Neural and Behavioral Consequences of Perceptual Organization using Proto-Objects

by

Daniel M. Jeck

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

March, 2018

Abstract

The human visual system utilizes attention to direct processing towards areas of interest. In particular, certain objects in a visual scene can be salient, meaning they attract attention rather than being the targets of some search process. Visual salience appears to be driven by the formation of visual proto-objects, which have been hypothesized to cause an increase in synchronous firing between neurons encoding parts of an object. This thesis approaches proto-objects both from a behavioral level and at a low level of analyzing synchrony. At the behavioral level, existing studies of visual salience rely on many repetitive trials or task instructions to tell study participants what to do, which can influence attentive behavior in a top-down manner, confounding the measurement of salience. I introduce an experimental paradigm that records attentional selections from subjects without any such information, and used this paradigm to analyze whether proto-objects interact in the determination of salience. The results show that uniqueness of an object does indeed attract attention, and I develop a model that normalizes among proto-objects to explain the measured data. At the neuronal level, I develop a more rapid method to perform jitter hy-

ABSTRACT

pothesis tests regarding detecting the presence of synchronous spiking between pairs of neurons. While the detection of synchrony does imply some connection between neurons, I also show that the inference of a change in common input from changes in synchrony is not possible.

Primary Reader: Ernst Niebur

Secondary Reader: Howard Egeth

Acknowledgments

Thanks to many people for your help through my graduate work. To Tiffany Jeck, thanks for pushing me to go to graduate school and believing in me more than I do myself at times. Thanks to Ernst and my labmates Grant and Brian for creating such a great lab environment, with an incredible amount of freedom available to explore computational neuroscience and great discussions. Thanks to MBI for the rest of that environment. I deeply enjoyed the journal club here and hope to be able to replicate it elsewhere in the future.

Thanks to my parents, brother, and extended family for your guidance and teaching me from a young age what mathematics and science are, both over the diner table and elsewhere.

Finally, thanks to the Biomedical Engineering department for the freedom to rotate in my first year to find this lab.

Dedication

This thesis is dedicated to Cathleen and Herbert Morawetz. Your memory will always be an inspiration to me.

Contents

Abstract **ii**

Acknowledgments **iv**

List of Tables **xi**

List of Figures **xii**

1 Introduction **1**

1.1 Proto-Objects and Visual Salience 3

1.2 Proto-Objects and Synchrony 6

2 Attentive Pointing in Natural Scenes Correlates with Other Measures of Attention **9**

2.1 Introduction 9

2.1.1 Determining saliency from behavior 10

2.1.2 Reporting attended locations by pointing to them 14

2.1.3 Limitations due to map estimation 16
2.2 Methods 17
2.2.1 Apparatus, participants, and procedure 18
2.2.2 Stimuli 19
2.2.3 Data Analysis 21
2.2.3.1 Correlations between maps 21
2.2.3.2 Null hypothesis: Correlations reflect no differences between images 23
2.2.3.3 Hypothesis: Correlations are limited by sampling error 24
2.2.3.4 Population averages 26
2.3 Results 27
2.3.1 Fixations vs. Interest Points 29
2.3.2 Fixations vs. Computed Saliency 29
2.3.3 Interest Points vs. Computed Saliency 30
2.3.4 Fixations *vs.* Tap Points 30
2.3.5 Interest Points vs. Tap Points 32
2.3.6 Tap Points vs. Computed Saliency 33
2.3.7 Coarse Scale Analysis 33
2.4 Discussion 36
2.4.1 A new experimental paradigm for quantitative characterization of attentional selection 36

2.4.2 Effects of sampling error on correlations 41

3 Utilizing the Salience of Unique Objects to Test Models of Salience 46

3.1 Introduction 46

3.1.1 Related Models 49

3.1.2 Psychophysical Support and Limitations 52

3.2 Methods 54

3.2.1 Experimental Paradigm 54

3.2.2 Stimuli 55

3.2.3 Proto-Object Comparison Model 56

3.3 Results 62

3.3.1 First View Only 62

3.3.2 All Presentations 64

3.3.3 POC Model 66

3.4 Discussion 69

3.4.1 Simple Scenes and Models of Salience 69

3.4.2 Top down Influences and Possible Experimental Confounds . . 71

3.4.3 Object-based Models 75

3.4.4 What Are Unique/Rare Objects? 76

3.5 Supplementary Information 78

3.5.1 Linear separability of the unique faint square in feature space 78

3.5.2 Additional Figures 79

4 Closed Form Jitter Analysis of Neuronal Spike Trains **82**

4.1 Introduction 82

4.2 Materials and Methods 85

4.2.1 The Monte Carlo Jitter Method 85

4.2.2 Closed Form Computation 90

4.2.2.1 Probability Distribution For One Interval 90

4.2.2.2 Jitter-Corrected Cross Correlation 92

4.2.2.3 Probability Distribution For Spike Train 94

4.3 Results 94

4.3.1 Computational Complexity 94

4.3.1.1 Monte Carlo Method 95

4.3.1.2 Closed Form Probability Distribution 96

4.3.1.3 Jitter-Corrected Cross Correlation 98

4.3.2 Computational Execution Time 99

4.4 Discussion 102

5 Neuronal Common Input Strength is Unidentifiable from average firing rates and synchrony **110**

5.1 Introduction 110

5.2 Methods 113

5.3 Results 116

5.4 Conclusion 120

Appendix A Appendix **121**

A1 Natural Scenes Tapping Experiment 121

A1.1 Demographics 121

A1.2 Error modes 122

A1.3 Statistical validation 123

Vita **142**

List of Tables

4.1 **Glossary.** Variables are listed in the order in which they are introduced. 86

List of Figures

2.1 Experimental procedure. The rectangles represent an approximation of what was shown to participants on the tablet screen. First, they saw an initialization screen and tapped on either of the small black squares at the bottom. This brought up a test image which alternated between natural scenes and simple scenes. They then tapped on the test image at a place of their choosing which was, according to instructions, the first place they looked at when the test image had appeared. Tapping position and reaction time were collected, the initialization screen reappeared, and the cycle re-commenced. 19

2.2 Data analysis method. (A) Example image overlaid with collected fixation points (blue dots) and tap points (yellow dots), and grid lines used to bin the data. (B) Corresponding fixation map and tap map. Both maps are binned in a 12×16 grid, with each bin showing the average of 64×64 pixels. (C) Surrogate maps generated from the fixation data used to approximate sampling error in the correlation between the fixation and tap data, see text. (D) Comparison of the measured value (red) to the histograms of the null hypothesis (blue) and the sampling error hypothesis (black). Means and standard deviations of the distributions generated from the null hypothesis and the sampling error hypothesis are shown above the distributions. For this image, fixation data and tap data correlate more than predicted by the null hypothesis ($p = 0.002$), and cannot be distinguished from predictions of the sampling error hypothesis ($p = 0.11$). 28

2.3 Aggregate results of natural scene analysis at 12×16 resolution. Each subplot shows a distribution of measured correlations between two types of maps compared against the null hypothesis and sample error hypothesis. Means of each distribution are shown above the histograms, with error bars indicating standard error given the 48 images used. Most error bars are smaller than the markers used. (A) Fixation and Interest maps. (B) Fixation and Computed saliency maps generated from Itti et al. (1998). (C)Interest and saliency maps. (D) Fixation and Tap maps. (E) Interest and Tap maps. (F) Computed saliency and Tap maps. All measured averages are significantly above the null hypothesis ($p < 0.05$). All measured averages are below the sample error hypothesis ($p < 0.05$), with the exception of the comparison between computed saliency and tap maps ($p = 0.08$), panel F. The legend in panel B applies to all panels. For color figures see the online version of the article. 34
2.4 Comparison of $R(\hat{F}, \hat{I})$ and $R(\hat{F}, \hat{T})$ when using only a portion of the interest points and tap points. All fixation data was used to generate $\hat{F}$ for all simulations. 100 Simulations were performed for each number of data points. Standard error is less than line width. For color figures see the online version of the article. 35
2.5 Aggregate results of correlation analysis at coarse resolution, when images were divided in a 3×4 grid. Symbols as in Figure 2.3. For color figures see the online version of the article. 35

3.1 Example salience processing of Itti et al. (1998).(A) One of the experimental stimuli used as input to the model. (B) Three example scales of center-surround (CS) response. At all scales, the gray square has the weakest response. (C) Final output of the model. Intensity represents saliency. Color and orientation channels are included in the computation but they do not make a substantial contribution for this image. 51
3.2 Output of models of salience for the input shown in Figure 3.1A with white regions indicating high salience. (A) Walther and Koch (2006) (B) Russell et al. (2014), (C) Perazzi et al. (2012), (D) Kümmerer et al. (2016). 51
3.3 Stimuli 57
3.4 Model 63

3.5 (A-D) Rates at which participants tapped on singleton (sing.) squares *vs.* non-singleton corresponding squares (Non-sing.) in a control image. Error bars represent standard error. (A) Gray/Black *vs.* All-Black comparison for each participant's first tap. (B) Gray/Black *vs.* All-Black comparison for all taps. (C) Pink/Red *vs.* All-Red comparison for each participant's first tap. (D) Pink/Red *vs.* All-Red comparison for all taps. (E) Rates at which participants tapped on the singleton squares (colored circles, see legend), and each of the various non-singleton squares in the All-Red and All-Black images (green circles). The horizontal axis is the Euclidean distance from the center of the image. Fit lines were generated for each singleton image type individually and for Non-singletons combined, colors same as for the corresponding circle symbols. (F) The vertical intercept of each fit line from (E) with standard error bars (G/B = Gray/Black, P/R = Pink/Red, B/G = Black/Gray, B/Y = Blue/Yellow, NS = Non-singletons). The symbol $\approx$ indicates that no pairwise difference was found ($p \geq 0.05$). All other intercept pairs differed significantly ($p < 0.05$). 67
3.6 Model behavior on an example simple scene (top row) and natural scene (bottom). (A) Input image. (B) Output of the Russell et al. (2014) model. (C) Output of the model without normalization, using S_i instead of P_i in equation 3.3. (D) Output of the POC model. (E) Image with tap data (top) and fixation map (bottom) overlaid. For the natural scene saliency and fixation maps are downsampled to the 12×16 resolution used in Chapter 2. 69
3.7 Raw Data 80
3.8 Raw Data 81

4.1 Computation time 108
4.2 Jitter Corrected Correlogram Computation time 109

5.1 Network Structure. Two LIF neurons (LIF1 and LIF2) receive input with Poisson statistics that is the sum of independent (P_1, P_2) and common (P_C) spike trains. 113
5.2 Example results. Two slices through the input parameter space are shown to intersect in output statistic space. Color indicates the value of R 117

5.3 Maps of parameters onto output space for additional parameter sets. (a) Each surface is generated by varying w_1 from 0.02 to 0.07 and R from 0.18 to 0.91. Different surfaces are generated by stepping w_C and w_2 from 0.02 to 0.07. R value replicated in color. Existence of intersections between surfaces of different colors indicates that R is not identifiable in the output space. (b) Slice through the surfaces in (a) where the firing fate of LIF1 is 24.6 Hz. Black circle indicates one region where R is not identifiable. 119

A1 Demographics of the 252 participants. 122
A2 Histograms of 1000 p-values under the null hypothesis (A) under the empirical p-value from equation A.1, and (B) under the Gaussian assumption from the main text. 126

Chapter 1

Introduction

Humans and many animals use their visual system to obtain an understanding of the world around them. The human visual system receives $\tilde{9}00$ kilobits of information per second from each eye (Koch et al., 2006). This information is more densely represented at the fovea than in peripheral vision, and as such the brain has an important role to play in determining where to point the eye to receive the most pertinent information from the environment. The selection of gaze direction is referred to as *overt attention* in the literature because the process is an externally visible indicator of what a person is paying attention to. This distinguishes overt attention from covert attention, where a person can fixate on one location but allocate attentional resources to something in their peripheral vision (James, 1890). Covert attention is usually measured by how well human or animal subject does at some visual task in the periphery. Cues that have a high probability of being informative can allow ex-

perimenters to measure the relative performance of some task in an attended location (where the cue indicates) and in an unattended location (away from the cue) at some lower frequency. In an example neurophysiological experiment (Martin and von der Heydt, 2015) a cue indicated that monkeys needed to discriminate motion 80% of the time in one location and 20% of the time the motion would occur elsewhere. Findings of differing behavioral performance can then be interpreted as a behavioral affect due to attention being directed towards the cued location, and neurophysiological correlates of attention can be measured.

