxidis, Costas Anastassiou, Nathan Faivre, Erik Schomburg, Adam Shai, Michael Hill, Yazan Billeh, Julien Dubois, Leonard Mlodinow, Zhongzheng Fu, Shengxuan Ye, Xun Wang, Peng He, Minghong Lin, Qi Zhao, Song Cao, Xiaochen Lian, Xuan Dai

and Wutu Lin. Like Californian sunshine, they made my stay at Caltech a heartwarming experience.

My gratitude also goes to Professor Brian Brophy, for offering me amazing opportunities to the stage play *Pasadena Babalon*, and the movie *PhD Comics*; and Professor Ken Pickar, for introducing me to the world of entrepreneurship, which eventually defines my career path.

Finally, I wish to thank my parents, for the unconditional love that they gave, and for the sacrifice that they have taken to foster my scientific worldview throughout the years. I wish to thank my wife Amanda Song for her full-hearted support to my spiritual goals. The delicious breakfast she prepares for me is always my best energy supply after a sleepless night of work.

Abstract

This thesis addresses a series of topics related to the question of how people find the foreground objects from complex scenes. With both computer vision modeling, as well as psychophysical analyses, we explore the computational principles for low- and mid-level vision.

We first explore the computational methods of generating saliency maps from images and image sequences. We propose an extremely fast algorithm called Image Signature that detects the locations in the image that attract human eye gazes. With a series of experimental validations based on human behavioral data collected from various psychophysical experiments, we conclude that the Image Signature and its spatial-temporal extension, the Phase Discrepancy, are among the most accurate algorithms for saliency detection under various conditions.

In the second part, we bridge the gap between fixation prediction and salient object segmentation with two efforts. First, we propose a new dataset that contains both fixation and object segmentation information. By simultaneously presenting the two types of human data in the same dataset, we are able to analyze their intrinsic connection, as well as understanding the drawbacks of the most popular but inappropriately labeled salient object segmentation dataset. Second, we also propose an algorithm of salient object segmentation. Based on our novel discoveries on the connections of fixation data and salient object segmentation data, our model significantly outperforms all existing models on all 3 datasets with large margins.

In the third part of the thesis, we discuss topics around the human factors of boundary analysis. Closely related to salient object segmentation, boundary analysis focuses on delimiting the local contours of an object. We identify the potential pitfalls of algorithm evaluation for the problem of boundary detection. Our analysis indicates that today's popular boundary detection datasets contain significant level of noise, which may severely influence the benchmarking results. To give further insights on the labeling process, we propose a model to characterize the principles of the human factors during the labeling process. The analyses reported in this thesis offer new perspectives to a series of interrelating issues in low- and mid-level vision. It gives warning signs to some of todays

"standard" procedures, while proposing new directions to encourage future research.

Contents

Chapter 1

Introduction

Coming along with the growing amount of image data is the pressing demand for algorithms that analyze these visual information from raw pixels. One aim of computer vision is to inspect an image and present a concise overview of the objects in the scene and their interactions. It requires the visual system to 1) localize the objects in the visual field, and 2) separate the objects from their background content. Known as the figure-ground separation problem, this challenge has been a perennial topic of discussion in computer vision, cognitive science, and neuroscience.

Given an image that contains one or several figures, the raw sensory information has to be perceptually grouped into clusters of foreground regions called *figures*, according to Gestalt psychology, and the background called *ground*. After perceptual organization, the visual information is further passed down to the high-level visual processing stream to fulfill functions such as object detection and face recognition. In computer vision, this process is related to a series of topics such as fixation prediction, salient object detection, and boundary detection. In psychophysics, figure-ground organization is a vital component that connects low-level sensations to high level cognitions. For my doctorate study, I have explored the computational models as well as the psychophysics around the topic of figure-ground separation in natural and synthetic images.

The major distinction of the figure-ground problem (e.g. against high-level vision) is the ill-defined nature of the ground-truth. For object-level tasks such as face identification or object recognition, the ground-truth comes from the physical properties of the visual world. This type of ground-truth - either the name of the person or the category of the object, are unambiguous, and often easy to obtain. In contrast, the ground-truth of the figure-ground separation is the intermediate representation hidden under the flow of consciousness. There is no physical reality with which we can define the absolute correct answer for a figure-ground problem. The ground-truth are often composed of behavioral measures from human subjects, such as eye fixations, or hand drawing

labels.

There are two types of inconsistencies in the ground-truth of a figure-ground problem: the inter-subject difference, and the intra-subject difference. First, data collected from different subjects may not always come to a consensus. For instance, Fig. 1.1 gives three proposals of the figure segmentation, and each one of them seems plausible. Moreover, different forms of behavioral data capture different aspects of the internal representation. Even though the vision community believes that both eye fixations and hand labels are related to the figure-ground problem, they cannot be used interchangeably due to many differences between these two forms of ground-truths.

Figure 1.1: In one of the lab meetings, Christof showed his new tattoo, which is from Ramon y Cajal's original drawing of the rodent neocortex. The neocortex tatoo, Christof's arm, and Christof himself can all be considered as the figure of the image. This example illustrates the ill-defined nature of the figure-ground problem.

1.1 My strategy

For primates, finding foods or predators rapidly is essential for their survival. Figure-ground separation is evolved as a computationally-efficient mechanism to cope with such incessant flow of visual information. In computer vision, figure-ground modules such as visual saliency are often used as a filter to find perceptually-meaningful structures from massive, disorganized visual data, and give real-time feedbacks. Therefore, run-time speed is a critical consideration for the algorithm design. Careful approximations are often introduced to solve this NP-hard problem. With every approximation also comes counter-examples – for instance, there might not be any efficient algorithm, even for humans, to find the proverbial needle in the haystack. At the early stage of the vision hierarchy, an algorithm should prioritize the efficiency at common scenarios over the accuracy in complex but rare scenes.

Psychophysical data plays a particular role in our analysis of the figure-ground problem. Different from today's common treatments in computer vision (e.g. averaging the data over all subjects), we perform careful analyses of the results of these behavioral experiments. The extra bits of information that we have gained could help us better understand the computational mechanisms of vision system in the following ways:

Defining the problem There are many examples in psychophysics, in which the human visual system fails to parse the information from surprisingly simple patterns. These "counter-examples" of vision provide valuable insights to an algorithm designer. It is critical to understand in what condition, under what kinds of limitation, and with which part of the visual information, humans are able to separate objects from their background in the scene.

Probing the limit of a dataset Like every other experiment, the psychophysical experiment of constructing the ground-truth of a dataset is noisy. The bias and variance of the ground-truth data determine the limits of benchmarking capability. As the competition among today's algorithms often goes to 3 digits after the decimal, it is urgent to make sure that the datasets are capable of measuring such tiny difference.

Proposing new directions The ground-truth data can be much more informative than just ranking algorithms based on their benchmark scores. By fully utilizing the rich descriptions from the psychophysical data, we can split the easy and hard sub-problems and propose new directions for future research.

1.2 Thesis Outline

The thesis is composed of several projects that I explored during my graduate study, which can be roughly clustered into the following three topics:

1.2.1 Computational modeling of visual saliency

We first explore the computational methods of generating saliency maps from images and image sequences. In Chapter 2, *Saliency Detection using Image Signature*, we designed an extremely fast algorithm called *Image Signature* [1] that detects the locations in the image that attract human eye gazes. In contrast to earlier models [2, 3, 4, 5] that either design or learn (using standard machine learning techniques) the properties of a salient feature, Image Signature exploits holistic properties

of the non-salient background. We formulate the figure-ground separation problem as a signal decomposition problem: $\mathbf{x} = \mathbf{f} + \mathbf{g}$, where $\mathbf{x}$ is the input image, $\mathbf{f}$ is the figure that is *spatially* sparse, and $\mathbf{g}$ is the ground, which has a sparse representation in the Discrete Cosine Transformed domain (*spectrally* sparse). Instead of pursuing the exact decomposition of $\mathbf{x}$, we further show that this decomposition depends implicitly on the binary quantization of the spectral coefficients of the Discrete Cosine Transform of $\mathbf{x}$. In other words, we can extract, in one line of MATLAB code, a surprisingly simple "bar-code" for each image, and use it to represent the spatial organization of the saliency of the image.

To detect the figures in video clips in Chapter 3, we propose the *Phase Discrepancy* model [6], which is a spatial-temporal extension to the original Image Signature. With additional theoretic analysis and experiments on natural image datasets, we show that the Phase Discrepancy model is an effective way of compensating camera self-motion as well as capturing the moving objects. In particular, the algorithm does not rely on prior training on particular features or categories of an image, and can be implemented in 9 lines of MATLAB code.

1.2.2 From fixations to salient objects

Fixation maps are the probabilistic distributions on the image plane. They are often blurry, and do not have clear boundaries of the objects. Therefore, for many computer vision applications, this representation is not as useful as object masks, which not only preserve the location but also the contour of the detected objects. In recent years, a new topic called salient object segmentation is getting the attention from the computer vision community. Despite its goal of attacking the same figure-ground separation problem, the salient object segmentation sub-field is evolved independently from the previous research in fixation prediction. It creates a discomforting segregation of the two research topics.

Another issue is the quality of the ground-truth. Hundreds of algorithms are benchmarked primarily on one dataset with only one labeler to annotate the images. There are no quantitative measures of the quality of such a ground-truth label.

In Chapter 4, we bridge the gap between fixation prediction and salient object segmentation with two efforts. First, we propose a new dataset [7] that contains both fixation and object segmentation information. By simultaneously presenting the two types of human data in the same dataset, we are able to analyze their intrinsic connection, as well as understanding the drawbacks of today's "standard" but inappropriately labeled salient object segmentation dataset.

Second, we also propose an algorithm of salient object segmentation. Based on our novel discoveries on the connections of fixation data and salient object segmentation data, we can effectively make use of the results of existing fixation prediction algorithms. Our model significantly outperforms *all* existing models on *all* 3 datasets with large margins.

1.2.3 Psychophysical analyses of boundary labeling and benchmarking

In the third part of the thesis, we discuss topics around the human factors of boundary analysis. Closely related to salient object segmentation, boundary analysis focuses on delimiting the local contours of an object. Defining the boundaries in the image is not an easy task. Previous approaches usually employ many labelers to draw boundaries on every image. Compared with the active research on finding a better boundary detector to refresh the performance record, there is surprisingly little discussion on the boundary detection benchmark itself. In Chapter 5, we address the issues in the benchmarking procedure in a popular dataset of boundary detection. The goal of this work [8] is to identify the potential pitfalls during algorithm evaluation. First, with a novel psychophysical experiment, we realize that certain types of human labels, called the *orphan labels*, are not capable of measuring the performance of today's algorithms. Second, we discover the issue of *precision bubble* in the benchmarking procedure that amplifies the effect of the orphan labels. Third, with a series of analyses, we conclude that the *strong boundaries* where all human labelers agree is a better-defined sub-problem than "detecting all boundaries in the image." However, when facing this new challenge, none of today's major algorithms are capable of detecting such boundaries better than a random method.

To give further insights on the labeling process of many labelers annotating the same image with different preferences (hidden variables), in Chapter 6, we characterize the computational principles of the human factors. Following the hypothesis of perceptual hierarchy of the image, we use a partially ordered set to unify multiple labels of the same image, and infer the label strengths based on their ordering of perceptual importance. This approach does not only give a clean-up view of the scene, but also tells of the relative relationship among labelers. It allows the model to detect defective labelers as well as unreliable labels that receive contradictory outputs from different labelers.

Chapter 2

Saliency Detection using Image Signature

Abstract

We introduce a simple image descriptor referred to as the *Image Signature*. We show, within the theoretical framework of sparse signal mixing, that this quantity spatially approximates the foreground of an image. We experimentally investigate whether this approximate foreground overlaps with visually conspicuous image locations by developing a saliency algorithm based on the Image Signature.

This saliency algorithm is capable of reproducing the famous search-asymmetry phenomenon on synthetic images. it also accurately predicts human fixation points on many datasets such as the Bruce and Tsotsos [3] benchmark dataset, and does so in much shorter running time. In a related experiment, we demonstrate with a *change-blindness* dataset that the distance between images induced by the Image Signature is closer to human perceptual distance than can be achieved using other saliency algorithms, pixel-wise or GIST [9] descriptor methods. Finally, we introduce an experiment of face-perspective analysis to further illustrative the perceptual properties that are captured by Image Signature.

2.1 Introduction

The problem of finding all objects in a scene and separating them from the background is known as *figure-ground separation*. The brain can perform this separation very quickly [10], and doing so on a machine remains a major challenge for engineers and scientists. The problem is closely related to many of the core applications of machine vision, including scene understanding, content-based

image retrieval, object recognition, and tracking. In this chapter, we provide an approach to the figure-ground separation problem using a binary, holistic image descriptor called the "Image Signature." It is defined as the sign function of the Discrete Cosine Transform (DCT) of an image. As we shall demonstrate, this simple descriptor preferentially contains information about the foreground of an image – a property that we believe underlies the usefulness of this descriptor for detecting salient image regions.

In Section 2.2, we formulate the figure-ground separation problem in the framework of sparse signal analysis. We prove that the Inverse Discrete Cosine Transform (IDCT) of the Image Signature concentrates the image energy at the locations of a spatially-sparse foreground, relative to a spectrally-sparse background. Then, in Section 2.3, we validate the theoretical limits of Image Signature with a series of experiments on synthetic images.

In Section 2.4, four experiments are presented to quantify the performance of Image Signature in various tasks. First, we use search asymmetry as an illustrative example to show that the Image Signature algorithm faithfully captures the characteristics of human vision behavior. Second, we demonstrate that the saliency maps derived from the Image Signature outperform many leading saliency algorithms on five datasets of eye-movement fixation points. In the same section, we also introduce the reaction-time data collected from nine subjects in a change-blindness experiment. We show that the distance between images induced by the Image Signature most closely matches the perceptual distance between images inferred from these data among competing measures derived from other saliency algorithms, the GIST descriptor, and simpler pixel measures. We conclude the section with a demonstration of the head-orientation-prediction experiment. The binary codes of Image Signature cluster faces with same head orientation with high accuracy, outperforming the best known face orientation detector on the FERET dataset [11].

2.1.1 Related work

Holistic image processing short-circuits the need for segmentation, key-point matching, and other local operations. Bolstered by growing general interest in large-scale image datasets such as 80-million tiny images [12], and ImageNet [13], the early-stage holistic descriptors have become a topic of intense study in computer vision. GIST [9] is an excellent example of such algorithms in this field. Other holistic scene models focus on the separation of foreground and background. For example, Candès et al. [14] introduced a sparse matrix factorization model.

The Image Signature discards amplitude information across the entire frequency spectrum,

storing only the sign of each DCT component, equivalent to the phase spectrum of a Fourier decomposition. The study of spectral properties of natural images dates back to 1980. In [15] and [16], the authors show that important visual information is stored in the phase spectrum of the image. More recently, statistical models of natural images [17, 18] also argued that natural images have distinctive spectra. Research from computer graphics [19] often categorizes the spectral components of an image as the low-frequency component, i.e. roughly the smoothed copy of the original image, and the high-frequency component, i.e. the residual between the original image and the low-frequency component. Since 2007, saliency detection based on spectral analysis of images has produced amazing, yet not fully-explained results. The first among the family of *spectral saliency* algorithms is Hou's *Spectral Residual*. In [20], they used the residual Fourier amplitude spectrum combined with the original phase spectrum to construct the saliency map of an image. This method gives surprisingly good results; however, its original theoretical explanation is flawed. Later studies point out that one can achieve similar saliency maps by discarding the residual amplitude spectrum, and only using the phase information. Since the initial discovery, follow-up studies have made several attempts to explain the computational mechanism of spectral saliency: [21, 22, 23, 24, 25]. In Section 2.3, we will use artificially-generated images to challenge the validity of these alternative theories, and test the necessity and sufficiency of the theory behind Image Signature.

2.2 Image Signature

$\hat{\mathbf{x}}$	$\text{DCT}(\mathbf{x})$.
$\text{sign}(\mathbf{x})$	The entrywise sign operator.
$\bar{\mathbf{x}}$	$\text{IDCT}\big[\text{sign}(\hat{\mathbf{x}})\big]$, the reconstructed image.
$T_{\mathbf{x}}$	Support set of $\mathbf{x}$.
$\Omega_{\mathbf{x}}$	Support set of $\hat{\mathbf{x}}$.
$\lvert x\rvert$	The absolute value of a real number x.
$\lvert\mathcal{S}\rvert$	The cardinality of a set $\mathcal{S}$.
$\lVert\mathbf{x}\rVert_p$	The ℓ^p norm of vector $\mathbf{x}$ ($p = 2$ if omitted).
$\langle\mathbf{x}, \mathbf{y}\rangle$	The inner-product of $\mathbf{x}$ and $\mathbf{y}$.
$E(X)$	The expectation of random variable X.
$\circ$	The Hadamard (entrywise) product operator.
$*$	The convolution operator.

Table 2.1: Important notation and terms used in this chapter.

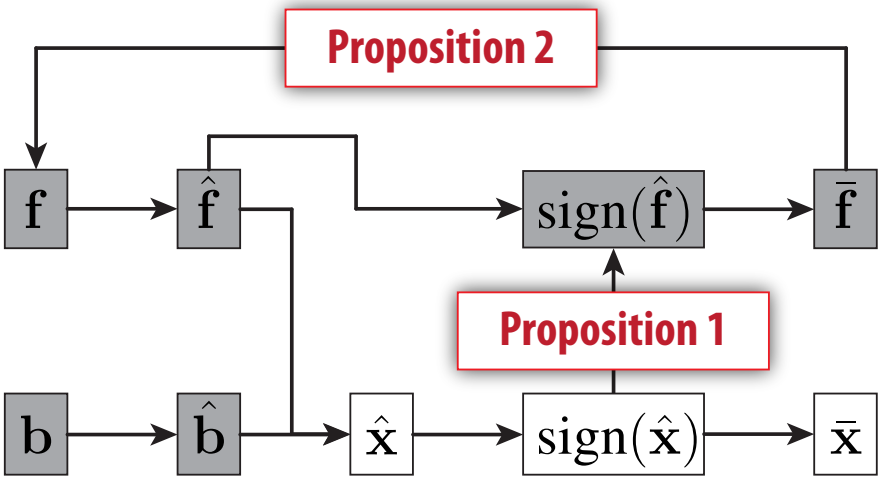


Figure 2.1: An illustration of the structure of the proof. Hidden variables are shown in gray boxes. The goal of Image Signature is to represent the spatial support of $\mathbf{f}$ with $\bar{\mathbf{x}}$. The two propositions can be interpreted as following: Prop. 1 proves that $\text{sign}(\hat{\mathbf{f}})$ can be approximated by $\text{sign}(\hat{\mathbf{x}})$. Prop. 2 proves that $\mathbf{f}$ can be approxmiated by $\bar{\mathbf{f}}$.

We begin by considering gray-scale images that exhibit the following structure:

$$\mathbf{x} = \mathbf{f} + \mathbf{b}, \quad \mathbf{x}, \mathbf{f}, \mathbf{b} \in \mathbb{R}^N. \tag{2.1}$$

$\mathbf{f}$ represents the foreground or figure signal, and is assumed to be sparsely supported in the standard spatial basis. $\mathbf{b}$ represents the background, and is assumed to be sparsely supported in the basis of the Discrete Cosine Transform (DCT). In other words, both $\mathbf{f}$ and $\hat{\mathbf{b}}$ have only a small number of non-zero components. Table 2.1 lists the important definitions used throughout the rest of this section.

Performing the exact separation between $\mathbf{b}$ and $\mathbf{f}$ given only $\mathbf{x}$ and the fact of their sparseness is, in general, very difficult. For the problem of figure-ground separation, we are only interested in the spatial support of $\mathbf{f}$ (the set of pixels for which $\mathbf{f}$ is non-zero). In this chapter, we show first analytically, then, empirically, that given an image that can be decomposed as Eq. 2.1, we can approximately isolate the support of $\mathbf{f}$ by taking the sign of the mixture signal $\mathbf{x}$ in the transformed domain, and then inversely-transform it back into the spatial domain by computing the reconstructed image $\bar{\mathbf{x}} = \text{IDCT}\big[\text{sign}(\hat{\mathbf{x}})\big]$. Formally, the image signature is defined as:

$$\text{ImageSignature}(\mathbf{x}) = \text{sign}(\text{DCT}(\mathbf{x})), \tag{2.2}$$

If we assume that an image foreground is visually conspicuous relative to its background, then we can form a *saliency map* $\mathbf{m}$ (see [2] for classic use) by smoothing the squared reconstructed image defined above:

$$\mathbf{m} = g * (\bar{\mathbf{x}} \circ \bar{\mathbf{x}}), \tag{2.3}$$

where g is a Gaussian kernel. Our experiments in Section 2.4 show that a simple Gaussian smoothing is necessary here, because the support $T_{\mathbf{f}}$ of a salient object is usually not only spatially sparse, but also localized in a contiguous region.

We also define a distance metric D between images $\mathbf{x}^1$ and $\mathbf{x}^2$ based on the ℓ^0 distance between image signatures (viz., the Hamming distance):

$$D(\mathbf{x}^1, \mathbf{x}^2) = \|\text{sign}(\hat{\mathbf{x}}^1) - \text{sign}(\hat{\mathbf{x}}^2)\|_0 \tag{2.4}$$

Building on the idea that the image signature preferentially contains foreground information, this subtraction compares the sparse foreground information in two images without explicitly first computing either $\mathbf{b}$ or $\mathbf{f}$. Later, we provide empirical evidence for the utility of this metric.

2.2.1 Image Signature: Foreground Properties

In this section, we provide evidence that, for an image that adheres to a certain mathematical structure, the image signature can be used to approximately obtain the location of the foreground.

Proposition 1 (Signature suppresses background) *The image reconstructed from the image signature approximates the location of a sufficiently sparse foreground on a sufficiently sparse background as follows:*

$$E\Big(\frac{\langle \bar{\mathbf{f}}, \bar{\mathbf{x}} \rangle}{\|\bar{\mathbf{f}}\| \cdot \|\bar{\mathbf{x}}\|}\Big) \geq 0.5, \quad \textit{for } |\Omega_{\mathbf{b}}| < \frac{N}{6}\,. \tag{2.5}$$

Proof 1 *Our proof is based on the Uniform Uncertainty Principle (UUP) proposed by Candès et al. [26]. Let Θ be a subset of $\{1, 2, \ldots, N\}$ of size $|\Theta|$. UUP states that if $\mathbf{f}$ is sufficiently spatially sparse, that is, if:*

$$|T_{\mathbf{f}}| \leq \alpha |\Theta| / \lambda\,, \tag{2.6}$$

where λ is the over-sampling factor, and α is a sufficiently small constant, then with an overwhelm-

ing probability, the energy of $\hat{\mathbf{f}}$ *supported on* Θ *is bounded:*

$$\frac{|\Theta|}{2N}\|\mathbf{f}\| \leq \|\hat{\mathbf{f}} \circ \mathbf{1}_\Theta\| \leq 3\frac{|\Theta|}{2N}\|\mathbf{f}\|\,, \tag{2.7}$$

where $\mathbf{1}_\Theta$ *is the vector with zeros at component indices not in* Θ *and ones at component indices in* Θ*.*

The over-sampling factor depends on the choice of transform. In [27], Rudelson et al. show that for Fourier transform, $\lambda = O(\log^5 N)$*. Because of the similarities between DCT and DFT, and that images are real-valued, this factor is the same for the DCT. In fact, one can construct a signal* $\mathbf{x}' \in \mathbb{R}^{4N}$ *from the original* $\mathbf{x} \in \mathbb{R}^N$ *as following:*

$$\begin{aligned} x'_{2n} &= x_n & x'_{2n-1} &= 0 \\ x'_{4N-2n+2} &= x_{N-n+1} & x'_{4N-2n+1} &= 0\,, \end{aligned}$$

such that DFT$(\mathbf{x})$ *exactly equals DCT*$(\mathbf{x}')$*.*

According to Plancherel's theorem, we have $\|\mathbf{f}\| = \|\hat{\mathbf{f}} \circ \mathbf{1}_{\Omega_\mathbf{f}}\|$*. Then the following inequality can be derived from UUP (Ineq. 2.7):*

$$\begin{aligned} & 3\frac{|\Omega_\mathbf{f}|}{2N}\|\mathbf{f}\| \geq \|\hat{\mathbf{f}} \circ \mathbf{1}_{\Omega_\mathbf{f}}\| \\ & 3\frac{|\Omega_\mathbf{f}|}{2N} \geq 1 \\ & |\Omega_\mathbf{f}| \geq \frac{2}{3}N. \end{aligned} \tag{2.8}$$

Recall that Ineq. 2.8 holds with overwhelming probability only if $\mathbf{f}$*, the foreground is sufficiently spatially sparse in the sense of Ineq. 2.6.*

From this, we estimate $\langle \bar{\mathbf{f}}, \bar{\mathbf{x}} \rangle$*:*

$$\begin{aligned} \langle \bar{\mathbf{f}}, \bar{\mathbf{x}} \rangle &= \big\langle \mathit{IDCT}[\mathit{sign}(\hat{\mathbf{f}})], \mathit{IDCT}[\mathit{sign}(\hat{\mathbf{x}})] \big\rangle \\ &= \mathit{IDCT}\big[\big\langle \mathit{sign}(\hat{\mathbf{f}}), \mathit{sign}(\hat{\mathbf{x}}) \big\rangle\big] \\ &= \big\langle \mathit{sign}(\hat{\mathbf{f}}), \mathit{sign}(\hat{\mathbf{x}}) \big\rangle \\ &= \sum_{i \in \Omega_\mathbf{b}} \mathit{sign}(\hat{f}_i) \cdot \mathit{sign}(\hat{x}_i) \\ &+ \sum_{j \notin \Omega_\mathbf{b}} \mathit{sign}(\hat{f}_j) \cdot \mathit{sign}(\hat{x}_j)\,. \end{aligned} \tag{2.9}$$

Since $\mathbf{f}$ *and* $\mathbf{b}$ *are independent from each other, we assume:*

$$P(sign(\hat{f}_i) = sign(\hat{b}_i) | i \in \Omega_{\mathbf{b}}) = 0.5\,,$$

where $P(\cdot)$ *stands for probability. Then:*

$$\begin{aligned} & P(sign(\hat{f}_i) = sign(\hat{x}_i) | i \in \Omega_{\mathbf{b}}) \\ = \; & P(sign(\hat{f}_i) = sign(\hat{b}_i) | i \in \Omega_{\mathbf{b}}) \\ & + P(|\hat{f}_i| > |\hat{b}_i|, sign(\hat{f}_i) \neq sign(\hat{b}_i) | i \in \Omega_{\mathbf{b}}) \geq 0.5\,. \end{aligned}$$

Therefore:

$$E\Big[\sum_{i \in \Omega_{\mathbf{b}}} sign(\hat{f}_i) \cdot sign(\hat{x}_i) \Big] \geq 0\,. \tag{2.10}$$

Since $sign(\hat{b}_j) = 0$*, for* $j \notin \Omega_{\mathbf{b}}$*, we have:*

$$\sum_{j \notin \Omega_{\mathbf{b}}} sign(\hat{f}_j) \cdot sign(\hat{x}_j) = \sum_{j \notin \Omega_{\mathbf{b}}} sign(\hat{f}_j)^2 \geq |\Omega_{\mathbf{f}}| - |\Omega_{\mathbf{b}}|\,. \tag{2.11}$$

Combining Ineq. 2.10, Ineq. 2.11, and Eq. 2.9, we have:

$$E\Big(\frac{\langle \bar{\mathbf{f}}, \bar{\mathbf{x}} \rangle}{\|\bar{\mathbf{f}}\| \cdot \|\bar{\mathbf{x}}\|} \Big) \geq \frac{|\Omega_{\mathbf{f}}| - |\Omega_{\mathbf{b}}|}{\sqrt{|\Omega_{\mathbf{f}}| \cdot |\Omega_{\mathbf{x}}|}} \geq \frac{|\Omega_{\mathbf{f}}| - |\Omega_{\mathbf{b}}|}{N}\,. \tag{2.12}$$

Given the bound provided by Ineq. 2.8,

$$E\Big(\frac{\langle \bar{\mathbf{f}}, \bar{\mathbf{x}} \rangle}{\|\bar{\mathbf{f}}\| \cdot \|\bar{\mathbf{x}}\|} \Big) \geq \frac{2}{3} - \frac{1}{N} |\Omega_{\mathbf{b}}| \geq 0.5\,.$$

if we assume that the background $\mathbf{b}$ *is sufficiently sparse:* $|\Omega_{\mathbf{b}}| < N/6$*.*

An important note is that Proposition 1 does not depend on the relative energies of the foreground and background, $\|\mathbf{f}\|$ and $\|\mathbf{b}\|$, only their sparseness. This will later be demonstrated empirically in Section 2.3 on synthetic data.

Proposition 1 does not establish a direct relationship between the reconstructed image $\bar{\mathbf{x}}$ and the actual foreground $\mathbf{f}$; instead, the relationship is to a function of the foreground, $\bar{\mathbf{f}}$. We will now show that $\bar{\mathbf{f}}$ contains important information about the spatial support of the foreground, $T_{\mathbf{f}}$.

Proposition 2 *For a foreground signal* $\mathbf{f}$ *with non-zero elements independently drawn from the unit Gaussian distribution, over* 79% *of* $\bar{\mathbf{f}}$ *is expected to be contained in the support of the foreground* $T_{\mathbf{f}}$*. Namely,*

$$\begin{aligned} E\Big(\frac{\alpha}{\sqrt{\alpha^2+\beta^2}}\Big) &\geq \sqrt{\frac{2}{\pi}} \approx 0.7979 \text{ where} \\ \alpha &= \sqrt{\sum_{i\in T_{\mathbf{f}}} \bar{f}_i^2}, \text{ and} \\ \beta &= \sqrt{\sum_{j\notin T_{\mathbf{f}}} \bar{f}_j^2}. \end{aligned}$$

Proof 2 *First, we quantify the expected correlation between* $\hat{\mathbf{f}}$ *and* $sign(\hat{\mathbf{f}})$*:*

$$\begin{aligned} E\Big(\frac{\langle \hat{\mathbf{f}}, sign(\hat{\mathbf{f}})\rangle}{\|\hat{\mathbf{f}}\|\cdot\|sign(\hat{\mathbf{f}})\|}\Big) &= E\Big(\frac{\sum_i \hat{f}_i \cdot sign(\hat{f}_i)}{\|\hat{\mathbf{f}}\|\cdot\|sign(\hat{\mathbf{f}})\|}\Big) \\ &= E\Big(\frac{\sum_i |\hat{f}_i|}{\|\hat{\mathbf{f}}\|\cdot\|sign(\hat{\mathbf{f}})\|}\Big). \end{aligned} \tag{2.13}$$

For zero-mean unit-variance normally distributed f_i*,*

$$E\Big(\frac{\langle \hat{\mathbf{f}}, sign(\hat{\mathbf{f}})\rangle}{\|\hat{\mathbf{f}}\|\cdot\|sign(\hat{\mathbf{f}})\|}\Big) = E(|\hat{f}_i|) = \sqrt{\frac{2}{\pi}}. \tag{2.14}$$

Then, we show that the amount of energy of $\bar{\mathbf{f}}$ *that falls into* $T_{\mathbf{f}}$ *is lower-bounded.*

Because the correlation between a pair of signals in the spatial domain is equal to their correlation in the DCT domain, we have:

$$\begin{aligned} \frac{\langle \hat{\mathbf{f}}, sign(\hat{\mathbf{f}})\rangle}{\|\hat{\mathbf{f}}\|\cdot\|sign(\hat{\mathbf{f}})\|} &= \frac{\langle \mathbf{f}, \bar{\mathbf{f}}\rangle}{\|\mathbf{f}\|\cdot\|\bar{\mathbf{f}}\|} \\ &= \frac{\langle \mathbf{f}, \bar{\mathbf{f}}\rangle}{\|\mathbf{f}\|\sqrt{\sum_{i\in T_{\mathbf{f}}} \bar{f}_i^2 + \sum_{j\notin T_{\mathbf{f}}} \bar{f}_j^2}} \\ &= \frac{\langle \mathbf{f}, \bar{\mathbf{f}}\rangle}{\|\mathbf{f}\|\sqrt{\alpha^2+\beta^2}}. \end{aligned} \tag{2.15}$$

Let $\mathbf{1}_{T_{\mathbf{f}}}$ *be the indicator function that has the value 1 for all elements of* $T_{\mathbf{f}}$ *and* 0 *elsewhere. From the Cauchy-Schwartz inequality:*

$$\langle \mathbf{f}, \bar{\mathbf{f}}\rangle = \langle \mathbf{f}\circ\mathbf{1}_{T_{\mathbf{f}}}, \bar{\mathbf{f}}\circ\mathbf{1}_{T_{\mathbf{f}}}\rangle \leq \alpha\|\mathbf{f}\|. \tag{2.16}$$

According to Eq. 2.14 and Eq. 2.16:

$$\begin{aligned} E\Big(\frac{\langle \mathbf{f}, \bar{\mathbf{f}} \rangle}{\|\mathbf{f}\| \cdot \|\bar{\mathbf{f}}\|}\Big) &= \sqrt{\frac{2}{\pi}} \leq E\Big(\frac{\alpha \|\mathbf{f}\|}{\|\mathbf{f}\| \cdot \sqrt{\alpha^2 + \beta^2}}\Big) \\ E\Big(\frac{\alpha}{\sqrt{\alpha^2 + \beta^2}}\Big) &\geq \sqrt{\frac{2}{\pi}} \approx 0.7979\,. \end{aligned} \tag{2.17}$$

2.2.2 Hamming distance captures the angular difference between structurally similar images

As we have suggested in Eq. 2.4, the Hamming distance D between two image signatures can be used as a distance metric. We show below how this distance is related to the angular difference between a pair of images $\mathbf{x}^1$ and $\mathbf{x}^2$.