Attention was originally thought of as a spotlight in which additional information was gathered (James, 1890). However, the spotlight metaphor has a number of drawbacks. Scholl (2001) reviewed a number of studies in which attention appears to be directed more towards objects in a scene than spatial regions, as the spotlight metaphor may imply. For example, Egly et al. (1994) showed that attending to one part of an object caused increased attention (*i.e.* increased accuracy in detecting a luminance change) in other parts of the same object. Moore et al. (1998) extended this result by showing that the parts of the same object could be occluded from one another and the attentional benefit would still persist. These results and others reviewed in (Scholl, 2001) led to the conclusion that there is some pre-attentional processing that forms an organization of the visual scene. However, Rensink (2000) pointed out that this pre-attentional processing has substantial limits. Without attention, we are subject to change blindness, where modifications to unattended portions of a scene

go unnoticed. As such, he proposes that the pre-attentional organization of a scene is in the form of proto-objects, which are temporally fleeting representations of portions of a scene that do not contain as much information as an attended object.

This thesis consists of theoretical contributions to the understanding of proto-objects, both at the behavioral level and at the level of interactions between individual neurons. In Section 1.1 I introduce the connection between proto-objects and visual salience. Analysis of existing methods of measuring visual salience led to a new experimental paradigm for recording attentional responses (Chapters 2), and a test of existing models of visual salience that led to the development of a new model (Chapter 3). In Section 1.2 I discuss neurophysiological data that correlates with proto-object representations. One key measure is jitter-corrected synchrony between neuronal spike trains of neurons that represent parts of a proto-object. In Chapter 4 I describe closed-form methods for computing relevant statistics, and in Chapter 5 I describe some limitations in interpreting these statistics.

1.1 Proto-Objects and Visual Salience

Factors that affect attention can be separated into two categories. The first category consists of top-down effects, which depend on the internal state of the observer. These influences can include the present goals or an assigned task, which have been shown to change viewing behavior (DeAngelus and Pelz, 2009; Yarbus, 1967). The

second category of affects on visual attention are bottom-up, driven by the contents of the visual scene. A region that attracts attention independently of any task is said to be visually *salient*. For example, a bright flash in an otherwise still scene will attract attention (*e.g.* Anderson et al., 2011) [1].

Visual salience has been connected to proto-objects by findings that show that the organization of elements of a scene can affect visual salience (Kimchi et al., 2007). They created scenes out of L shapes, which could be organized so that the orientation of four L's would make up a square (one L at each corner) or not. Study participants were then asked to make a discrimination about the color of one of the L's that was either part of the square, outside of the square, or when no square was present. The authors observed lower reaction times when the target L was inside a square compared to the no-square condition, and higher reaction times when the target was outside the square. These finding indicate that the proto-object organization of elements of a scene (*i.e.* the organization of the L's into the proto-object of a square) are not only the basis of attentional selection, but they also can be visually salient.

Visual salience gives us an opportunity to explore the formation and features of proto-objects in a computational manner. A good model of visual salience can be evaluated against experiments where eye fixation is controlled like Kimchi et al. (2007) as well as eye-tracking experiments where the participant is able to freely view a natural scene (Parkhurst et al., 2002), and the features of such a model should be

[1] Though I should note that a top-down signal to ignore salient stimuli can avoid attentional capture (Bacon and Egeth, 1994)

predictive of the features the brain assigns to proto-objects.

However, the measurement of visual salience independent of top-down influences on attention is difficult. If we are interested in proto-objects, then the these top-down influences are important to remove, as higher-level representations of objects and their relations may begin to influence behavior. In existing controlled studies (*e.g.* Kimchi et al., 2007; Nothdurft, 2000), participants are shown the same or similar visual scenes repeatedly and often with task instructions that inform the viewer of what they will see. Such information will induce expectations about the upcoming stimulus in the participant, which may cause unwanted top-down influences on attention. Therefore, we cannot conclude from existing evidence that the measured attentional effects are driven by bottom-up cues.

This confound led us to create a new paradigm using a tablet computer, inspired by Firestone and Scholl (2014), where participants are naïve to what they are about to view. Chapter 2, following the work done in Jeck et al. (2017), describes the new paradigm and validates it by measuring the correlation between existing measures of attention and the new method. Chapter 2 also contains an interesting bootstrap hypothesis test to determine how similar two empirically measured probability distributions would be if they were in fact identical. Unlike standard permutation tests, this method is applicable even in cases where the probability distributions are estimated in different ways.

With the new paradigm validated, it was then possible to test existing models

of visual salience directly. In Chapter 3, following Jeck et al. (2018), proto-object based models of visual salience rely on contrast with the background to determine the presence of a proto-object. However, as we show, visual salience is not always correlated with such contrast. A unique low-contrast object in a scene can stand out from the rest because of its relationship to others. For example, a unique gray square on a white background will have a higher salience than a black square on the same background if the black squares are numerous (see Figure 3.1 for an example). We used our new paradigm to gather data illustrating this problem with existing models, and developed a new one to address the problem.

1.2 Proto-Objects and Synchrony

At the opposite description level, another approach to studying proto-objects is to analyze the firing behavior of neurons in the visual system. Neurons in the visual system have regions of the visual field that drive their response. These regions are referred to as receptive fields in the literature (Hubel and Wiesel, 1968). In general, if the retinal image is uniform over the receptive field, then the firing rate of the neuron will be at baseline. This region is sometimes referred to as the *classical* receptive field (CRF) because if there is a stimulus in the receptive field, the response of the neuron can still be modulated by signals outside that field. An example of effects caused by stimuli outside the CRF is the encoding of border ownership in visual area V2 (Zhou

et al., 2000). So called border ownership neurons in V2 modulate their responses to an edge in their CRF depending on which side of the CRF contains a figure, as opposed to the background. These neurons are similarly modulated by selective attention (Sugihara et al., 2004), indicating that they may be part of a proto-object representation.

Zhou et al. (2000) also found that modulation due to side-of-figure in border ownership neurons arises less than 25 ms after the neurons' initial firing increase, independent of how large the figure is. If area V2 were to compute border ownership by itself, larger figures would cause border ownership to arise more slowly. This is because V2 is retinotopically organized, and given the slow conduction velocity of horizontal fibers, integration of information over a larger area would be easily detected. In contrast, feedback connections from an external area would be very rapid, explaining the size invariance of the border ownership signals. Craft et al. (2007) proposed a model of figure-ground assignment in which border ownership neurons receive common feedback from hypothetical grouping cells, whose activity represents the presence of a proto-object. If such feedback is from a shared source then the relevant border ownership neurons will be modulated upwards by the same figure, and potentially spike at similar times, exhibiting what is known as synchrony. This synchrony has been found among border ownership neurons whose firing rates are both modulated upwards by the same figure Dong et al. (2006).

Since then, methods for the detection of synchronous firing have been made more

mathematically rigorous (Amarasingham et al., 2012) and been utilized in the study of border ownership and attention (Martin and von der Heydt, 2015). In Chapter 4 I describe improvements to the computational efficiency of these methods, summarized previously in Jeck and Niebur (2015a). In Chapter 5 I point out that while the detection of non-zero synchronous firing is mathematically rigorous, detecting a *change* in the level of synchrony does not have intuitive implications about the strength of common input, described previously in Jeck and Niebur (2015b).

Chapter 2

Attentive Pointing in Natural Scenes Correlates with Other Measures of Attention

2.1 Introduction

As noted in 1.1, factors that influence visual attention a separated into top-down influences and bottom-up influences, referred to as visual salience. While the definitions of top-down and bottom-up attention are clear, it is in practice difficult to dis-entangle their effects. For instance, observers who repeatedly perform tasks designed to measure bottom-up attentional effects may form expectations of what the next trial may be. These expectations will change their internal state and therefore

add a top-down component to their responses. This chapter describes work done in Jeck et al. (2017). Specifically,the goals of that study were to:

- Introduce open ended self reports as a new experimental assay for selective attention and show that it can be measured efficiently using a pointing/tapping paradigm
- Develop a new experimental design in which each participant views only a small numbers of scenes. This reduces the contamination of bottom-up attentional effects by top-down expectations due to participants viewing similar stimuli many times
- Compare the results of this experiment with three other measures of attention and salience: fixations, interest points, and computed saliency
- Analyze the effects of sample size on estimating correlation between maps. The small number of samples from the pointing/tapping paradigm results in a statistical effect that causes the correlation between different maps to be systematically underestimated. We will clarify the influence of finite numbers of samples on the correlation between maps

2.1.1 Determining saliency from behavior

There are several methods that allow researchers to characterize items or regions that observers direct their attention to. One very influential approach has been vi-

sual search. Search for targets that differ from distractors by one of several low-level features (*e.g.* luminance, color, orientation contrast) takes a (generally short) time that is nearly independent of the number of distractors in the display (Egeth et al., 1972; Treisman and Gelade, 1980). In contrast, targets that could be distinguished from distractors only by combinations of such features require search times that increased roughly linearly with the number of distractors (Egeth et al., 1984; Treisman and Gelade, 1980). These and related results were fundamental in the construction of computational models for visual search (Wolfe, 1994, 2007; Wolfe et al., 1989) and for saliency determination and attentional selection (Itti and Koch, 2001; Itti et al., 1998; Niebur and Koch, 1996).

Given past success in utilizing features that promote efficient search, it is tempting to continue using visual search as a way to test models of visual salience. However, search tasks are limited in their applicability to measuring salience because participants are typically informed about the types of images they are about to see (*e.g.* "an image in which there is a single target and many distractors"), and the target and distractors are often described before the task begins. This information generates top-down influences that are likely to interact with bottom-up selection mechanisms. Even when participants are only told to look for a unique target, without being informed how it will differ from other objects ("odd-man out" tasks), they are still being informed about the structure of the image. It is then difficult to decide whether the participants find the target due to its bottom-up saliency features, or because of

its uniqueness (Bacon and Egeth, 1994). Results therefore may reflect a mixture of bottom-up (saliency) and top-down components of unknown composition.

This concern applies also to measurements of salience where participants give their subjective assessment of which of two stimuli is more salient (*e.g.* Nothdurft, 2000). These experiments require that participants know that a stimulus will appear made up of oriented bars where two of them (one to the left and one to the right of fixation) will differ from the rest. As with search tasks, this information potentially biases the response of the participant. Indeed Nothdurft refers to needing additional concentration (clearly a top down process) to make difficult salience assessments. Furthermore, even if participants are not informed explicitly about the nature of the visual scene they are observing, the process of performing a task many times will likely give them information about what to expect.

While top-down influences can probably never be excluded entirely, our goal in this project is to reduce them. One possible way to mitigate top-down influences is to use "overt attention" in a free viewing task as an indicator for covert attention. In this approach, introduced by Parkhurst et al. (2002) and used in many subsequent studies (for a review see Borji and Itti, 2013), observers look at images (or videos) which can be natural or abstract scenes while their eye movements are tracked. Areas of the scene that are fixated are taken to be attended, a conclusion supported by findings from Deubel and Schneider (1996) that visual discrimination performance is enhanced at saccade targets. In the absence of a specific task ("free viewing"), it seems reason-

able to assume that at least for the first few images, and for the first few fixations in these images, observers let themselves be guided by the visual input, rather than by some more complex strategy. This assumption becomes less plausible, however, the longer the sequence of images becomes and the longer the duration becomes that observers view any given image. Indeed, Parkhurst et al. (2002) found that the agreement between eye fixation data and predictions of a purely bottom-up computational model of saliency (Itti et al., 1998) decreased with viewing time/fixation number for a given image. It is not known whether the level of agreement depended on how many images had been viewed previously.

In principle it is possible to use the eye tracking method, with naïve participants viewing only a small number of scenes. In practice, the overhead of setting up an eye tracker system for each participant would make gathering fixation data for a small number of images per participant a very cumbersome task. We recruited 252 participants in this study, an order of magnitude more than participated in the latest saliency benchmark by Borji and Itti (2015), making eye-tracking each subject prohibitive.

To counteract this difficulty, we developed a novel experimental paradigm with the goal of gathering data from many participants where each participant only performed a small number of trials. The new paradigm is centered on showing subjects a short sequence of images and recording the response of each subject to each image. Some of the images are simple displays (similar to typical visual search arrays like those used

by Treisman and Gelade, 1980) that are designed to test a specific hypothesis about what features of an image affect salience. Future work will discuss the structure of these images and the results gathered. Alternating with these images are natural scenes, the focus of this report. The goal in presenting these scenes to participants is to determine the extent to which salience as measured in our new experimental paradigm comports with salience data from previous studies. The natural scenes were therefore a subset of those used in a previous study (Masciocchi et al., 2009), and we will compare results obtained in our new paradigm with those from that study.