Let ϕ_i denote the i^{th} basis function of the DCT. We assume that ϕ_i is independent of both $\mathbf{x}^1$ and $\mathbf{x}^2$. From [28] (Lemma 2.2), we know that:

$$P\Big[\text{sign}(\langle \mathbf{x}^1, \phi_i \rangle) \neq \text{sign}(\langle \mathbf{x}^2, \phi_i \rangle)\Big] = \frac{1}{\pi} cos^{-1}\Big(\frac{\langle \mathbf{x}^1, \mathbf{x}^2 \rangle}{\|\mathbf{x}^1\| \cdot \|\mathbf{x}^2\|}\Big)\,.$$

Let $d_i(\mathbf{x}^1, \mathbf{x}^2)$ be the indicator function:

$$d_i(\mathbf{x}^1, \mathbf{x}^2) = \begin{cases} 0 & \text{if sign}(\langle \mathbf{x}^1, \phi_i \rangle) \neq \text{sign}(\langle \mathbf{x}^2, \phi_i \rangle) \\ 1 & \text{if sign}(\langle \mathbf{x}^1, \phi_i \rangle) = \text{sign}(\langle \mathbf{x}^2, \phi_i \rangle)\,, \end{cases}$$

and $D(\mathbf{x}^1, \mathbf{x}^2) = \|\text{sign}(\hat{\mathbf{x}}^1) - \text{sign}(\hat{\mathbf{x}}^2)\|_0 = \sum_i^N d_i(\mathbf{x}^1, \mathbf{x}^2)$, since $\langle \mathbf{x}, \phi_i \rangle = \hat{x}_i$. Then, the Chernoff bounds guarantee that:

$$\begin{aligned} \forall \epsilon > 0, \qquad & P\big(D > (1+\epsilon) N\mu\big) < e^{-\frac{1}{4} N \mu \epsilon^2} \\ \forall 0 < \epsilon < 1, \qquad & P\big(D < (1-\epsilon) N\mu\big) < e^{-\frac{1}{2} N \mu \epsilon^2}\,, \end{aligned}$$

where $\mu = E(d_i)$. This result indicates that for large enough N, the following statement is true with high probability:

$$(1-\epsilon)\frac{D}{N} \leq \frac{1}{\pi} cos^{-1}\Big(\frac{\langle \mathbf{x}^1, \mathbf{x}^2 \rangle}{\|\mathbf{x}^1\| \cdot \|\mathbf{x}^2\|}\Big) \leq (1+\epsilon)\frac{D}{N}\,. \tag{2.18}$$

Suppose that the pair of images $\mathbf{x}^1$ and $\mathbf{x}^2$ share the same structures, i.e. $D(\mathbf{x}^1, \mathbf{x}^2) \ll 0.5$. Then, the distance changes are more likely to be accounted by foreground change. However, if the

two images' structure has changed (for instance, the location of the foreground object has changed), then $D(\mathbf{x}^1, \mathbf{x}^2) \approx 0.5$, where no further changes in either foreground or background will push the expectation of D further close to 0.5.

2.3 Image Signature on synthetic images

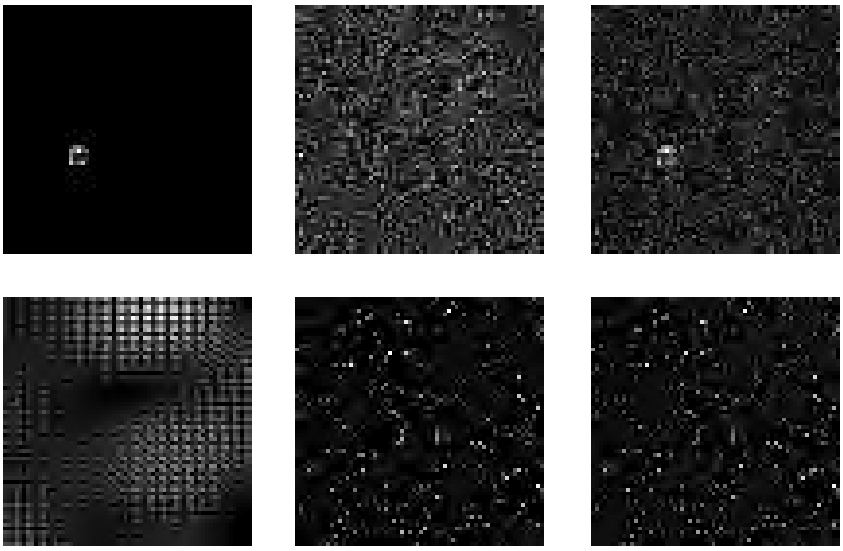


Figure 2.2: An illustration of the randomly generated images. The first row: $\mathbf{f}$, $\mathbf{b}$, and $\mathbf{x}$ in the spatial domain. The second row: The same signals represented in the DCT domain: $\hat{\mathbf{f}}$, $\hat{\mathbf{b}}$, and $\hat{\mathbf{x}}$.

In the previous section, we provided theoretical arguments connecting the Image Signature to the spatial support of a sparse foreground. In this section, instead of using natural images, we employ a series of synthetic image experiments. These carefully-constructed cases are the test bed to examine the necessary and sufficient conditions of the algorithm.

Let $\mathbf{f}, \mathbf{b}, \mathbf{x} \in \mathbb{R}^{64\times 64}$. The support of the foreground is a 5×5 block ($|T_{\mathbf{f}}| = 25$) that appears at a random location. The support for $\hat{\mathbf{b}}$ is randomly selected in the DCT domain, with $|\Omega_{\mathbf{b}}| = 500$. For $i \in T_{\mathbf{f}}$, the amplitude of each pixel f_i is drawn from a normal distribution. Similarly, for $j \in \Omega_{\mathbf{b}}$, each $\hat{b}_j$ is drawn from normal distribution. Fig. 2.2 shows $\mathbf{f}$, $\mathbf{b}$ and $\mathbf{x}$ in both the spatial and the DCT domains.

The Image Signature reconstruction is illustrated in Fig.2.3. Note that a Gaussian blurring is used to suppress the noise introduced by the sign quantization. Ideally, the standard deviation σ of the Gaussian kernel should be proportional to the size of the object of interests. We here choose $\sigma = 0.05$ of the image width (in other words, we implicitly assume that the width of the object is

Figure 2.3: An example of the input image $\mathbf{x}$, the reconstructed image $\bar{\mathbf{x}}$, and the saliency map $\mathbf{m}$.

about 10% of the image width).

From Proposition 1, it follows that the reconstructed image $\bar{\mathbf{x}}$ should be insensitive to the amplitude of the foreground. Instead, only its spatial support should be affected. We tested this by multiplying the amplitude of the foreground $\mathbf{f}$ by a factor of 10^{-5}, 10^{-10} and 10^{-15}, while holding the background completely constant. As predicted by the theory, the reconstructed signal $\bar{\mathbf{x}}$ is not changed by the foreground energy until it approaches a minimal numerical value in MATLAB, 2.2×10^{-16}.

Figure 2.4: The reconstructed $\bar{\mathbf{x}}$ with foreground re-weighted as 10^{-5}, 10^{-10} and 10^{-15}. The saliency maps of these signals remain almost the same despite a huge difference of the foreground amplitude.

The result in Fig. 2.4 directly contradicts the theory of [22], which erroneously explains the spectral saliency as a biologically-plausible process of Gabor linear filtering followed by normalization. In fact, the strong invariance to foreground signal strength makes Image Signature immune to today's rampant and often uncontrolled attempts to impose biological plausibility on any computational model.

Proposition 2 guarantees that the majority of the foreground energy stays in the support of the foreground after the sign quantization. Our theoretical justification is based on a Gaussian distribution. However, it has been suggested by [17] that the histogram of pixel intensity of natural images

follows a power law (that is, the pixel intensity follows a Pareto distribution). We generated the foreground pixels based on three different distributions – normal distribution, uniform distribution, and Pareto distribution with the PDF $f(x) = (1+x)^{-2}$ – and tested whether the energy of $\bar{\mathbf{x}}$ was constrained in the foreground region. For fair comparisons, the foreground was normalized to $[0, 1]$. The proportion (in the sense of Proposition 2) that fell into $T_{\mathbf{f}}$ was: 79.8%, 75.6%, and 79.3% for the three distributions, respectively.

In some scenarios, the background may not be ideally sparse. In Fig. 2.5, we provide an empirical demonstration to test the robustness of this method with respect to non-sparse backgrounds. We observe a clear trend that the energy within $T_{\mathbf{f}}$ drops as the complexity of the background increases. It is worth mentioning that even with fairly complex backgrounds (with $|\Omega_{\mathbf{b}}| = 3000$), the saliency map still clearly shows the shape of the foreground support of the image.

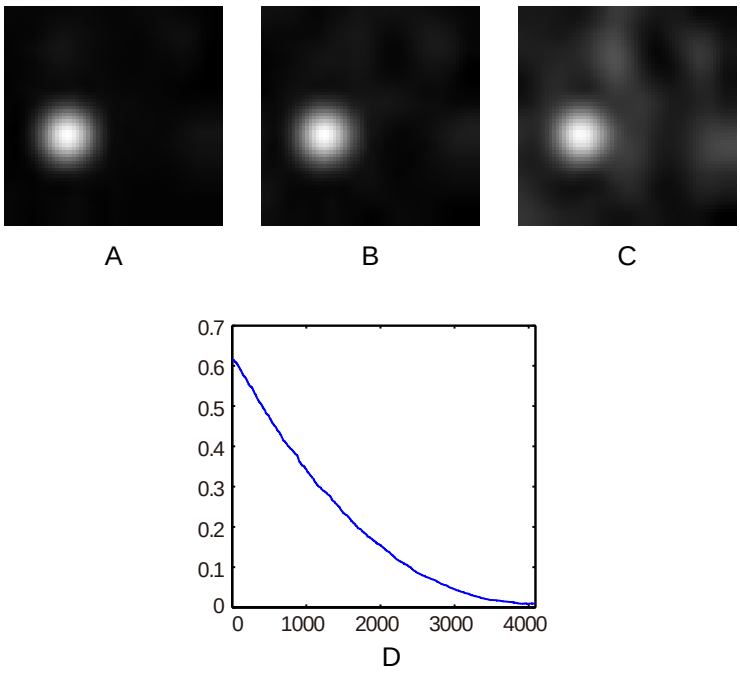


Figure 2.5: Performance of the Signature algorithm with different background complexity on a 64×64 image ($N = 4096$). A. $|\Omega_{\mathbf{b}}| = 1600$. B. $|\Omega_{\mathbf{b}}| = 2400$. C. $|\Omega_{\mathbf{b}}| = 3200$. D. Proportion of energy concentrated in the foreground support $T_{\mathbf{f}}$, as a function of background cardinality $|\Omega_{\mathbf{b}}|$. The proportion of energy is the square of the fraction provided in Proposition 2.

It is worth noting that Image Signature does not make any assumption about the frequency range of the non-zero components of $\hat{\mathbf{b}}$. Wether the background is made of low-frequency or high-frequency DCT components makes *no difference* to the result saliency map. In Fig. 2.6, we construct two special cases to illustrate this property.

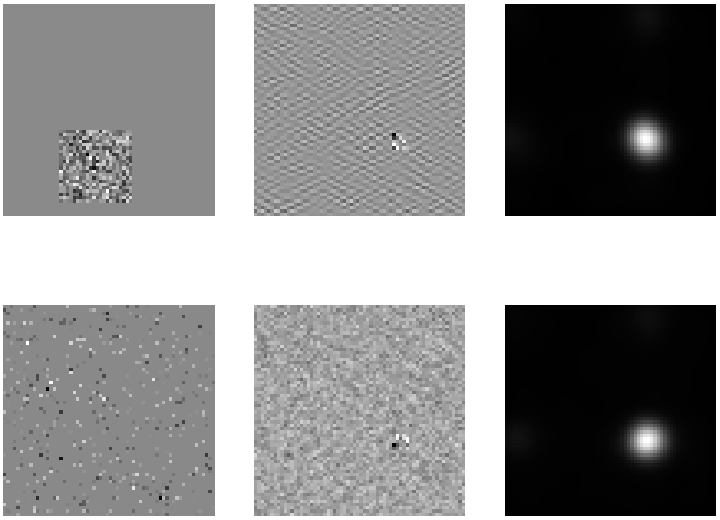


Figure 2.6: Two examples of different frequency components of the background. The first row shows $\hat{\mathbf{b}}^1$, $\mathbf{x}^1$, and the saliency map $\mathbf{m}^1$. The second row shows $\hat{\mathbf{b}}^2$, $\mathbf{x}^2$, and $\mathbf{m}^2$. The two images share the same $\mathbf{f}$. $\hat{\mathbf{b}}^1$ has its support a 20×20 rectangle frequency band, whereas the support of $\hat{\mathbf{b}}^2$ is 400 randomly-selected pixels such that $\Omega_{\mathbf{b}^1} = \Omega_{\mathbf{b}^2} = 400$. Despite different compositions of their background, the two saliency maps look extremely similar.

Fig. 2.6 is a counter example to *any* theory that relates spectral saliency to operations on neighboring frequency components. In addition to the original Spectral Residual [20], more recent theories such as [25] revisited the amplitude spectrum and incorrectly concluded that the spikes in the amplitude spectrum determine the foreground-background configuration.

2.4 Experiments

2.4.1 Search asymmetry

Visual search asymmetry [29] is a well-known phenomenon in which the human performance can be very counter-intuitive. In Fig. 2.7, a subject can easily locate one C in many O's. However, in the reversed problem of finding the O among C's, the observer loses her/his ability to detect the figure in the first sight. This striking phenomenon should not be considered as a defect of the cortical algorithm for figure-ground separation. Instead, it reflects some of the "empirical designs" that work for most of the natural scenes.

If we use existing algorithms to generate saliency maps of these synthetic images, *none* of them except for Image Signature is able to pass the test of search asymmetry. Fig. 2.7 shows the detail.

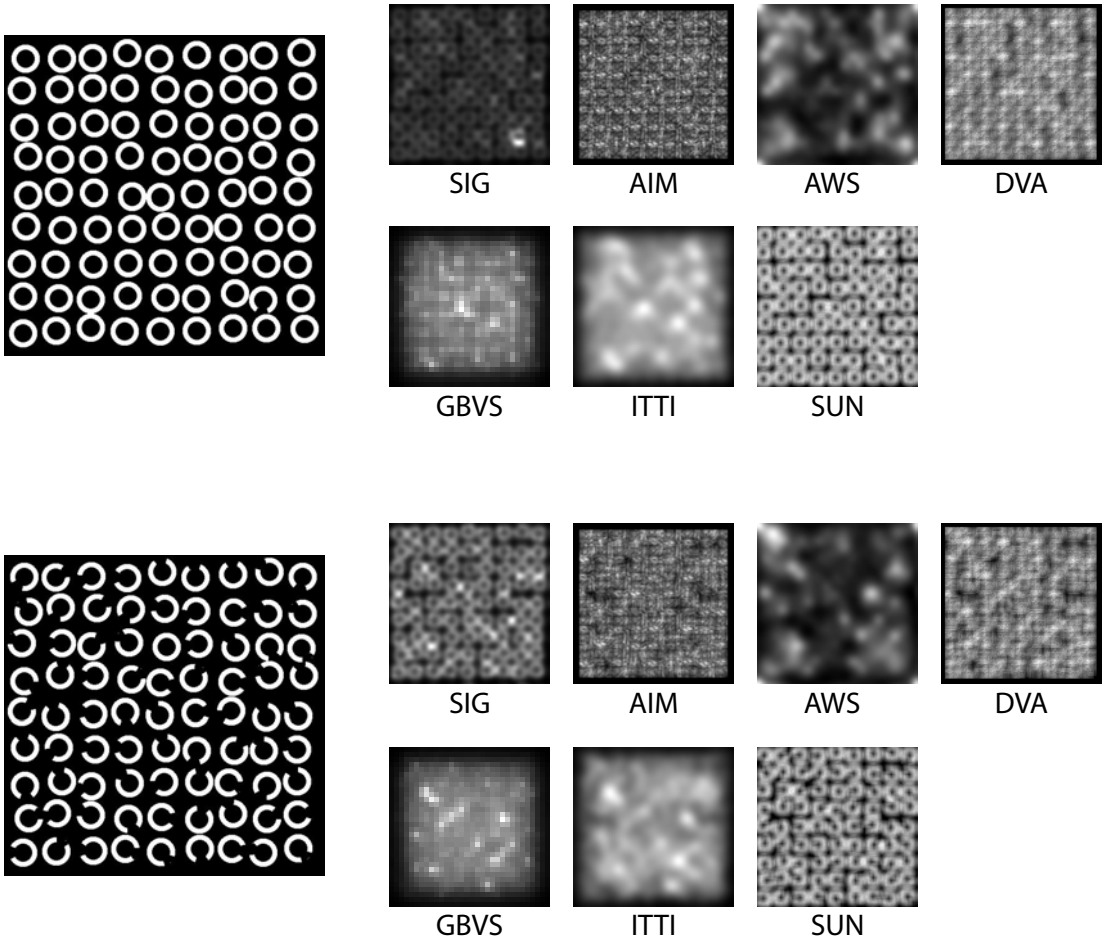


Figure 2.7: Visual search asymmetry. There are two scenarios in this visual search task. The easy task is to find the only 'C' among numerous 'O's, whereas the hard task is to find the unique 'O' in 'C's. We expect a successful algorithm to detect the easy task but fail in the hard task as humans do. Among all 7 algorithms, Image Signature is the only one that passes this test.

2.4.2 Generating the saliency map of an image

In this section, we report our experimental findings in saliency detection using the Image Signature. As we demonstrated earlier, the reconstructed image detects spatially-sparse signals embedded in spectrally-sparse backgrounds. We will show that the saliency map (Eq. 2.3), which is the Image Signature reconstruction of the foreground spatial support, greatly overlaps with regions of human overt attentional interest, measured as fixation points on the image.

The exact details of the saliency algorithm are as follows: first, a color image is resized to a coarse 64×48 pixel representation. Then, for each color channel $\mathbf{x}^i$, the saliency map is formed

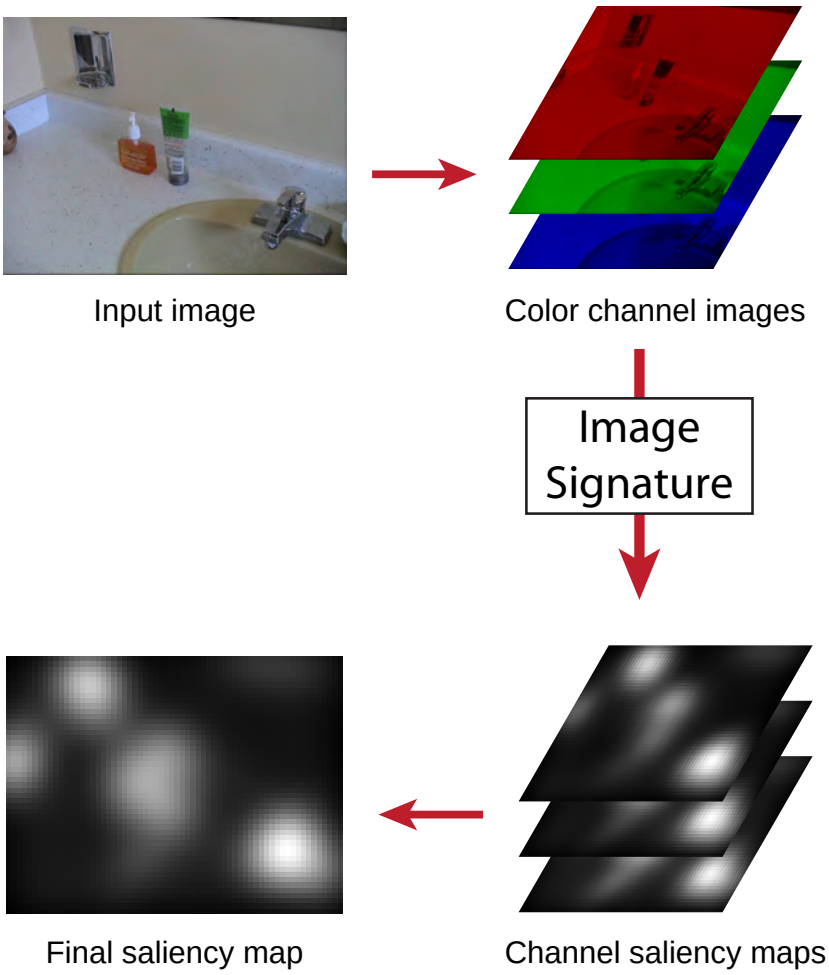


Figure 2.8: An illustration of the Image Signature algorithm pipeline. The input color image is decomposed into 3 channels. A saliency map is computed for each color channel independently, and the final saliency map is simply the sum across the 3.

from the image reconstructed from the Image Signature:

$$\mathbf{m} = g * \sum_i (\bar{\mathbf{x}}^i \circ \bar{\mathbf{x}}^i) \,. \tag{2.19}$$

The standard deviation of the Gaussian blurring kernel g will be discussed in greater detail in the following section.

For the choice of color channels, we use CIELab color spaces. An illustration of this algorithm is shown in Fig. 2.8.

2.4.2.1 Predicting human fixation

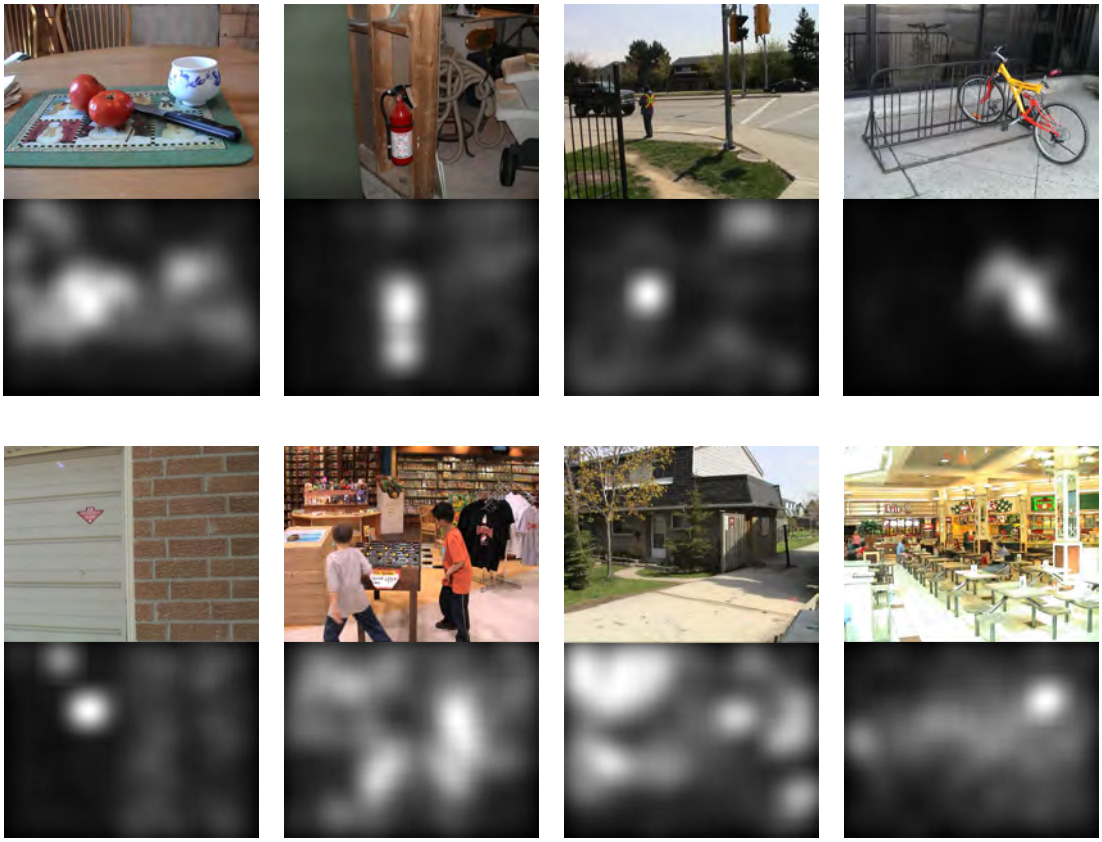


Figure 2.9: Sample images from the Bruce dataset and their corresponding saliency maps using the Image Signature algorithm with $\sigma = 0.05$.

To validate the saliency maps generated by our algorithm, we use 5 datasets of human eye-tracking data. One of the first publicly-available datasets is introduced by Bruce and Tsotsos [3] (denoted **Bruce**). It consists of 20 subjects free-viewing 120 color images (681×511 pixels) for 4 seconds each. Some sample images are shown in Fig. 2.9. The dataset created by Cerf et al. [30] (denoted **Cerf**) is a dataset focusing on human faces. We also used [31] (denoted as **Judd**), the largest generic dataset of human eye fixations. Moreover, we added the small dataset of [32] (denoted **ImgSal**). Lastly, the new dataset [7] created by us (denoted **PASCAL-S**) is also used in the evaluation. Background information of the 5 datasets is summarized in Tab. 2.2.

In order to evaluate the consistency between a particular saliency map and a set of fixations of the image, we computed an ROC Area Under the Curve (AUC) score for each image. As [33] and [34] have pointed out, human fixations have strong center-bias,which may affect the performance

Name	Subject#	Image #	Year
Bruce	20	120	2006
Cerf	13	200	2008
Judd	15	1003	2009
IS	21	235	2013
PASCAL-S	8	850	2014

Table 2.2: The background information of all 5 datasets used to evaluate algorithms.

of a saliency algorithm. To remove this center bias, we follow the procedure of [33]: for one image, the positive sample set is composed of the fixation points of all subjects on that image, whereas the negative sample set is composed of the union of all fixation points across all images from the same dataset, except for the positive samples. Each saliency map generated by the algorithm is thresholded and then considered as a binary classifier to separate the positive samples from negative samples. At a particular threshold level T, the true positive rate is the proportion of the positive samples that fall in the positive (white) region of the binary saliency map (Fig. 2.10-B). The false-positive rate can be computed in a similar way by using the negative sample set. Sweeping over thresholds yields an ROC curve, of which the area beneath provides a good measure of the power of the saliency map to accurately predict where fixations occurred on an image. Chance level is 0.5, and perfect prediction is 1.0.

We compare our saliency maps generated from the Image Signature (denoted **SIG**) to the following published saliency algorithms: the original Itti-Koch saliency model [2] (denoted **Itti**), Dynamic Visual Attention model [4] (denoted **DVA**), Graph-Based Visual Saliency [35] (denoted **GBVS**), Attention based on Information Maximization [3] (denoted **AIM**), Adaptive Whitening Saliency [5] (denoted **AWS**), and Saliency Using Natural image statistic [34] (denoted **SUN**) for comparison. All of the algorithms are based on the original MATLAB implementations available on the authors' websites.

An important note about these experiments is that the AUC score is quite sensitive to blurring a saliency map. Some kind of smoothing has been explicitly or implicitly included in most of the algorithms. In order to make a fair comparison, we parameterize the standard deviation of the blurring kernel, and evaluate the performance of an algorithm under different blurring conditions, applied to the final master saliency maps.

In Fig. 2.11, we show how the AUC score of all these 7 algorithms depends on the standard deviation of a Gaussian smoothing kernel applied to the final saliency maps. We see that the per-

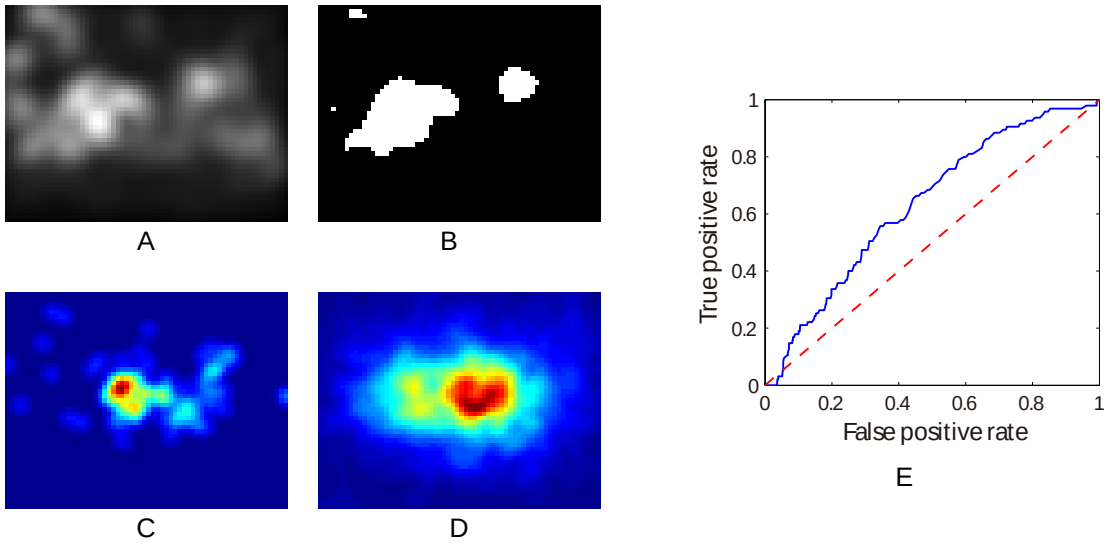


Figure 2.10: An illustration of the AUC computation on the first image in Fig. 2.9. A. The saliency map generated by Lab-Signature algorithm. B. The binary map (thresholded at $T = 0.5$). C. The positive sample set of human fixations on this image (represented as a heat map). D. The negative sample set of human fixations, containing all fixations across the entire dataset, except those contained in the positive sample set (represented as a heat map). Both Figure C and D are smoothed for display clarity, but the AUC computation uses the exact fixation points. E. The blue curve shows the ROC curve of Image Signature algorithm on this image, with the red reference line indicating the chance level. The area under the blue curve is 0.6329.

formance of Image Signature is very competitive. The regions highlighted by the Image Signature saliency algorithm overlap to a surprisingly large extent with those image regions looked at by humans in free viewing. It is also interesting to observe that the optimal blurring factor σ is quite stable across different algorithms. In other words, we can choose one σ that works well for many algorithms. In Table 2.3, we list the AUC score of each algorithm under its optimal σ on all datasets.

Importantly, not only is Image Signature one of the most accurate algorithms, it also runs hundreds times faster than AIM and AWS, both of which parred with Image Signature in terms of accuracy. The excellent speed advantage is due to the small number of channels used in Image Signature compared to other saliency algorithms (see [34] for a comparison of saliency algorithms by computational components). Fig. 2.12 reports each algorithm's MATLAB run-time measurements averaged over the dataset. Compared to the Image Signature, which uses only three color channels at a single spatial scale, Itti and GBVS rely on seven feature channels and multiple spatial scales; DVA uses 192 filters of 192 dimensions, AIM uses 25 filters of 1323 dimensions, and SUN uses 362 filters of 363 dimensions.

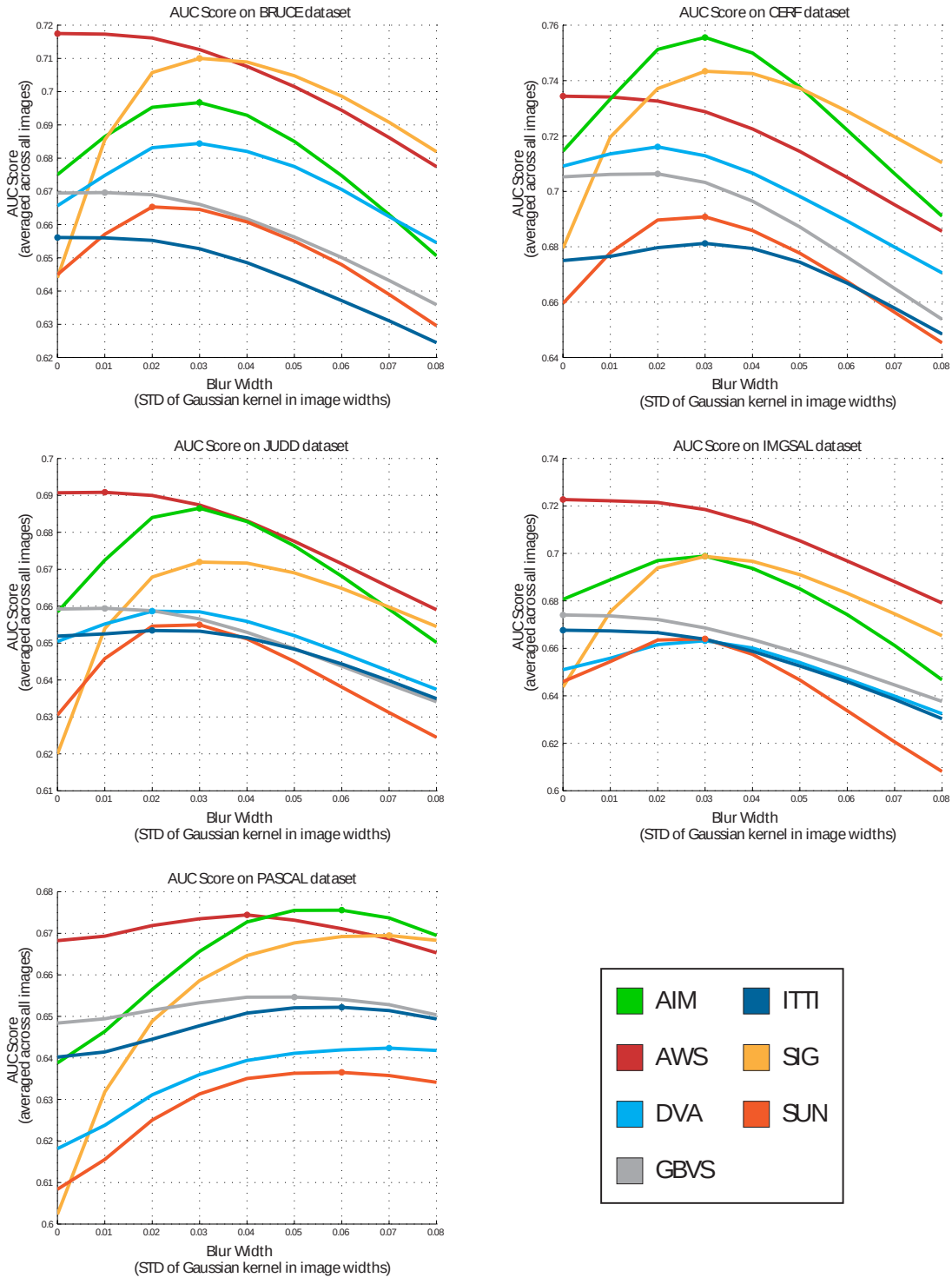


Figure 2.11: The AUC metric reported herein and in other papers is quite sensitive to blurring. Parameterized by a Gaussian's standard deviation in image widths, this factor is explicitly analyzed to provide a better understanding of the comparative performance of an algorithm. For each algorithm, its optimal blurring factor is labeled as a dot on the plot. Solid line is used for algorithms whose average computing time is less than 1 second. For the computationally more expensive algorithms (GBVS, AIM-original, SUN-original), a dashed line is used to draw their performance curve.

Algorithm	AUC Scores				
	Bruce	Cerf	Judd	ImgSal	PASCAL-S
AWS	0.718	0.734	0.691	0.723	0.674
AIM	0.697	0.756	0.687	0.699	0.676
DVA	0.684	0.716	0.659	0.663	0.642
GBVS	0.670	0.706	0.659	0.674	0.655
ITTI	0.656	0.681	0.653	0.668	0.652
SIG	0.710	0.743	0.672	0.699	0.669
SUN	0.665	0.691	0.655	0.664	0.637

Table 2.3: The AUC scores of all 7 algorithms on 5 different fixation datasets.

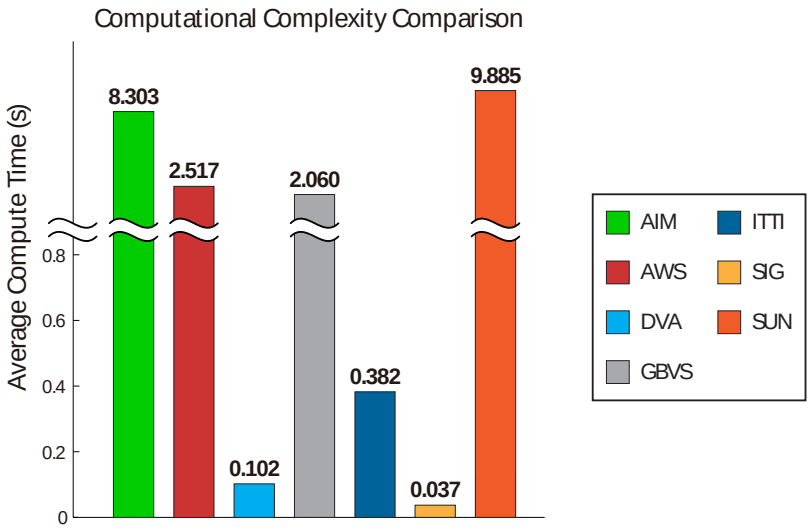


Figure 2.12: Average compute time for each algorithm. All algorithms are implemented in the single-thread MATLAB environment on an 8 core 2.5GHz Xeon workstation.

2.4.3 Correlations to change-blindness

Change-blindness [36] is a striking phenomenon in which a subject fails to notice otherwise obvious changes in a pair of images, even when the viewing time extends over a minute or longer. In such an experiment, the original image and a modified version of it alternate repeatedly, but, critically, with a brief masking inserted in between. The ordinary perceptual motion or flicker which would accompany such a change is eliminated by the intervening interval which acts as a sort of mask. The observer must thus rely on his visual memory to identify the change. This is surprisingly difficult.

The phenomenon has inspired a rich literature in visual scene and object perception. Rensink et al. [37] suggests that an observer has to encode the image into an internal scene representation, which is sparse and incomplete. This very narrow bottleneck of representation has been demon-

Figure 2.13: The experimental paradigm for change-blindness. In image 2, the window on the adobe wall has been removed. The subject has to report detection by clicking on the changed area of either image.

strated to be tightly related to the deployment of visual attention. There have been studies [38] that suggest that attended objects are more likely to be encoded in the working memory than non-attended ones.

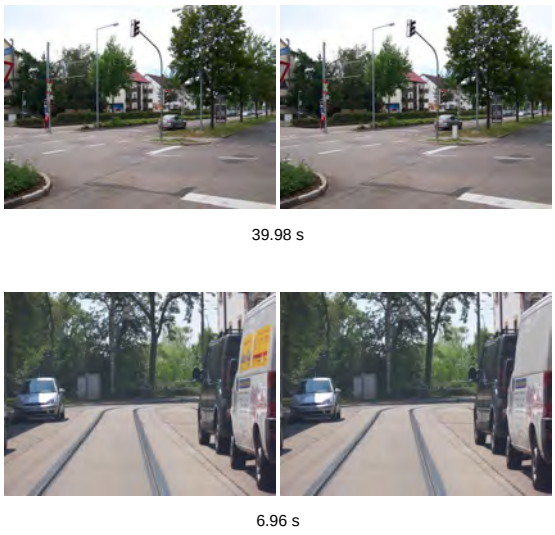


Figure 2.14: Two sample image pairs. Labels indicate the median reaction time of 9 subjects. Top: the difference is a small white post in the center divider (absent left, present right). Bottom: The difference is the yellow sign on the van (present left, absent right).

Below, we use behavioral data from human subjects as an alternative ground-truth to test the efficacy of our image signature. Results demonstrate that the signature distance of two image signatures is strongly (inversely) correlated with the reaction time of human subjects in detecting the change. To our knowledge, there have been no previous attempts to correlate a computational representation of a visual scene with the reaction time in a change-blindness experiment.

2.4.3.1 Experiment setup

In an experiment conceived by one of the authors (C.K.) and Claudia Wilimizig[1], 60 color images of real world scenes from personal albums were selected. For each original image, 2 modified versions were made, each with one object removed and retouched manually using Adobe Photoshop. The artifacts caused by image processing were kept minimal (Fig. 2.14 gives several examples of the stimuli). During each trial, the original image was displayed for $480ms$, followed by $160ms$ black masking, and then $480ms$ for the modified image, and then $160ms$ masking. The trial stops after 60s, or when the subject responds by clicking on the image. If the selected location was far away from the true modification, or if the subject did not respond within $60s$, or if the response time was less than $640ms$ (before the first onset of the second image), the trial was discarded. 9 naive subjects with normal vision participated in the experiment. Subjects correctly identified the change (or signaled no change) in 1011 (93.6%) of the $9 \times 2 \times 60 = 1080$ trials.

Because the reaction time distribution among subjects is highly non-linear, we instead compute the log reaction time. The inter-subject correlation (correlating one subject's reaction time against the remaining 8 subjects) of the reaction times improves from 0.3558 to 0.5305 when moving from a linear to a log reaction time, suggesting that the log reaction time correlation is a more meaningful metric than the linear reaction time correlation.

2.4.3.2 Correlate algorithm output with reaction time

As the consequence of a complex cognitive process, the reaction time of a subject in a change-blindness experiment is influenced by many factors. We here correlate such reaction times with various measures derived from the original image and its modified version.

First, reaction times are compared with the saliency of the modified objects. For a good saliency algorithm, we expect the saliency value of an object to be inversely correlated with the reaction time, since the more salient an object is, the more easily a subject can spot it, and thus detect its removal. The saliency value of a removed object is computed by the mean (or sum) pixel intensity of the object region in the saliency map of the original image.

Second, reaction times are compared to the Hamming distance (Eq. 2.4) between the image signature descriptor of the original image and that of the modified image. As described in Sec. 2.2.2, this distance is a sensitive one when images share a background, as they do in the case of

[1]Images were prepared by Amy Chung-Yu Chou, and data was collected by Tom Laudes.

a change-blindness pair. The distance between the descriptors should be related to the extent of difference in their salient, or foreground, regions.

Third, the widely used GIST descriptor [9] is used to describe each image in a change-blindness pair, and reaction times are compared to the GIST distance. [39] showed that perceptually similar images are usually close together in GIST descriptor space. GIST uses 8 orientations, 4 scales for each 4×4 grid of an RGB color channel, mapping an image to a $8 \times 4 \times 16 \times 3 = 1536$ dimensional real-valued descriptor.

Lastly, we use the pixel-wise distances between the images in the change-blindness pair, and compare these with reaction times. We actually use two pixel-wise measures: the ℓ_0 and ℓ_2 distances between the original and modified image. The ℓ_0 distance is exactly equal to the modified area size.

Let $\mathbf{h}_i$ be the log reaction times of the i^{th} subject (a vector with a component for each image in the dataset), and $\mathbf{v}$ be the image pair distances according to one of the methods described above; then, the normalized correlation c is given by correlating $\mathbf{v}$ with each $-\mathbf{h}_i$, normalized by the mean inter-subject correlation, and averaging over 9 subjects:

$$c = \frac{1}{9} \sum_{i=1}^{9} \frac{\text{corr}(-\mathbf{h}_i, \mathbf{v})}{E_{j \neq i}\left[\text{corr}(\mathbf{h}_j, \mathbf{h}_i)\right]}. \tag{2.20}$$

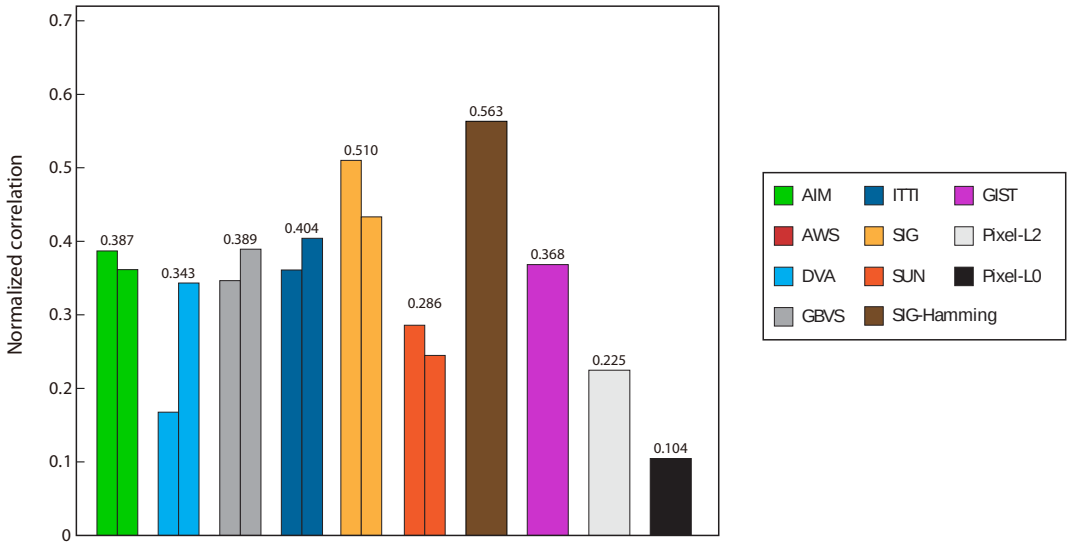


Figure 2.15: The average normalized correlation between reaction time and algorithm outputs. For the first 9 saliency algorithms, the left bar is the performance using the mean pixel value of the object region, whereas the right bar is the result of the sum of pixel saliency value (object size is variable). The score above each pair is the maximum correlation value among the two.

The results are summarized in Fig. 2.15. Among all 10 algorithms, the Hamming distance between Image Signature descriptors correlates best with reaction times. That is, among the methods tried here, the perceptual distance between change-blindness pairs is best explained by the image signature descriptor. Given our understanding of the connection between foreground information and the signature, a difficult change-blindness trial is likely one in which the removed object is perceived as part of the background, because in such a trial, we expect a small signature distance.

2.4.4 Image Signature and face orientation

To further illustrate the Image Signature as a compact descriptor of the image, we use the FERET face database [11] as the corpus for analysis. This database contains 1400 images of 200 individuals. For each individual, 7 different images were taken, among which, 5 involves head-orientation change ($-20°, -10°, 0°, 10°, 20°$, respectively), and 2 other images taken in $0°$ pose contain facial expression and illumination changes. In our experiments, these images are considered as front-face ($0°$).

We split the dataset into 700 training images where the labels are readily available for the algorithm, and 700 testing images where the labels will be estimated using K-NN algorithm. The core idea of this experiment is to illustrate the neighborhood structure of Image Signature defined by Eq.2.4. By choosing $K = 20$ and using majority voting to determine the head orientation for a testing image, this simple algorithm achieved 98.86% accuracy. That is, only around 8 images out of 700 images in the test set were classified as wrong. To the best of our knowledge, the best available result on the FERET database is done by [40] with an accuracy of 97%, which means over 20 misclassifications.

By comparing the distance metric of Image Signature against that of another famous descriptor, GIST, we can obtain even more interesting results. As shown in Fig.2.16, the difference between Signature and GIST is prominent: on one hand, signature neighborhood is much more consistent in head perspective than the GIST neighborhood. On the other hand, however, GIST is much more successful in extracting identity information. This result suggests that GIST captures the identity information, whereas the Signature captures the perspective information.

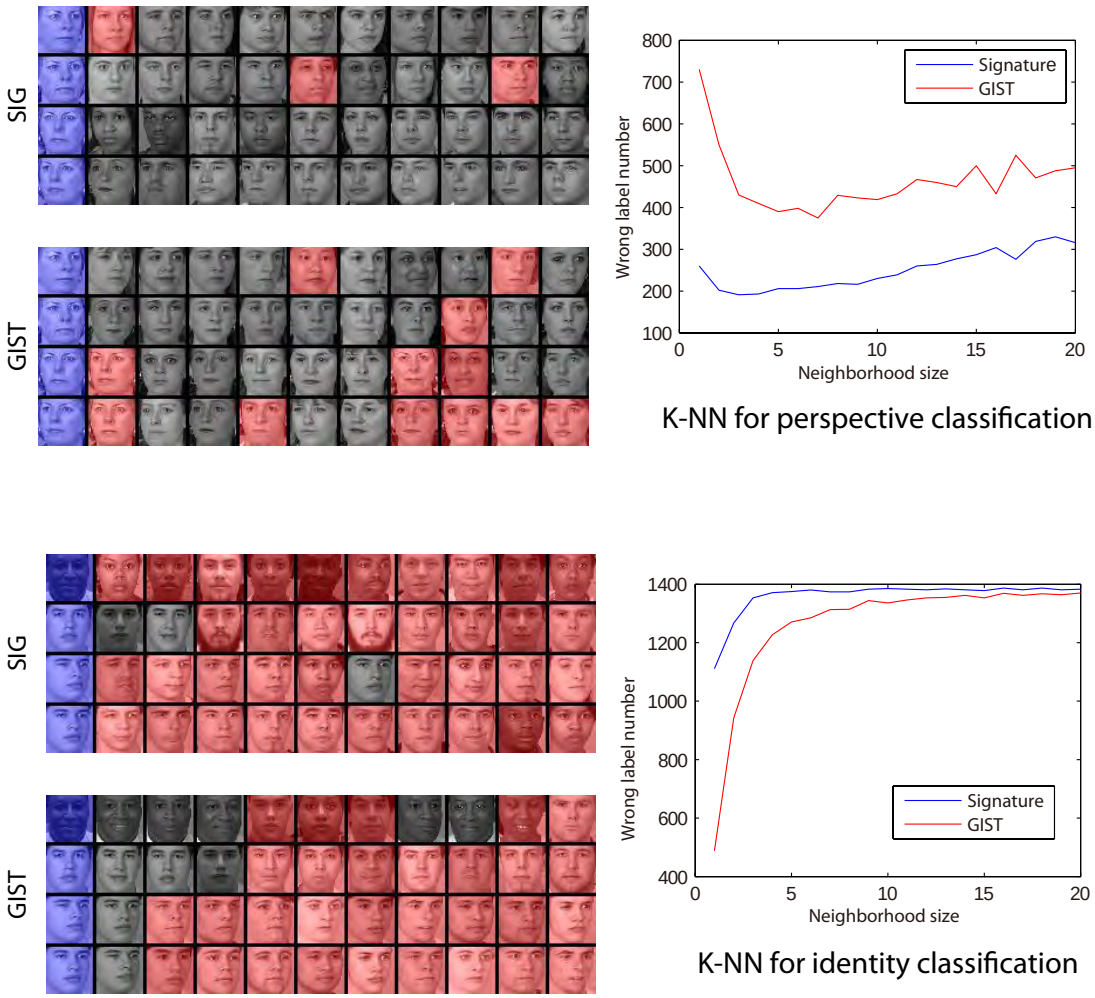


Figure 2.16: The neighbors of faces under different metric functions. In each row, the blue image is the query image, followed by 10-nearest neighbors. Red squares signal mismatches, which depends on different tasks (e.g. if the task is perspective classification, faces with different identity but the same orientation will be considered correct).

2.5 Conclusion

We introduced the Image Signature as a simple yet powerful descriptor of natural scenes. We proved on the basis of theoretical arguments that this descriptor can be used to approximate the spatial location of a sparse foreground hidden in a spectrally sparse background. We provided synthetic experiments to test the necessity and sufficiency of the assumptions behind Image Signature. We use psychophysical experimental data to show that the approximate foreground location highlighted by the Image Signature was remarkably consistent with both search asymmetry patterns, as well as the locations of human eye-movement fixations, predicting them as good as, or even better than, leading saliency algorithms at a fraction of the computational cost. We provided results from a *change- blindness* experiment in which the perceptual distance between slightly different images was predicted most accurately by the Image Signature descriptor. We also illustrate the neighborhood structure by performing Image Signature descriptor K-NN on a face-orientation dataset.

Chapter 3

A Phase Discrepancy Analysis for Object Motion

Abstract

Detecting moving objects in dynamic backgrounds remains a challenge in computer vision and robotics. This chapter presents a surprisingly-simple algorithm to detect objects in such conditions. Based on theoretic analysis, we show that 1) the displacement of the foreground and the background can be represented by the phase change of Fourier spectra, and 2) the motion of background objects can be extracted by *Phase Discrepancy* in an efficient and robust way. The algorithm does not rely on prior training on particular features or categories of an image, and can be implemented in 9 lines of MATLAB code.

In addition to the algorithm, we provide a new database for moving-object detection with 20 video clips, 11 subjects and 4785 bounding boxes to be used as a public benchmark for algorithm evaluation.

3.1 Introduction

Detecting moving objects in a complex scene is a problem in computer vision of many practical interests. It is closely related to a variety of critical applications such as tracking, video analysis, content retrieval, and robotics. Generally speaking, motion-detection methods can be categorized into the following main approaches: background modeling, view geometry, detection by recognition, and saliency-based detection.

Many models try to attack the problem of detection under controlled situations. For instance, some algorithms assume a stationary camera. This assumption leads to a branch of techniques

called background subtraction. The main idea is to learn the appearance model of the background [41] [42]. A moving object in the scene is then detected by subtracting the background image from the current image. However, scene appearance captured by a moving camera, with foreground and backgrounds in arbitrary depths and viewpoints, can be very complicated. Thus, most of the background models perform poorly on moving camera recordings.

To circumvent these problems, some other algorithms detect motion via camera geometry [43] [44]. This approach estimates the camera parameters under certain geometric constraints, uses these parameters to compensate for camera-induced motion, and separate the moving object from the residual motion in the scene [45].

Another branch of popular algorithms stems from object detection, either based on pre-trained detectors, or visual saliency. Some algorithms can detect objects from particular categories, such as faces [46] or pedestrians [47]. These algorithms usually require offline training, and can only handle a very limited number of object categories. Moreover, finding an invariant object detector that overcomes illumination/view-point changes and occlusion is already a challenge in computer vision.

On the other hand, a saliency-based detector relies on the assumption that the object is statistically different from its background. For most of the saliency detection algorithms, unique features of the foreground, such as color, orientation [2], sparse filter responses [4], and temporal cues [48] are utilized to generate a saliency map that predicts the location of an object. The advantage of a saliency-based approach is that it has few assumptions about the appearance of the object, and is therefore capable of detecting unspecified objects without pre-training. Nevertheless, for efficiency considerations, most of today's models do not incorporate motion cues in an explicit way, which leads to poor performance in video surveillance (as we will see in Section.3.3.4).

In principle, a visual system needs *only* motion cues to detect an moving object – even if the scene is disturbed by camera's ego-motion. With full knowledge of the optical flow, the mission of object detection is to find the cluster of consistent motion that is induced by the foreground. Nevertheless, the computational burden of an optical flow algorithm is usually very heavy.

3.1.1 Related work

In 2001, Vernon [49] proposed using a Fourier transform to untangle the complexity of object motions. In his theory, object segmentation and exact velocity recovery can be achieved by solving a linear system. Based on the translation property of the Fourier transform, a moving object

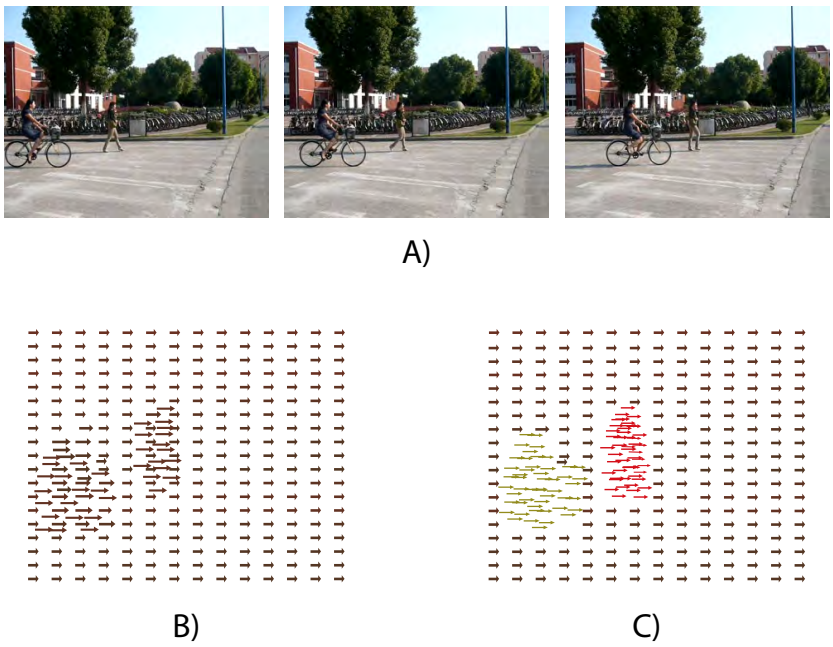


Figure 3.1: An illustration of moving object detection from a perspective of optical flow analysis. **A**): A video sequence with both object motions and camera motion. **B**): The corresponding optical flow. **C**): The segmentation result that detects the moving objects.

corresponds to a phase change in the Fourier spectrum. For a scene composed of m objects, exact recovery is achieved by solving a linear equation with $2m$ unknowns. The drawback of this approach is that the number m of objects must be specified beforehand. Moreover, the segmentation and velocity recovery require observing $2m$ frames, which contain every object moving at a constant speed. These constraints preclude Vernon's approach from real-world applications.

3.1.2 An outline of our approach

We start from a similar perspective to that of Vernon: spatially distributed information can be efficiently accumulated in the Fourier spectrum. However, instead of finding the exact solution for a constrained problem, we seek an approximate solution with a minimal number of assumptions.

To extract moving objects from dynamic backgrounds, our model follows the idea of predictive coding. First, we predict the next frame only considering background movements. Then, by comparing our prediction against the actual observation, pixels representing the foreground emerge due to the large reconstruction error. With rigorous analysis, we show that a 9-line MATLAB script can approximately recover the camera motion with bounded error.

3.2 The Theory

We denote $f(\mathbf{x}, t)$ as our observation at time t[1], where $\mathbf{x} = [x_1, x_2]^\top$ is the 2-dimensional vector of a spatial location. Let $\mathcal{I}$ be the ensemble of pixels. For any image, we have the partition $\mathcal{I} = \{\mathcal{F}_t, \mathcal{B}_t\}$. Every pixel belongs to the foreground $\mathcal{F}_t$ or the background $\mathcal{B}_t$.

For typical sampling rates, the ego-motion of the camera is well approximated by a uniform translation of the background. If we know this displacement $\mathbf{v} = [v_1, v_2]^\top$, we can predict the appearance of the background in the next frame based on the *intensity constancy* assumption [50] that the spatial translation does not change pixel values:

$$f(\mathbf{x}, t) = f(\mathbf{x} + \mathbf{v}, t + 1), \quad \text{where } \mathbf{x} \in \mathcal{B}_t \bigcap \mathcal{B}_{t+1} \tag{3.1}$$

This assumption requires that pixels $\mathbf{x}$ at t and $\mathbf{x} + \mathbf{v}$ at $t + 1$ belong to the background. We further denote $\check{\mathcal{B}}_t = \hat{\mathcal{B}}_{t+1} = \mathcal{B}_t \bigcap \mathcal{B}_{t+1}$.

Once we have the ground-truth of the ego-motion, we can reconstruct the next frame by shifting every pixel from $\mathbf{x}$ to $\mathbf{x} + \mathbf{v}$. This reconstruction is expected to perform poorly for pixels in $\mathcal{I} - \check{\mathcal{B}}_t$, the foreground. Thus, we can take the error as a likelihood function of the appearance of moving objects at certain locations. In other words, the reconstruction error map $s(\mathbf{x}, t)$ can be considered as a *saliency map* [2] for moving objects:

$$s(\mathbf{x}, t) = \Big[f(\mathbf{x} + \mathbf{v}, t + 1) - f(\mathbf{x}, t) \Big]^2. \tag{3.2}$$

3.2.1 Phase discrepancy and ego-motion

In order to generate the saliency map, we need to know the displacement vector $\mathbf{v}$. In the Fourier domain, the spatial displacement in Eq. 3.1 can be efficiently represented by the phase of the Fourier spectrum.

Let $F_{\mathbf{x},t}(\boldsymbol{\omega}) = \mathcal{F}[f(\mathbf{x}, t) \cdot \delta_{\mathbf{x}_i}(\mathbf{x})]$ denote the 2-D Discrete Fourier transform of a single pixel, where $\boldsymbol{\omega} = [\omega_1, \omega_2]^\top$, and the indicator function $\delta_{\mathbf{x}_i}(\mathbf{x})$ is defined as:

$$\delta_{\mathbf{x}_i}(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{x} \in \mathbf{x}_i, \\ 0 & \text{otherwise.} \end{cases}$$

[1] For simplicity, we only consider gray-scale images in this section. A simple extension to color images is provided in Section 3.3.

The Fourier spectrum of the entire image $F_t(\boldsymbol{\omega})$ can be obtained by:

$$F_t(\boldsymbol{\omega}) = \sum_{\mathbf{x}_i \in \mathcal{I}} F_{\mathbf{x}_i,t}(\boldsymbol{\omega})$$

Known as the translation property [51], a spatial displacement entails a phase change, yet leaves the Fourier amplitudes intact:

$$F_{\mathbf{x}+\mathbf{v},t+1}(\boldsymbol{\omega}) = F_{\mathbf{x},t}(\boldsymbol{\omega}) e^{-i\cdot\Phi(\mathbf{v})}, \tag{3.3}$$

where $\Phi(\mathbf{v}) = \boldsymbol{\omega}^\top \mathbf{v} = \omega_1 v_1 + \omega_2 v_2$, which we call the *Phase Discrepancy* in the following discussions.

Because the entire background has approximately the same displacement $\mathbf{v}$, Eq. 3.3 has a compact form for $\check{\mathcal{B}}_t$:

$$\sum_{\mathbf{x}_i \in \hat{\mathcal{B}}_{t+1}} F_{\mathbf{x}_i,t+1}(\boldsymbol{\omega}) = \sum_{\mathbf{x}_i \in \check{\mathcal{B}}_t} F_{\mathbf{x}_i,t}(\boldsymbol{\omega}) e^{-i\cdot\Phi(\mathbf{v})}. \tag{3.4}$$

We have the following decomposition:

$$\begin{aligned} F_{t+1}(\boldsymbol{\omega}) &= \sum_{\mathbf{x}_i \in \mathcal{I}} F_{\mathbf{x}_i,t+1}(\boldsymbol{\omega}) = \sum_{\mathbf{x}_i \in \check{\mathcal{B}}_t} F_{\mathbf{x}_i,t}(\boldsymbol{\omega}) e^{-i\cdot\Phi(\mathbf{v})} + \sum_{\mathbf{x}_i \in \mathcal{I} - \hat{\mathcal{B}}_{t+1}} F_{\mathbf{x}_i,t+1}(\boldsymbol{\omega}) \\ &= F_t(\boldsymbol{\omega}) e^{-i\cdot\Phi(\mathbf{v})} - \sum_{\mathbf{x}_i \in \mathcal{I} - \check{\mathcal{B}}_t} F_{\mathbf{x}_i,t}(\boldsymbol{\omega}) e^{-i\cdot\Phi(\mathbf{v})} + \sum_{\mathbf{x}_i \in \mathcal{I} - \hat{\mathcal{B}}_{t+1}} F_{\mathbf{x}_i,t+1}(\boldsymbol{\omega}). \end{aligned}$$

Although it seems impossible to calculate $\Phi(\mathbf{v})$ without the foreground/background partition, in the next section, we show that good approximation of phase discrepancy is achievable in some cases.

3.2.2 Approximating the phase discrepancy

Since it is impossible to quantify the appearance and location of the pixels in $\mathcal{I} - \check{\mathcal{B}}_t$, we assume $F_{\mathbf{x}_i,t}(\boldsymbol{\omega})$ follows an independent normal distribution in the complex domain; that is,

$$\text{Real}\{F_{\mathbf{x}_i,t}(\boldsymbol{\omega})\} \sim N(0,1); \qquad \text{Imag}\{F_{\mathbf{x}_i,t}(\boldsymbol{\omega})\} \sim N(0,1). \tag{3.5}$$

For a simpler notation, we define a complex variable $z_i = F_{\mathbf{x}_i,t}(\boldsymbol{\omega})$. Let $Z_n = \sum_{i=1}^{n} z_i$ be the

sum of this sequence. We have the following:

$$\begin{aligned} \text{Real}\{Z_n\} &\sim N(0,n) \\ \text{Imag}\{Z_n\} &\sim N(0,n) \end{aligned}$$

Because $|Z_n| = \sqrt{\text{Real}\{z_i\}^2 + \text{Imag}\{z_i\}^2}$, it follows a χ distribution with 2 degrees of freedom:

$$p(|Z_n| = x) = \sqrt{n}\sigma x e^{-x^2/2}. \tag{3.6}$$

Thus, the expectation of the spectral amplitude is determined by the number of pixels in the summation. More specifically:

$$\frac{E(|F_t(\boldsymbol{\omega})|)}{E(|\sum_{\mathbf{x}_i \in \check{\mathcal{B}}_t} F_{\mathbf{x}_i,t}(\boldsymbol{\omega})|)} = \frac{\sqrt{\#(\mathcal{I})}}{\sqrt{\#(\check{\mathcal{B}}_t)}}. \tag{3.7}$$

The number of pixels in the foreground and background are estimated from our hand labeled database (see Section 3.3). On average, our bounding box of the foreground (an over-estimation of the actual foreground) occupies 5% pixels of the frame [2]. Thus, we approximate the phase discrepancy in Eq. 3.5 by:

$$\tilde{\Phi}(\mathbf{v}) = \angle F_{t+1}(\boldsymbol{\omega}) - \angle F_t(\boldsymbol{\omega}). \tag{3.8}$$

The estimation error comes from the pixels of the foreground and occluded parts of the background. The cumulative effect of these pixels at frequency $\boldsymbol{\omega}$ can be considered as added noise to variable η to the original variable $F_t(\boldsymbol{\omega})e^{-i\cdot\Phi(\mathbf{v})}$ in Eq. 3.5, where:

$$\eta = -\sum_{\mathbf{x}_i \in \mathcal{I} - \check{\mathcal{B}}_t} F_{\mathbf{x}_i,t}(\boldsymbol{\omega})e^{-i\cdot\Phi(\mathbf{v})} + \sum_{\mathbf{x}_i \in \mathcal{I} - \hat{\mathcal{B}}_{t+1}} F_{\mathbf{x}_i,t+1}(\boldsymbol{\omega}).$$

From Eq. 3.7, we set $F_t(\boldsymbol{\omega})e^{-i\cdot\Phi(\mathbf{v})}$ to 1 to determine the distribution of η:

$$E(|\eta|) = \frac{\sqrt{2\#(\mathcal{I} - \check{\mathcal{B}}_t)}}{\sqrt{\#(\check{\mathcal{B}}_t)}} \approx \sqrt{0.1}; \qquad \angle\eta \sim U(0, 2\pi). \tag{3.9}$$

[2] In other databases such as [52] and [53], objects are in a similar size.

The upper bound of error in $\tilde{\Phi}(\mathbf{v})$ is therefore:

$$\max\left[\Phi(\mathbf{v}) - \tilde{\Phi}(\mathbf{v})\right] = \max\left\{\tan^{-1}\left[E(|\eta|)\right]\right\} \approx 0.31. \tag{3.10}$$

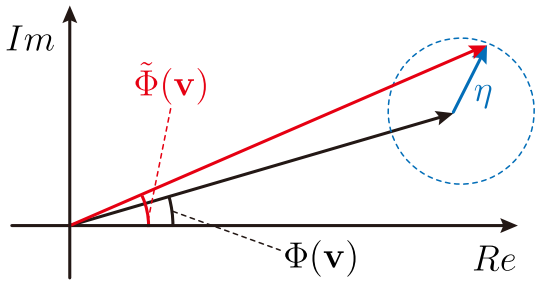


Figure 3.2: A diagram of the angular error calculation. Given $E(|\eta|) = \sqrt{0.1}$, the upper bound of the angular error is 0.31 $(17.6°)$, and the mean angular error is 0.21 . $(12.3°)$

As long as the approximation in Eq. 3.8 holds, we can construct the estimated spectrum $\tilde{F}_{t+1}(\boldsymbol{\omega})$ from $F_t(\boldsymbol{\omega})$ and $\tilde{\Phi}$:

$$\begin{aligned}\tilde{F}_{t+1}(\boldsymbol{\omega}) &= F_t(\boldsymbol{\omega})e^{-i\cdot\tilde{\Phi}(\mathbf{v})} = |F_t(\boldsymbol{\omega})| \cdot e^{-i[\angle F_t(\boldsymbol{\omega}) + \tilde{\Phi}(\mathbf{v})]} \\ &= |F_t(\boldsymbol{\omega})| \cdot e^{-i[\angle F_{t+1}(\boldsymbol{\omega})]}\end{aligned}$$

Finally, the saliency map has the simple form:

$$\begin{aligned}s(\mathbf{x}, t) &= \left\{\mathcal{F}^{-1}\left[F_{t+1}(\boldsymbol{\omega})\right] - \mathcal{F}^{-1}\left[\tilde{F}_{t+1}(\boldsymbol{\omega})\right]\right\}^2 \\ &= \left\{\mathcal{F}^{-1}\left[\left(|F_{t+1}(\boldsymbol{\omega})| - |F_t(\boldsymbol{\omega})|\right) \cdot e^{-i\angle F_{t+1}(\boldsymbol{\omega})}\right]\right\}^2\end{aligned} \tag{3.11}$$

3.2.3 Eliminating boundary effects

The 2-D Discrete Fourier Transform implicitly implies periodicity of the signal. This property invalidates Eq. 3.1, since pixels around the edge of the frame do not have their correspondences in the next frame. As a result, these frame-edges often have very large reconstruction errors and mislead the saliency maps (see Fig. 3.3.C).