The data being compared here are attentional maps aggregated over a pool of participants. Such maps have been used in the study of salience extensively (Borji and Itti, 2013), and because they are population averages we can gather data to make attentional maps from a similar population without needing to gather new fixation data from the same subjects.

2.1.2 Reporting attended locations by pointing to them

Our new experimental paradigm for fast assessment of attentional selection was inspired by a study by Firestone and Scholl (2014) although those authors used a very different stimulus set and had a different motivation. The main idea is that, instead of recording eye movements, we ask participants to communicate their selections in a

natural way by tapping on a screen with their (index) finger. Specifically, we ask the subjects to "tap the first place you look when the image appears." This instruction gives us a quick way to communicate in a non-technical manner that the participant should select the first attended location on the image, rather than an arbitrary point as requested by Firestone and Scholl (2014). Even though instructions refer to where the participants look first, we do not attempt to determine whether any single individual is able to report their eye movements successfully. Instead, we are concerned with whether the population-level attentional maps we derive from the responses reflect previous measures of attention. We will validate our method by comparing these maps on when gathered for the same set of images.

We view this method of obtaining attentional maps as an alternative read-out of attention consisting of two (possibly interacting) components: self-report, and manual selection by finger tapping. Self reports have previously been taken as valid assessments of attentional selection when reporting attended locations in an experiment (*e.g.* Nothdurft, 2000). Responding by tapping allows participants to indicate any location on the screen, rather than a pre-defined set of locations via a key press, or a less easily quantified verbal report. While it has been shown that planning manual movements can draw attention independently of eye movements (Jonikaitis and Deubel, 2011) in carefully controlled experiments,it is much more common for eye movements to guide hand movements when no experimental restrictions are in place (Fisk and Goodale, 1985; Neggers and Bekkering, 2000), minimizing the probability

that a manual read out interferes with the self-report. Self reports also allow for the possibility of participants reporting the location of their covert attention rather than the location where they fixate, which may differ.

From a practical point of view, the method we use to record pointing behavior makes it a very fast, intuitive and simple process for collecting large amounts of selection data from a large and diverse participant population. Images were presented on an electronic tablet, and participants were instructed to tap on the first location that they looked at in the image, allowing for easy and precise recording of tap locations. In addition to allowing us to gather data from a large number of participants, the process reduces the information the participants were likely to have about the nature of the stimulus. We could then compare the responses of these relatively uninformed subjects to previously obtained measures of salience.

2.1.3 Limitations due to map estimation

We will follow the approach by Masciocchi et al. (2009) for computing correlations between different selection responses over the image. In that study, participants were asked to select interesting points on an image with a mouse. The distribution of selected points on the image was then interpreted as an estimate of the "interest map" internal to the participants that generated the data. Similarly, the distribution of recorded fixations from a free viewing task was turned into an estimate of a "fixation map." Both were compared with computed saliency maps. In the present study, we

will introduce a third set of human response maps, defined by the pointing/tapping locations which we call "tap maps."

When comparing any two of these estimated maps, their measured correlation is determined by the nature of the two tasks and data types, as well as the amount of data collected to form the estimate. As we show in Section 2.2.3.3, the finite amount of collected data biases the computed correlation between maps toward zero. We develop a bootstrap procedure to estimate how large the bias would be if the two maps were drawn from the same underlying distribution. This procedure gives us insight into how correlated the data types could be and helps determine which comparisons between maps may benefit from further data collection.

2.2 Methods

All methods were approved by the Johns Hopkins Institutional Review Board and carried out in accordance with the Code of Ethics of the World Medical Association (Declaration of Helsinki). Alpha for all significance tests was set to 0.05. All data and code used for the analysis described in this section are available at https://github.com/dannyjeck/Attention-maps-comparison.

2.2.1 Apparatus, participants, and procedure

Participants were 252 passers-by on the Johns Hopkins University Homewood Campus (151 female; see Figure A1 for demographic information). They were approached by the experimenter and asked if they were interested in performing a short psychology experiment. If they answered in the affirmative, they were given instructions, as follows.

Participants were asked to give their gender (male/female) and age group (18-22, 23-30, 31-40, 41-50, and 51+). On a tablet computer (Apple Computers, iOS 8.3 operating system, screen 9.7" with 1024×768 resolution, occupying approximately $15° - 35°$ of visual angle depending on how far away it was held), participants were then shown a white screen with two small black squares (see Figure 2.1), which we call the initialization screen. They were informed that tapping on either one of the squares would bring up a test image, and were instructed, "When the image appears, tap the first place you look." After the participant had tapped first the initialization screen and then the location selected by him or her on the test image, the latter was immediately replaced by the initialization screen, and the cycle recommenced. This sequence of events continued until all images had been shown, with participants responding at their own pace. The position of the tap on the test image and the time between the taps on the initialization screen and on the test image were recorded. Test images strictly alternated between a natural scene and a simple scene consisting of colored squares on a white background, see section 3.2.2 and Figure 2.1. Each

participant saw a total of 12 images of which the first always was a natural scene.

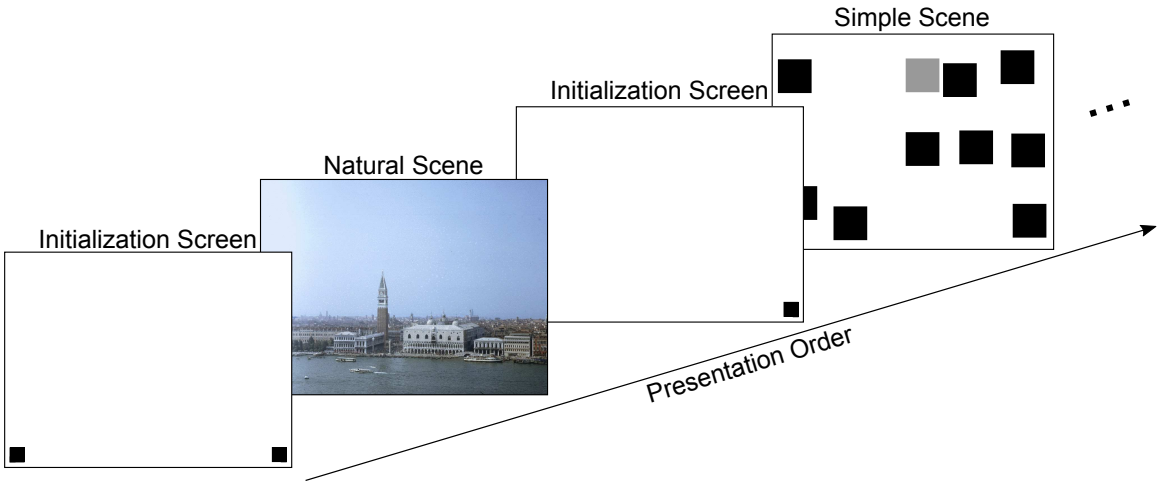

Figure 2.1: Experimental procedure. The rectangles represent an approximation of what was shown to participants on the tablet screen. First, they saw an initialization screen and tapped on either of the small black squares at the bottom. This brought up a test image which alternated between natural scenes and simple scenes. They then tapped on the test image at a place of their choosing which was, according to instructions, the first place they looked at when the test image had appeared. Tapping position and reaction time were collected, the initialization screen reappeared, and the cycle re-commenced.

2.2.2 Stimuli

The stimulus set consisted of 48 natural scenes and 30 simple scenes. The natural scenes were taken from a previous study by Masciocchi et al. (2009) in which participants performed two tasks. One was free-viewing the scenes while their eye movements were recorded. In the other task, participants clicked with a mouse on locations on the scenes that they considered the most interesting; these locations were called "interest points." The size of the original images was 640×480, and they were resized for our purposes using MATLAB's (The MathWorks, Inc., Natick, MA)

default image resizing function to fit the 1024×768 resolution of the tablet screen. Out of the four image classes in the Masciocchi et al. (2009) study we only used two, consisting of images of buildings and landscapes. Out of this set of 50 images, we randomly removed two to make the total number of natural scenes a multiple of six (the number of natural scenes each participant saw). The chosen 48 images were then separated into eight groups of six. The natural scenes for each participant rotated through these groups of six, such that every eighth participant saw the same six natural scenes. These scenes were presented in randomized order and always alternated with the simple scenes. No participant saw the same image twice.

The simple scenes consisted of a white background with randomly placed colored or gray-level squares, as shown in Figure 2.1. For the purposes of this study, they only served to interrupt the sequence of natural scenes and to decrease potential interactions between tapping locations on subsequent natural scenes. For more information about the motivation behind the generation of the simple scenes and results, see Chapter 3. We note that the strict alternation of simple and natural scenes may allow participants to develop an expectation of the *type* of the subsequently presented image (simple or natural). Neither simple nor natural scenes are, however, predictive in any way about the *contents* of the next presented image, therefore no information about salient locations in an upcoming image is predicted by the sequence of images. Furthermore, no prediction is possible until at least one repetition has occurred, *i.e.* the second natural scene, which applies to one-third of the data collected.

2.2.3 Data Analysis

2.2.3.1 Correlations between maps

Selections of image areas by human observers (fixations, interest points, and taps) were first transformed into maps of the same dimension as the images. Computing the pairwise correlations between such maps as well as between the maps and the results of computational models of salience provide a measure of similarity between the different data collection methods and the models used. We reduced the resolution of the maps by binning the data. The reduction in resolution mitigates the possibility that fixations, taps, or interest selections that are near to each other are being counted as entirely distinct, though this is not the case for responses near the edge of the selected bins. We chose a 12×16 grid to tile the image (for an example see Figure 2.2B), therefore, each bin covers 64×64 image pixels. We chose this level of reduction in resolution since it is comparable to the eye tracker error used in obtaining fixation data (see Masciocchi et al., 2009, for details) and also roughly matches the size of a human finger pad when collecting tapping data. We also analyzed a coarser image resolution to examine the effects of resolution on the different correlations, results are shown in Figure 2.5. Similar findings between these two bin sizes confirm that the results are robust to bin size selection.

Tap maps were generated by weighing each tap on the appropriate image equally and binning them as described above. Interest maps were generated from from the

data of Masciocchi et al. (2009), by taking each subjects first interest selection, the most interesting point per the instructions in the experiment, with each subject weighed equally. Fixations maps were generated by weighing each fixation by its duration. We also compared the distributions of fixations, interest points, and taps with saliency maps that were generated from the Itti et al. (1998) computational model of saliency at the same resolution.

Here we analyze the relationships between four processes: the three unknown processes, F generating fixation data, I generating interest point selections, T generating taps, and the known process S generating computed salience. If we assume each subject response is independent, then for a specific image, each unknown process can be described by a multinomial probability distribution (similar to a dice roll) from which data are drawn. We indicate the image number by adding a subscript to the process. For instance, for the k-th image I_k is a distribution from which each new interest point selection (by a different participant) is drawn. When we gather data, we are able to form estimates of these processes $\hat{F}_k, \hat{I}_k$, and $\hat{T}_k$ by computing the fraction of data points that fall in each bin for the k-th image. Since we are estimating a multinomial distribution using counts of the data, the resulting estimates of the rate of responses falling in a given bin are unbiased. However, as we will show in Section 2.2.3.3, the correlation values in comparing these maps are biased. Finally, as S is a known computational model, there is no need to form estimates of this process.

The measured covariation between any two processes P and Q on the k-th image,

indexed in their horizontal and vertical dimensions by (i, j), with M bins total is,

$$\begin{aligned} C(\hat{P}_k, \hat{Q}_k) &= \frac{1}{M} \sum_{i,j} \hat{P}_k(i,j)\hat{Q}_k(i,j) - \frac{1}{M^2} \sum_{i,j} \hat{P}_k(i,j) \sum_{i,j} \hat{Q}_k(i,j) \\ &= \frac{1}{M} \sum_{i,j} \hat{P}_k(i,j)\hat{Q}_k(i,j) - \frac{1}{M^2} \end{aligned} \tag{2.1}$$

where the last equality holds because $\hat{P}_k$ and $\hat{Q}_k$ are probability distributions and therefore sum to unity.