Assume we have two adjacent image frames. We use $\mathcal{C}_1$ and $\mathcal{C}_2$ to denote the pixels that lead to boundary effects. As such:

$$\mathcal{C}_1 = \{\mathbf{x}_i \mid \mathbf{x}_i \in \mathcal{B}_1, \mathbf{x}_i + \mathbf{v} \notin \mathcal{I}\}; \qquad \mathcal{C}_2 = \{\mathbf{x}_i \mid \mathbf{x}_i \in \mathcal{B}_2, \mathbf{x}_i - \mathbf{v} \notin \mathcal{I}\} \tag{3.12}$$

If we predict frame 2 based on frame 1 (as Eq. 3.11 states), we will have a large error at $\mathcal{C}_1$. However, using Eq. 3.11, we have no problem in recovering pixels in $\mathcal{C}_2$. Reciprocally, if we reverse the temporal order – reconstructing frame 1 from frame 2 – only $\mathcal{C}_2$ has the boundary effect.

In a more rigid format, we denote the temporally-ordered saliency map that compares the predicted frame 2 with observed frame 2 as $\overrightarrow{s}(\mathbf{x}, t)$, and the saliency map using reversed sequence (predicting frame 1 from frame 2) as $\overleftarrow{s}(\mathbf{x}, t+1)$. We have:

$$\begin{aligned} &\overrightarrow{s}(\mathbf{x}_i, t) > \varepsilon, \quad \text{where } \mathbf{x}_i \in \mathcal{C}_1; \qquad &&\overleftarrow{s}(\mathbf{x}_i, t+1) \leq \varepsilon, \quad \text{where } \mathbf{x}_i \in \mathcal{C}_1 \\ &\overrightarrow{s}(\mathbf{x}_i, t) \leq \varepsilon, \quad \text{where } \mathbf{x}_i \in \mathcal{C}_2; \qquad &&\overleftarrow{s}(\mathbf{x}_i, t+1) > \varepsilon, \quad \text{where } \mathbf{x}_i \in \mathcal{C}_2, \end{aligned}$$

where ε is bounded by Eq. 3.10.

In an elegant form, we finally eliminate the boundary effect by combining the two maps:

$$s(\mathbf{x}, t) = \sqrt{\overrightarrow{s}(\mathbf{x}, t) \cdot \overleftarrow{s}(\mathbf{x}, t+1)} \tag{3.13}$$

For $\forall \mathbf{x}_i \in \mathcal{C}_1 \bigcup \mathcal{C}_2$, it is easy to see that $s(\mathbf{x}_i, t) \to 0$ as either $\overrightarrow{s}(\mathbf{x}_i, t) \to 0$, or $\overleftarrow{s}(\mathbf{x}, t+1) \to 0$. The saliency map generated by Eq. 3.13 is shown in Fig. 3.3-D.

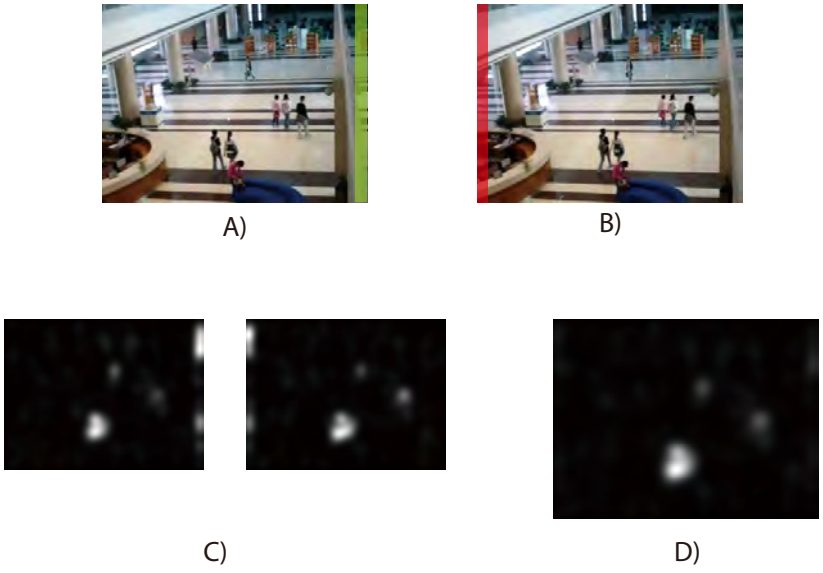


Figure 3.3: An illustration of the boundary effect. **A**) & **B**): Two adjacent frames. Green and red shadows in each frame indicate $\mathcal{C}_1$ and $\mathcal{C}_2$, respectively. **C**): The saliency map based on single-sided temporal order. Note that the border effect is as strong as the moving pedestrian in the center. **D**): The final saliency map.

3.3 Experiments

3.3.1 Implementing the phase discrepancy algorithm

In MATLAB, the phase discrepancy algorithm is:

```
FFT1=fft2(Frame1);
FFT2=fft2(Frame2);
Amp1=abs(FFT1);
Amp2=abs(FFT2);
Phase1=angle(FFT1);
Phase2=angle(FFT2);
mMap1=abs(ifft2((Amp2-Amp1).*exp(i*Phase1)));
mMap2=abs(ifft2((Amp2-Amp1).*exp(i*Phase2)));
mMap=mat2gray(mMap1.*mMap2);
```

`Frame1` and `Frame2` are consecutive frames. In our experiment, the size of image is grayscaled and shrunk to 120×160. On a 2.2GHz Core 2 Duo personal computer, this code performs at refresh rates as high as 75 frames per second.

One natural way to extend this algorithm to color images is to process each color channel separately, and combine saliency maps for each channel linearly. However, by tripling computational cost, the foreground pixels of color images do not seem to violate the intensity constancy assumption three times stronger than the grayscale image. Indeed, our observation is corroborated by experiments. A comparison experiment of color image detection is in Section.3.3.3. Since our algorithm emphasizes processing speed, we use grayscale images in most of our experiments.

We also notice that in real-world scenes, the intensity constancy assumption is subject to noises, such as background perturbation (moving leaves of a tree), sampling alias, or CCD noise. One way to reduce such noise is to combine the results from adjacent frames. However, we can only do so when the sampling rate is high enough such that the object motion in the saliency map is tolerable. In our experiments, we produce a reliable saliency map from 5 consecutive frames. At 20Hz, 5 frames takes about 0.25 second; this approach reduces the noise effectively without causing a drift in the salient region (see Fig.3.4).

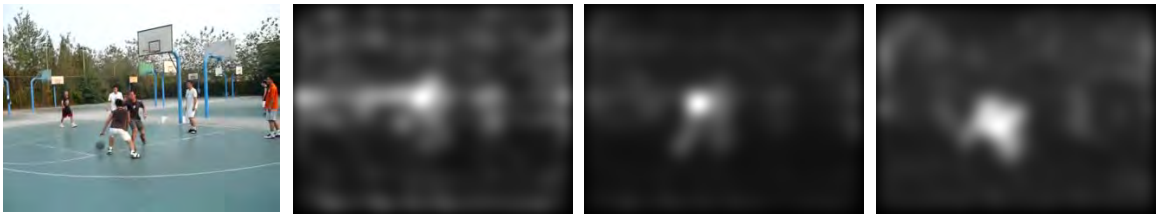


Figure 3.4: A comparison of combining the saliency maps of different frames. **A**): The saliency map computed by 2 frames. **B**): The saliency map by combining 5 frames (0.25 second). **C**:) The saliency map by combining 20 frames (1 second).

3.3.2 A new database for moving object detection

There are several public databases for evaluating motion detectors and trackers, such as PETS [52] and CAVIAR [53]. However, very few of them considered camera motion. In this section, we introduce a new database to evaluate the performance of a moving-object-detection algorithm.

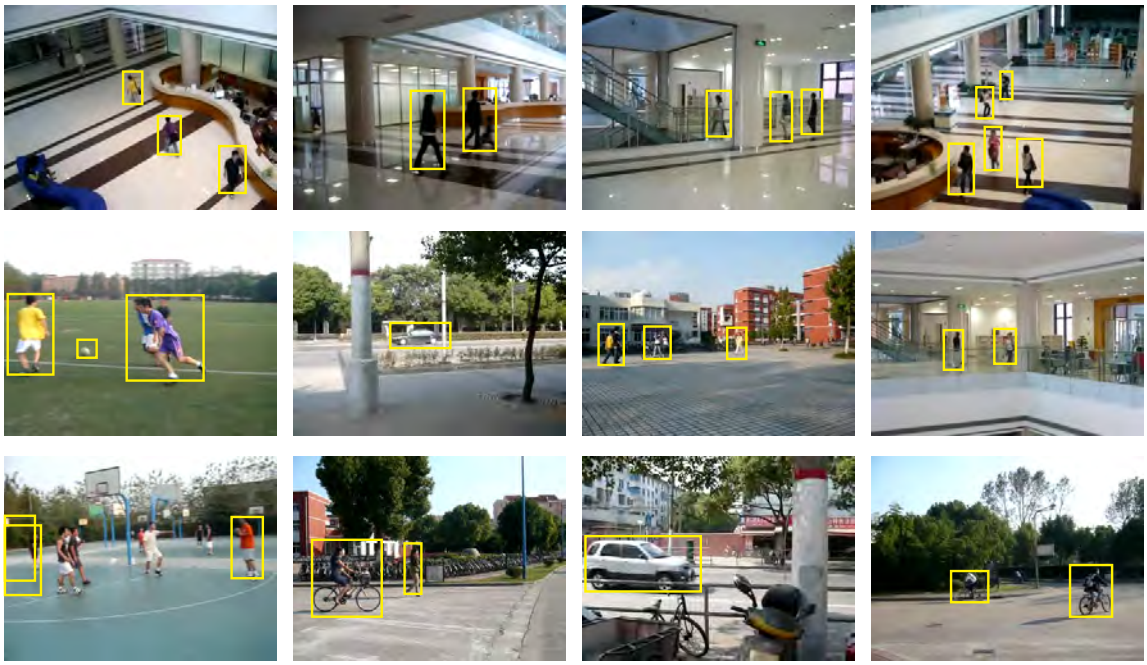


Figure 3.5: Sample frames of clips in the database of object-motion detection. Both scenes and moving objects vary from clip to clip.

Our database consists of indoor/outdoor scenes (see Fig.3.5). All clips were collected by a moving video camera under a 20-FPS sampling rate. Different categories of objects are included in

the video clip, such as walking pedestrians, cars and bicycles, and sports players. Given the high refresh rate, motion in adjacent frames is very similar. Therefore, it is unnecessary to label every frame. The original 20-FPS videos are given to our subjects for motion detection. For labeling, we asked each subject to draw bounding boxes on a small number of key frames by sub-sampling the sequence on a 0.5-second interval. Eleven naive subjects labeled all moving objects in the video. Some numbers from this database are in Table 3.1.

Items	Clips	Frames	Labelers	Key frames	Bounding boxes
Number	20	2557	11	297	4785

Table 3.1: A summary of our database.

The evaluation metric of the database is the same as PETS [54]. Although we have data from multiple subjects, the output of an algorithm is compared to one individual at a time. Let R_{GT} denotes the ground truth from the subject. The result generated by the algorithm is denoted as R_D. A detection is considered a true positive if:

$$\frac{Area(R_{GT} \cap R_D)}{Area(R_{GT} \cup R_D)} \geq Th, \tag{3.14}$$

The threshold Th defines the tolerance of a post-system that is connected to an object detector. If we use a loose criterion (Th is small), even a minimal overlap between the generated bounding box and ground-truth is considered a success. However, for many applications, a much higher overlap, equivalent to a much tighter criterion and a larger value of Th, is needed. In our experiments, we use $Th = 0.5$.

For the n^{th} clip, using the i^{th} subject as the ground-truth, we use GT_n^i, TP_n^i, FP_n^i to denote the number of ground-truth, true-positive, and false-positive bounding boxes, respectively. The Detection Rate (DR) and False Alarm Rate (FAR) are determined by:

$$DR_n = \frac{\sum_i TP_n^i}{\sum_i GT_n^i} \qquad FAR_n = \frac{\sum_i FP_n^i}{\sum_i TP_n^i + FP_n^i}. \tag{3.15}$$

In a frame where multiple bounding boxes are presented, finding the correct correspondence for Eq.3.14 can be very hard. Given a test bounding box, we simply compare it against every ground-truth bounding box, and pick the best match. Although this scheme does not guarantee that one ground-truth bounding box is used only once, in practice, confusions are rare.

3.3.3 Performance evaluation

To determine bounding boxes from the saliency map, an algorithm needs to know certain parameters such as spatial scale and sensitivity. To achieve a good performance without being trapped by parameter tuning, we use Non-Maximal Suppression [55] to localize the bounding boxes from the saliency map. This algorithm has three parameters $[\theta_1, \theta_2, \theta_3]$.

First, the algorithm finds all local maxima within a radius θ_1. Every local maximum greater than θ_2 is selected as the seed of a bounding box. Then, the saliency map is binarized by threshold θ_3. Finally, the rectangular contour that encompasses the white region surrounding every seed is considered as a bounding box.

It is straightforward to assume that the parametrization is consistent over different clips in our database, and the locations of objects are independent among different clips. Therefore, we use cross-validation to avoid over-fitting the model. In each iteration, we take 19 clips as the training set to find the parameters that maximize:

$$\sum_{m \in \{training\}} DR_m(1 - FAR_m),$$

And use the remaining clip to test the performance. The final results of DR and FAR are the average among different clips. Samples of detected objects are shown in Fig.3.6. The quantitative result of our model is listed in Table 3.2.

3.3.4 Comparison to other methods

To evaluate the performance of our algorithm, four representative algorithms are introduced to give comparative results on our database: the Mixture of Gaussian model [41], the Dynamic Visual Attention model [4], the Bayesian Surprise model [48], and the Bottom-up Saliency model with flicker channel [2]. MATLAB/C++ implementations of all these algorithms are available on authors' websites. Examples of the generated saliency maps are shown in Fig. 3.7. As for the quantitative experimental part, the parameters of Non-Maximal Suppression is trained in the same way as we described in Section 3.3.3 to generate bounding boxes from the saliency maps. The quantitative results are shown in Table 3.2. Our phase discrepancy model is the best in detecting moving objects.

It is worth noting that not all of these algorithms are designed to detect moving objects in a dynamic scene. In fact, the performance of an algorithm is determined by how well its underlying

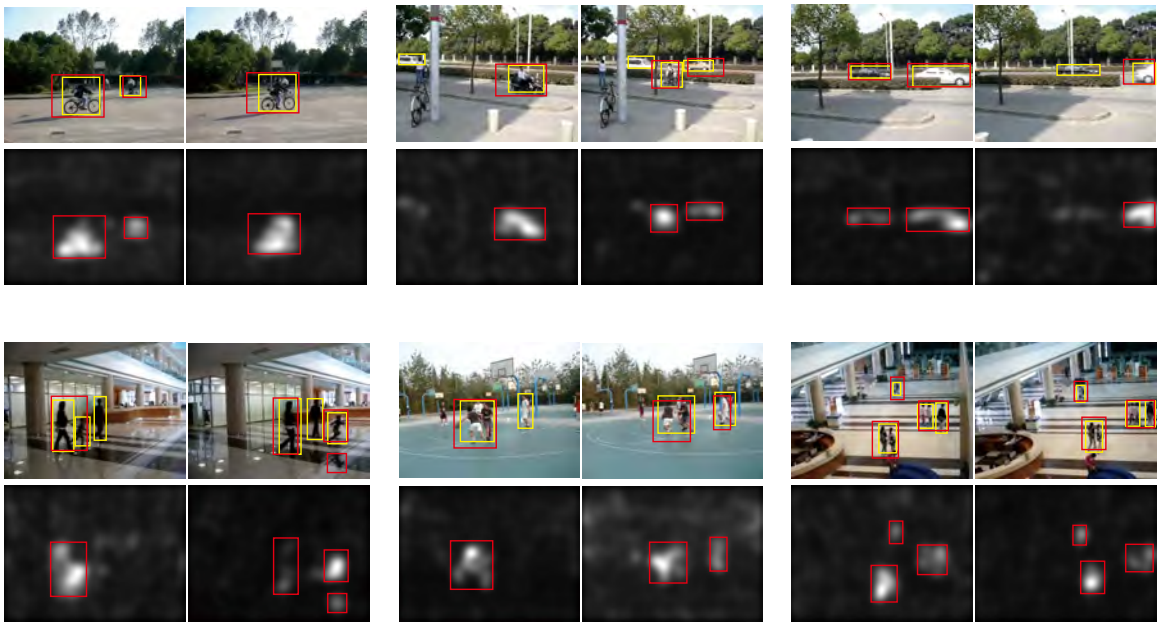


Figure 3.6: Saliency maps and the ground-truth bounding boxes. In each image/saliency map pair, red bounding boxes are generated by our algorithm. Yellow bounding boxes are the ground-truth drawn by a human subject.

hypothesis is consistent with the data. In our database, an "object" is defined by its motion in contrast to the background. There is no assumption such as objects possessing unique features, or the background being monotonous. Therefore, it is not surprising that some algorithms did not perform very well in this experiment.

	Detection Rate	False Alarm Rate
Human average	0.84 ± 0.08	0.15 ± 0.08
Our model	0.46 ± 0.14	0.58 ± 0.24
Our model (color)	0.48 ± 0.18	0.57 ± 0.24
Dynamic Visual Attention [4]	0.32±0.22	0.86±0.10
Bayesian Surprise [48]	0.12±0.09	0.92±0.04
Saliency [2]	0.09±0.08	0.98±0.01
Mixture of Gaussian [41]	0.00±0.00	1.00±0.00

Table 3.2: Detection performance of human average and our model.

3.3.5 Database consistency

The motivation behind the analysis of database consistency comes from the fact there is no objective "ground-truth" for moving object detection. Although the ground-truth consistency issue is not

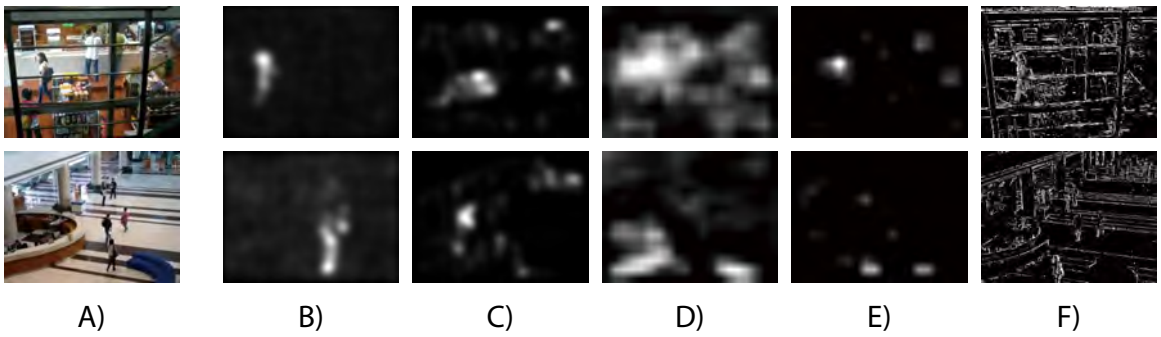


Figure 3.7: Saliency maps generated by different algorithms. **A**): Original image. **B**): Our model. **C**): Dynamic Visual Attention [4]. **D**): Bayesian Surprise [48]. **E**): Saliency [2]. **F**): Mixture of Gaussian [41].

widely concerned in the object detection and tracking databases, List et al. [56] analyzed the statistical variation in the hand label data of CAVIAR [53], and showed that inter-subject variability can compromise benchmark results. In our database, we also observed that the same video clip can be interpreted in different ways. For instance, in Fig. 3.8.A, some subjects label multiple players as one group, yet other subjects label every individual as one object.

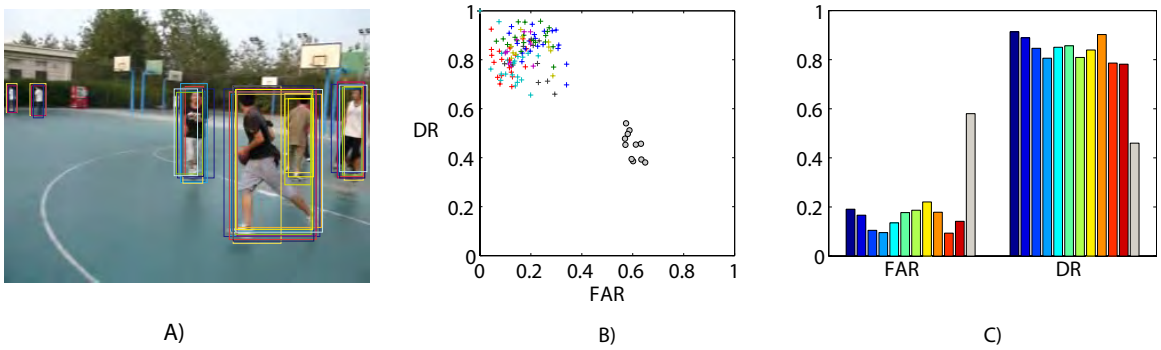


Figure 3.8: **A**): Different interpretations of moving objects by different subjects. This image overlays the bounding boxes of 11 subjects. Boxes in the same color are drawn by the same person. We see that the incongruence among different subjects is not negligible. **B**): The FAR-DR plot of all subjects and our algorithm. Each + in the same color represents the assessment of the same subject. Each $\circ$ indicates the performance of our algorithm. Among different subjects, the DR fluctuates from 0.65 to 1, whereas the FAR fluctuates from 0 to 0.4. The average human performance is $FAR = 0.15 \pm 0.08$, $DR = 0.84 \pm 0.08$.**C**): Color bars indicate the FAR and DR for the subjects. The gray bars are the performance of our algorithm.

A good benchmark should have consistent labels across subjects. To evaluate the consistency of our database, we assess the performance of the i^{th} subject based on the j^{th} subject's ground-

truth. Therefore, for each individual, we have 10 points on the FAR-DR plot. As a comparison, the performance of our algorithm is also provided. Each data point is generated by selecting one individual as the ground-truth, and perform cross-validation over 20 trials. The result is shown in Fig.3.8.

From these results, we see that even a human subject cannot achieve perfect detection. In other words, a computer algorithm is "good enough" if its performance has the same distribution as humans' on the FAR-DR plot.

3.3.5.1 Threshold and accuracy tolerance

Note in Eq. 3.14, the choice of $Th = 0.5$ is arbitrary. This parameter determines the detection tolerance. To evaluate Th's influence, FAR and DR are computed as functions of Th (see Table 3.3).

Th		0	0.1	0.2	0.3	0.4
Human Detection Rate		0.92	0.91	0.91	0.90	0.88
Human False Alarm Rate		0.07	0.07	0.07	0.08	0.11
Model Detection Rate		0.83	0.82	0.80	0.75	0.63
Model False Alarm Rate		0.18	0.20	0.24	0.31	0.43
Th	0.5	0.6	0.7	0.8	0.9	1.0
Human Detection Rate	0.84	0.77	0.62	0.37	0.15	0.00
Human False Alarm Rate	0.15	0.22	0.37	0.62	0.85	1.00
Model Detection Rate	0.46	0.20	0.07	0.02	0.00	0.00
Model False Alarm Rate	0.58	0.80	0.93	0.98	1.00	1.00

Table 3.3: Human average (DR,FAR) and model average (DR,FAR) with respect to threshold.

3.3.5.2 The influence of object sizes

As we have shown in Eq. 3.10, the upper bound of error is a function of object size. To provide a empirical validation of our algorithm performance on large objects, we selected 2 clips in our database that contained the biggest objects, and tested our algorithm. The average area of the foreground objects is 10% of the image size (compared to 5% of the original experiment). The new performance is shown in Table 3.4.

	Original experiment	Clips with large objects
Detection Rate	0.46 ± 0.14	0.41 ± 0.14
False Alarm Rate	0.58 ± 0.24	0.65 ± 0.08

Table 3.4: The algorithm performance over a large object database. The performance drop is small.

3.4 Discussions

3.4.1 Sources of errors

One of the challenges is to estimate the bounding boxes for adjacent, sometimes occluded objects that move in the same direction (such as in Fig. 3.6.F). To unravel the complexity of multiple moving objects, either long term tracking, or a more powerful segmentation from saliency map to bounding boxes is required.

In some cases, we also need to incorporate top-down modulations from a level of object recognition. Since the saliency map is a pixel-based representation, it favors moving parts of an object (such as a waving hand) over the entire object. A canonical interesting example is in Fig. 3.6.D: our algorithm identifies the reflection on the floor as an object, yet none of our subjects labeled the reflection as an object.

3.4.2 Connections to spectral saliency

In 2007, Hou et al. proposed the Spectral Residual theory [20]. In 2012, the same topic is revisited by Hou et al. in [1], which is discussed in Chapter 2 of this thesis. This algorithm takes the phase part of the spectrum of an image, and does the inverse transform. In other words, the saliency map generated by the Spectral Residual is the asymptotic limit of Phase Discrepancy when the second frame has $\mathbf{v} \rightarrow 0^+$. However, $\mathbf{v} \rightarrow 0^+$ is ill-defined in the problem of Phase Discrepancy; as the displacement approaches infinitesimal, no motion information is available, and the problem of moving-object detection degrades to the problem of still-image figure-ground separation.

3.4.3 Concluding remarks

In this chapter, we propose a new algorithm for motion detection with a moving camera in the Fourier domain. We define a new concept named Phase Discrepancy to explore camera motions. The spectrum energy of an image is generally dominated by its background. Based on this observation, we derive an approximation algorithm to extract the Phase Discrepancy. A simple motion saliency map generation algorithm is introduced to detect moving foreground regions. The saliency map is constructed by the Inverse Fourier Transform of the difference of two successive frames' spectrum energies, keeping the phase of two images invariant. The proposed algorithm does not rely on prior training on a particular feature or categories of an image. A large number of computer simulations

are performed to show the strong performance of the proposed method for motion detection.

Chapter 4

From Fixations to Salient Object Segmentation

Abstract

In this chapter, we provide an extensive evaluation of fixation prediction and salient object segmentation algorithms, as well as statistics of major datasets. Our analysis identifies serious design flaws of existing salient object benchmarks, called the *dataset design bias*, by over-emphasising the stereotypical concepts of saliency. The dataset design bias does not only create the discomforting disconnection between fixations and salient object segmentation, but also misleads the algorithm designing.

Based on our analysis, we propose a new high-quality dataset that offers both fixation and salient object segmentation ground-truth. With fixations and salient objects being presented simultaneously, we are able to bridge the gap between fixations and salient objects, and propose a novel method for salient object segmentation. Finally, we report significant benchmark progress on 3 existing datasets of segmenting salient objects.

4.1 Introduction

Bottom-up visual saliency refers to the ability to select important visual information for further processing. The mechanism has proven to be useful for human as well as computer vision. Unlike other topics such as object detection/recognition, saliency is not a well-defined term. Most of the works in computer vision focus on one of the following two specific tasks of saliency: fixation prediction and salient object segmentation.

In a fixation experiment, saliency is expressed as eye gaze. Subjects are asked to view each

image for seconds while their eye fixations are recorded. The goal of an algorithm is to compute a probabilistic map of an image to predict the actual human eye gaze patterns. Alternatively, in a salient object segmentation dataset, image labelers annotate an image by drawing pixel-accurate silhouettes of objects that are believed to be salient. Then, the algorithm is asked to generate a map that matches the annotated salient object mask.

Various datasets of fixation and salient object segmentation have provided objective ways to analyze algorithms. However, existing methodology suffers from two major limitations: 1) algorithms focusing on one type of saliency tend to overlook the connection to the other side, and 2) benchmarking primarily on one dataset tends to overfit the inherent bias of that dataset.

In this chapter, we explore the connection between fixation prediction and salient object segmentation by augmenting 850 existing images from the PASCAL 2010 [57] dataset with eye fixations and salient object segmentation labeling. In Sec. 4.3, we argue that by making the image acquisition and image annotation *independent* from each other, we can avoid *dataset design bias*, a specific type of bias that is caused by experimenters' unnatural selection of dataset images.

With fixations and salient object labels simultaneously presented in the same set of images, we report a series of interesting findings. First, we show that salient object segmentation is a valid problem because of the high consistency among labelers. Second, unlike fixation datasets, the most widely used salient object segmentation dataset is heavily biased. As a result, all top-performing algorithms for salient object segmentation have poor generalization power when they are tested on more realistic images. Finally, we demonstrate that there exists a strong correlation between fixations and salient objects.

Inspired by these discoveries, in Sec. 4.4 we propose a new model of salient object segmentation. By combining existing fixation-based saliency models with segmentation techniques, our model bridges the gap between fixation prediction and salient object segmentation. Despite its simplicity, this model significantly outperforms state-of-the-arts salient object segmentation algorithms on all 3 salient object datasets.

4.2 Related Works

In this section, we briefly discuss existing models of fixation prediction and salient object segmentation. We also discuss the relationship of salient object to generic object segmentation such as CPMC [58, 32]. Finally, we review relevant research pertaining to dataset bias.

4.2.1 Fixation prediction

The problem of fixation based bottom-up saliency is first introduced to the computer vision community by [2]. The goal of this type of models is to compute a "saliency map" that simulates the eye movement behaviors of human. Patch-based [2, 3, 35, 4, 34] or pixel-based [1, 5] features are often used in these models, followed by a local or global interaction step that re-weights or re-normalizes features saliency values.

To quantitatively evaluate the performance of different fixation algorithms, ROC Area Under the Curve (AUC) is often used to compare a saliency map against human eye fixations. One of the first systematic datasets in fixation prediction was introduced in [3]. In this paper, Bruce *et al.* recorded eye fixation data from 21 subjects on 120 natural images. In a more recent paper [31], Judd *et al.* introduced a much larger dataset with 1003 images and 15 subjects.

Due to the nature of eye tracking experiments, the error of recorded fixation locations can go up to 1°, or over 30 pixels in a typical setting. Therefore, there is no need to generate a pixel-accurate saliency map to match human data. In fact, as pointed out in [1], blurring a saliency map can often increase its AUC score.

4.2.2 Salient object segmentation

It is not an easy task to directly use the blurry saliency map from a fixation prediction algorithm. As an alternative, Liu *et al.* [59] proposed the MSRA-5000 dataset with bounding boxes on the salient objects. Following the idea of "object-based" saliency, Achanta *et al.* [60] further labeled 1000 images from MSRA-5000 with pixel-accurate object silhouette masks. Their paper showed that existing fixation algorithms perform poorly if benchmarked F-measures of PR curve. Inspired by this new dataset, a line of papers has proposed [61, 62, 63] to tackle this new challenge of predicting full-resolution masks of salient objects. An overview of the characteristics and performances of salient object algorithms can be found in a recent review [64] by Borji *et al.*

Despite the deep connections between the problems of fixation prediction and object segmentation, there is a discomforting isolation between major computational models of the two types. Salient object segmentation algorithms have developed a set of techniques that have little overlapping with fixation prediction models. This is mainly due to a series of differences in the ground-truth and evaluation procedures. A typical fixation ground-truth contains several fixation dots, while a salient object ground-truth usually has one or several positive regions composed of thousands of

pixels. Having different priors of sparsity significantly limited the model of one type in having good performance on tasks of the other type.

4.2.3 Objectness, object proposal, and foreground segments

In the field of object recognition, researchers are interested in finding objects independent of their classes [65]. Alexe *et al.* [66] used a combination of low/mid-level image cues to measure the "objectness" of a bounding box. Other models, such as CPMC [58, 32] and Object Proposal [67], generate segmentations of candidate objects without relying on category specific information. The obtained "foreground segments," or "object proposals," are then ranked or scored to give a rough estimate of the objects in the scene.

The role of a scoring/ranking function in the aforementioned literature shares a lot of similarities with the notion of saliency. In fact, [66] used saliency maps as a main feature for predicting objectness. In Sec. 4.4, we propose a model based on the foreground segmentations generated by CPMC. One fundamental difference between these methods in regard to visual saliency is that an object detector is often exhaustive – it looks for *all* objects in the image irrespective of their saliency value. In comparison, a salient object detector aims at enumerating a subset of objects that exceed a certain saliency threshold. As we will discuss in Sec. 4.4, an object model, such as CPMC offers a ranking of its candidate foreground proposals. However, the top ranked (e.g. first 200) segments do not always correspond to salient objects or their parts.

4.2.4 Datasets and dataset bias

Recently, researchers started to quantitatively analyze dataset bias and its detrimental effect in benchmarking. Dataset bias arises from the selection of images [68] as well as the annotation process [33]. In the field of visual saliency analysis, the most significant bias is *center bias*. It refers to subjects' tendency to look more often at the center of the screen [33]. This phenomenon might be partially due to experimental constraints such as a subject's head being on a chin-rest during the fixation experiment, and partially due to the photographer's preference to align objects at the center of the photos.

Center bias has been shown to have a significant influence on benchmark scores [31, 34]. Fixation models either use it explicitly [31] or implicitly by padding the borders of a saliency map [35, 4, 34]. To make a fair evaluation of the algorithm's true prediction power, Tatler [33] proposed

a shuffled-AUC (s-AUC) score to normalize the effect of center-bias. In s-AUC, positive samples are taken from the fixations of the test image, whereas the negative samples are from all fixations across all other images.

4.3 Dataset Analysis

In this chapter, we will benchmark on the following datasets: Bruce [3], Judd [31], Cerf [30], FT [60], and IS [25]. Among these 5 datasets, only Judd and Cerf provide fixation ground-truth. FT only provides salient object ground-truth. IS provides both fixations[1] as well as salient object masks. While the Bruce dataset was originally designed for fixation prediction, it was recently augmented by [69] with 70 subjects under the instruction to *label the single most salient object in the image*. In our comparison experiment, we include the following fixation prediction algorithms: ITTI [2], AIM [3], GBVS [60], DVA [4], SUN [34], SIG [1], AWS [5]; and the following salient object segmentation algorithms: FT [60], GC [61], SF [62], and PCAS [63]. These algorithms are top-performing ones in major benchmarks [64].

4.3.1 Psychophysical experiments on the PASCAL-S dataset

Our PASCAL-S dataset is built on the validation set of the PASCAL VOC 2010 [57] segmentation challenge. This subset contains 850 natural images. In the fixation experiment, 8 subjects were instructed to perform the "free-viewing" task to explore the images. Each image was presented for 2 seconds, and eye-tracking re-calibration was performed on every 25 images. The eye gaze data was sampled using the Eyelink 1000 eye-tracker at $125Hz$. In the salient object segmentation experiment, we first manually perform a full segmentation to crop out all objects in the image. An example segmentation is shown in Fig. 4.1.B. When we build the ground-truth of full segmentation, we adhere to the following rules: 1) we do not intentionally label parts of the image (e.g. faces of a person); 2) disconnected regions of the same object are labeled separately; 3) we use solid regions to approximate hollow objects, such as bike wheels.