The Pearson correlation coefficient R between estimates $\hat{P}_k$ and $\hat{Q}_k$ is then computed as,

$$R(\hat{P}_k, \hat{Q}_k) = \frac{C(\hat{P}_k, \hat{Q}_k)}{\sqrt{C(\hat{P}_k, \hat{P}_k)}\sqrt{C(\hat{Q}_k, \hat{Q}_k)}} \tag{2.2}$$

This quantity can vary between $R = -1$ for perfectly anticorrelated data and $R = 1$ for perfectly correlated data. We compare its value against two hypotheses, discussed in the following two subsections, 2.2.3.2 and 2.2.3.3. We refer to the average correlation coefficient over all images by dropping the subscripts in the argument.

2.2.3.2 Null hypothesis: Correlations reflect no differences between images

We consider first the (null) hypothesis that the contents of specific images do not affect the participants' responses. Under this hypothesis, for instance $R(\hat{F}_i, \hat{T}_i)$, the correlation between the fixation map from image i and the tap map from the same

image is drawn from the same distribution as $R(\hat{F}_i, \hat{T}_j)$, the correlation between the fixation map from image i and the tap map from image j, for all i and j. We can approximate this null hypothesis distribution using a bootstrap technique to compute correlations between two types of maps (*e.g.* tap maps and fixation maps) using permutations of the image orders. Note that under this null hypothesis, image contents can still exert systematic influences on the selections but these influences do not differ systematically between different images. Therefore, the hypothesis includes correlations due to influences like center bias, "photographer's bias" (systematically placing objects of perceived importance in specific locations in the image), similarities due to similar image content, or other spatial preferences in common between participants. The null hypothesis does, however, exclude correlations caused by salient features of specific images.

2.2.3.3 Hypothesis: Correlations are limited by sampling error

At the other extreme, even for strong influences of image contents on correlations, estimating correlation from noisy estimates of the true processes generating the data create a bias in the measured correlation between any two types of maps. We illustrate this effect in a simple example. Consider two very simple one-dimensional identical distributions $P_k = Q_k = [0.5, 0, 0.5]$. If we draw an infinite number of samples from these (identical) distributions and use equation 2.2 to compute the correlation between

the measured estimates, we obtain $R(\hat{P}_k, \hat{Q}_k) = 1$, as expected. But now consider the case of finite numbers of samples, and in the extreme, that only one sample from each distribution is drawn. Then, the estimate of the each distribution will either be $[1, 0, 0]$ or $[0, 0, 1]$. If they are the same, then $R(\hat{P}_k, \hat{Q}_k) = 1$ but if they are different $R(\hat{P}_k, \hat{Q}_k) = -\frac{1}{2}$. Therefore, the expected correlation is $\frac{1}{4}$. This bias towards zero will be non-zero for any finite number of samples drawn.

We want to gain an intuitive understanding of the bias in correlation for the unknown distributions underlying our data that is caused by the limited number of samples drawn. For this purpose, we developed a procedure in which we resample one of the maps with the same number of data points measured in the other to approximate how correlated the data could be under the hypothesis that the underlying processes were identical. Let P_k and Q_k be two processes with $\hat{P}_k$ estimated using n_P data points and $\hat{Q}_k$ estimated using n_Q data points, and let $n_P > n_Q$. First we select the type of map with the most data points, $\hat{P}_k$, and treat it as a perfect estimate of its underlying process. We then draw n_Q data points from $\hat{P}_k$ (with replacement) and compute a surrogate, $\tilde{P}_k^Q$. The tilde is used to indicate that the value is a resampling of the data from $\hat{P}_k$ and the superscript indicates the source of the number of data points used in the resampling. We then compute $R(\hat{P}_k, \tilde{P}_k^Q)$, the correlation between the surrogate data and the original map (see Figure 2.2C). For example, if the two maps in this procedure were fixations and taps and there were more fixations than taps, we would draw (with replacement) a number of surrogate data points from the

fixation data set that was the same as that of recorded taps, and compute R between the surrogates and the original fixation map, $R(\hat{F}_k, \tilde{F}_k^T)$. For the reasons discussed in the previous paragraph, this value will be less than unity and it provides an intuitive estimate for how much the sampling error biases the measured correlations, $R(\hat{F}_k, \hat{T}_k)$. This procedure of generating surrogates and correlating with the original data can be repeated many times to refine the estimate of the bias in the correlation measurement under this hypothesis and to build a distribution against which to perform a hypothesis test (Figure 2.2D). We call this hypothesis the "sample error hypothesis," which assumes that a non-unity correlation measurement is due entirely to finite sample size. We note that this hypothesis is not truly an upper bound on the measured correlation (see Section 2.4.2 for a counterexample). We also note that, while this hypothesis is technically a null hypothesis against which we perform statistical tests, for the sake of clarity we will reserve the name "null hypothesis" for the hypothesis described in Section 2.2.3.2.

All resampling procedures were repeated with 1000 surrogates compared against the original.

2.2.3.4 Population averages

We analyzed the mean correlations between types of maps (*e.g.* taps and fixations) across all images (see Figure 2.3), which, as before, we denote by dropping the image number subscript. For example $R(\hat{F}, \hat{I})$ is the correlation between measured fixation

and interest data averaged over all images. Similarly the average correlation under the assumption that the underlying distributions are actually identical (being the distribution of the interest data, which is the larger data set) and sampled with the number of fixations is given by $R(\hat{I}, \tilde{I}^F)$. The distributions of the null hypothesis differ between the combinations of maps but are identical for all image pairs of a given combination, *e.g.* Fixation and Interest maps in Figure 2.3B. Since many correlation values are averaged and we are measuring the difference between two mean values, hypothesis testing against the null becomes a two-sample Z-test. When testing against the sample error hypothesis we also perform a two-sample Z-test (see Section A1.3 for validation of this method). Because both the final tests of significance average over all images and because the null and sample error hypotheses are relatively easy to reject (even though they are non-trivial), small p-values are expected. Beyond hypothesis testing, the mean correlation values provided by the null and sample error hypotheses also give points of reference against which we can compare the measured correlation values.

2.3 Results

We recorded 1510 taps from 252 participants (151 female; see Figure A1 for demographic information). The median of the reaction time (RT), defined as the time from tapping on the initialization screen to tapping on the test image, was about 1.4

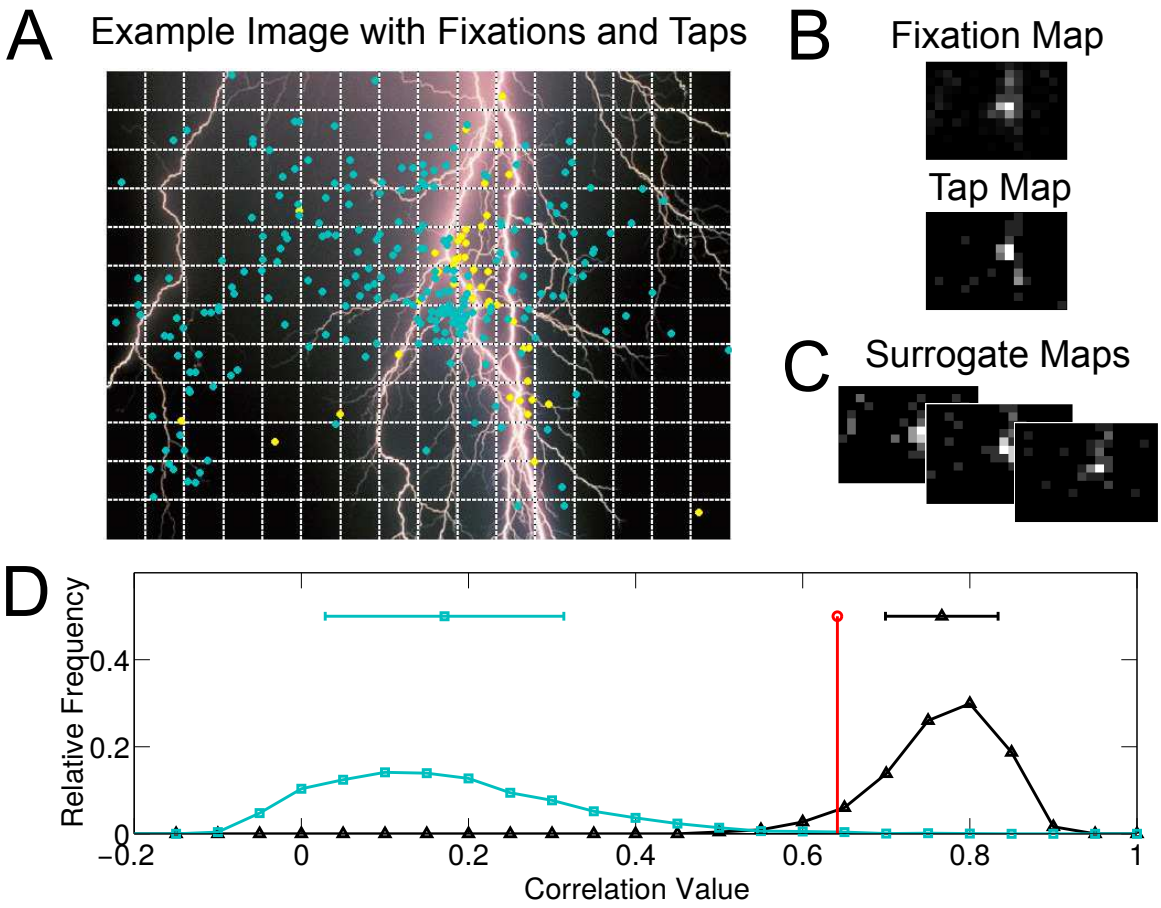

Figure 2.2: Data analysis method. (A) Example image overlaid with collected fixation points (blue dots) and tap points (yellow dots), and grid lines used to bin the data. (B) Corresponding fixation map and tap map. Both maps are binned in a 12×16 grid, with each bin showing the average of 64×64 pixels. (C) Surrogate maps generated from the fixation data used to approximate sampling error in the correlation between the fixation and tap data, see text. (D) Comparison of the measured value (red) to the histograms of the null hypothesis (blue) and the sampling error hypothesis (black). Means and standard deviations of the distributions generated from the null hypothesis and the sampling error hypothesis are shown above the distributions. For this image, fixation data and tap data correlate more than predicted by the null hypothesis ($p = 0.002$), and cannot be distinguished from predictions of the sampling error hypothesis ($p = 0.11$).

seconds. Reaction times were skewed to the right (mean 1.6 seconds). We did not analyze RTs in detail because our data collection system did not allow precise control of the timing of image presentation. Data collection was completed after seven days of full time data collection.

2.3.1 Fixations vs. Interest Points

Aggregate results of our analysis for all images are shown in Figure 2.3. First, we re-analyzed the data from the Masciocchi et al. (2009) study with our methods. The analysis confirmed their result that interest and fixation data are correlated beyond the null hypothesis, $R(\hat{F}, \hat{I}) = 0.53$, Z-test $p = 1.3 \times 10^{-73}$; see Figure 2.3A. In addition, we now extend their results by showing that sufficient data was collected in that study so that the correlation under the sample error hypothesis between interest and fixations is very high, $R(\hat{F}, \tilde{F}^I) = 0.98$, indicating that the measure of correlation $R(\hat{F}, \hat{I}) = 0.53$ likely has very little bias. Differences between fixation and interest maps were not due to sampling error, Z-test $p = 1.5 \times 10^{-74}$.

2.3.2 Fixations vs. Computed Saliency

For the comparison of fixations and computed saliency from the Masciocchi et al. (2009) study (see Figure 2.3B) we found that the measured correlation exceeded the null hypothesis, $R(\hat{I}, S) = 0.19$, Z-test $p = 1.6 \times 10^{-16}$. Correlation under the sample

error hypothesis is low for this comparison, $R(S, \tilde{S}^I) = 0.58$, though clearly higher than the measured correlation, Z-test $p = 6.8 \times 10^{-23}$.

2.3.3 Interest Points vs. Computed Saliency

We also compared interest points and computed saliency from Masciocchi et al. (2009), see Figure 2.3C. We found that the measured correlation exceeded the null hypothesis, $R(\hat{F}, S) = 0.30$, Z-test $p = 1.1 \times 10^{-18}$. Here the correlation under the sample error hypothesis is much lower than unity, $R(S, \tilde{S}^F) = 0.55$, indicating a potential bias in the measured correlation, though again higher than the measured values, Z-test $p = 9.5 \times 10^{-67}$.