We then conduct the experiment of 12 subjects to label the salient objects. Given an image, a subject is asked to select the salient objects by clicking on them. There is no time limitation or constraints on the number of objects one can choose. Similar to our fixation experiment, the

[1] IS provides raw gaze data at every time point. We use the following thresholds to determine a stable fixation: min fixation duration: $160ms$, min saccade speed: $50px/100ms$.

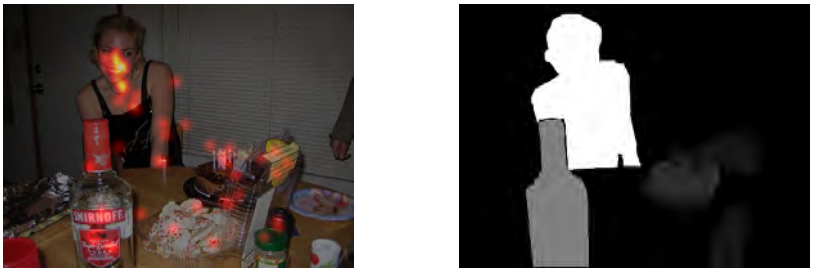


Figure 4.1: An illustration of PASCAL-S dataset. Our dataset provides both eye fixation (Fig. C) and salient object (Fig. D) mask. The labeling of salient objects is based on the full segmentation (Fig. B). A notable difference between PASCAL-S and its predecessors is that each image in PASCAL-S is labeled by multiple labelers without restrictions on the number of salient objects.

instruction of labeling salient objects is intentionally kept vague. The final saliency value of each segment is the total number of clicks that it receives, divided by the number of subjects.

4.3.2 Evaluating dataset consistency

Quite surprisingly, many of today's widely used salient object segmentation datasets do not have any guarantee on the inter-subject consistency. To compare the level of agreement among different labelers in our PASCAL-S dataset and other existing dataset, we randomly select 50% of the subjects as the test subset. Then, we benchmark the saliency maps of this test subset by taking the rest subjects as the new ground-truth subset. For the fixation task, the test saliency map for each image is obtained by first plotting all the fixation points from the test subset, and then filtering the saliency map by a 2D Gaussian kernel with $\sigma = 0.05$ of the image width. For the salient object segmentation task, the test/ground-truth saliency maps are binary maps obtained by first averaging the individual segmentations from the test/ground-truth subset, and then threshold with $Th = 0.5$[2] to generate the binary masks for each subset. Then, we compute either the AUC score or the F-measure of the test

[2] At least half of the subjects within the subset agree on the mask.

subset, and use this number to indicate the inter-subject consistency.

We notice that the segmentation maps of the Bruce dataset are significantly sparser than maps in PASCAL-S or IS. Over 30% of the segmentation maps of the Bruce dataset are completely empty. This is likely a result of the labeling process. In Borji *et al.*'s experiment [69], the labelers are forced to choose only one object for each image. Images with two or more equally salient objects are very likely to become empty after thresholding. Although Bruce is one of the very few datasets that offer both fixations and salient object masks, it is not suitable for our analysis.

For datasets PASCAL-S and IS, we benchmark the F-measure of the test subset segmentation maps by the ground-truth subset. The result is shown in Tab. 4.1.

AUC scores				
PASCAL-S	Bruce	Cerf	IS	Judd
0.835	0.830	0.903	0.836	0.867

F-measures	
PASCAL-S	IS
0.972	0.900

Table 4.1: Inter-subject consistency of 2 salient object segmentation datasets and 5 fixation datasets.

Similar to our consistency analysis of salient object dataset, we evaluate the consistency of eye fixations among subjects (Tab. 4.1). Even though the notion of "saliency" under a context of a complex natural scene is often considered as ill-defined, we observe highly consistent behaviors among human labelers in both eye-fixation and salient object segmentation tasks.

4.3.3 Benchmarking

In this section, we benchmark 7 fixation algorithms: AWS [5], AIM [3], SIG [1], DVA [4], GBVS [35], SUN [34], and ITTI [2] on 5 datasets: Bruce [3], Cerf [30], IS [25], Judd [31], and our PASCAL-S. For salient object segmentation, we bench 4 algorithms: SF [62], PCAS [63], GC [61], and FT [60] on 4 datasets: FT [60], IS [25], and our PASCAL-S. For all algorithms, we use the original implementations from the authors' websites. The purposes of this analysis are: 1) to highlight the generalization power of algorithms, and 2) to investigate inter-dataset difference among these independently constructed datasets. The benchmark results are presented in Fig. 4.2. Sharply contrasted to the fixation benchmarks, the performance of all salient object segmentation algorithms drops significantly when migrating from the popular FT dataset. The *average* performance

of all 4 algorithms has dropped, from FT's 0.8341, to 0.5765 (30.88% drop) on IS, and 0.5530 (33.70% drop) on PASCAL-S. This result is alarming, because the magnitude of the performance drop from FT to any dataset by any algorithm, can easily dwarf the 4-year progress of salient object segmentation on the widely used FT dataset. Moreover, the relative ranking among algorithms also changes from one dataset to another.

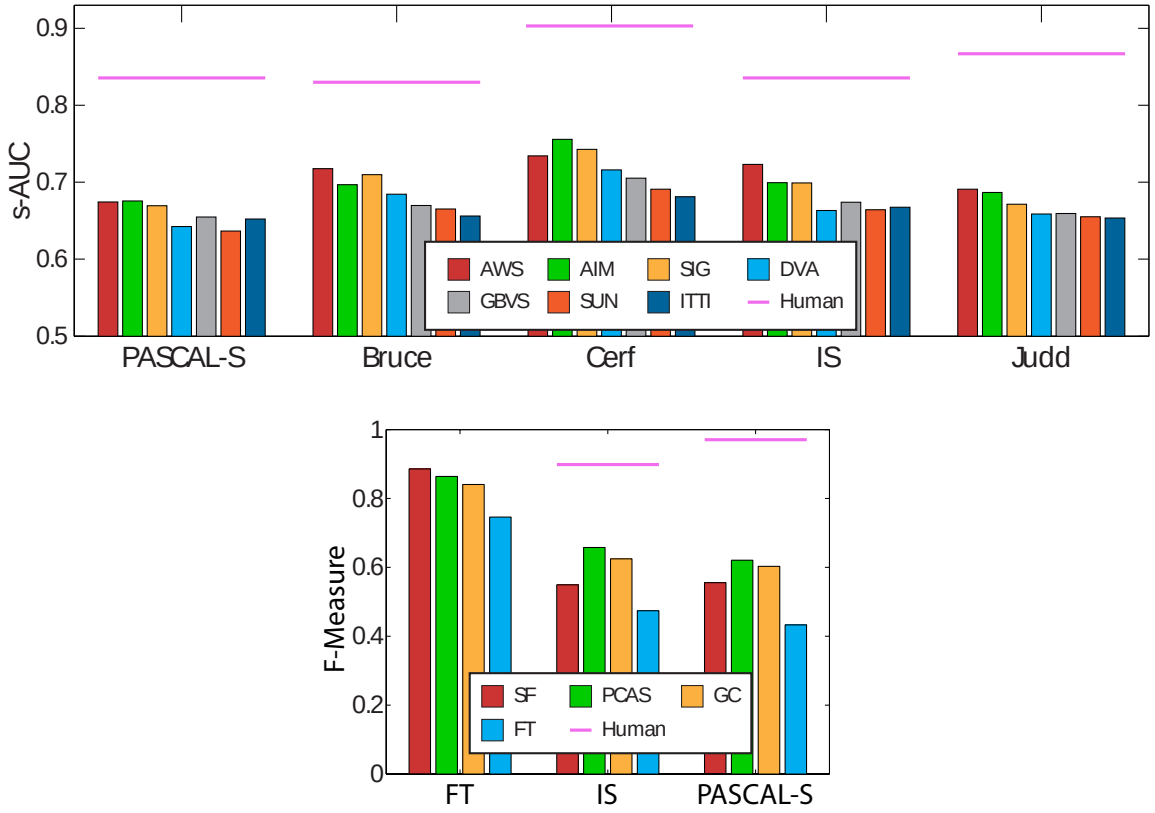


Figure 4.2: **Left**: The s-AUC scores of fixation prediction. **Right**: The F-Measure scores of salient object segmentation. According to [60], we choose $\beta = 0.3$ to calculate the F-measure from PR curve. In both figures, magenta lines show the inter-subject consistency score of these datasets. These numbers can be interpreted as the upper-bounds of algorithm scores.

4.3.4 Dataset design bias

The performance gap among datasets clearly suggests new challenges in salient object segmentation. However, it is more important to pinpoint the cause to the performance degradation rather than to start a benchmark race on another new dataset. In this section, we analyze the following image statistics in order to find the similarities and differences of today's salient object segmentation

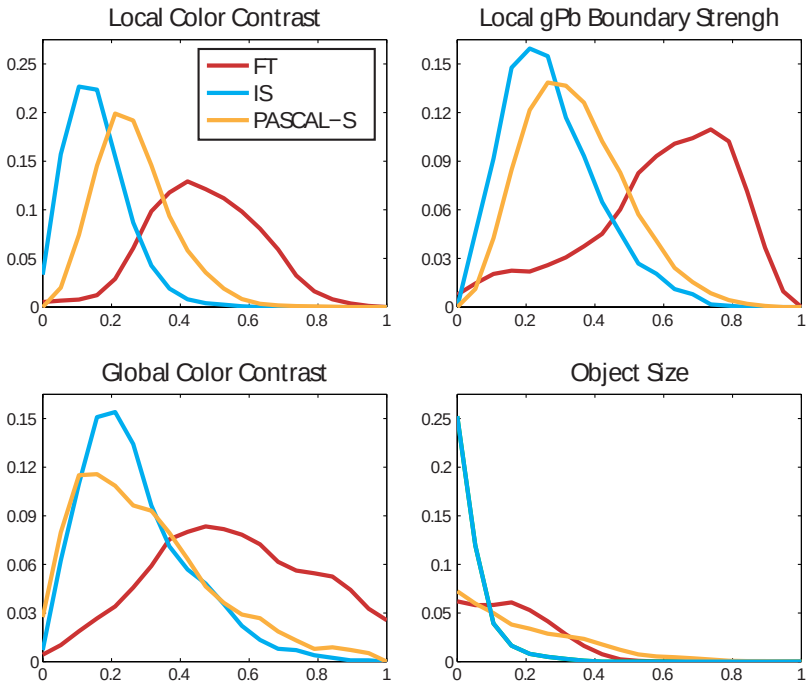


Figure 4.3: Image statistics of the salient object segmentation datasets. The statistics of the FT dataset is different from other datasets in local/global color contrast and boundary strength. As for object size, the PASCAL-S contains a balanced mix of large and small objects.

datasets:

Local color contrast: Segmentation or boundary detection is an inevitable step in most salient objects detectors. It is important to check whether the boundaries are "unnaturally" easy to segment. To estimate the strength of the boundary, we crop a 5×5 image patch at the boundary location of each labeled object, and compute RGB color histograms for foreground and background separately. We then calculate the χ^2 distance to measure the distance between two histograms. It is worth noting that some of the ground-truths of the IS dataset are not perfectly aligned with objects' boundaries, resulting in an underestimate of local contrast magnitude.

Global color contrast: The term "saliency" is also related to the global contrast of the foreground and background. Similar to the local color contrast measure, for each object, we calculate the χ^2 distance between its RGB histogram and the background RGB histogram.

Local gPB boundary strength: While color histogram distance captures some low-level image features, an advanced boundary detector, such as gPB [70], combines local and global cues

to give a more comprehensive estimate of the presence of a boundary. As a complementary result to our local color contrast measure, for the object boundary pixel, we compute the mean gPB response of a 3×3 local patch.

Object size: In each image, we define the size of an object as the proportion of pixels in the image. Most of the objects in IS are very small.

As shown in Fig. 4.3, the FT dataset stands out in local/global color contrast as well as the gPB boundary strength statistics. At first glance, our observation that FT contains unnaturally strong object boundaries seems acceptable, especially for a dataset focusing on salient object analysis. Strong boundaries are linked to the core concept of saliency: a foreground object with discernable boundaries being surrounded by a background that has contrastive colors. In fact, many images in the FT dataset are textbook examples to demonstrate the definition of saliency. The notion of "saliency" in FT is much less ambiguous than in other datasets. However, such reduction of ambiguity during the *image selection process* is more destructive rather than constructive, for the purpose of testing saliency. The confusion between the image-selection process and the image annotation process introduces a special kind of bias by over-expressing the desired properties of the target concept, and reducing the presence of negative examples. We call this type of bias the *dataset design bias*:

> During the design of a dataset, the image annotation process should be independent of the image selection process. Otherwise, the *dataset design bias* will arise as a result of disproportionate sampling of positive/negative examples.

4.3.5 Fixations and F-measure

Previous methods of salient object segmentation have reported a big margin of F-measures over all fixation algorithms. However, most of these comparisons are done on the FT dataset, which has shown to have non-negligible dataset bias. Another factor that contributes to the inferior performance of fixation algorithms is the center-bias. Major salient object segmentation algorithms, such as SF, PCAS, and GC, have implemented their own treatments for the center-bias. In contrast, many fixation prediction algorithms, such as AWS and SIG, do not implement center-bias as they expect to be benched by s-AUC score, which cancels the center-bias effect. To discount the influence of center-bias, we add a fixed Gaussian ($\sigma = 0.4$ of the image width) to the saliency maps generated

by all fixation algorithms, and then benchmark these algorithms on all 3 salient object datasets. The result is shown in Fig. 4.6.

In addition, we also tested the F-measure of the ground-truth human fixation maps on IS and PASCAL-S. Each fixation map is blurred by a Gaussian kernel with $\sigma = 0.03$ of the image width. No center-bias is superimposed, because the human fixations are already heavily biased towards the center of the image.

When we remove the effect of center-bias and dataset design bias, the performance of fixation algorithms becomes very competitive. We also notice that the F-measure of the ground-truth human fixation maps is rather low compared to the inter-subject consistency scores in the salient object labeling experiment. This performance gap could be either due to a weak correlation between the fixation task and the salient object labeling task, or the incompatibility of the representation of fixations (dots) versus the representation of objects (regions). In the next section, we will show that the latter is more likely to be true. Once equipped with appropriate underlying representation, the human fixation map, as well as its algorithm approximations, generates accurate results for salient object segmentation.

4.4 From Fixations to Salient Object Detection

Many of today's well-known salient object algorithms have the following two components: 1) a suitable representation for salient object segmentation, and 2) computational principles of feature saliency, such as region contrast [61] or element uniqueness [62]. However, neither of these two components alone is new to computer vision. On one hand, detecting boundaries of objects has been a highly desired goal for segmentation algorithms since the beginning of computer vision. On the other hand, defining rules of saliency has been studied in fixation analysis for decades. In this section, we build a salient object segmentation model by combining existing techniques of segmentation- and fixation-based saliency. The core idea is to first generate a set of object candidates, and then use the fixation algorithm to rank different regions based on their saliency. This simple combination results in a novel salient object segmentation method that outperforms *all* previous methods by a large margin.

4.4.1 Salient object, object proposal and fixations

Our first step is to generate the segmentation of object candidates by a generic object proposal method. We use CPMC [58] to obtain the initial segmentations. CPMC is an unsupervised framework to generate and rank plausible hypotheses of object candidates without category-specific knowledge. This method initializes foreground seeds uniformly over the image and solves a set of min-cut problems with different parameters. The output is a pool of object candidates as overlapping figure-ground segments together with their "objectness" scores. The goal of CPMC is to produce a over-complete coverage of potential objects, which could be further used for tasks such as object recognition.

The representation of CPMC-like object proposal is easily adapted to salient object segmentation. If all salient objects can be found from the pool of object candidates, we can reduce the problem of salient object detection to a much easier problem of salient segment ranking. Ranking the segments also simplifies the post-processing step. As the segments already preserve the boundary of the image, no explicit segmentation (e.g. GraphCut [61]) is required to obtain the final binary object mask.

To estimate the saliency of a candidate segment, we utilize the spatial distribution of fixations within the object. It is well known that the density of fixation directly reveals the saliency of the segment. The non-uniform spatial distribution of fixations on the object also offers useful cues to determine the saliency of an object. For example, fixation at the center of a segment will increase its probability of being an object. To keep our frame work simple, we do not consider class-specific or subject-specific fixation patterns in our model.

4.4.2 The model

We use a learning-based framework for the segment selection process. This is achieved by learning a scoring function for each object candidate. Given a proposed object candidate mask and its fixation map, this function estimates the overlapping score (intersection over union) of the region with respect to the ground-truth, similar to [32]. We use a random regression forest to learn the scoring function. A random regression forest is an ensemble of decision trees. For each branch node, a feature is selected from a random subset of all features, and a decision boundary is set by minimizing the Minimum Square Error (MSE). The leaf nodes keep the mean value of all training samples that end up in the node. The final result is a weighted average of all leaf nodes that a testing

Shape Features	Dims
Area	1
Centroid	2
Convex Area	1
Euler Number	1
Perimeter	1
Major/Minor Axis Length	2
Eccentricity	1
Orientation	1
Equivalent Diameter	1
Solidity	1
Extent	1
Width/Height	2
Fixation Features	Dims
Min/Max Fixation Energy	2
Mean Fixation Energy	1
Weighted Fixation Centroid	2
Fixation Energy Ratio	1
Histogram of Fixations	12

Table 4.2: Shape and fixation features used in our model.

sample reaches. We choose random forest since our feature vector contains discrete values (Euler Number), which can be easily handled in a decision tree. For all our experiments, we train a random forest using 30 trees, where each node uses 6 feature dimensions.

We extract two types of features: shape features and fixation distribution features within the object. The shape features characterize the binary mask of the segment, which includes major axis length, eccentricity, minor axis length, and the Euler number. For fixation distribution features, we first align the major axis of each object candidate, and then extract a 4×3 histogram of fixations density over the aligned object mask. This histogram captures the spatial distribution of the fixations within a object. Finally, the 33-dimensional feature vector is extracted for each object mask. The details of the 33-dimensional feature can be found in Table 4.2. In particular, the *Fixation Energy Ratio* is the defined as the sum of fixation energy within the segment, divided by the sum of fixation energy of the whole image.

For each dataset, we train a random forest with 30 trees, using a random sampling of 40% of the images. The rest of the images are used for testing. The results are averaged on a 10-fold random split of the training and testing set. We use a random regression forest to predict the saliency score of an object mask. In the testing phase, each segment is classified independently. We generate the

salient object segmentation by averaging the top-K segments at pixel level. We then use simple thresholding to generate the final object masks. As our saliency scores are defined over image segments, this simple strategy leads to fairly good object boundaries.

Note that no appearance feature is used in our method, because our goal is to demonstrate the connection between fixation and salient object segmentation. Our algorithm is independent of the underlying segmentation and fixation prediction algorithms, allowing us to switch between different fixation algorithms or even human fixations.

4.4.3 Limits of the model

Our model contains two separate parts: a segmenter that proposes regions, and a selector that gives each region a saliency score. In this section, we explore the limitation of our model by replacing each part at a time. First, we quantify the performance upper-bound of the selector, and then, the best achievable results of the segmenter are also presented.

To test the upper-bound of the selector, we train our model on the ground-truth segments of PASCAL-S (e.g. Fig. 4.1.B) with human fixation maps. With a perfect segmenter, this model can accurately estimate the saliency of segments using fixation and shape information. On the test set, it achieves a F-Measure of 0.9201 with $P = 0.9328$ and $R = 0.7989$. This result is a strong validation to our motivation, which is to bridge the gap between fixations and salient objects. It is worth mentioning that this experiment requires a full segmentation of all objects of the entire dataset. Therefore, PASCAL-S is the only dataset that allows us to test the selector with an ideal segmenter.

Second, we test the upper-bound performance of the CPMC segmentation algorithm. We match each segment from CPMC to the ground truth object annotations, and greedily choose the segments (out of the first 200 segments) with the best overlapping scores. Again, the result is very positive. With the first 200 segments, the best CPMC segments achieved an F-Measure of 0.8699 ($P = 0.8687$, $R = 0.883$) on PASCAL-S dataset. Similar results are observed in FT (F-Measure = 0.9496, $P = 0.9494$, $R = 0.9517$) and IS (F-Measure = 0.8416, $P = 0.8572$, $R = 0.6982$) datasets.

Salient Object	FT	IS	PASCAL-S
FT	0.7427	0.4736	0.4325
GC	0.8383	0.6261	0.6072
PCAS	0.8646	0.6558	0.6275
SF	0.8850	0.5555	0.5557
CPMC + Fixation	FT	IS	PASCAL-S
AIM	0.8920	0.6728	0.7204
AWS	0.8998	0.7241	0.7224
DVA	0.8700	0.6377	0.7112
GBVS	0.9097	0.7264	0.7457
ITTI	0.8950	0.6827	0.7288
SIG	0.8908	0.7255	0.7214
SUN	0.8635	0.6249	0.7058
Orig. Fixation	FT	IS	PASCAL-S
AIM	0.6858	0.4804	0.6267
AWS	0.7228	0.6033	0.5084
DVA	0.6534	0.4795	0.5426
GBVS	0.6899	0.5333	0.6383
ITTI	0.6544	0.4431	0.6228
SIG	0.6741	0.6110	0.5897
SUN	0.6692	0.3881	0.5482
Fix	N/A	0.6972	0.6781
Baseline Models	FT	IS	PASCAL-S
CPMC Ranking	0.4421	0.5287	0.6339
CPMC + Human	N/A	0.7863	0.7756
CPMC Best	0.9496	0.8416	0.8699
GT Seg. + Human	N/A	N/A	0.9201

Table 4.3: Results of our model compared to existing salient object algorithms. Our model achieves better F-measure than all major salient object segmentation algorithms on all **three** datasets, including the heavily biased FT dataset. Results are reported on the testing set (60% of the images) over 10 random splits in three datasets.

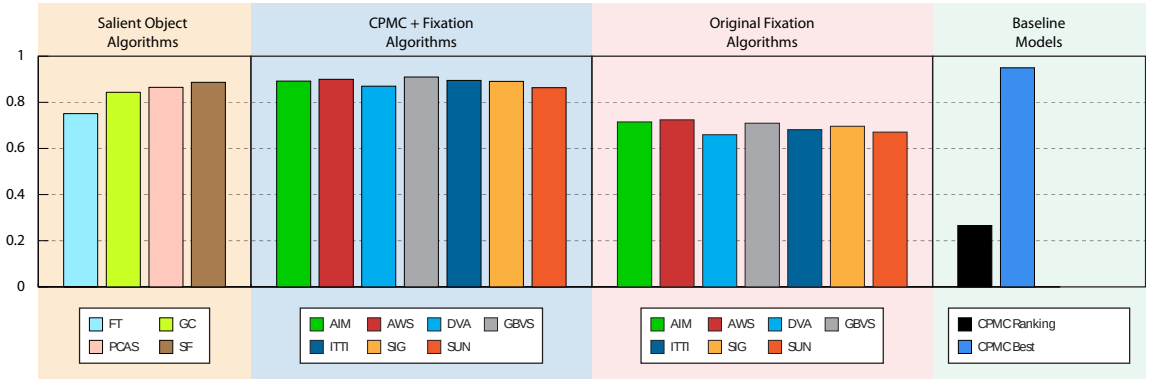


Figure 4.4: The F-measures of all algorithms on the FT dataset. Due to the absence of human fixations on this dataset, as well as the full segmentation ground-truth, we are unable to evaluate CPMC + Human Fixations and GT Seg + Human Fixations.

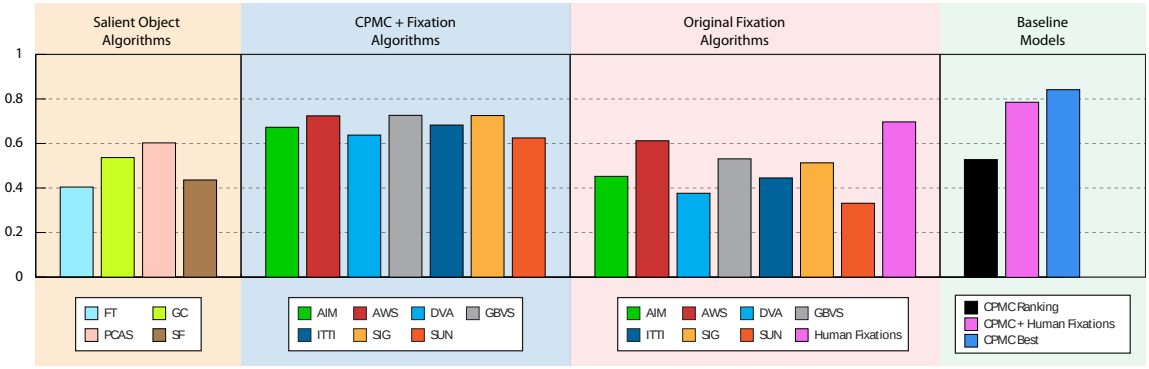


Figure 4.5: The F-measures of all algorithms on the IS dataset. Due to the absence of full segmentation ground-truth of this dataset, we are unable to evaluate GT Seg + Human Fixations.

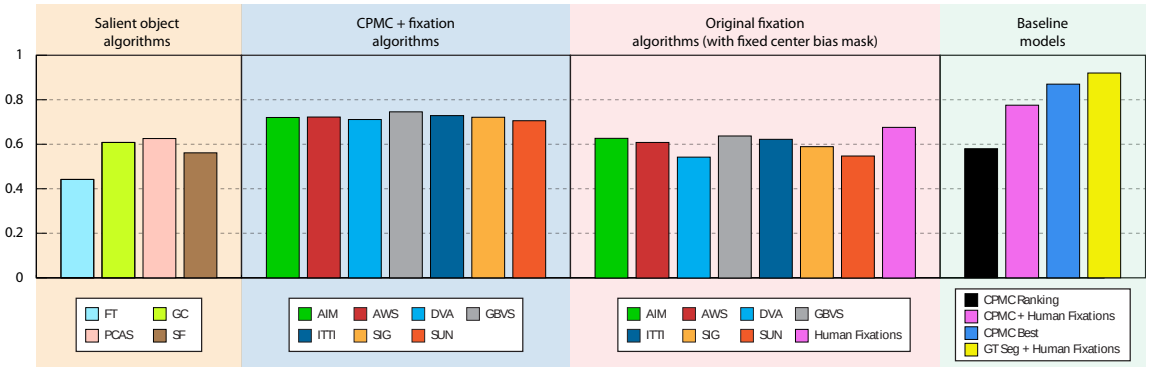


Figure 4.6: The F-measures of all algorithms on the PASCAL-S dataset.

4.4.4 Results

The full benchmarking results of all algorithms on all 3 datasets are shown in Fig. 4.4, Fig. 4.5, and Fig. 4.6. In particular, for all CPMC-related algorithms, we choose the top $K = 20$ segments. The results are grouped into 4 categories:

Salient Object Algorithms refer to the 4 algorithms `FT`[60], `GC`[61], `PCAS`[63], and `SF`[62] that are originally proposed for salient object segmentation.

CPMC + Fixation Algorithms refer to our model. We choose the top 10 segments for each image, and assign scores based on the fixation map of 7 algorithms: `AIM`[3], `AWS`[5], `DVA`[4], `GBVS`[35], `ITTI`[2], `SIG`[1], and `SUN`[34].

Original Fixation Algorithms refer to the 7 fixation algorithms. To cancel the effect of center bias, we add a fixed center bias with $\sigma = 0.4$ of the image width to each generated fixation map.

Baseline Models refers to 4 other models. `CPMC Ranking` refers to the original rankings of CPMC, with the same choice of $K = 10$. `CPMC+Human Fixations` refers to a variation of our model that replaces the algorithm fixation maps with human fixations – supposedly, the human map should reflect the saliency of the scene more accurately than algorithms. `CPMC Best` refers to the salient object maps generated by greedily selecting the best CPMC segments with respect to the ground truth. This score estimates the upper limit of any algorithm that is based on CPMC segmentation. Finally `GT Seg + Human Fixations` refers to the method that uses ground-truth segmentations plus human fixations. This score validates the strong connection between the fixation task and the salient object segmentation task.

From Fig. 4.6, we make two key observations. First, our method consistently outperforms the original CPMC ranking function by a large margin, independently of the underlying fixation prediction algorithm. Second, the performance of our model converges much faster than the original CPMC with respect to K. Our model provides decent F-measure with moderate $K = 20$ segments, while the CPMC ranking function does not converge even with $K = 200$ (see Fig. 4.7). The result suggests that our method can effectively segment salient objects using a small number of segments.

In Fig. 4.6, we compare our results with the state-of-the-art salient object algorithms with $K = 20$. Our method outperforms the state-of-the-art salient object segmentation algorithms by a large

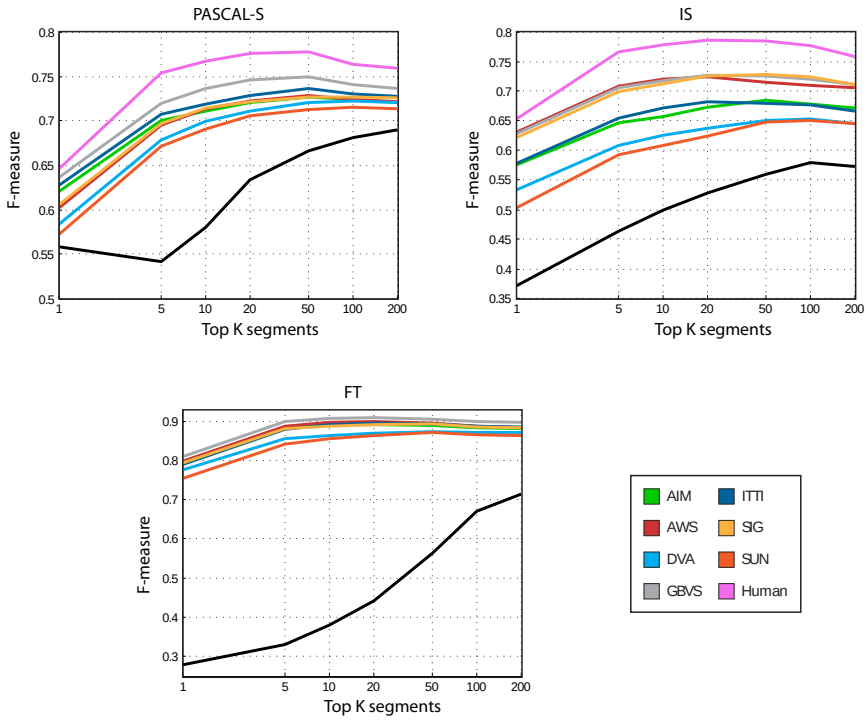


Figure 4.7: F-measures of our salient object segmentation method under different choices of K, in comparison to the CPMC ranking function. Results are reported on the testing set (60% of the images) of 3 datasets over 10 random splits. Compared to the original CPMC ranking function, our method obtains satisfactory F-measures with small $K = 20$.

margin. In this dataset, we achieved an improvement of 11.82% with CPMC+GBVS in comparison to the best performing salient object algorithm PCAS. The details of the experimental results are reported in Tab. 4.3. The combination of CPMC+GBVS leads to best results in all 3 datasets.

Finally, some example results of all algorithms are shown in Fig. 4.8, Fig. 4.9, and Fig. 4.10.

4.5 Conclusion

In this chapter, we explore the connections between fixation prediction and salient object segmentation by providing a new dataset with both fixations and salient object annotations. We conduct extensive experiments on the dataset for both tasks, and compare the results with major benchmarks. Our analysis suggests that the definition of a salient object is highly consistent among human subjects. We also point out significant dataset design bias in major salient object benchmarks. The bias is largely due to deliberately emphasising the concept of saliency. We argue that the problem of salient object segmentation should move beyond the textbook examples of visual saliency. A possible new

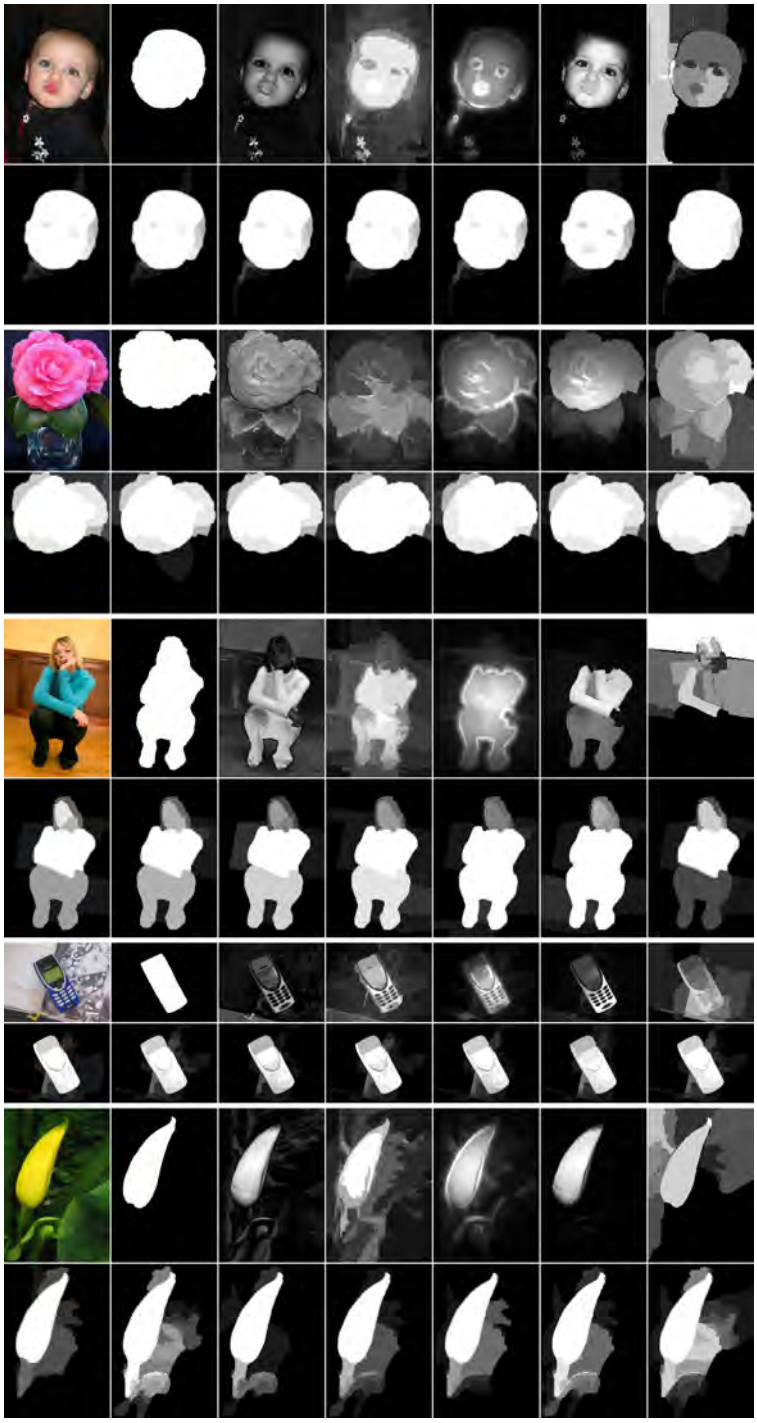


Figure 4.8: Visualization of salient object segmentation results on FT. Each two-row set compares the results of one image. The first row includes results from existing methods (Left to Right): Original image, Ground-truth mask, FT, GC, PCAS, SF and CPMC ranking; The second row shows results of our method using different fixations (Left to Right): AIM, AWS, DVA, GBVS, ITTI, SIG and SUN. We are not able to report results using human fixations. The images are selected by sorting the F-measure of our results in a decreasing order.

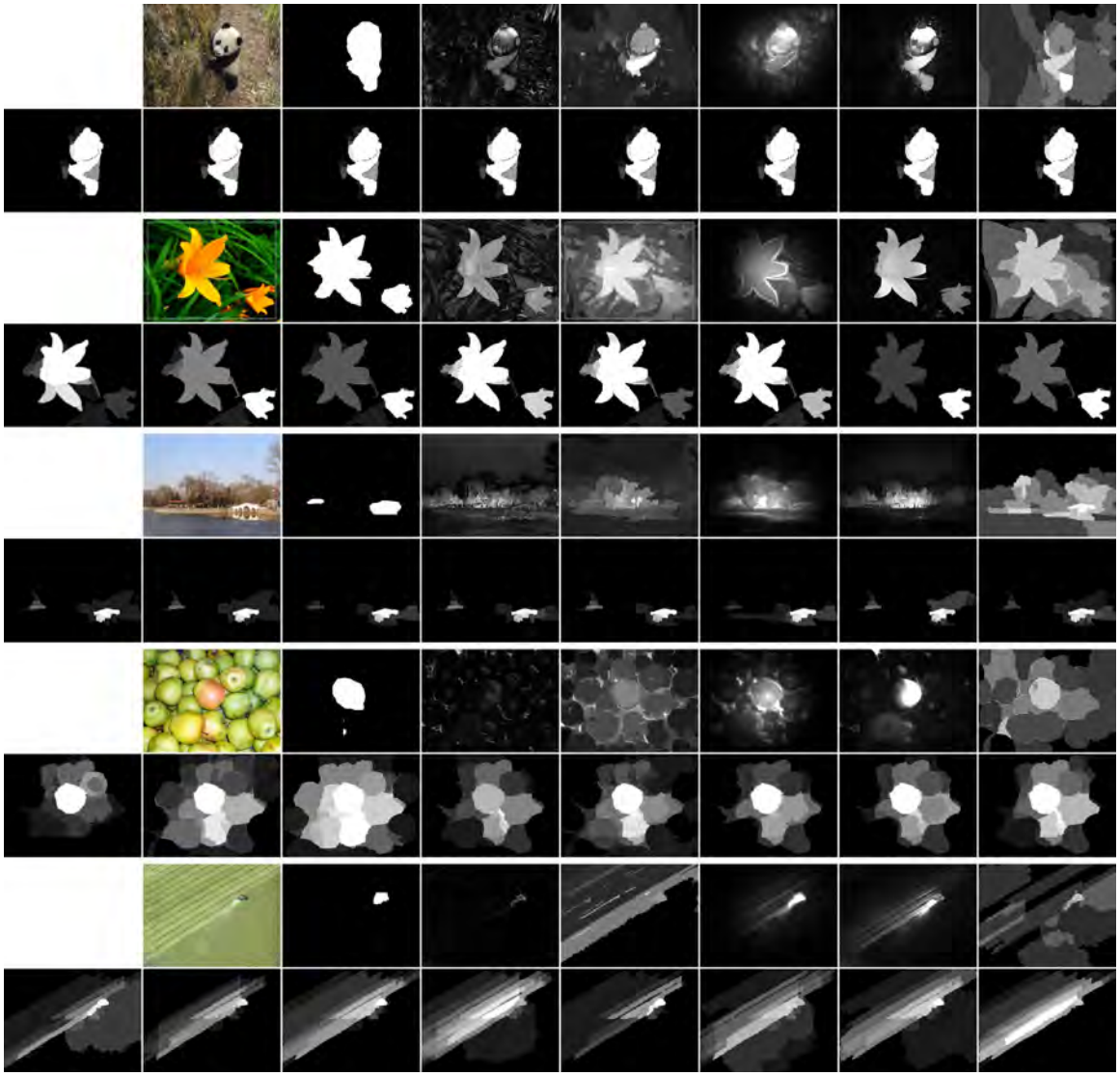


Figure 4.9: Visualization of salient object segmentation results on IS. Each two-row set compares the results of one image. The first row includes results from existing methods (Left to Right): Original image, Ground-truth mask, FT, GC, PCAS, SF and CPMC ranking; The second row shows results of our method using different fixations (Left to Right): Human Fixation, AIM, AWS, DVA, GBVS, ITTI, SIG and SUN. The images are selected by sorting the F-measure of our results in a decreasing order. We notice that IS favors sparse saliency maps, since it contains a significant portion of small salient objects.

Figure 4.10: Visualization of salient object segmentation results on PASCAL-S. Each two-row set compares the results of one image. The first row includes results from existing methods (Left to Right): Original image, Ground-truth mask, FT, GC, PCAS, SF and CPMC ranking; The second row shows results of our method using different fixations (Left to Right): Human Fixation, AIM, AWS, DVA, GBVS, ITTI, SIG and SUN. The images are selected by sorting the F-measure of our results in a decreasing order.

direction is to look into the strong correlation between fixations and salient objects. Built on top of this connection, we propose a new salient object segmentation algorithm. Our method decouples the problem into a segment generation process, followed a saliency scoring mechanism using fixation prediction. This simple model outperforms state-of-the-art salient object segmentation algorithms on all major datasets. Our dataset, together with our method, provides a new insight into the challenging problems of both fixation prediction and salient object segmentation.

Chapter 5

An Analysis of Boundary Detection Benchmarking

Abstract

As we analyzed in the previous chapter, accurate detection of object contours are essential to the figure-ground problem. In this chapter and Chapter 6, we discuss various issues around the problem of boundary detection.

For an ill-posed problem like boundary detection, human labeled datasets play a critical role. Compared with the active research on finding a better boundary detector to refresh the performance record, there is surprisingly little discussion on the boundary detection benchmark itself.

The goal of this chapter is to identify the potential pitfalls of today's most popular boundary benchmark, BSDS 300. In the chapter, we first introduce a psychophysical experiment to show that many of the "weak" boundary labels are unreliable, and may contaminate the benchmark. Then, we analyze the computation of f-measure and point out that the current benchmarking protocol encourages an algorithm to bias towards those problematic "weak" boundary labels. With this evidence, we focus on a new problem of detecting strong boundaries as one alternative. Finally, we assess the performances of 9 major algorithms on different ways of utilizing the dataset, suggesting new directions for improvements.

5.1 Introduction

Boundaries in an image contain cues that are very important to high level visual tasks such as object recognition and scene understanding. Detecting boundaries has been a fundamental problem since the beginning of computer vision. In the development of boundary detection, datasets [71, 57, 72,

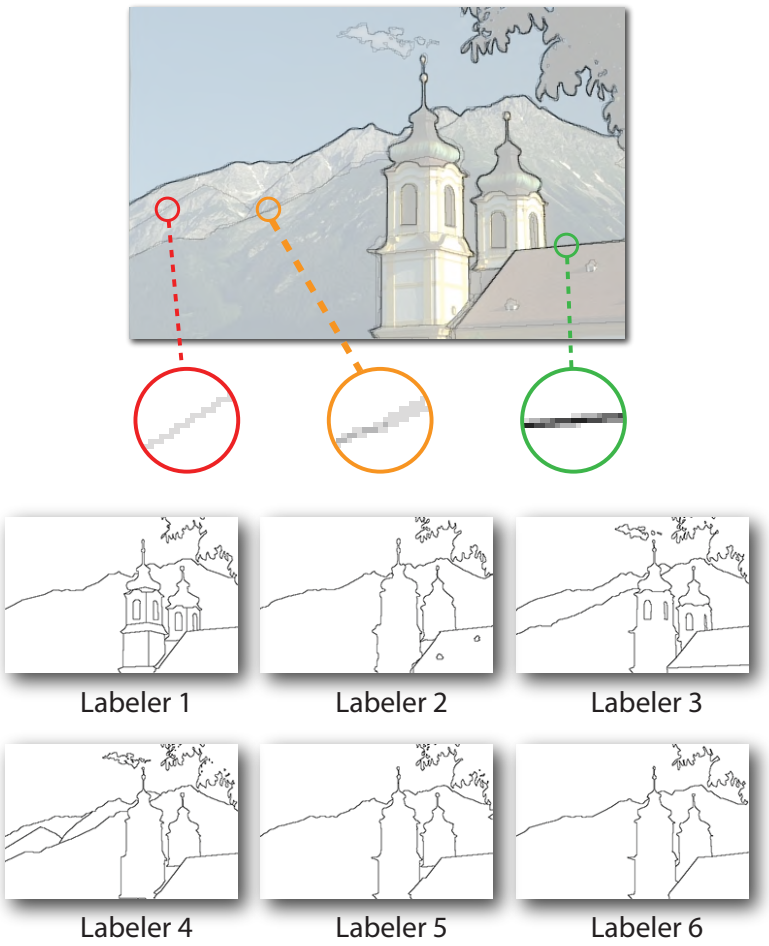


Figure 5.1: An example image, and the corresponding labels from BSDS 300. The top figure shows the original image overlapping with all 6 boundary maps from labelers. There is a clear difference among different labelers. The red circle gives an example boundary segment that is labeled by only one out of 6 labelers (labeler 4). The boundary segment in the orange circle is labeled by two labelers (labeler 3 and 4). The boundary segment in the green circle is unanimously labeled by all 6 labelers.

73] – along with their evaluation criteria[1] – have played critical roles. These datasets are responsible for our progress in the problem of boundary detection, not only because they provide an objective quantity to judge the value of each newly proposed algorithm, but also because the images, the labels, and the evaluation standards they set forth have heavily influenced the researchers during the development of a boundary detection algorithm.

5.1.1 Boundary detection is ill-defined

What is a boundary? A universally accepted definition of a boundary may not exist. No matter how the definition is made, one can always find counter-examples on which people disagree. In today's most popular benchmark BSDS 300 [71], 28 human labelers contributed a total number of 1667 high quality boundary maps on 300 images of natural scenes (200 training, 100 testing). Within the entire dataset, it is hard to find a image where different people have perfectly matched labels.

In high-level vision tasks such as object recognition or scene classification, human annotation has been traditionally considered reliable. However, the ill-posed nature of boundary detection makes this problem a different scenario. There is surprisingly little discussion about ground-truth data reliability for boundary detection. It is commonly held that human annotations from BSDS 300 are reliable. Previously, [71, 74] have the following observations regarding to the reliability of BSDS 300:

1. Labelers are well trained and correctly instructed. Examined separately, each boundary seems to be aligned to some underlying edge structure in the image. The effect of an adversarial labeler (labelers with totally irrelevant output) is minimal.

2. Label variability can be explained by a perceptual organization hierarchy. Even though different labelers may annotate boundaries in different levels of details, they are consistent in a sense that the dense labels "refine" the corresponding sparse labels without contradicting them. In other words, the same image always elicits the same perceptual organization across different labelers.

Nevertheless, none of these observations are strong enough to legitimatize the BSDS 300 as a benchmark. To be able to evaluate an algorithm faithfully, the benchmark has to be free from both type I (false alarm) and type II (miss) statistical errors. Aforementioned observation $\#1$ rules out

[1] In this chapter, we refer to the images and the labels as *datasets*, while the term *benchmark* includes images and labels as well as the corresponding evaluation criteria.

type I errors. However, the risk of type II remains as unchecked errors in human labels. It is possible that the labelers may miss some equally important boundaries. Once we benchmark an algorithm, the incomplete data may incorrectly penalize an algorithm that detects true boundaries.

As for observation $\#2$, the hierarchical organization of boundaries raises more fundamental questions: Can we give equal weights to the strong boundaries where everyone agrees, and the weak boundaries where only one or two labelers have noticed? When we say "boundary detection," are we trying to solve one single problem with different thresholds? Or different problems at different levels of the perceptual hierarchy?

5.1.2 The perceptual strength of a boundary

In this chapter, the *perceptual strength* of a boundary segment refers to the composite effect of all factors that influence personal decision during boundary annotation. Such factors may include border contrast, object type, or line geometry. One simple way to approximate the perceptual strength of each boundary segment is to take the proportion of labelers who have labeled that specific segment. To get rid of local alignment noise, we match each pair of human boundary maps using the assignment algorithm proposed in [75], with the same parameter set [76] used for algorithm evaluation. For instance, given an image with N labelers, if a boundary pixel from one subject matches with M other labelers, it has a perceptual strength of $\frac{M+1}{N}$. The weakest boundary labels are the ones annotated by only one labeler. These boundaries are referred to as *orphan labels*. In BSDS 300, 29.40% of the boundary labels are orphan labels. In comparison, the second largest population (28.99%) are *consensus labels* that are labeled by everyone.

Clearly, the orphan labels and the consensus labels are not equal. In Sec. 5.3, we use a psychophysical experiment to assess the statistical difference of weak/strong boundaries. Our experimental results indicate that weak (especially orphan) labels are not capable of evaluating today's algorithms.

Based on this novel discovery, in Sec. 5.4, we investigate the impact of these weak boundaries on the current evaluation system. A disappointing yet alarming result is that all of the 9 algorithms experience significant performance drops if we test them on strong boundaries only. Furthermore, we pinpoint a mechanism called *precision bubble* in the original BSDS 300 benchmarking algorithm. This mechanism tends to exaggerate the precision of an algorithm, especially when the weak labels are included in the ground-truth.

We raise an important, yet largely neglected, question: *are we ready to detect strong bound-*

aries? Our analysis shows that none of the 9 algorithms is capable of discovering strong boundaries significantly better than random selection. The output values of the algorithms are either independent or weakly correlated with the perceptual strength. This result is in sharp contrast to many of today's popular practices of using the output of a boundary detector algorithm as an informative feature in high-level boundary analysis. We conclude our discussion with a comparison of pB v.s. retrained-pB and BSDS 300 v.s. BSDS 500.

5.2 Related works

Over the last 12 years, a great number of boundary detection algorithms have been proposed. The benchmark's F-measure, according to the measurements proposed in [76], has increased 7 percent, from 64.82% [76] to 71.43% [77]. In this chapter, we focus on 9 major boundary detection algorithms (shown in Tab. 5.1).

All of these algorithms, except cCut, provided very competitive F-measures at the time when they were first introduced. F-measure, also known as F-score, or the harmonic mean of precision and recall, is recommended in [76] as a summary statistic for the precision-recall property. Over the past 10 years, it has been accepted as the most important score to judge a boundary detector.

Along with boundary detection, a parallel line of work [78, 79, 80] focuses on the detection of "salient boundaries." These works emphasize on finding salient 1-D structures from the ensemble of line segments discovered by a boundary detector. The stated advantage of these algorithms is to gain extra precision scores at low-recall regions. Therefore, it is interesting to include cCut [80], one of the latest algorithms in this line, and evaluate it under our quantitative framework.

Name	F-measure	Year
pB [76]	0.65	2002
UCM [81]	0.67	2006
Mincover [82]	0.65	2006
BEL [83]	0.66	2006
gPB [70]	0.70	2008
XRen [84]	0.67	2008
NMX [85]	0.71	2011
cCut [80]	0.45	2011
SCG [77]	0.71	2012

Table 5.1: The list of boundary detection algorithms referred in this chapter. Their F-measures increase over time.

5.2.1 Relevant theories on dataset analysis

In contrast to the perennial efforts in breaking benchmark performance records, theoretical analysis on benchmark reliability has been brought to people's attention only in recent years. These studies can be roughly categorized into either human annotation analysis, or benchmark design analysis. The first problem of human annotation comes with the recent trends of obtaining annotation data via crowdsourcing [86]. Many seminal models [87, 88] have been proposed to analyze the crowdsourced annotation process in general. Specifically, [89] has proposed strategies to estimate the quality of crowdsourced boundary annotation. On the other hand, [68] has raised a series of interesting questions in regard to the design philosophy of today's object recognition benchmarks. Their alarming results suggest the potential pitfalls of some widely adopted benchmarks.

5.3 A psychophysical experiment

While collecting the human annotation, BSDS 300 [71] gave the following instructions to each of the labelers:

> *Divide each image into pieces, where each piece represents a distinguished thing in the image. It is important that all of the pieces have approximately equal importance. The number of things in each image is up to you. Something between 2 and 20 should be reasonable for any of our images.*

The instruction is intentionally made vague in order to minimize potential labeling bias towards any specific sub-type of boundaries. However, the absence of precise instruction also leads to a considerable labeling variation. As we have discussed in Sec. 5.1, 31.39% of the boundary labels are *orphan labels*. On one hand, we know that these boundaries are labeled by well-educated Berkeley students chosen from a graduate level computer vision class. On the other hand, we are also aware that the annotation of these orphan labels is due to a pure random assignment of labelers. How well can we trust these relatively weak labels?

In this section, we introduce a two-way forced choice paradigm to test the reliability of a boundary dataset. In each trial, a subject[2] is asked to compare the relative perceptual strength of two local boundary segments with the following instruction:

[2]We refer to *labelers* as the people who originally labeled the BSDS300 dataset, while *subjects* refers to people we recruited to perform our two-way forced choice experiment.

Boundaries divide each image into pieces, where each piece represents a distinguished thing in the image. Choose the relatively stronger boundary segment from the two candidates.

One of the two boundary segments is chosen from the human label dataset, and the other is a boundary segment produced by an algorithm. The advantage of this two-alternative experiment is that it cancels out most of the cognitive fluctuations, such as spatial attention bias, subject fatigue, and decision thresholds that are different among subjects. Moreover, compared to the tedious labeling process, this paradigm is much simpler and cheaper to implement via crowdsourcing. In our experiment, the average response time for each trial is 5 seconds. One caveat is that the comparison experiment requires the algorithm-generated candidate segment to have a similar appearance to the human labels. Among the 9 benched algorithms, BEL is the only algorithm that does not produce thinned edges, and is therefore skipped for the experiment.

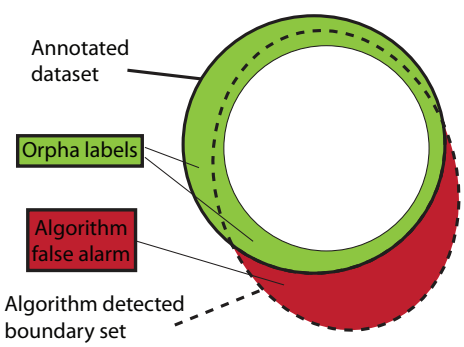
5.3.1 Easy and hard experiments for boundary comparison

Using different boundary sampling strategies, we can design two experiments: hard and easy. In the hard experiment, each algorithm is first thresholded at its optimal F-measure, and then matched to the original human labels to find false alarms – boundary segments that are considered weaker than human labels. Then, for each testing image, we randomly draw one instance of algorithm false alarm, and compare it against another randomly selected human orphan label. Fig. 5.2 gives a detailed illustration of this process. This experiment is called the "hard experiment" because the relative order between human labeled orphan labels and algorithm detected false alarms is not easy to determine (as one can see in Fig. 5.2.C).

Similarly, we also design an easy experiment. First, we remove all the human labels that are not unanimously labeled by everyone. This leaves us with a very small but strong subset of labels (perceptual strength equals 1). Then, with this new dataset, we re-benchmark all 8 algorithms, determining their optimal F-measures and thresholds (higher than their original thresholds), and find each algorithm's false alarms under its new optimal threshold. Finally, the competition is made between strong human labels and confident output of algorithm false alarms.

For each algorithm on either easy/hard experiment, we produced one trial per image for all 100 test images. Five subjects participated in the experiment, and a total number of 8000 responses were collected. The final ordering for each trial was determined by majority voting of all 5 subjects. To

A)



B) C)

Figure 5.2: An illustration of the two-way, forced choice experiment (hard mode). **A) The experiment interface**: In each trial, a subject is presented with two images. On the left image, two boundary segments (high contrast squares with red lines) are superimposed onto the original photo. The subject is asked to click on one of two boundary segments that she/he feels stronger. At the same time, the original image is also presented in a separate window. **B) The Venn diagram of sets of boundary segments**: The thick circle encompasses the full human labeled boundary set of the dataset. The subset of orphan labels is shown in the green area. The algorithm detected boundary set is the dotted ellipsoid. The subset of algorithm false alarms is highlit in red. In each trial, we randomly select one boundary segment from the green area, and one from the red area. **C) Orphan labels v.s. algorithm-false-alarms**: Some example images with both human orphan labels (shown in green lines) and false alarms of PB algorithm (shown in red lines). In many examples, the relative strength between algorithm false alarm and human orphan labels is very hard to tell.

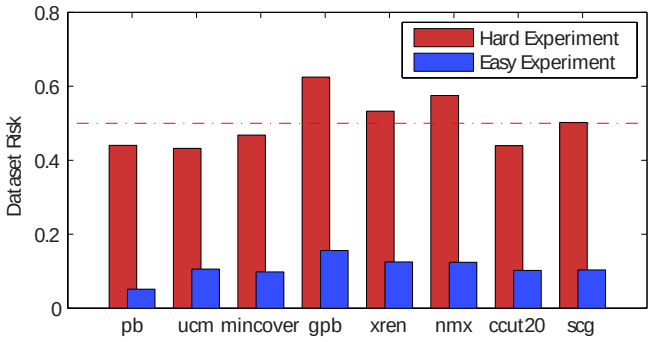


Figure 5.3: Results of our hard and easy experiments. The average risks over all algorithms are 0.5017 and 0.1082 for hard and easy experiments, respectively. Dotted red line indicates the 50% chance performance. The average result of the hard experiment is even greater than chance level.

interpret the result, we introduce a term called *dataset risk*. This value measures the probability that an algorithm false alarm wins over a human label. Ideally, a perfectly constructed dataset should have zero risk, because it does not miss any strong boundary segments, and algorithm false alarms are always weaker than any instance from the perfect boundary dataset. However, our experiment results in Fig. 5.3 show that the BSDS 300 – especially those orphan labels – are far away from being perfect.

5.3.2 Interpreting the risk of a dataset

From Fig. 5.3, we observe high risks in the hard experiment for all algorithms that we have tested. The first conclusion one can draw from this observation is rather depressing – the orphan labels are extremely unreliable, since they falsely classify good algorithm detections into false alarms (or falsely include weak algorithm detections as hits, depending on the thresholds). Yet, we can also interpret the results of hard experiments in a more optimistic way: the computer vision algorithms have performed so well that their results look as good as some of the humans'. In other words, these algorithms have passed a restricted Turing test if the dataset risk is equal to or greater than 0.5.

No matter whether we choose the pessimistic or the optimistic perspective, it is clear that the orphan labels are not appropriate to serve as a benchmark – or even parts of a benchmark. Instead, we should focus more on the consensus boundaries, because the risk is much lower.

It is worth mentioning that our results on the easy experiment do not necessarily imply that the consensus boundaries is a perfect dataset. However, as long as the missed boundaries of consensus

labels cannot be accurately detected by an algorithm, this data remains as valid for a benchmark. In other words, given the performance of today's top algorithms, detecting strong boundaries is a meaningful Turing test that is not yet solved.

5.4 F-measures and the precision bonus

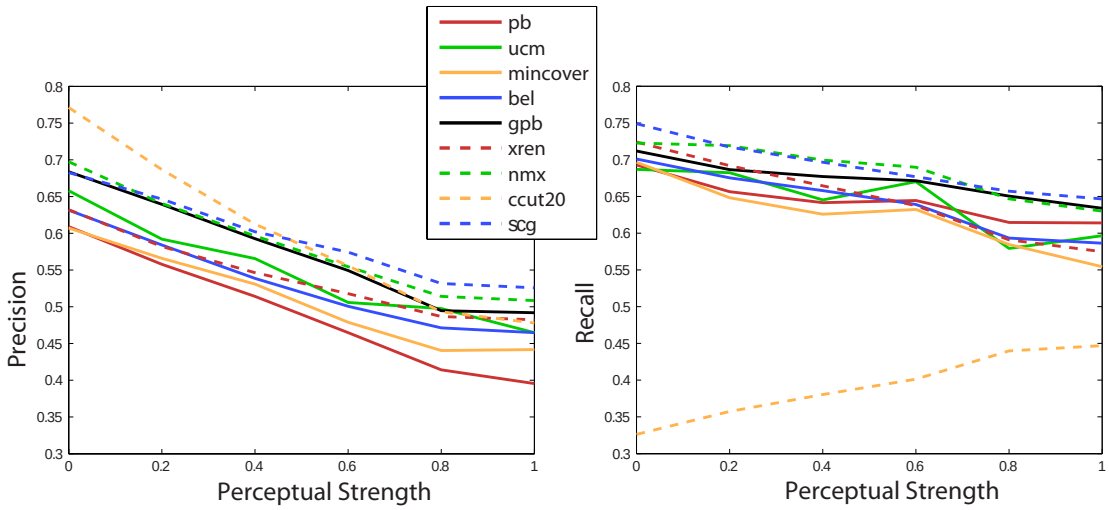


Figure 5.4: The optimal precision and recall values for all algorithms, benchmarked under different label strengths. By increasing the perceptual strength, we transform the problem from "boundary detection" to "strong boundary detection." The precision values for algorithms dropped 28.7% on average. In contrast, the recall values, which are not affected by the precision bonus, only dropped 9% on average.

Given the fact that the orphan labels are unreliable, what role do those labels play in the benchmarking process? How much can they affect the result of F-measure? In this section, we show that the orphan labels can create a "precision bonus" during the calculation of the F-measure.

In the original benchmarking protocol of BSDS 300, the false negative is defined by comparing *each* human boundary map with the thresholded algorithm map, and counting the unmatched human labels. In comparison, the false positive is defined by comparing the algorithm map with *all* human maps, and then counting the algorithm labels that are not matched by *any* human. In other words, the cost of each algorithmic missing pixel is proportional to the human labelers who have detected that boundary, whereas the cost of each false alarm pixel is just one. This protocol exaggerates the importance of the orphan labels in the dataset, and encourages algorithms to play "safely" by enumerating an excessive number of boundary candidates. Strategically, detecting

strong boundaries has become a much more risky endeavor under the current framework of F-measure.

We can better evaluate the impact of such *precision bonus* by re-benchmarking the algorithms on different levels. First, we threshold the human labels by different perceptual strengths, from $0, 0.2, 0.4 \ldots$ to 1. And then use each of these subsets of the human labels as the ground-truth to benchmark all 9 algorithms. At each perceptual strength, an algorithm finds its optimal threshold that produces the maximal F-measure. Fig. 5.4 plots the precision and recall values at the optimal algorithm thresholds for all 9 algorithms.

Despite its strong influence on the benchmark scores, the precision bubble by itself should not be considered as a "mistake" in the design. What makes today's benchmarking practice questionable is the joint cause due to the following factors: 1) the weak boundaries in BSDS 300 are not reliable enough to evaluate today's algorithms; and 2) precision bonus gives extra credits to algorithms working on the low perceptual strength boundaries – which, according to factor 1, is not a good practice.

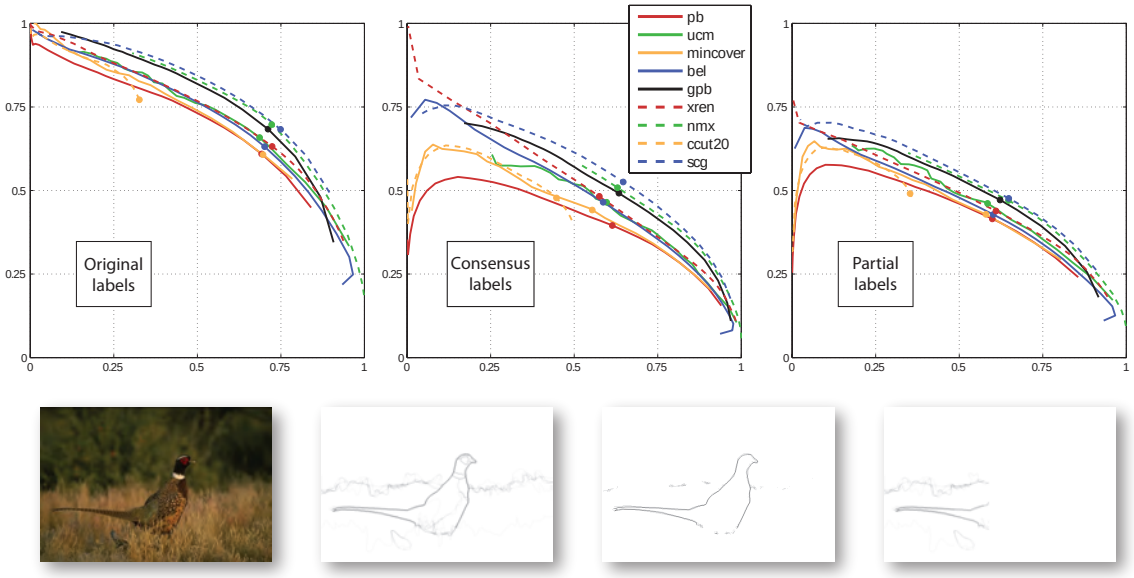


Figure 5.5: Benchmarking all 9 algorithms using different ground truths. The top figure shows the precision recall curves, with solid dots indicating the maximal F-measure location. The bottom figure gives an example image and the ground-truth labels: original labels, consensus labels, and partial labels. The partial label (bottom right figure) of this image is clearly an unrealistic ground-truth, because the majority of the bird boundary is discarded.

5.5 Detecting strong boundaries

The simplest way to avoid the problem of weak labels is to benchmark the algorithms using consensus labels only, as shown in Fig. 5.5. However, the performances of the tested algorithms have dropped so significantly that it stimulates us to ask another question: *are we detecting strong boundaries better than random?*

To compute the baseline performance of a null hypothesis, we design a control experiment called *partial labels*. In this experiment, we crop out a part of each human boundary map to make the total number of pixels in the remaining map equal to that of a strong boundary map (see Fig. 5.5). Because such a cropping operation is completely independent of the image content, it can be considered as a random sub-sampling from an algorithmic perspective.

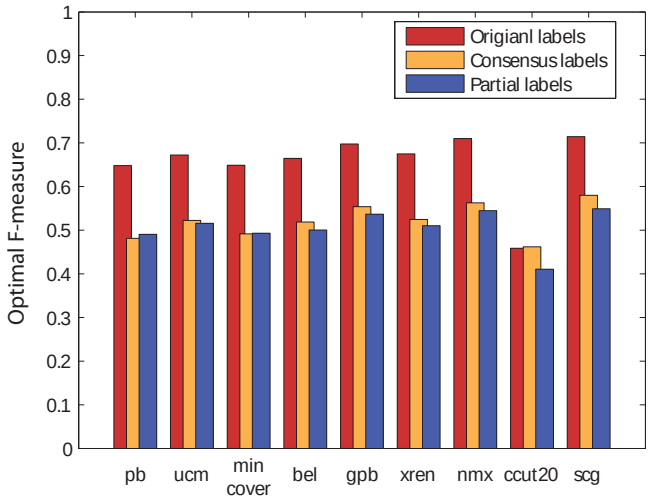


Figure 5.6: Algorithm performances (optimal F-measures) evaluated under different ground-truths.

With the PR curves shown in Fig. 5.5, the optimal F-measures of all three experiments are compared in Fig. 5.6. Except for cCut, all other algorithms have suffered severe performance decreases when shifting from detecting all labels to detecting consensus labels only. Such a performance drop is so devastating that the F-measures are no better (even worse for pB algorithm) than the control experiment with randomly contaminated ground-truth.

In this experiment, the salient boundary algorithm cCut has the most significant performance drop on partial labels. However, the overall performance of cCut is not comparable with the state-of-the-arts detectors (such as gPB, NMX, or SCG), even if we benchmark them on the consensus labels.

The comparative results of consensus and partial labels contradict our intuition that algorithm detection strength is correlated with the perceptual strength of a boundary. It also questions the practices in computer vision that use boundary detector output as a feature for high-level visual tasks. For instance, intervene contour [90, 91] is a well-established method that computes the affinity of two points in the image by integrating the boundary strengths along the path that connects those two points. Many other works such as [92, 93, 94] also included pB (or gPB) boundary intensity in their feature design. To understand the relationship between algorithm output and the perceptual strength of a boundary, we further plot the perceptual strength distribution with respect to algorithm detector output for all 9 algorithms. In Fig. 5.7, we can see that the correlation between algorithm output and perceptual strength of the boundary is rather weak.

5.5.1 Retrain on strong boundaries

Another useful test to evaluate our current progress on strong boundary is to retrain an algorithm. Because of its great popularity, we focus on pB algorithm for the retraining experiment. Using the publicly available MATLAB codes from the authors' websites, we re-generate the training samples with consensus boundaries, and learn a new set of parameters. This retrained-pB is then compared against the original pB in the original as well as the consensus label test sets. The retrained-pB does not gain superior F-measure, even if we use consensus labels as the ground-truth.

5.5.2 BSDS 300 and BSDS 500

Recently, BSDS 300 has been enriched to BSDS 500 with 200 additional testing images. According to [70], the protocol used to collect new human labels remains the same as in BSDS 300. According to our analysis, the populations of orphan and consensus labels of these 200 new images are 30.58% and 30.15%, respectively. Not only do the statistics of BSDS 500 look very similar to those of the original BSDS 300, the performance of algorithms on this new dataset is also very close. Since BSDS 500 is fairly new, not many algorithms have provided their results on this new dataset. We choose the two most representative algorithms, SCG and gPB, for our analysis. The optimal F-measure of these algorithms under all boundaries, or under consensus boundaries are reported in Fig. 5.9.