2.3.4 Fixations *vs.* Tap Points

In the remaining three panels of Figure 2.3 we compare the correlations between the tap data collected in the present study with other attentional selection quantities. Correlations between fixation and tap data are shown in Figure 2.3D. The correlation level is similar to that between fixations and interest points in the Masciocchi et al. (2009) study, $R(\hat{F}, \hat{T}) = 0.45$, and it is again significantly above the null hypothesis ($p = 1.0 \times 10^{-39}$). Because fewer taps were collected than fixation points, the correlation under the sampling error hypothesis is $R(\hat{F}, \tilde{F}^T) = 0.64$. This is still significantly above the measured value ($p = 6.5 \times 10^{-16}$) but substantially below unity,

indicating that the correlation may be substantially biased by the limited amount of data gathered.

It is unclear whether gathering more data would cause the measured correlation to increase or not. It may be that the "true" tap map T (which would be obtained if unlimited amounts of data were collected) is less diffuse than the measured fixation map $\hat{F}$, in which case the measured tap map $\hat{T}$ is a good estimate of the T map and the measured $R(\hat{F}, \hat{T})$ value is close to $R(F, T)$. Alternatively, the T map could be much more correlated with fixations than our measured map, in which case gathering more data will increase the correlation. We can say with high confidence that $R(F, T)$ is less than unity and greater than 0.41 (two standard errors below $R(\hat{F}, \hat{T}) = 0.45$).

We investigated the relationship between $R(F, T)$ and $R(F, I)$ further by computing $R(\hat{F}, \hat{T})$ and $R(\hat{F}, \hat{I})$ with subsets of the data collected for $\hat{T}$ and $\hat{I}$. We did this by drawing a number of data points without replacement from the tap data and interest data, and forming new estimates of the tap and interest maps. These were then correlated with $\hat{F}$ to qualitatively see whether the correlations $R(\hat{F}, \hat{T})$ and $R(\hat{F}, \hat{I})$ are converging as data is collected and to compare the two measures when equal numbers of data points are gathered. Results for various sizes of subsamples (up to the number of taps and interest points gathered per image) are shown in Figure 2.4. It is seen that for equal numbers of data points, $R(\hat{F}, \hat{T})$ and $R(\hat{F}, \hat{I})$ track each other closely, with both correlations increasing approximately logarithmically (about linearly in the semi-logarithmic plot) with the number of data points. For

example, $R(\hat{F},\hat{T}) = 0.44$ and $R(\hat{F},\hat{I}) = 0.46$ when 29 interest points/taps are used per image. This is the largest number of taps available for all images. The number of data points available for fixations is larger than for taps and it can be seen that for much larger numbers (above ≈ 100), $R(\hat{F},\hat{I})$ starts to plateau. The observation that $R(\hat{F},\hat{I})$ plateaus agrees with our previous analysis that $R(\hat{F},\hat{I})$ has very little bias since $R(\hat{F},\tilde{F}^I)$ is nearly 1 and the asymptotic value in Figure 2.4 approaches the mean of $R(\hat{F},\hat{I})$ shown in Figure 2.3B, about 0.53.

2.3.5 Interest Points vs. Tap Points

Tap data was also found to be significantly correlated with interest point data beyond the null hypothesis, $R(\hat{I},\hat{T}) = 0.50$, $p = 1.2\times10^{-58}$, and correlation under the sample error hypothesis was significantly higher than the measured value, $R(\hat{I},\tilde{I}^T) = 0.85$, $p = 1.3\times10^{-34}$, Figure 2.3E. The difference between $R(\hat{I},\tilde{I}^T)$ and $R(\hat{F},\tilde{F}^T)$ indicates that there is some difference between interest points and fixations that can not be explained by the smaller number of tap data. Despite drawing the same amount of data (the number of tap points) from the interest maps as we did from the fixation maps, the correlation under the sample error hypothesis is higher for interest maps because they are more focused than fixation maps (*i.e.* participants selected interest points in tighter clusters than was found in their fixations). Therefore, these clusters can be estimated more accurately with a smaller amount of tap data than for the more diffuse fixation maps.

2.3.6 Tap Points vs. Computed Saliency

Finally, saliency maps computed from the Itti et al. (1998) model were compared against the tap data and found to correlate beyond the null hypothesis, $R(S, \hat{T}) = 0.21$, $p = 4.3 \times 10^{-15}$, though not significantly below the sample error hypothesis, $R(S, \tilde{S}^T) = 0.25$, $p = 0.075$. This relatively low value of $R(S, \tilde{S}^T)$ is obtained because the computed saliency maps were relatively diffuse.

2.3.7 Coarse Scale Analysis

We also repeated the above analysis using fixation, interest, tap and salience maps at a coarser 3×4 resolution (the coarsest resolution possible with square bins). Results are shown in Figure 2.5. At this resolution all measured R values and resampled R values were higher, with measured R always falling between the null hypothesis and the sample error hypothesis (all $p < 0.05$). The level of measured correlation is thus dependent on the resolution used but the main results for the finer resolution hold. Because the measured correlations are still above the null hypothesis we can conclude that even for a very coarse grid, the image content is still informative beyond center bias, photographer's bias, or other structures common to a large fraction of images.

In summary, we found that tapping locations are correlated with the locations selected by each of the three measures considered previously: fixations, interest, and computed saliency (Masciocchi et al., 2009). The null hypotheses of lack of correlation

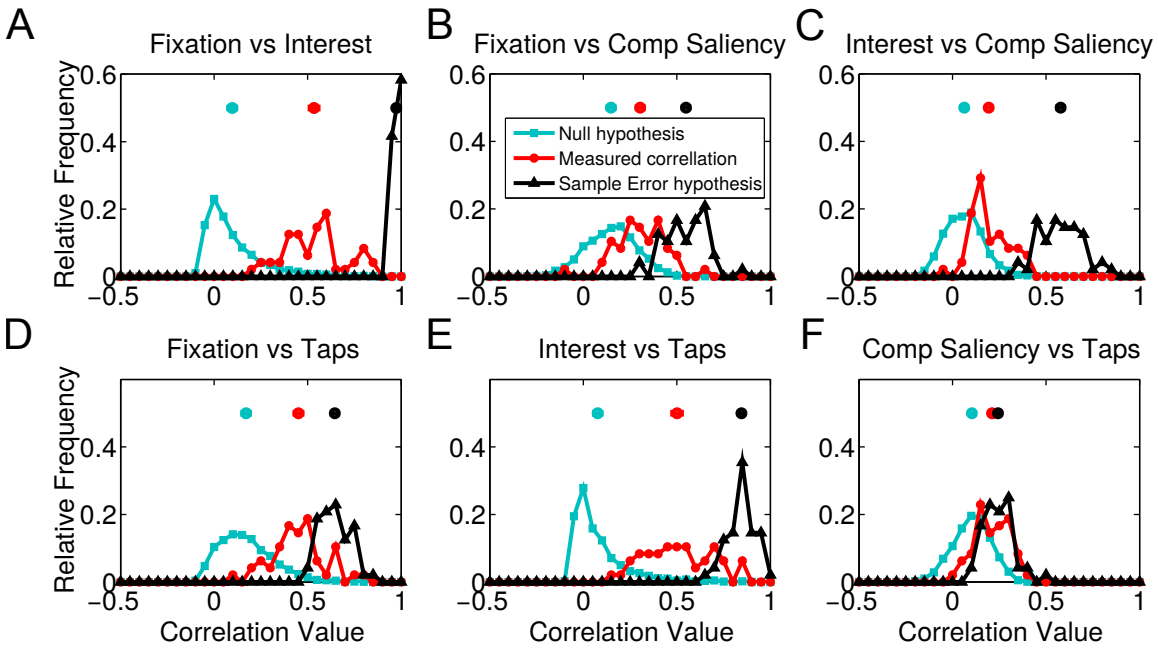

Figure 2.3: Aggregate results of natural scene analysis at 12×16 resolution. Each subplot shows a distribution of measured correlations between two types of maps compared against the null hypothesis and sample error hypothesis. Means of each distribution are shown above the histograms, with error bars indicating standard error given the 48 images used. Most error bars are smaller than the markers used. (A) Fixation and Interest maps. (B) Fixation and Computed saliency maps generated from Itti et al. (1998). (C)Interest and saliency maps. (D) Fixation and Tap maps. (E) Interest and Tap maps. (F) Computed saliency and Tap maps. All measured averages are significantly above the null hypothesis ($p < 0.05$). All measured averages are below the sample error hypothesis ($p < 0.05$), with the exception of the comparison between computed saliency and tap maps ($p = 0.08$), panel F. The legend in panel B applies to all panels. For color figures see the online version of the article.

between tap locations and these three measures could all be rejected with high significance. Furthermore, we identified an important source of systematic downward shift (bias) of correlations between maps which is due to the finite numbers of selection points.

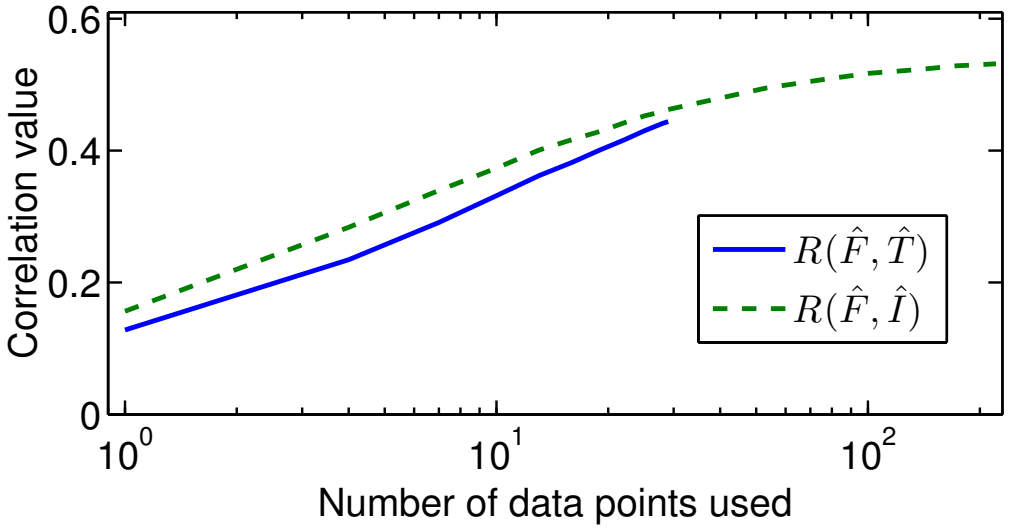

Figure 2.4: Comparison of $R(\hat{F},\hat{I})$ and $R(\hat{F},\hat{T})$ when using only a portion of the interest points and tap points. All fixation data was used to generate $\hat{F}$ for all simulations. 100 Simulations were performed for each number of data points. Standard error is less than line width. For color figures see the online version of the article.

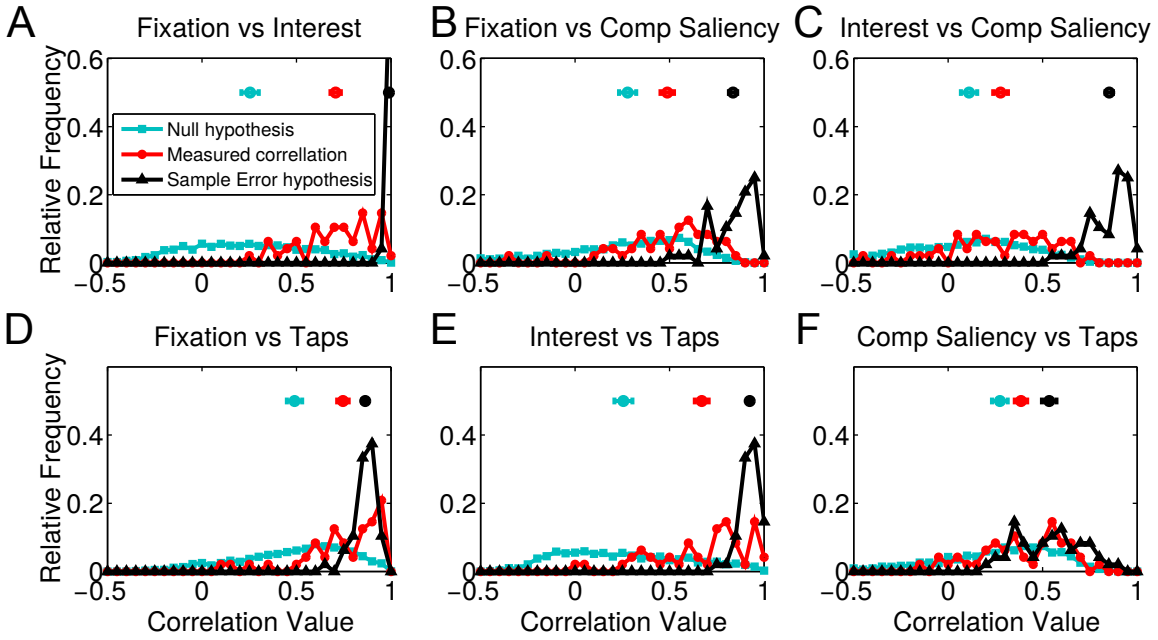

Figure 2.5: Aggregate results of correlation analysis at coarse resolution, when images were divided in a 3×4 grid. Symbols as in Figure 2.3. For color figures see the online version of the article.