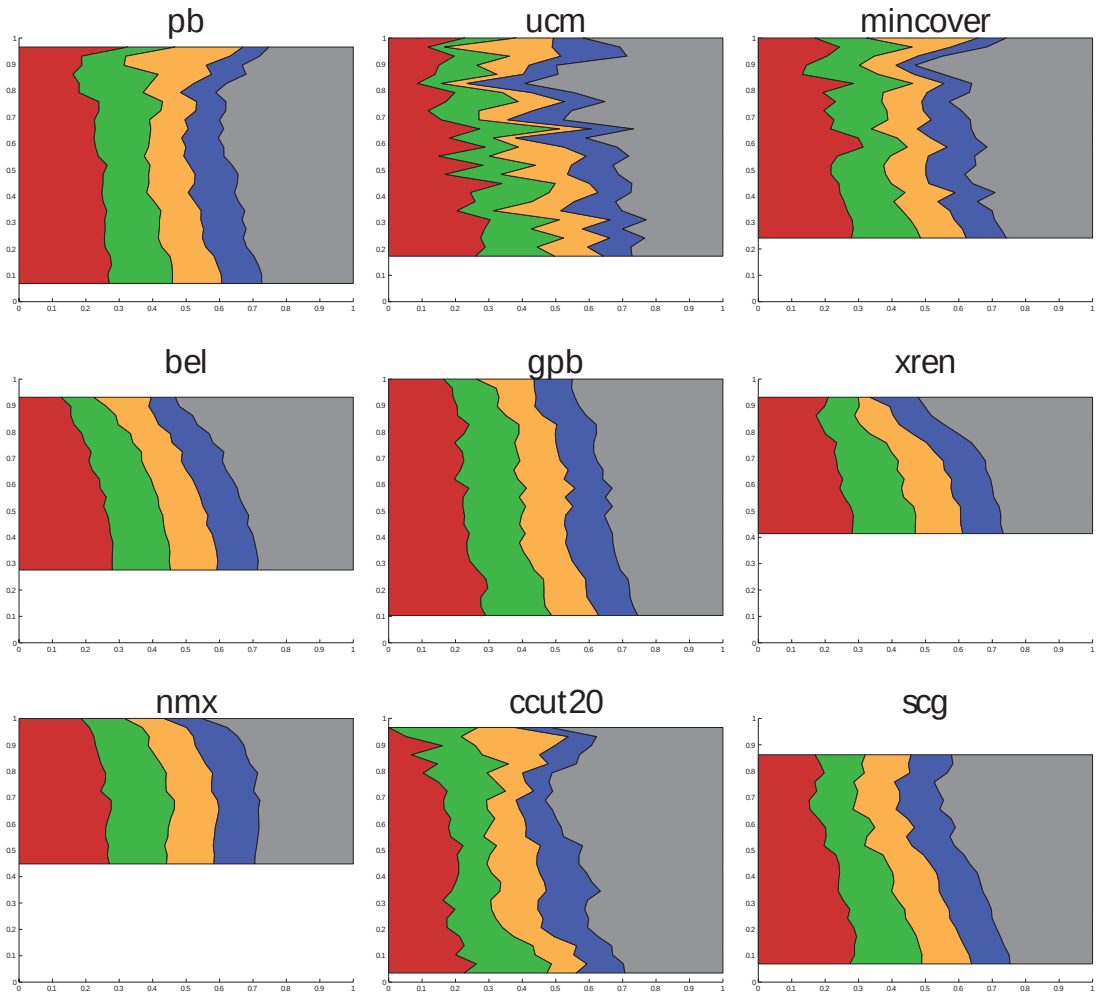


Figure 5.7: Boundary perceptual strength distribution. This experiment is done with the original (full) labels. In each sub-figure, the X-axis is the percentage of matched human label strength (always summing to 1), and the Y-axis is the algorithm output value. If we extract one row with $y = k$ in a sub-figure, the color strips represent the distribution of the human labels that are matched to all algorithm pixels where detection output is equal to k. The red area represents human labels with perceptual strength in $[0, 0.2)$, whereas green represents perceptual strengths in $[0.2, 0.4) \ldots$, and finally, the gray area shows the population of consensus labels. Ideally, the gray area should have a upper triangular shape (XREN is the closest) – that is, the algorithm output being correlated with human perceptual strength.

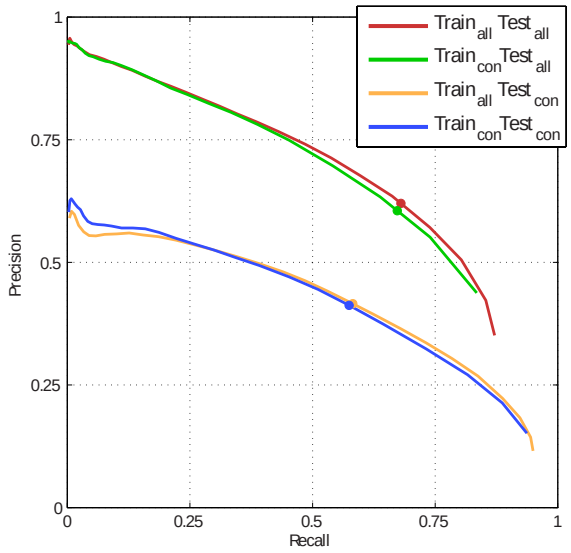


Figure 5.8: Retrain pB algorithm using consensus labels, and compare the results on original (all) and consensus (con) boundaries respectively.

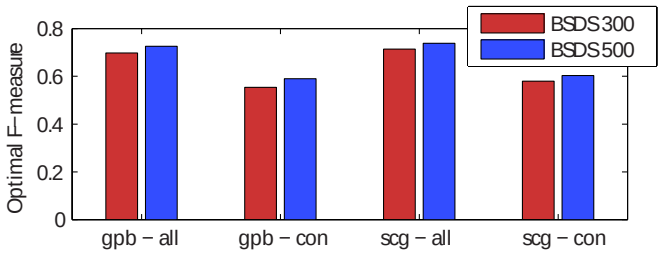


Figure 5.9: Comparison of SCG and gPB algorithms on BSDS 300 and BSDS 500 datasets. The comparison is also made by either using original (all) boundaries or consensus (con) boundaries only. The difference between BSDS 300 and BSDS 500 is small (mean difference is 0.028) and consistent (STD over all 4 different settings is 0.0058).

5.6 Discussion

In this chapter, we have raised doubts in regard to the current way of benchmarking an algorithm on the most popular dataset of boundary detection. With a psychophysical experiment, we show that the weak labels, especially the orphan labels, are not suitable for benchmarking algorithms. However, if we shift from the original problem of boundary detection, to the new problem of strong boundary detection, we are on one hand blessed with a more reliable dataset, but on the other hand, we are disappointed by the experimental results showing that none of the current algorithms have evidence of good performance.

Our results in Fig. 5.7 do not conclude that the current algorithms' output value is a useless

feature for high-level tasks. The validity of using boundary detector output to reveal high-level semantic information may not have a one-line answer. It depends critically on the specific scenarios as well as the design of the high-level vision algorithms. At present, researchers from different topics have not yet converged to one common framework.

Chapter 6

Modeling of Human Labeler Behaviors

Abstract

This chapter presents an analysis of a boundary detection dataset. We focus on the human factors in this boundary detection task. Our observation contradicts traditional understandings that human labels are either consistent, or reconcilable within a simple parametric model. In order to make quantitative predictions of the consistency of any human label, we introduce a framework of partially ordered sets that determines the fundamental relations of the interplay among perceptual states of the subjects, and the perceptual quantity inherent with the image. We designed an algorithm that effectively recovers the interactions of all 28 subjects in the Berkeley Segmentation Dataset. Moreover, the ordering information allows us to estimate the lower and upper bounds of the perceptual strength at an image location, such that we can quantify the "boundary strength" of the image based on the ensemble of human labels.

6.1 Introduction

Boundary detection is a fundamental problem in computer vision. Thanks to the availability of public datasets, boundary detection algorithms have learned from images labeled by multiple human observers. The Berkeley Segmentation Data Set (BSDS300) [71] is the most widely used dataset in this field. Over the last 10 years, a great number of boundary detection algorithms with increasing computational complexity have been tested on this benchmark. Their performance has increased a mere 7 percent, from 65% [95] to 72% [85], where 50% is chance performance. This relatively slow progress persuades us to look back and re-examine the human factors of this challenging problem.

As we have discussed in Chapter 5, the human labeling process is not perfect. Noisy human

data plus the current evaluation framework has many detrimental effects on the algorithms. In this section, we will go into a deeper analysis on modeling the human labeling process.

There are two basic criteria to evaluate the quality of a dataset: 1) The dataset has to be *precise* enough such that hand drawing variance between two human labelers with the same intention are within some acceptable threshold. 2) The dataset has to be *consistent* enough such that different human subjects agree on similar intentions.

The precision of human drawing, given no ambiguity in the intention, is quantitatively discussed in [71] and [96]. Despite different experiment setup, both studies reach the same conclusion that in the controlled condition, human drawing is very reliable, and the variance can be controlled within several pixels.

However, if we look into a dataset with multiple labelers, such as BSDS300, we often see huge variance among different labelers, way beyond several pixels. These differences are mainly due to the different interpretations of the labeling task. In this chapter, we analyze the label inconsistency issues from a perspective of human factors. We present a framework based on partially ordered sets that explores the possibility of putting the human labels into a hierarchy, as well as giving a network among the relationship of different labelers.

6.1.1 Related works

Besides the well-known BSDS300, there are several other datasets of boundary detection, such as PASCAL-VOC [57] and Weizmann horses [72]. However, these datasets are not well suited for human label analysis, because they usually employ only one labeler per image. Recently, the BSDS300 dataset has been extended by BSDS500 [70] with 200 new images and 1000 more label images. This extension intends to preserve the data consistency with the original BSDS300. Therefore, we expect that our conclusion in BSDS300 will also be applicable to BSDS500. One issue with the BSDS500 dataset is that it does not include the labeler/image correspondence (e.g. Subject 1013 is the 3rd labeler to image 887253). Therefore we cannot build a model for individual subjects. Throughout this chapter, we focus on the original BSDS300 dataset.

In the previous literature, Cole *et al.* presented in [96] a quantitative study of human line drawings. They recruited trained artists and examined their line sketches of 3-D models. The main motivation of their paper was to establish the connections among properties of a computer graphics model and human responses. One intriguing observation from [96] is that even in this well-controlled environment, quite different styles of drawing lines persist, indicating a non-negligible

variance in line patterns. As we will show in the rest of our discussion, such variance can be a severe source of noise in a less constrained environment, such as BSDS300.

The analytical approach that we adopt in this paper shares similarities with the recent trend of *crowdsourcing* [86], [88]. As large-scale human annotation (using, for example, Amazon's *Mechanical Turk*) becomes a popular source of ground truth, the quality control of human output receives more and more attention. Typically, a crowd-source model aims to facilitate the annotation process by producing more accurate labels based on noisy output of many individuals. However, as we have discussed in the previous chapter, simply merging all human data with a high level of inconsistency could be problematic while benchmarking. When confronted with natural scenes, different human observers have quite different responses, resulting in labeling inconsistencies. The source of such inconsistency is due to a variety of top-down factors, including different knowledge background, cognitive reasoning, and vicissitudes of top-down attention. Yet, today's generic boundary detection algorithms are not sophisticated enough to reliably model any of the above mentioned high-level processes. To bridge the gap between algorithms and the black box of cognitive process, we refine the noisy human data from a perspective of an algorithm, with an emphasis on data consistency. As a result, inconsistent data are excluded from the training and testing sets.

This chapter is organized as follows. Section 6.2 begins with a qualitative, conceptual model based on our current understanding of human label characteristics. We then introduce a series of examples that illustrate the problem of finding a consistent subset of human labels. In Section 6.3, we formulate the problem of consistency as the construction of a partially ordered set based on noisy observations. We propose an algorithm that not only reconstructs the hierarchy of the partially ordered set, but also infers the hidden variables of each labeler at every location. Finally, we discuss the potential limits of boundary detection.

6.2 Heuristics of boundary labels

In BSDS300, 28 human labelers contributed a total number of 1667 boundary maps associated with 300 images of natural scenes (200 training, 100 testing) in this dataset. During labeling, the software interface always superimposes the subject's annotations onto the input image. This ensures that there is no global distortion or offset. The main difference between two subjects can be well analyzed by comparing subjects' drawing patterns around a small neighborhood. In our analysis, the basic unit is the local patch. If we take a closer look at the local difference among subjects, we

can roughly categorize them into two classes: *imprecision* and *inconsistency*.

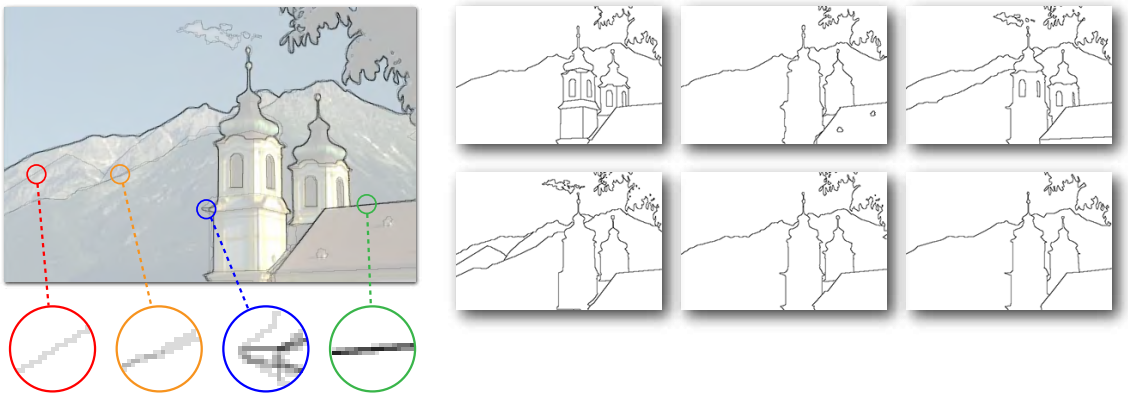


Figure 6.1: The left image contains 6 boundary maps, superimposed onto the original input image. The density of each line reflects the number of overlapping line drawings. The right figure shows individual boundary maps for 6 subjects.

Fig. 6.1 gives both examples of imprecise and inconsistent labels. The red circle in the image (zoomed-in at the bottom on the left) contains one boundary labeled by a sole subject, No. 4. This is an extreme example of *inconsistency*. According to Chapter 5, instances in which a single subject labels a boundary are referred to as *orphan labels*. The orange circle has two labels (subject 3 and 4), which is mildly inconsistent. The blue circle contains boundaries of all 6 subjects, but some of the boundaries are misaligned with each other. This is an example of *imprecision*. Finally, the green circle encompasses a remarkably strong edge, which is unanimously labeled by all 6 subjects. As we will discuss in Section 6.3.1, inter-subject difference due to imprecise labels is not a severe problem. The main focus of the analysis is to propose a possible remedy for label consistency.

6.2.1 Label Consistency

By comparing orphan labels (the red patch) with orange and green patches in Figure 6.1, we can clearly see a continuous transition of consistency among label patches. Surprisingly, there has been little quantitative study on label consistency over the past ten years. Today's most popular benchmark, BSDS300, adopts an extremely lenient policy on label consistency: every boundary pixel weighs equally. For both training and testing, the ground-truth boundary map is the *union* of all subjects' maps.

In addition to the *precision bubble* that we have discussed in the previous section, this straightforward procedure suffers from two more problems:

1. Interesting and potentially useful label information remains unexploited. The consistency of a label is a natural reflection of the *perceptual strength* of the boundary. As we will discuss in Section 6.3.4, this property is related to the identification of salient boundaries.

2. The total number of labelers is not evenly distributed among images. Labelers per image ranges from 4 to 9. Moreover, different images are assigned to different sets of labelers. Under the above mentioned procedure, a single, diligent individual can change the label statistics of the entire image. Fig. 6.2.A shows such an example.

Figure 6.2: A) has 9 labelers. Notably, this image includes subject 1130, one of the most cautious labeler of the entire dataset. In comparison, none of the 4 labelers in B) is patient enough to label the individual logs of the hut or rocks in the wall behind the log house.

6.2.2 Hierarchy of perceptual organization

To understand inconsistent labels, [97, 71, 81] have hypothesized a model of perceptual organization hierarchy. This model assumes a perceptual tree over an image. Deeper nodes (far away from the root) correspond to weak, detailed boundaries; top nodes are the strongest contours in a scene. The same image always elicits the same perceptual organization across different viewers. When labeling, subjects choose their own scales and draw boundary maps of different levels of granularity.

This perceptual hierarchy makes the problem of boundary detection tractable. One algorithm is expected to give one output boundary map for each image. Within the single boundary map, the hierarchy is represented by the output value at each boundary location. In previous literature, the continuous valued boundary detections are called *probability of boundary* in [76], or boundary saliency in [81, 80]. These terms share similar insights to our model.

6.2.3 In search for globally consistent labels

The idea of perceptual hierarchy explains a considerable portion of inconsistent boundaries of the image, and has influenced the designs of many algorithms [81, 70]. Nevertheless, counter-examples do exist. For instance, Fig. 6.3 gives a few examples: these inconsistent labels (red and green) cannot be explained as different choices of perceptual scales. The label process seems to be influenced by a combination of many factors, such as object tracer (in Fig. 6.3.1, tracing the bough even when strong edge is absent), top-down attention (labeling one side of the image more carefully than the other side), line geometry (preference of straight lines of the architecture detail or curved natural lines), or even object types (shadow and pebble detectors).

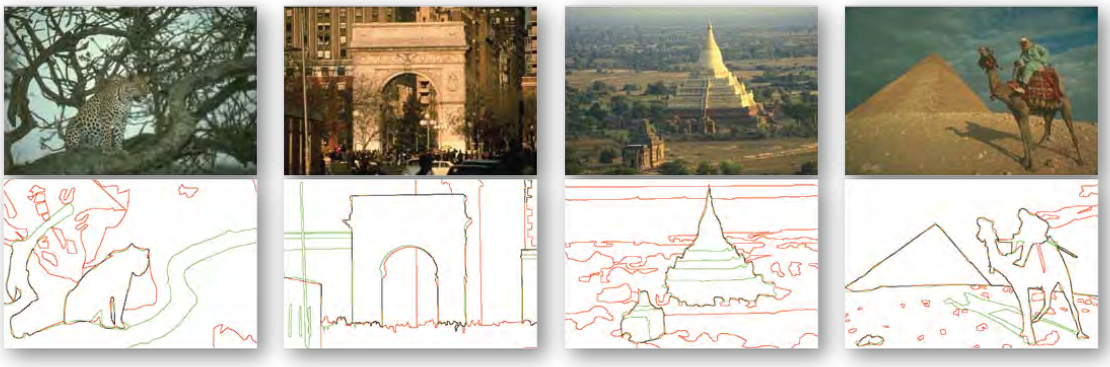


Figure 6.3: In each image, we manually select 2 labelers (each image has 5 or 6 labelers), and use red/green lines to distinguish their boundary maps. If two labels overlap with each other, the line is shown in black color.

Instead of giving a fine-grain categorization on types of boundaries, the analysis in this section aims at using the hierarchical perception model to explore a subset of the data while maintaining a globally consistent and constant threshold for each subject. In the next section, we will present an iterative algorithm that compares all the images that two labelers have mutually labeled, determines their partial order, constructs a full graph that demonstrates subject relationships, and finally exploits the hidden variable of perceptual strength at each boundary location.

Prior to our work, Martin *et al.* in [71] briefly mentioned global inconsistency. They conclude that the perception of different subjects is reconcilable.[1]. The evidence to support their argument is: 1.) inter-subject error in the same image is lower than that of a different random image; and 2.) human error is smaller than the error of the *Normalized Cut* algorithm [98]. Even with

[1]In the original paper, they used the term "consistency." We avoid using the same term, since in our context, consistency refers to a very specific type of similarity.

impeccable quantitative analysis, this evidence is not strong enough to prove the global consistency of the data. In the next section, with more accurate analysis, we will quantify the ratio of precise/consistent/globally consistent labels throughout the entire BSDS300.

6.3 Modeling boundary labeling as a partially ordered set

At each location l of the image, we assign a value $\tau(l)$ that represents the *perceptual strength* of the boundary. For each subject j, a different *perceptual threshold* τ_j determines her/his threshold during labeling, such that any boundary that satisfies $\tau(l) \geq \tau_j$ will be annotated.

We assume a constant non-negative perceptual threshold $\tau_j \in \mathbb{R}_{\geq 0}$ for subject $j \in \mathcal{J}$, where $\mathcal{J}$ is the set of all 28 labelers. $\mathcal{J}_m$ is the set of subjects that have labeled image m. Similarly, we define the total set of images as $\mathcal{M}$, and the set of images labeled by subject j as $\mathcal{M}_j$. Since the set of perceptual threshold is defined on $\mathbb{R}_{\geq 0}$, it has a total order. Without loss of generality, we define $\tau_1 \leq \tau_2 \leq \tau_3 \ldots \leq \tau_{28}$.

In image m, the set of subject j's label pixels (the support) is denoted as ψ_j^m. The upper index m can be omitted if there is no confusion. A local patch of label by subject j on image m, location l is denoted as $x_j^m(l)$. The number of label pixels in a patch is denoted as $|x_j(l)|$. In our experiments, each local patch is 33×33 pixels. The objective of our analysis is to find the maximal subset $\bigcup \phi_j^m \subseteq \bigcup \psi_j^m$ for all $j \in \mathcal{J}$ and $m \in \mathcal{M}$, such that the consistency criterion we mentioned in Section 6.2 is properly met.

Each location l has a perceptual boundary strength $\tau(l)$, which is the same constant for all subjects. During labeling, a subject perceives the hierarchical representation of the entire scene, and compares it with her/his internal perceptual threshold τ_j. Once $\tau(l) \geq \tau_j$, the subject annotates the boundary. Throughout this implicit mental reasoning, the only observed variable is $x_j(l)$. The strategy of our analysis takes 3 steps. First, we establish the connection from τ_j and $\tau(l)$ to $x_j(l)$. This is to define a pairwise relation between $x_i(l)$ and $x_j(l)$. On one hand, such relation preserves the implicit partial order among $\{\tau_j\}$, so that $\{x_j(l)\}$ form an ordered set. On the other hand, the relation between $x_i(l)$ and $x_j(l)$ can be inferred from their appearance (the pixels). Then, in the second step, we simultaneously filter out samples from $\{x_j(l)\}$ that do not follow the 3 properties of a partial set, and reconstruct a graph based on observations. Because such a reconstruction is order-isomorphic to $\{\tau_j\}$, we will be able to determine the orders of the implicit perceptual variables τ_j. Finally, with the partial order structure and consistent labels, we can further estimate the bounds of

$\tau^m(l)$ of an image independent of the subjects.

6.3.1 Human labels are precise

Before we proceed to the formal analysis of consistency, it is necessary to check whether the human labels are precise enough for further analysis. It is worth mentioning that label precision should be examined independently from label consistency – patch pairs can be precise but not consistent (e.g. one patch is the proper subset of the other).

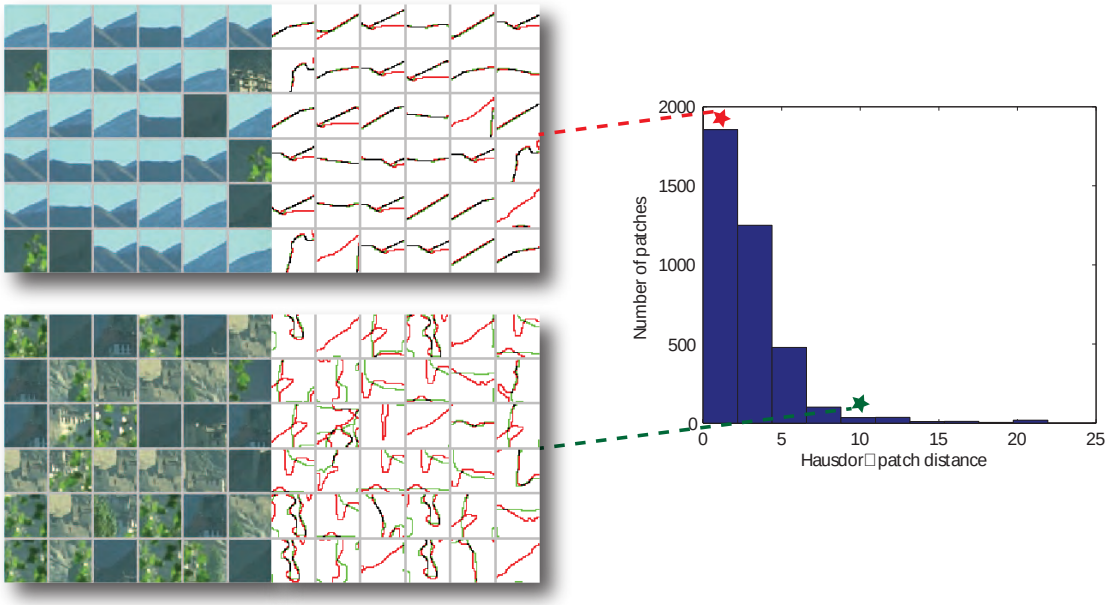


Figure 6.4: The left two figures show patches of precise (top) and imprecise (bottom) labels, and their corresponding image patches. The right figure gives the histogram of the Hausdorff patch distance. Stars on each bin show the bin on which the left figures are sampled.

Given two patches $x_i(l)$ and $x_j(l)$, where $|x_i(l)| \leq |x_j(l)|$, we define pairwise Hausdorff patch distance as:

$$d_H\big(x_i(l), x_j(l)\big) = \sup_{p \in x_i(l)} \inf_{q \in x_j(l)} d(p, q), \tag{6.1}$$

where $d(\cdot, \cdot)$ is the Euclidean distance.

Eq. 6.1 is defined in a slightly different way from the Hausdorff distance in metric space. Correspondence is asymmetrically computed from the sparse patch $x_i(l)$ to the dense patch $x_j(l)$. This property guarantees that for $x_i(l) \subseteq x_j(l)$, $d_H\big(x_i(l), x_j(l)\big)$ is always equal to 0. At the top left of Fig. 6.4, we can see several instances of precise but inconsistent label patches.

Our results on label precision correlate well with that of [96]. They reported that 75% of the label pixel distances are less than 1mm[2]. To keep the model simple, in the rest of this section, we will neglect the influence of imprecise labels.

6.3.2 Label consistency and the partial order set

The first step toward order analysis is to define the pairwise relation between $x_i(l)$ and $x_j(l)$. Because the label patches are noisy, we use a soft likelihood function instead of hard decision boundary to characterize the inequality relation between $x_i(l)$ and $x_j(l)$:

$$p(x_i(l) \leq x_j(l)) = \begin{cases} 1 & \text{if } |x_i(l)| \leq |x_j(l)| \\ |x_j(l)|/|x_i(l)| & \text{otherwise.} \end{cases} \tag{6.2}$$

We call $x_i(l)$ and $x_j(l')$ *comparable* (denoted as $x_i(l) \sim x_j(l')$), iff $x_i(l) \leq x_j(l')$, or $x_j(l') \leq x_i(l)$. Throughout our analysis, $\leq$ is defined at the same location where at least one subject has labeled. In other words:

$$x_i(l) \sim x_j(l') \quad \text{iff } l = l', l \in \phi_i \cup \phi_j \tag{6.3}$$

To improve the clarity, we use Φ to denote the proper union of consistent labels in which $\sim$ holds. Similarly, Ψ is the proper union of ψ_j of comparable raw labels. For instance, in Eq. 6.3, $\Phi = \phi_i \cup \phi_j$.

In an abstract way, we can formulate the process of labeling as a mapping $f : \{\tau_j\} \times \{\phi_j\} \rightarrow \{x_j(l)\}$. Without taking the risk of speculating a detailed functional form of this mapping, the only assumption of f is to be reverse-order-preserving, such that:

$$\tau_i \leq \tau_j \Rightarrow x_i(l) \geq x_j(l), \tag{6.4}$$

where $f(\tau_i, l) = x_i(l)$, $f(\tau_j, l) = x_j(l)$, and $l \in \Phi$.

The implicit total order relation of $\{\tau_j\}$ and the reverse-order-preserving mapping of Eq. 6.4 in together guarantee the partial order of the consistent subset of $\{x_j(l)\}$, where $l \in \Phi$. For each pair of subjects with non-empty intersect $\mathcal{K}_i \cap \mathcal{K}_j$, we can plot the 2D histogram of pairwise comparison of $|x_i(l)|$ and $|x_j(l)|$ at any location $l \in \Psi$ (see Fig. 6.5), and make inference of the ordering of τ_i

[2] 1mm is about 4 pixels in a 96 DPI display device.

and τ_j by:

$$p(\tau_i \leq \tau_j) \propto \mathcal{L}(\tau_i \leq \tau_j) = \prod_{l \in \Phi} p(x_i(l) \leq x_j(l)). \tag{6.5}$$

where $\mathcal{L}(\tau_i \leq \tau_j)$ is the un-normalized likelihood function.

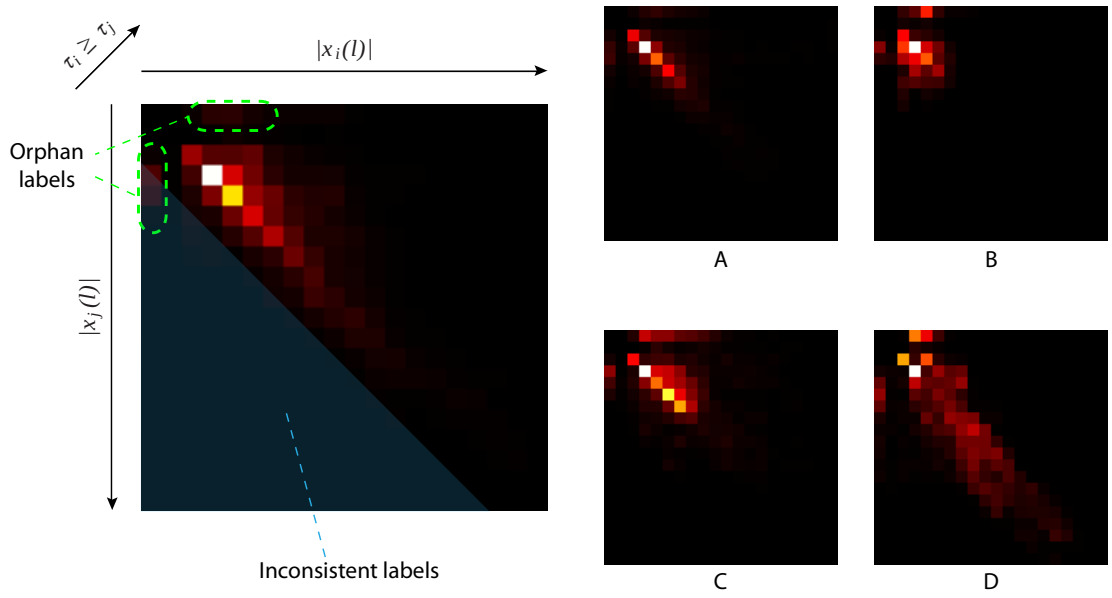


Figure 6.5: This figure shows several examples of the joint 2D histogram of $|x_i(l)|$ and $|x_j(l)|$. Green ellipses indicate bins that correspond to orphan labels (either $|x_i(l)|$ or $|x_j(l)|$ equals zero). From global statistics, we find that subject i's labels are in general more dense than subject j's. With the hypothesis $\tau_i \geq \tau_j$, we conclude that all of the samples from the lower triangular matrix are inconsistent (shown in transparent blue triangle). The right side of the figure shows more pairwise histograms. In particular, A is an example where 2 subjects give highly-correlated labels. In C, even though the labels are not perfectly matched, it is likely that subject j (y-axis) is a subset of subject i, and their correspondent labels can be well-explained by the perceptual threshold model. Histograms B and D both contain a considerable number of inconsistent samples.

In principle, the samples from Eq. 6.5 should be drawn from the consistent subset Φ in order to estimate the likelihood of the ordering. However, the consistent subset construction, in turn, depends on the ordering determined by Eq. 6.5. To solve this chicken-egg problem, we propose a greedy algorithm that verifies the order of one pair of subjects at a time, and update the consistent subset incrementally.

6.3.3 Analyzing global ordering among subjects

Without loss of generality, we can represent the partial ordering among all subjects using a Directed Acyclic Graph, denoted as $\mathcal{G}$. Each node in $\mathcal{G}$ represents a subject, and the directed edge indicates

their partial order. For subject i, j with $\tau_i \leq \tau_j$, there is a path in $\mathcal{G}$ such that $i \to j$.

Despite the relatively straightforward estimation of pairwise ordering between two subjects, the reconstruction of a global ordering faces two more issues worth mentioning. First, $\mathcal{G}$ has to be acyclic; otherwise, we will have $\tau_i \leq \tau_j, \tau_j \leq \tau_k$, but $\tau_k \leq \tau_i$ that ends up with the trivial solution of $\tau_i = \tau_j = \tau_k$. Second, In BSDS300, some pairs of subjects i, j have $\mathcal{M}_i \cap \mathcal{M}_j = \emptyset$. This fact makes $\mathcal{G}$ more complex than a chain. In other words, the observation given by this dataset leaves some nodes non-comparable.

Our procedure is to estimate Φ with the best of our knowledge of $\mathcal{G}$. In each iteration, we add one edge to $\mathcal{G}$ and update the sampling weights for Φ, until we encounter a cycle in the graph. At t^{th} iteration, we denote the sampling weight of $x_j(l)$ as $w_j^t(l)$ and the graph as $\mathcal{G}^t$. The algorithm is as follows:

Algorithm: DAG construction

---

1. To initialize, set $w_j^0(l) = 1$ and $\mathcal{G}^0$ as an edgeless graph with 28 nodes.
2. Sample $\hat{\Phi}$ from Ψ, with probability $w_j^t(l)$.
3. $\forall \{i, j\} \notin edge\{\mathcal{G}^t\}$, compute $\mathcal{L}_{\hat{\Phi}}(\tau_i \leq \tau_j)$ and $\mathcal{L}_{\hat{\Phi}}(\tau_j \leq \tau_i)$ where $\mathcal{L}_{\hat{\Phi}}(\tau_i \leq \tau_j) = \prod_{l \in \Phi} p(x_i(l) \leq x_j(l))$.
4. Compute the confidence $C(i, j)$ and $C(j, i)$, where $C(i, j) = \mathcal{L}_{\hat{\Phi}}(\tau_i \leq \tau_j) / \mathcal{L}_{\hat{\Phi}}(\tau_j \leq \tau_i)$.
5. Find $i^\star, j^\star = \arg\max C(i, j)$.
6. Add $\{i^\star, j^\star\}$ to $edge\{\mathcal{G}^t\}$.
7. Find all descendant $desc\{i^\star\}$ of $i^\star$.
8. For each $k \in desc\{i^\star\}$, update $w_k^t(l) = w_k^{t-1}(l) \cdot p(x_k(l) \leq x_{i^\star}(l))$.
9. If $\mathcal{G}^t$ is DAG, goto step 2. Otherwise, terminate with $\mathcal{G}^\infty = \mathcal{G}^t$, $w_j^\infty(l) = w_j^t(l)$.

Once terminated, this greedy algorithm returns $w_j^t(l)$ as its probability of being a consistent sample. Fig. 6.8 contains several comparisons of the result. Fig. 6.7 shows the Hasse diagrams of the partially ordered set of all 28 subjects. Each arrow represents a $\leq$ relation. For instance, $7 \to 11$ means that subject 7 is more rigorous than subject 11 ($\tau_{11} \leq \tau_7$). The number on each arrow indicates the size of commonly labeled images of two subjects ($|\mathcal{M}_7 \cap \mathcal{M}_{11}| = 22$). This quantity can be considered as the strength of the edge. A Hasse diagram simplifies the edges by removing superfluous edges induced by transitivity. For example, since $\tau_8 \leq \tau_{11}$ and $\tau_{11} \leq \tau_7$, we

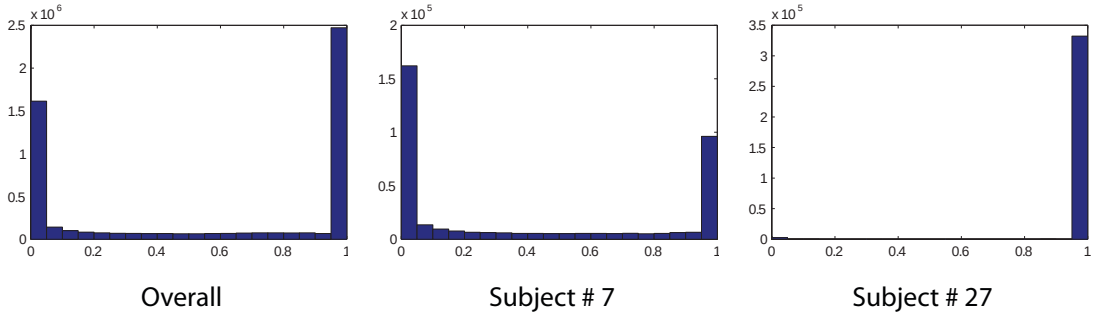


Figure 6.6: The histograms of w_j^∞. According to $\mathcal{G}^\infty$, subject # 7 has the highest τ among all subjects, which makes him a subset in almost every pairwise comparison. As a result, any offset labels that subject #7 made are considered inconsistent, and will be greatly suppressed. In comparison, subject 27 is among the most lenient labelers. Being a superset of almost every one, this subject is rarely penalized for inconsistency, thus $w_{27}^\infty(l) \approx w_{27}^0(l)$.

have $\tau_8 \leq \tau_7$.

The stability of our greedy DAG construction algorithm can be evaluated by analyzing the Hasse diagrams of different runs. We run the algorithm 30 times, and Fig. 6.7.A shows the common graph that is identical throughout all runs. Fig. 6.7.B-D plots three individual runs of the algorithm. To illustrate the difference, we highlight the difference between each single-run Hasse diagram and the Hasse diagram of common graph (Fig. 6.7.A) by illustrating the numbers of common images in red. The difference among these figures are very small, and usually due to the unstable statistics from small number of images.

Our numerical experiments show that this algorithm is extremely stable. All of our five experiments converge after 230 iterations. This number is remarkably high, considering that BSDS200 only has 271 pairs of comparable $\{i, j\}$ where $\mathcal{M}_i \cap \mathcal{M}_j \neq \emptyset$. In comparison, if we directly compare all samples, and build the DAG based on the confidence value ranking, the graph contains 115 edges. Therefore, the most likely explanation to this favorable DAG-preserving greedy algorithm is that it accurately captures properties of the total order set $\{\tau_j\}$, which is acyclic by itself.

It is also helpful to look into the distribution of $w_j^\infty(l)$. Fig. 6.6 shows 3 histograms of all $w_j^\infty(l)$ of the entire dataset, as well as two representative subjects. Our observation throughout the overall distribution of $w_j^\infty(l)$ is that the separation of "consistent" and "inconsistent" labels is extremely sharp, suggesting that the conceptual dualism of label consistency is a very accurate description of the real data. To answer the question we set forth in the introduction, about 43% of the labels have $w_j^\infty(l) < 0.5$.

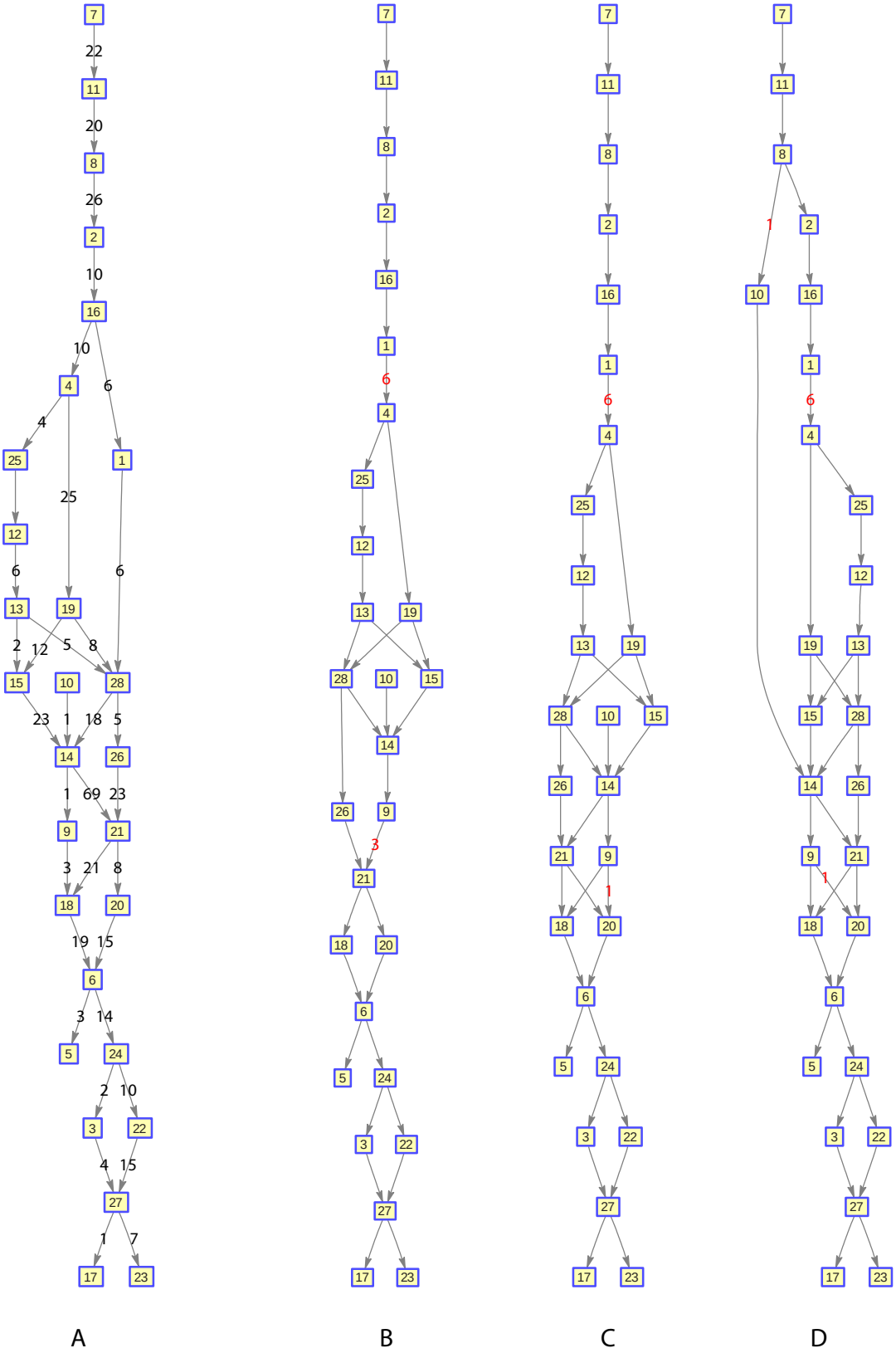


Figure 6.7: The Hasse diagrams of subjects relationships. A: the common graph that is identical throughout all 30 runs. B-D: Hasse diagram of individual runs.

6.3.4 Determine the saliency of boundary labels

Because $\mathcal{G}^{\infty}$ is not fully connected, we cannot obtain the total order for every $\{\tau_j\}$. Therefore, it is guaranteed that we cannot quantify τ_j, which is to map τ_j onto the total ordered set of $\mathbb{R}_{\leq}$. However, it is easy to show that on each image, the subgraph $\{i, j\}$, where $i, j \in \mathcal{J}_m$ is fully connected.

For image m, the total number of consistent labels of subject j is $|\phi_j^m|$, which can be estimated as follows:

$$E(|\hat{\phi}_j^m|) = \sum_{l \in \psi_j^m} w_j^{\infty}(l). \tag{6.6}$$

We can rank the subjects based on Eq. 6.6. By definition, we have $\tau_i \leq \tau_j$ if $E(|\hat{\phi}_j^m|) \leq E(|\hat{\phi}_i^m|)$. Without loss of generality, we assume the subjects in $\mathcal{J}_m$ form a chain with order: $\tau_1 \leq \tau_2 \leq \ldots \leq \tau_j \leq \ldots$. Given the total order[3], we can further infer the rank of each label $x_j^m(l)$ and consequently infer the perceptual strength $\tau(l)$:

$$\begin{aligned} &\sum_{i=1}^{j-1} |\hat{\phi}_j^m| \leq rank\big[x_j^m(l)\big] \leq \sum_{i=1}^{j} |\hat{\phi}_j^m| \\ &\inf rank\big[x_j^m(l)\big] \leq \tau(l) \leq \sup rank\big[x_j^m(l)\big]. \end{aligned} \tag{6.7}$$

Finally, a comparative illustration of our perceptual strength analysis is shown in Fig. 6.8.

6.4 Discussions

Today, the standard approach for boundary detection algorithm benchmarking is using the PR-curve. Similar to the perceptual threshold τ_j, an algorithm changes its threshold and generates a single binary map of boundary detection. As a convention, the "optimal" threshold is determined by the geometric mean of precision and recall, and this measure is called "F-measure". Given the fact that 43% of the labels are inconsistent, it will be interesting to see how the threshold selection is correlated with label consistency.

In the boundary map produced by an algorithm, each non-zero pixel of detection corresponds to a human label. Since we have obtained the consistency at each human label location, it is straightforward to compute the proportion of inconsistent labels at each step of algorithm label change. Because of its popularity and superior performance, we examined the algorithm output

[3]In a probabilistic setting where ϕ_j is substituted by $\hat{\phi}_j$, we take the probability as a linear weight multiplied to the ascending ranking of each label.

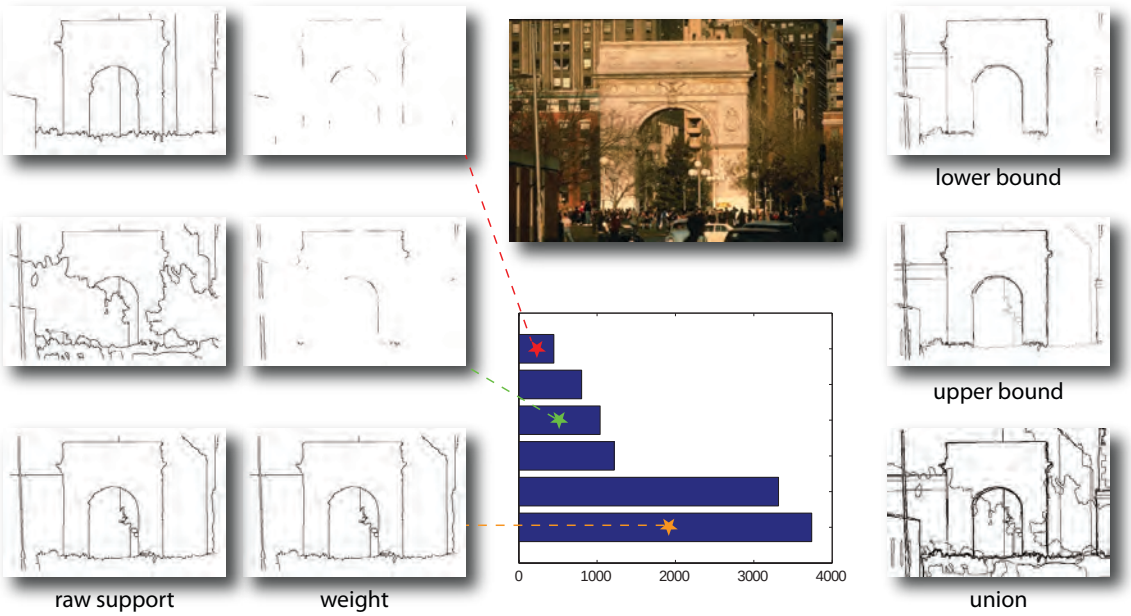


Figure 6.8: The Arc de Triomphe is labeled by 6 subjects. The leftmost column shows the supports ψ_j for 3 subjects. In the second column, the corresponding $\hat{\phi}_j$ are presented. The histogram in the third column shows the expected consistent label size $E(|\hat{\phi}_j^m|)$ for all 6 subjects. The star on each bar illustrates the order of each subject on the left 2 columns. The second subject (green star) is an excellent example of how dramatically the partial order suppresses inconsistent labels. With the totally ordered set, we can bound each label. The lower and upper bound of the boundary perceptual strength, as well as the union of all subjects (which is the current implementation of the ground truth) are presented in the rightmost column.

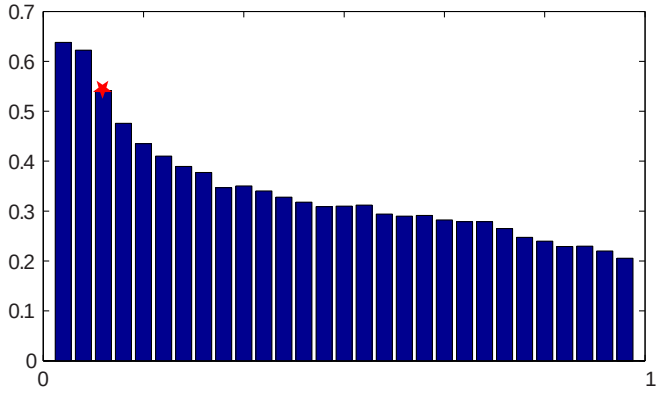


Figure 6.9: Proportion of matching to inconsistent labels as a function of threshold. The red circle shows the "optimal" threshold where the geometric mean of precision and recall is maximized.

of [70] on 100 test images of BSDS300. Using a similar protocol to [76], we split the algorithm threshold into 30 bins, and perform sparse alignment using [75]. Our statistical result is presented in Fig. 6.9. Even though the dataset contains a considerable 43% of inconsistent labels, the "optimal" threshold matches an even bigger portion (54%). The trend of "overfitting" to inconsistent data is observed in most of the major algorithms, such as [76, 81]. It seems that all of these algorithms are motivated to fit on the less important boundaries in order to maximize the benchmark scores. This phenomenon encourages us to rethink the value of today's boundary detection benchmarks, and their relative ranks.

6.4.1 Conclusion

In this chapter, we present a meta-analysis for the human labeled dataset for boundary detection. We use a quantitative approach to investigate the representational limit of the "perceptual hierarchy". With the partially ordered set as a conservative yet powerful tool, we propose a reconstruction algorithm that recovers the ordering relation of the scene. The reconstructed partial order not only helps us to determine the consistency of the data, but also empowers us to bound the perceptual strength, or the "saliency", of a human labeled boundary. Finally, our algorithm also points out the potential bias of using F-measure as the only score to evaluate boundary detection algorithms.

It is commonly believed that human labelers differ only by their degree of granularity, and hence correspond to a partially ordered set. We formalize a model to test this hypothesis using concepts like perceptual strength and perceptual threshold. We analyze the Berkeley dataset using this parametric hierarchical model. Our results show that, by contrast, there are many inconsistencies

between labelers that cannot be explained by this model. Indeed, nearly half of the label data cannot be explained by our parametric hierarchical model. These inconsistencies may be due to attentional or other causes (see Arc de Triomphe figure). Then, we use the our model to reconstruct the partially ordered set of labels most consistent with the data. This reconstructed partial order not only helps us determine the consistency of the data, but it also enables us to bound the perceptual strength, or "saliency", of human labeled boundaries. We also examine the performance of the most popular edge detection algorithms and show that they may be "overfitting the data" by paying too much attention to the inconsistent and less important edges. Our analysis also points out the potential dangers of using F-measures as the only way to evaluate boundary detection algorithms.

Chapter 7

Discussions

In Chapter 2, we analyze the problem of figure-ground separation, and introduce the Image Signature theory as an efficient algorithm to highlight important regions of the entire image. The theoretical foundation of Image Signature relies on sparsity of the image signal, which is universal in most of the natural images, making this algorithm robust under many scenarios. With extensive experiments, we show the efficacy of spectral analysis in predicting human fixations of natural images. Moreover, the search asymmetry on synthetic images, the change-blindness experiments on natural image pairs, and the face orientation prediction experiment on aligned face photos further corroborate our conclusion that the Image Signature algorithm has a close resemblance to the perceptual organizations of the human in tasks related to figure-ground separation. Based on the same spectral analysis technique, in Chapter 3, we extend the Image Signature to video and propose the Phase Discrepancy algorithm that further utilizes the phase information of the Fourier transform. This algorithm gives a simple treatment to compensate for camera ego-motion, and provides excellent results in detecting figures in complex and dynamic scenes.

Whereas Chapter 2 and Chapter 3 focus on the first step of figure-ground separation, in Chapter 4, we move forward from localization to segregation by proposing a new dataset with simultaneously recorded eye fixation data and salient object segmentation data. Thanks to this new dataset, our model bridges the gap between fixation prediction and salient object segmentation. It allows us to build today's best performing salient object segmentation algorithm without reinventing the wheel.

The dissociation of saliency and object segmentation persuade us to investigate the problem of segmentation and contour detection in a more generic sense. In Chapter 5, we analyze one of the most popular datasets, and estimates its reliability in benchmarking computer vision algorithms. Our conclusion is quite astonishing: with a novel psychophysical experiment, we found that over

30% of the labels are invalid, and should not be considered as part of the benchmark. As a result, the goal of the benchmark goes awry from its original purpose. To most boundary detection algorithms, their outputs seem less relevant to the high-level tasks than they were supposed to be. The conclusion of our analysis in this section suggests new directions for the boundary detection community.

Finally, in Chapter 6, we analyze the human labeling process of the ground-truth of a boundary detection dataset to give a quantitative estimate the intrinsic properties of boundaries. Using the partially ordered set as a simple yet powerful tool, we formulate the labeling process as a perception hierarchy, where different people have different levels of granularity. The ordering information allows us to estimate the lower and upper bounds of the perceptual strength at an image location, such that we can 1) exclude the "unreasonable" labels that an algorithm is guaranteed not able to recover, and 2) quantify the "boundary strength" of the image based on the ensemble of human labels.

Bibliography

[1] X. Hou, J. Harel, and C. Koch, "Image signature: Highlighting sparse salient regions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 1, pp. 194–201, 2012.

[2] L. Itti, C. Koch, E. Niebur, *et al.*, "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.

[3] N. Bruce and J. Tsotsos, "Saliency based on information maximization," *Advances in neural information processing systems*, vol. 18, p. 155, 2006.

[4] X. Hou and L. Zhang, "Dynamic visual attention: searching for coding length increments," in *Advances in Neural Information Processing Systems*, vol. 5, p. 7, 2008.

[5] A. Garcia-Diaz, V. Leborán, X. R. Fdez-Vidal, and X. M. Pardo, "On the relationship between optical variability, visual saliency, and eye fixations: A computational approach," *Journal of vision*, vol. 12, no. 6, p. 17, 2012.

[6] B. Zhou*, X. Hou*, and L. Zhang, "A phase discrepancy analysis of object motion," in *Computer Vision–ACCV 2010*, pp. 225–238, Springer, 2011.

[7] Y. Li*, X. Hou*, C. Koch, J. Rehg, and A. Yuille, "The secrets of salient object segmentation," in *Computer Vision and Pattern Recognition, 2014. CVPR'14*, vol. 2014, pp. 1–8, IEEE, 2014.

[8] X. Hou, A. Yuille, and C. Koch, "Boundary detection benchmarking: Beyond f-measures," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pp. 2123–2130, IEEE, 2013.

[9] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International journal of computer vision*, vol. 42, no. 3, pp. 145–175, 2001.

[10] H. Zhou, H. S. Friedman, and R. Von Der Heydt, "Coding of border ownership in monkey visual cortex," *The Journal of Neuroscience*, vol. 20, no. 17, pp. 6594–6611, 2000.

[11] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The feret evaluation methodology for face-recognition algorithms," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 10, pp. 1090–1104, 2000.

[12] A. Torralba, R. Fergus, and W. T. Freeman, "80 million tiny images: A large data set for nonparametric object and scene recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 30, no. 11, pp. 1958–1970, 2008.

[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 248–255, IEEE, 2009.

[14] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?," *Journal of the ACM (JACM)*, vol. 58, no. 3, p. 11, 2011.

[15] M. H. Hayes, J. S. Lim, and A. V. Oppenheim, "Signal reconstruction from phase or magnitude," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 28, no. 6, pp. 672–680, 1980.

[16] A. V. Oppenheim and J. S. Lim, "The importance of phase in signals," *Proceedings of the IEEE*, vol. 69, no. 5, pp. 529–541, 1981.

[17] D. L. Ruderman, "The statistics of natural images," *Network: computation in neural systems*, vol. 5, no. 4, pp. 517–548, 1994.

[18] D. L. Ruderman, "Origins of scaling in natural images," *Vision research*, vol. 37, no. 23, pp. 3385–3398, 1997.

[19] A. Oliva, A. Torralba, and P. G. Schyns, "Hybrid images," in *ACM Transactions on Graphics (TOG)*, vol. 25, pp. 527–532, ACM, 2006.

[20] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pp. 1–8, IEEE, 2007.

[21] C. Guo, Q. Ma, and L. Zhang, "Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1–8, IEEE, 2008.

[22] P. Bian and L. Zhang, "Biological plausibility of spectral domain approach for spatiotemporal visual saliency," in *Advances in Neuro-Information Processing*, pp. 251–258, Springer, 2009.

[23] B. Schauerte and R. Stiefelhagen, "Quaternion-based spectral saliency detection for eye fixation prediction," in *Computer Vision–ECCV 2012*, pp. 116–129, Springer, 2012.

[24] Y. Fang, W. Lin, B.-S. Lee, C.-T. Lau, Z. Chen, and C.-W. Lin, "Bottom-up saliency detection model based on human visual sensitivity and amplitude spectrum," *Multimedia, IEEE Transactions on*, vol. 14, no. 1, pp. 187–198, 2012.

[25] J. Li, M. D. Levine, X. An, X. Xu, and H. He, "Visual saliency based on scale-space analysis in the frequency domain," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 4, pp. 996–1010, 2013.

[26] E. J. Candes and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?," *Information Theory, IEEE Transactions on*, vol. 52, no. 12, pp. 5406–5425, 2006.

[27] M. Rudelson and R. Vershynin, "On sparse reconstruction from fourier and gaussian measurements," *Communications on Pure and Applied Mathematics*, vol. 61, no. 8, pp. 1025–1045, 2008.

[28] M. X. Goemans and D. P. Williamson, "Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming," *Journal of the ACM (JACM)*, vol. 42, no. 6, pp. 1115–1145, 1995.

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

[30] M. Cerf, J. Harel, W. Einhäuser, and C. Koch, "Predicting human gaze using low-level saliency combined with face detection," *NIPS*, vol. 20, 2008.

[31] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Computer Vision, 2009 IEEE 12th international conference on*, pp. 2106–2113, IEEE, 2009.

[32] F. Li, J. Carreira, and C. Sminchisescu, "Object recognition as ranking holistic figure-ground hypotheses," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pp. 1712–1719, IEEE, 2010.

[33] B. W. Tatler, R. J. Baddeley, and I. D. Gilchrist, "Visual correlates of fixation selection: Effects of scale and time," *Vision research*, vol. 45, no. 5, pp. 643–659, 2005.

[34] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "Sun: A bayesian framework for saliency using natural statistics," *Journal of vision*, vol. 8, no. 7, p. 32, 2008.

[35] J. Harel, C. Koch, P. Perona, *et al.*, "Graph-based visual saliency," *Advances in neural information processing systems*, vol. 19, p. 545, 2007.

[36] D. J. Simons and R. A. Rensink, "Change blindness: Past, present, and future," *Trends in cognitive sciences*, vol. 9, no. 1, pp. 16–20, 2005.

[37] R. A. Rensink, J. K. O'Regan, and J. J. Clark, "To see or not to see: The need for attention to perceive changes in scenes," *Psychological science*, vol. 8, no. 5, pp. 368–373, 1997.

[38] T. A. Kelley, M. M. Chun, and K.-P. Chua, "Effects of scene inversion on change detection of targets matched for visual salience," *Journal of Vision*, vol. 3, no. 1, p. 1, 2003.

[39] A. Torralba, R. Fergus, and Y. Weiss, "Small codes and large image databases for recognition," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1–8, IEEE, 2008.

[40] S. Z. Li, Q. Fu, L. Gu, B. Scholkopf, Y. Cheng, and H. Zhang, "Kernel machine based learning for multi-view face detection and pose estimation," in *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, vol. 2, pp. 674–679, IEEE, 2001.

[41] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," in *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*, vol. 2, IEEE, 1999.

[42] A. Mittal and N. Paragios, "Motion-based background subtraction using adaptive kernel density estimation," in *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, vol. 2, pp. II–302, IEEE, 2004.

[43] T. Y. Tian, C. Tomasi, and D. J. Heeger, "Comparison of approaches to egomotion computation," in *Computer Vision and Pattern Recognition, 1996. Proceedings CVPR'96, 1996 IEEE Computer Society Conference on*, pp. 315–320, IEEE, 1996.

[44] M. Han and T. Kanade, "Reconstruction of a scene with multiple linearly moving objects," *International Journal of Computer Vision*, vol. 59, no. 3, pp. 285–300, 2004.

[45] M. Irani and P. Anandan, "A unified approach to moving object detection in 2d and 3d scenes," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 20, no. 6, pp. 577–589, 1998.

[46] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004.

[47] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: A benchmark," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 304–311, IEEE, 2009.

[48] L. Itti and P. Baldi, "Bayesian surprise attracts human attention," *Advances in neural information processing systems*, vol. 18, p. 547, 2006.

[49] D. Vernon, *Fourier vision: segmentation and velocity measurement using the Fourier transform*. Springer, 2001.

[50] M. J. Black and P. Anandan, "A framework for the robust estimation of optical flow," in *Computer Vision, 1993. Proceedings., Fourth International Conference on*, pp. 231–236, IEEE, 1993.

[51] S. Mallat, *A wavelet tour of signal processing*. Academic press, 1999.

[52] "http://ftp.pets.rdg.ac.uk."

[53] "http://homepages.inf.ed.ac.uk/rbf/caviar."

[54] F. Bashir and F. Porikli, "Performance evaluation of object detection and tracking systems," in *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS)*, vol. 5, 2006.

[55] M. Sonka, V. Hlavac, R. Boyle, *et al.*, *Image processing, analysis, and machine vision*, vol. 3. Thomson Toronto, 2008.

[56] T. List, J. Bins, J. Vazquez, and R. B. Fisher, "Performance evaluating the evaluator," in *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*, pp. 129–136, IEEE, 2005.

[57] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results."

[58] J. Carreira and C. Sminchisescu, "Constrained parametric min-cuts for automatic object segmentation," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pp. 3241–3248, IEEE, 2010.

[59] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 2, pp. 353–367, 2011.

[60] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 1597–1604, IEEE, 2009.

[61] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu, "Global contrast based salient region detection," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pp. 409–416, IEEE, 2011.

[62] F. Perazzi, P. Krahenbuhl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 733–740, IEEE, 2012.

[63] R. Margolin, A. Tal, and L. Zelnik-Manor, "What makes a patch distinct?," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pp. 1139–1146, IEEE, 2013.

[64] A. Borji, D. N. Sihite, and L. Itti, "Salient object detection: A benchmark," in *Computer Vision–ECCV 2012*, pp. 414–429, Springer, 2012.

[65] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. H. S. Torr, "BING: Binarized normed gradients for objectness estimation at 300fps," in *Computer Vision and Pattern Recognition, 2014. CVPR 2014. IEEE Conference on*, IEEE, 2014.

[66] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the objectness of image windows," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 11, pp. 2189–2202, 2012.

[67] I. Endres and D. Hoiem, "Category independent object proposals," in *Computer Vision–ECCV 2010*, pp. 575–588, Springer, 2010.

[68] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pp. 1521–1528, IEEE, 2011.

[69] A. Borji, D. N. Sihite, and L. Itti, "What stands out in a scene? a study of human explicit saliency judgment," *Vision research*, vol. 91, pp. 62–77, 2013.

[70] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 5, pp. 898–916, 2011.

[71] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, vol. 2, pp. 416–423, IEEE, 2001.

[72] E. Borenstein and S. Ullman, "Class-specific, top-down segmentation," *Computer Vision–ECCV 2002*, pp. 639–641, 2002.

[73] S. Alpert, M. Galun, R. Basri, and A. Brandt, "Image segmentation by probabilistic bottom-up aggregation and cue integration," in *CVPR 2007. IEEE Conference on*, pp. 1–8, IEEE, 2007.

[74] D. Martin, J. Malik, and D. Patterson, *An Empirical Approach to Grouping and Segmentaqtion*. Computer Science Division, University of California, 2003.

[75] A. Goldberg and R. Kennedy, "An efficient cost scaling algorithm for the assignment problem," *Mathematical Programming*, vol. 71, no. 2, pp. 153–177, 1995.

[76] D. R. Martin, C. C. Fowlkes, and J. Malik, "Learning to detect natural image boundaries using local brightness, color, and texture cues," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, no. 5, pp. 530–549, 2004.

[77] X. Ren and L. Bo, "Discriminatively trained sparse code gradients for contour detection," *Advances in Neural Information Processing Systems*, vol. 25, 2012.

[78] S. Wang, T. Kubota, and J. Siskind, "Salient boundary detection using ratio contour," *Advances in Neural Information Processing Systems*, vol. 16, 2003.

[79] Q. Zhu, G. Song, and J. Shi, "Untangling cycles for contour grouping," in *Computer Vision–ICCV 2007. IEEE Conference on*, pp. 1–8, IEEE, 2007.

[80] R. Kennedy, J. Gallier, and J. Shi, "Contour cut: identifying salient contours in images by solving a hermitian eigenvalue problem," in *CVPR, 2011. IEEE Conference on*, pp. 2065–2072, IEEE, 2011.

[81] P. Arbelaez, "Boundary extraction in natural images using ultrametric contour maps," in *CVPR Workshop, 2006. IEEE Conference on*, pp. 182–182, IEEE, 2006.

[82] P. Felzenszwalb and D. McAllester, "A min-cover approach for finding salient curves," in *CVPR Workshop, 2006. IEEE Conference on*, pp. 185–185, IEEE, 2006.

[83] P. Dollar, Z. Tu, and S. Belongie, "Supervised learning of edges and object boundaries," in *CVPR, 2006 IEEE Conference on*, vol. 2, pp. 1964–1971, IEEE, 2006.

[84] X. Ren, "Multi-scale improves boundary detection in natural images," *Computer Vision–ECCV 2008*, pp. 533–545, 2008.

[85] I. Kokkinos, "Boundary detection using f-measure-, filter-and feature-(f 3) boost," *Computer Vision–ECCV 2010*, pp. 650–663, 2010.

[86] A. Sorokin and D. Forsyth, "Utility data annotation with amazon mechanical turk," in *CVPR Workshops, 2008. IEEE Conference on*, pp. 1–8, IEEE, 2008.

[87] V. Raykar, S. Yu, L. Zhao, A. Jerebko, C. Florin, G. Valadez, L. Bogoni, and L. Moy, "Supervised learning from multiple experts: Whom to trust when everyone lies a bit," in *ICML, 2009. ACM Conference on*, pp. 889–896, ACM, 2009.

[88] P. Welinder, S. Branson, S. Belongie, and P. Perona, "The multidimensional wisdom of crowds," in *In Proc. of NIPS*, pp. 2424–2432, 2010.

[89] S. Vittayakorn and J. Hays, "Quality assessment for crowdsourced object annotations," in *Proceedings of the British machine vision conference*, pp. 109–1, 2011.

[90] T. Leung and J. Malik, "Contour continuity in region based image segmentation," *Computer Vision–ECCV 1998*, pp. 544–559, 1998.

[91] T. Cour, F. Benezit, and J. Shi, "Spectral segmentation with multiscale graph decomposition," in *CVPR, 2005. IEEE Conference on*, vol. 2, pp. 1124–1131, IEEE, 2005.

[92] X. Ren, C. C. Fowlkes, and J. Malik, "Figure/ground assignment in natural images," in *Computer Vision–ECCV 2006*, pp. 614–627, Springer, 2006.

[93] V. Ferrari, T. Tuytelaars, and L. Van Gool, "Object detection by contour segment networks," *Computer Vision–ECCV 2006*, pp. 14–28, 2006.

[94] P. Arbeláez, B. Hariharan, C. Gu, S. Gupta, L. Bourdev, and J. Malik, "Semantic segmentation using regions and parts," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 3378–3385, IEEE, 2012.

[95] D. R. M. C. C. Fowlkes and J. Malik, "Learning to detect natural image boundaries using brightness and texture," in *Neural information processing systems*, 2002.

[96] F. Cole, A. Golovinskiy, A. Limpaecher, H. S. Barros, A. Finkelstein, T. Funkhouser, and S. Rusinkiewicz, "Where do people draw lines?," in *ACM Transactions on Graphics (TOG)*, vol. 27, p. 88, ACM, 2008.

[97] L. Najman and M. Schmitt, "Geodesic saliency of watershed contours and hierarchical segmentation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 18, no. 12, pp. 1163–1173, 1996.

[98] J. Shi and J. Malik, "Normalized cuts and image segmentation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 8, pp. 888–905, 2000.