2.4 Discussion

2.4.1 A new experimental paradigm for quantitative characterization of attentional selection

We have developed a new experimental paradigm to evaluate what parts of an image attract the attention of observers. We do so by asking the study participants to report where they look and read out that report with a finger tap on the selected location. As far as we are aware, this is the first study in which open ended self-reports of attended locations are gathered. Unlike previous methods, this paradigm is particularly well suited to collecting data from participants who are not informed about the nature of what will be presented, mitigating top down effects related to expecting certain stimulus types. We therefore interpret this new paradigm as a supplement to existing paradigms (free viewing, visual search, *etc.*) that can used to reduce top-down expectations that might bias participants' performance. Due to the simplicity of the experimental design, we were able to gather data from 252 subjects in seven days of data collection.

Pointing with a finger (similar to tapping a location) is a very natural and universal human behavior (Kita, 2003) which already appears during infancy, at about one year of age (Leavens et al., 2005; Tomasello et al., 2007). The purpose of finger pointing is typically to direct attention (either that of the tapping person or more

commonly that of another person) towards a specific part of the world. This behavior is thus often a direct, voluntary expression of attentional selection. It is more closely related to guiding the attentional direction of others than eye movements, although eye movements can also be used for directing attention in certain situations. While the term "overt attention" is traditionally used for eye movements (because they make the outcome of the covert attention process visible to the outside), pointing can therefore be seen as another form of overt attention, one that makes the outcome of the agent's attentional selection process explicit and instructs the observer to generate a "joint attentional frame" (Tomasello and Carpenter, 2007). This strong connection with attentional selection makes this process not only attractive by itself, for the purpose of deducing the outcome of the covert selection process, but also for comparison with other correlates of attention, like eye movements and conscious selection of interesting parts of a scene. It thus complements the classical eye tracking method (Parkhurst et al., 2002; Yarbus, 1967) and the selection of interest points (Masciocchi et al., 2009).

The high levels of correlation between the four measures used in this study (fixations, interest points, taps and computed salience; see Figure 2.3) support the conclusion that the tapping paradigm is a valid measure of salience. For instance, the high correlation between taps and fixations ($R(\hat{F}, \hat{T}) = 0.45$) indicates that the taps are capturing an aspect of salience seen in previous fixation studies. In fact, the value of $R(\hat{F}, \hat{T})$ is likely biased downwards by the limited sample size,like all the correlations

between maps. We have shown that if the fixations and taps were *perfectly* correlated, given the available number of data points the sampling error would still only result in a correlation coefficient of $R(\hat{F}, \tilde{F}^T) = 0.64$. See Section 2.4.2 for further discussion of the sample error hypothesis

There are further factors that are expected to reduce the correlation of the measurements between taps and fixations, bolstering our result. The set of participants, screen, image resolution, and viewing conditions all varied between paradigms, and the outdoor conditions of the tap experiment allowed for multiple sources of possible distractions, including other passers-by. The fact that we find significant correlations in the presence of all of these confounding variables indicates that the responses given by participants are robust to a variety of low-level manipulations even though the measured correlations are likely decreased by these effects. Our finding suggests that attention is deployed based on invariant representations that are shared by the various participants and invariant to changes in viewing conditions.

Another difference between paradigms was the duration of presentation. While the tapping paradigm may be considered deliberative, the fixation data we used (from Parkhurst et al. (2002) and Masciocchi et al. (2009)) were gathered over a five second viewing period for each subject, more than three times the median reaction time during the tapping experiment (1.4 seconds). Free viewing periods of five second duration are in common use also for fixation datasets such as the widely used CAT2000 dataset (Borji and Itti, 2015). Note that for the tapping study, the reaction time

includes the time after the subject has decided where to tap, the movement of the hand, as well as the (relatively short) delay between the tap on the initialization screen and the presentation of the image. We therefore estimate that the majority of subjects performed three or fewer saccades before deciding where to tap. In principle, one could compare the tapping locations with only the first fixations from the studies that presented the same images Masciocchi et al. (2009); Parkhurst et al. (2002). However, given the small number of participants in those studies, this analysis would not provide a meaningful map of fixated locations to compare against taps.

Finally, the process of making a hand movement may modify by itself the deployment of a participant's visual attention (Baldauf and Deubel, 2008; Jonikaitis and Deubel, 2011) thereby possibly changing the selected location. However, previous studies (Baldauf and Deubel, 2008; Deubel and Schneider, 1996, 2003; Jonikaitis and Deubel, 2011) all study conditions in which the reaching movements and saccades are planned in response to a cued location rather than indicating a salient stimulus. While more controlled research would be required to properly elucidate the interaction between manual selection and attention, we find it highly likely that the participant's selection is driven by their initial response to the image before the hand movement. If this were not the case, we would expect our measured correlations to be substantially lower.

In comparing the interest points and tap points (Figure 2.3 B-D), the results indicate that the correlation between our tap data and fixation data is approximately

as strong as the correlation between fixations and interest points ($R(\hat{F}, \hat{T}) = 0.45$ *vs.* $R(\hat{F}, \hat{I}) = 0.53$). The correlation between interest and fixations is not subject to sample size bias to the same extent described above because the correlation under the sampling error hypothesis ($R(\hat{F}, \tilde{F}^I) = 0.98$) is so close to unity. Given these results, we speculate that the responses for the tap experiment lie somewhere in between the more involuntary fixation responses and the more deliberative responses given in the interest points task.

The level of correlation between taps and computed salience ($R(S, \hat{T}) = 0.21$) in the natural scenes was lower than previous findings indicated for other correlates of attentional selection. Masciocchi et al. (2009) found the correlation coefficients between fixations and computed salience to be $R = 0.32$, and between interest and computed salience to be $R = 0.37$ using slightly different methods. The results of Masciocchi et al. (2009) are in closer agreement with our low-resolution analysis, which found $R(S, \hat{T}) = 0.38$ and $R(S, \tilde{S}^T) = 0.53$. These results indicate that the salience model from Itti et al. (1998) which was used in both the previous study and this one captures a substantial aspect of the bottom up processes that influence attention. However given the low correlation value, it is likely that other aspects of those processes are not being captured.

Overall, our results show highly significant correlations between attentional selections executed by the oculomotor system [Parkhurst et al, 2002, and many other more recent studies; for a review see Borji and Itti, 2013] and by the skeletomuscular

system. For the latter, this is the case both when conscious deliberation is encouraged (Masciocchi et al., 2009) and when it is discouraged (this study). Remarkably, these measures also correlate well with predictions of a very simple computational model of bottom-up attention (Itti et al., 1998). Without doubt, this simple model has limitations, *e.g.* in the representation of objects (Einhäuser et al., 2008, but see Borji et al, 2013), even though they can be overcome at least partially by more sophisticated proto-object based models (Mihalas et al., 2011; Russell et al., 2014). However, the fact that even a very basic model captures human behavior over such a large range of tasks illustrates the fundamental role of attentional selection for behavior.

2.4.2 Effects of sampling error on correlations

Another contribution of this study is a new way of analyzing correlations between maps of different types, such as fixations or taps, although our method should apply to many other kinds of maps. These maps are generated by accumulating many individual measurements into a "heat map," which can be interpreted as an estimate of the probability distribution of the data. The measured correlation between the maps (*e.g.* $R(\hat{F}, \hat{T})$) and the estimates of those probability distributions (here $\hat{F}$ and $\hat{T}$)) will depend on both the underlying distributions (F and T) and the quality of the estimates. The differences between the true distributions are of scientific interest. For the case of the maps considered in this study, these differences may be useful in determining what aspects of a scene draw attention, and their correlation is useful in

determining the validity of the tap experiment as a measure of salience.

Estimates of the true distributions based on finite amounts of data will, however, bias our estimate of the correlation. With an infinite number of data points, the true distributions could be measured to perfect accuracy. Given a fixed limited sample size, increasing the resolution of the maps increases the number of parameters in the distribution to be estimated and therefore decreases the accuracy. Similarly, if the true distribution is spread widely across the image, the accuracy of the estimate will be reduced much in the same way that, everything else being equal, the standard error of the mean for a distribution with high variance is greater than the standard error of the mean for one with low variance.

This source of bias in correlation measurements differs from the reduction ("attenuation") in correlation described by Spearman (1904) when measuring the correlation between two signals in noise. While both effects bias the observed correlation towards zero, the underlying mechanisms are quite different between our effect and Spearman's, making his method for correcting the bias inappropriate in our case. Spearman observed that the correlation between two processes is attenuated if noise is added to one or both of them, and in his 1904 study he developed a method to correct for the bias found in correlating noisy measurements. In contrast, in the effect described in the present study, no noise is added. The bias in the correlation here is due to the finite number of observations of the underlying distributions (for tap, fixation, and interest selection). In the example in Section 2.2.3.3 of the two simple

distributions, the correlation is biased because we only sample from a small number of points (in the extreme case discussed, just one), but there is no noise in the samples. The two effects are independent, one could have one or the other or both, and each contributes its own bias to the total decrease of the correlation. For instance, while the bias due to the limited sample size described in Section 2.2.3.3 disappears if the sample size goes to infinity, this is not the case for the noise-induced attenuation effect discovered by Spearman (1904).

One may still be tempted to apply the method from Spearman (1904) to correct for the bias found in correlating noisy measurements of probability distributions. After all, the estimates of probabilities can be thought of as a measurement of the true distributions plus noise. However, the noise characteristics are entirely different in the present case. Spearman (1904) assumes independent identically distributed additive noise, while the estimation error resulting from drawing a finite number of samples from a multinomial distribution is dependent on the value measured and exhibits covariation between bins (since the error must sum to zero) Spearman's method is therefore not a valid solution to this problem.

Given the potential sources of error in estimating correlation, we have developed a simulation-based method (Section 2.2.3 and Figure 2.2) to compute the correlation between maps assuming that the true maps are perfectly correlated. Note that, although one might think that the correlation of a map with itself is an upper bound on the correlation of the map with other maps, even for finite numbers of samples, this

is not the case. For a counter example, if $\hat{P} = [0, 1, 0]$ is measured with one sample, and $\hat{Q} = [0.3, 0.4, 0.3]$ is measured with (infinitely) many samples, then $R(\hat{P}, \hat{Q}) = 1$, but the expected value of $R(\hat{Q}, \tilde{Q}^P)$ is 0.1 because there is a probability of 0.6 that the single sample drawn from Q will be from either the first or last bin. In this case, the correlation is $-\frac{1}{2}$ because the peak in one distribution aligns with one of the two equal troughs in the second.

The use of Pearson correlation (R) is useful in gaining a qualitative measure of the similarity between the distributions. Overlapping peaks and troughs in distributions will result in positive R values. However, R is invariant to linear scaling. If one distribution is relatively uniform while another has high peaks and troughs, the R function may find them to be highly correlated so long as their peaks and troughs align. As such, the correlations measured in this study show that interest points, taps, and fixations all seem to fall on similar locations, though the distributions may have substantial differences under another metric.

The method of estimating the sampling error effect that we introduce is applicable to any correlation computation between estimates of a true distribution. In fact, the method can be extended to any metric of similarity between distributions or maps. For example, if Kullback-Leibler divergence (KLD) is believed to be a more appropriate metric of similarity, the sample error hypothesis can be used to generate surrogate data under the hypothesis that the two types of data are drawn from the same distribution. Then the KLD between the surrogate data and the original map

can be used to determine the size of the sampling error effects.

We also note that there may be methods to reduce the bias in the measured correlation using a Jackknife procedure (Efron, 1982), though it is unknown to what extent such a procedure would introduce unwanted variance into the estimation procedure.

Chapter 3

Utilizing the Salience of Unique Objects to Test Models of Salience

3.1 Introduction

Having developed a simple paradigm for recording attentional selections from naïve participants, we wished to test existing models of visual salience. Computational models of visual salience (*e.g.* Itti et al., 1998, see Section 3.1.1 for a discussion of more models) predict that regions of large center-surround contrast are salient. This contrast can arise in several visual submodalities (*e.g.* color, intensity) and at several spatial scales but the magnitudes of center-surround contrast and saliency are always positively correlated: higher contrast in a given submodality and scale contributes more to saliency than lower contrast. This leads to an interesting and

counterintuitive prediction that we test in this contribution.

Consider the image in Figure 3.1A, a number of black squares and one gray square on a white background. As long as the unique gray square is easily discriminated from both the background and the black squares, our intuition suggested to us that its uniqueness makes it the most salient stimulus. However, saliency models predict the opposite. First, we observe that only intensity contributes to saliency in this simple scene (no color *etc.*). Second, the center-surround contrast of the gray square is smaller than that of the black squares since it's intensity is closer to the background than that of the black squares. Therefore, the models predict that the black-on-white squares have higher saliency than the gray-on-white square. This is illustrated in Figure 3.1B which shows center-surround responses to the intensity channel of the image in Figure 3.1A at different spatial scales for the center and surround. For each of these center-surround computations, the gray square produces a weaker response than the black squares. Because of the lowered center-surround responses, the resulting saliency map (Figure 3.1C, computed from the model in Itti et al., 1998) associates a lower saliency level to the unique gray square than to the black squares. Indeed, no linear combination of these center surround maps can generate a saliency map in which the gray square has a higher value than both the black squares and the background (see Section 3.5.1).

One might expect that at large spatial scales the center-surround operation would compare the intensity of the squares with that of their neighbors, enhancing the gray

square. Salience of a unique stimulus is, indeed, enhanced if the distance between this stimulus and the other stimuli is small enough that the latter are located in the surround (as defined by the model) of the former (see *e.g.* Niebur et al., 2002, Figure 4). However, for the stimulus in Figure 3.1A, even though the surround of each square at larger spatial scales includes the black squares, it also includes much of the white background. The latter dominates in all cases, resulting in a mostly-white surround for all squares. When the center-surround operation computes the difference between the center (black or gray) and the surround (mostly-white), the black squares produce a larger difference than the gray one. We hypothesize that what makes the gray square unique, and therefore salient, is the difference of intensity between it and the set of black squares. Thus, saliency is still determined by a center-surround difference but this difference is computed in "feature space," with comparisons between *objects* rather than between spatially defined regions of the visual field. In Section 3.3.3 we propose a novel model of inter-object comparison that assigns high salience to unique objects. This raises a conflict between predictions of saliency map models (e.g. Itti et al., 1998) and our intuition that the gray square is salient. The goal of this paper is to test whether human behavior agrees with our intuition or with this and other models of salience. We also propose a novel model of inter-object comparison that assigns high salience to unique objects.

3.1.1 Related Models

Is the failure to assign higher saliency to a unique object limited to the original saliency map studies (Itti et al., 1998; Niebur and Koch, 1996), or does it affect a larger class of models? Since, as we discuss in Section 4.4, we believe that higher salience is assigned to unique objects because visual scenes are processed in terms of objects rather than of elementary visual features, we were first particularly interested in models that involve the formation of perceptual objects (or proto-objects, discussed in section 3.2.3). This is the case for the models developed by Walther and Koch (2006) and by Russell et al. (2014). We therefore ran these models on the input shown in Figure 3.1A, with results shown in Figure 3.2A and 3.2B, respectively. Both models assign lower salience to the gray square compared to the black squares, in disagreement with our intuition.

Given that we believe it is the uniqueness of the gray square that makes it stand out (or salient) in perception, we then searched for a model that is specifically designed to detect unique elements. We narrowed the choice down further by demanding that the model only takes into account the one feature, namely color (of which intensity is a special case) that distinguishes the unique stimulus from all others in our images, rather than integrating different submodalities (color, orientation, motion, *etc.*), as the saliency map models typically do in order to make them applicable to large classes of stimuli. Both of these conditions are fulfilled by a study by Perazzi et al. (2012). In their model, an image is decomposed into compact regions of roughly similar colors,

and the salience of a region is driven by the relation of two factors, the "uniqueness" and the "distribution" of a color. Colors that are far away in color space from others present in the image (such as gray in Figure 3.1A) are considered "unique" and therefore salient. Colors that are spatially distributed in the image (black and white in Figure 3.1A) are given a high "distribution" level, resulting in lowered salience.

The output of the Perazzi et al. (2012) model for the scene in Figure 3.1A is shown in Figure 3.2C. Given its design, we were surprised that it does not label the gray square as salient. We ran the Perazzi et al. (2012) model on all of the ten stimulus arrays with faint objects that we used in our behavioral experiments, described below and shown in the left panels of Figures 3.7 and 3.8. For eight of them, the unique square was not labeled as salient. We analyzed the internal model function and by fine-tuning one of its model parameters (σ_c), we managed to have the unique square labeled as salient for each of the ten images. However, we do not know how the modified model performs on images other than those ten for which its parameters were specifically tuned. We are also uncertain how this parameter change affects the behavior of the model on natural scenes. A further consideration is that the Perazzi et al. (2012) model is defined in purely functional terms and its relationship with what we know about early visual processing in biological systems is remote at best.

Finally, we tested our hypothesis on a model that predicts salience which was developed by Kümmerer et al. (2016). The model uses features from VGG-16 (Simonyan and Zisserman, 2015), a deep network trained for object recognition. This

model achieves high performance on natural scenes and is currently ranked highly on the MIT300 saliency benchmark which was introduced by Judd et al. (2012). For compatibility with other tested models, we used a version without center bias. As seen in Figure 3.2D, this model also assigns a lower salience to the unique object, indicating that the training of the model does not generalize well to comparing objects.

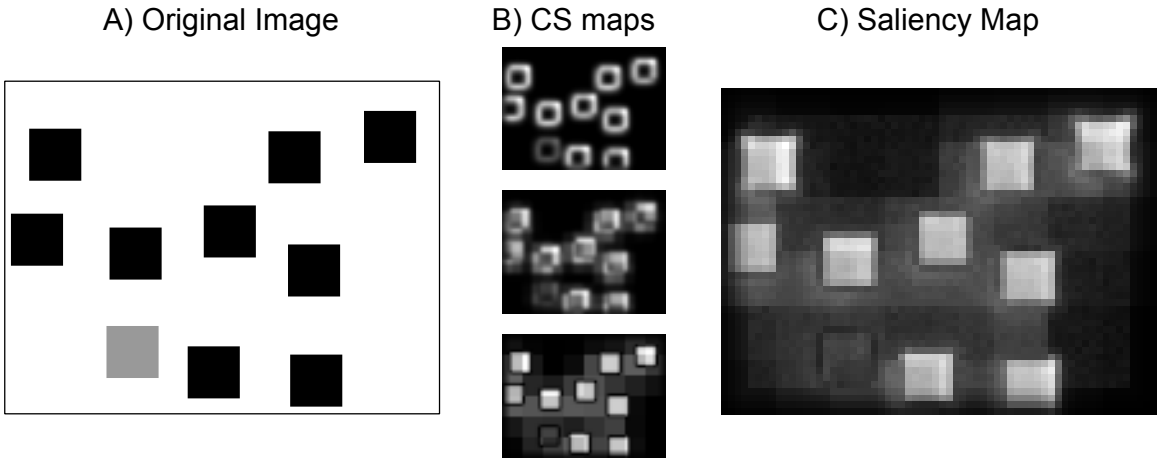

Figure 3.1: Example salience processing of Itti et al. (1998).(A) One of the experimental stimuli used as input to the model. (B) Three example scales of center-surround (CS) response. At all scales, the gray square has the weakest response. (C) Final output of the model. Intensity represents saliency. Color and orientation channels are included in the computation but they do not make a substantial contribution for this image.

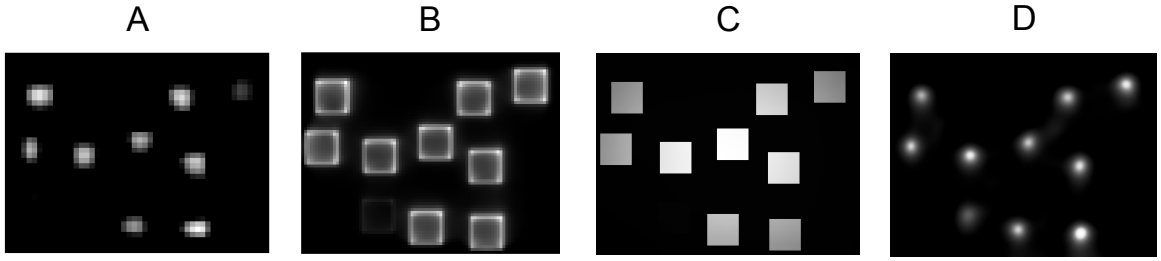

Figure 3.2: Output of models of salience for the input shown in Figure 3.1A with white regions indicating high salience. (A) Walther and Koch (2006) (B) Russell et al. (2014), (C) Perazzi et al. (2012), (D) Kümmerer et al. (2016).

3.1.2 Psychophysical Support and Limitations

Empirical data from the visual search literature are consistent with our intuition that the gray square in Figure 3.1A is salient. Treisman and Gormican (1988) showed that search for a gray target among black distractors is efficient (independent of the number of distractors) when the distractors, background and target colors are all easily discriminated. Efficient search can be explained most easily by assuming that attention is directed to the target. Bauer et al. (1996, their Figures 13 and 15) support a similar intuition for a case with distractors of two colors and the target falling between them in color space. Nothdurft (2006) also reports on several psychophysical experiments in which the participants assess the salience of different stimuli, and report the salience of a single low-contrast target among multiple high-contrast ones.

A substantial concern with these experiments is that all of them inform participants about the upcoming visual scene either directly, through the task instructions, or indirectly, by repeatedly presenting similar scenes (*e.g.* circular displays of dots where a few differ from the rest). Top-down influences of the task to be performed have been shown to affect participants' eye movements (DeAngelus and Pelz, 2009; Yarbus, 1967) and thus likely their attentional state. Therefore, we cannot conclude from existing evidence that the measured attentional effects are driven by bottom-up cues.

In Chapter 2, following Jeck et al. (2017), we addressed the difficulty of minimizing the contamination of behavioral measurements by top-down effects. The basic

ideas are that (a) task instructions are kept to a bare minimum, (b) spontaneity of responses is encouraged over long deliberations, and (b) every participant performs the task only a very limited number of times. These features are designed to minimize expectation which stimuli will be delivered next and anticipation of responses that participants might believe they are expected to provide. More specifically, we described a method of obtaining attentional responses from participants who were only minimally informed about the upcoming stimulus. Inspired by an experimental paradigm developed by Firestone and Scholl (2014), participants were shown a short sequence of complex natural scenes on a tablet computer and asked to "tap the first place you look when the image appears." Tapping responses were found to be significantly correlated with other measures of attention, specifically eye movements (Parkhurst et al., 2002), conscious selections of interesting image parts (Masciocchi et al., 2009), and the computational model by Itti et al. (1998).

In the following, we use the methods from Chapter 2 to empirically address the question of whether unique faint stimuli appear salient among a set of stronger stimuli, like the gray square among black squares in Fig 3.1A. We will use simplified images, similar to that figure, that are designed specifically to address this question by minimizing other possible influences although we believe that the effect is also present in more complex images. The scenes we use are described in detail in Section 3.2.2.

3.2 Methods

All methods were approved by the Johns Hopkins Institutional Review Board. Participants were passers-by on the Johns Hopkins University Homewood Campus. The significance level for all statistical tests was set to $\alpha = 0.05$. Code for the model described below will be made available at `https://github.com/dannyjeck/Proto-Object-Comparison`

3.2.1 Experimental Paradigm

We utilize the method described in Section 2.2 where participants are approached and asked to do a quick psychology experiment on a tablet computer. They are first shown a white screen with two small black squares (see Figure 2.1), which we call the initialization screen. They were instructed to tap on either one of the squares to bring up a test image, and were told "When the image appears, tap the first place you look". The image then appeared and, after the participant had tapped his or her selected location on it, the position of the tap and the reaction time (time between this tap and the tap on the initialization screen) were recorded. Test images strictly alternated between a natural scene and a "simple" scene (as defined in section 3.2.2). Each participant saw a total of 12 images of which the first always was a natural scene.

3.2.2 Stimuli

The stimulus set consisted of 30 images, with each showing a set of colored squares on a white background. We refer to these as "simple scenes" to distinguish them from the natural scenes that were presented in alternation with them. Responses to the natural scenes have been analyzed previously (Jeck et al., 2017), for the purpose of the current study they only serve to separate the simple scenes and to possibly reduce the predictability of the image sequence. On each of the simple scenes, the screen was separated equally into a 5×3 grid. In ten of the grid locations (randomly chosen) a square (120×120 pixels) was placed. Each square was placed at a random location (uniform distribution) within the central 80% of the grid cell in the horizontal and vertical directions. This placement pattern spaced out the squares evenly on average without creating a percept of a predictable pattern. The color of the squares varied among the six square image types generated, Gray/Black, All-Black, Black/Gray, Blue/Yellow, Pink/Red, and All-Red; an example of each is shown in Figure 3.3. Five images were generated for each image type. Each of the All-Black images was identical to one of the Gray/Black images except that the color of the single gray square was changed to black. This design allowed for a direct comparison between the gray square and the corresponding black square since the geometries of one All-Black and one Gray/Black image were identical. Likewise, each of the All-Red images was identical to one of the Pink/Red images except that the pink square was changed to red. Otherwise, all images were independent of each other. Note that the Itti

et al. (1998) model predicts that in a pair of Gray/Black and All-Black images with the squares at the same positions, the black square in the All-Black image at the same position of the single gray square is more salient (see Figure 3.1C) and therefore should be tapped more than the gray square. The analogous argument applies to the All-Red and Pink/Red pairs of images.

The simple scenes were separated into five groups of six, with each group containing one image from each type, and each of the matched images in the same groups. Thus each participant saw exactly one Gray/Black image, and the matched All-Black image was always shown to the same participant. Likewise, a Pink/Red image was shown with its matched All-Red image. Images were presented in randomized order, with the constraint that the first simple scene of a matched pair was always chosen such that each of the two members of a matched pairs of images had an equal number of participants see it first. For instance, the number of participants that saw the first Gray/Black image was the same as that of participants that saw the matched All-Black image as their first simple scene. This allowed us to perform a direct comparison of data gathered from the first time a participant saw a simple scene with matched sample sizes.

3.2.3 Proto-Object Comparison Model

One strategy how humans and other animals cope with the complexity of their environments is to transform raw sensory input into representations that match more

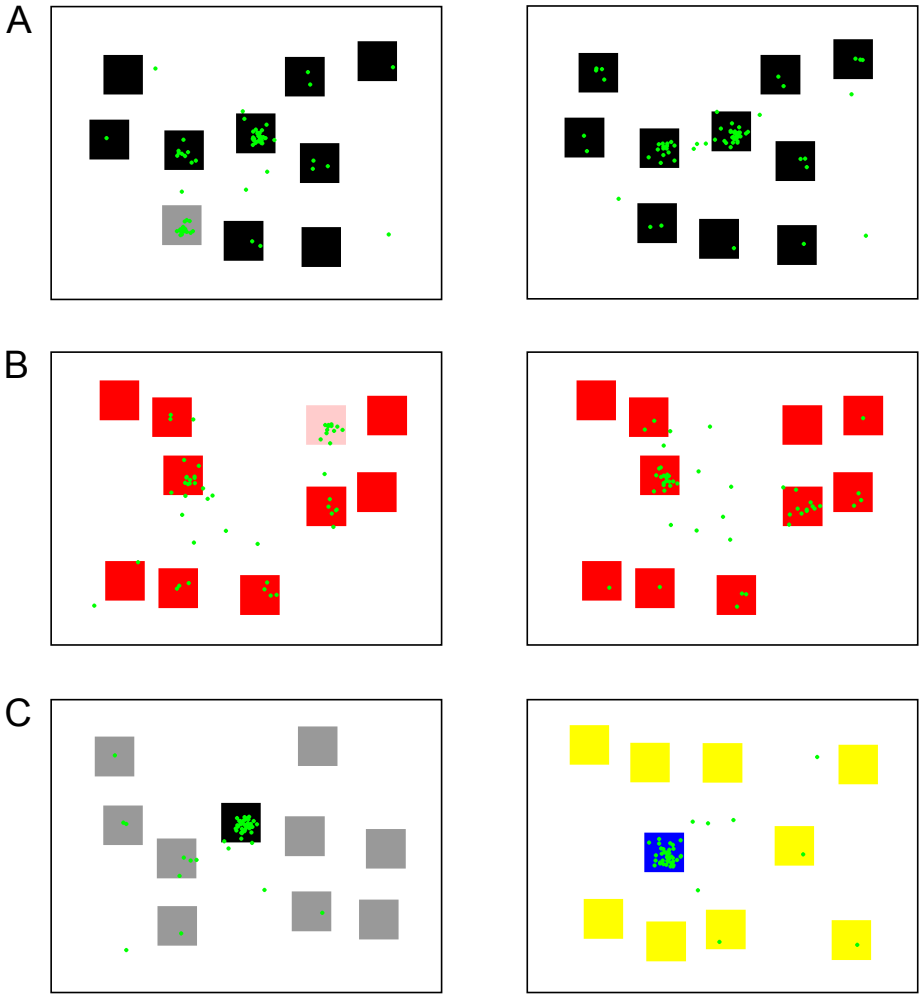

Figure 3.3: Example set of simple scenes, overlaid (green dots) with tap locations of all participants who saw this set. All taps shown are in response to the first time participants saw these scenes. (A) Gray/Black (left) and corresponding All-Black (Right) images. Note that the corresponding image has an identical spatial arrangement of squares. (B) Pink/Red (left) and corresponding All-Red (right) images. Again, the spatial arrangement is identical. (C) Black/Gray (left) and Blue/Yellow (right) images. The Black/Gray and Blue/Yellow images had independent spatial arrangements.

closely the functional relationships in the world. In the visual and auditory modalities this process is called perceptual organization Bregman (1990); Kimchi et al. (2003). In the visual system, the fundamental units of this representation are no longer activity

levels of retinal ganglion cells but their correlated patterns that correspond to visually perceived objects. We and others have developed quantitative models to understand the underlying computations Ardila et al. (2012); Craft et al. (2007); Hu et al. (2015, 2016, 2017); Mihalas et al. (2011); Pentland (1986); Ramenahalli et al. (2014); Russell et al. (2014); Walther and Koch (2006). As was observed by Rensink (2000) and Zhou et al. (2000), perceptual organization does not require the formation of fully-formed objects as would be needed for tasks like object recognition or discrimination. Instead, it is sufficient that the scene is segmented into entities that are characterized by a few elementary features, like their position, size *etc.*. Following Rensink (2000), we call these entities proto-objects. For the sake of simplicity, from now on we will use the terms "object" and "proto-object" interchangeably.

The behavioral results reported in section 3.3 suggest that humans compute relative saliency of simultaneously present (proto-)objects by comparing the features of these objects, rather than properties of regions that are defined by simple spatial relationships, as in center-surround contrast computations. To understand these behavioral results, we develop a computational model of visual saliency based on comparisons between objects. The model generalizes the idea that objects that differ from other objects are salient, while repeated objects are not salient. While in early models (Itti and Koch, 2001; Itti et al., 1998; Koch and Ullman, 1985; Niebur and Koch, 1996) the elements to be compared were defined purely spatially, newer approaches are based on proto-objects (Russell et al., 2014; Walther and Koch, 2006).

However, as discussed previously, these models cannot explain that humans assign higher salience to unique objects over repeated objects. In the following, we develop a model of this competitive interaction between objects.

To obtain a representation of proto-objects in a visual scene, we compute grouping cell activity as in the Russell et al. (2014) model but we remove the normalization procedure that follows in that model. These grouping cells tile the entire image with overlapping proto-objects of many different radii, and are computed on different submodalities (intensity, color, and orientation). A grouping cell in the Russell et al. (2014) model will have a strong response if it is at the center of a set of co-circular edges at the preferred radius of the grouping cell. Grouping cells in our model have a minimum preferred radius of 32 pixels and a maximum of 256 pixels.

Proto-objects in our model are defined by their position (X,Y) and radius (r). We do not, however, assume a binary distinction between the presence and absence of proto-objects at any location. Instead, the activity of grouping cell responses in the Russell et al. (2014) model provides a graded measure of the "belief" that a proto-object with a specific radius is present at a specific location. Let (X_i, Y_i), and r_i represent position and radius for the i-th proto-object. We define its strength S_i as the product of r_i^2 with the i-th grouping cell response. Since proto-objects are calculated by contrast-based mechanisms, the S values of proto-objects representing the gray square in Figure 3.1A are lower than those of black squares. An example of a set of grouping cell responses with a radius of 32 pixels to the stimulus in Figure 3.4A

is shown in Figure 3.4B.

In order to compare between proto-objects in our new normalization step below, we must first compute a set of features for each proto-object. For each proto-object, the POC model computes features over the image region defined by the circle with center at (X_i, Y_i) with radius r_i (see 3.4C for an example). We compute histograms of L, a^*, and b^* values from the CIELAB color space (Ibraheem et al., 2012). Each of these histograms has nine bins and is normalized to sum to 1 so that patches of different radii can be compared appropriately. We also compute histograms with nine bins for the potential radii of proto-objects in the patch. Eight of the bins have the value zero and the one which corresponds to the actual object radius having the value unity. For the i-th proto-object, this gives us a feature vector F_i whose components are the values of 36 different bins (nine bins for each of the four features L_i, a_i^*, b_i^*, and r_i). We refer to the value for the i-th proto-object in the j-th bin as F_{ij}. In the brain, we presume that these features are computed simultaneously with the computation of proto-objects themselves. The entries in the histograms correspond to activity patterns of separate neuronal populations that are tuned to the features represented by the histogram bins. We chose those features in our model partly for reasons of computational efficiency rather than as detailed implementations of biological processes. For instance, we do not claim that there are neuronal populations that one-by-one code L, a^*, and b^* values from the CIELAB color space but we do believe that color is represented explicitly in neuronal activity patterns.

Interaction between proto-objects is introduced through a normalization process,

$$N_{ij} = \frac{S_i F_{ij}}{\sum_k S_k F_{kj}} \tag{3.1}$$

if $\sum_k S_i F_{kj} \neq 0$, otherwise $N_{ij} = 0$. In this equation, the value in each histogram bin is multiplied by S_i so that strongly detected objects are given more weight in the normalization process, and strongly detected objects with the same feature values will suppress one another. This is illustrated in Figure 3.4D for three proto-objects: two with high S_i that share the same feature values and one with a lower S_i value that is unique. The strength of each proto-object is then computed as

$$P_i = \sum_j N_{ij} \tag{3.2}$$

and a saliency map Q is defined as a sum of Gaussians with weight given by P_i and locations given by the proto-object.

$$Q(x, y) = \sum_i P_i \exp\left(-\frac{(x - X_i)^2 + (y - Y_i)^2}{2\sigma_i^2}\right) \tag{3.3}$$

where the spread $\sigma_i = \frac{r_i}{2}$ of the Gaussian ensures that most of a proto-object's activation is near its center. A saliency map for the three proto-objects and the full output using all proto-objects is shown in Figure 3.4E.

We noted that the saliency maps generated by the POC model are blurrier than

the Russell et al. (2014) model, which has in the past correlated with improved performance on natural scenes (Judd et al., 2012). To illustrate the importance of the normalization process rather than the other implementation details of the model (*e.g.* the creation of the saliency map using a sum of Gaussians), we also generate a saliency map using equation 3.3 but replacing P_i with S_i. Note that this is equivalent (up to a scaling factor) of computing equation 3.1 without the denominator, since $\sum_j F_{ij} = 1$.

3.3 Results

We recorded 1512 taps on simple scenes from 252 participants (101 male, 151 female). Population results are shown in Figure 3.5. Reaction time (RT) was defined as the time from tapping on the initialization screen to tapping on the test image. Median RT was 1.3 seconds (mean 1.4 seconds). We did not analyze RTs in detail because our data collection system (iPad) did not allow control about the exact timing of image presentation.

3.3.1 First View Only

As noted, All-Black images and Gray/Black images came in pairs with identical layouts, with the only difference between the two members of a pair being that one of the black squares in the former had been changed to a gray square in the latter.

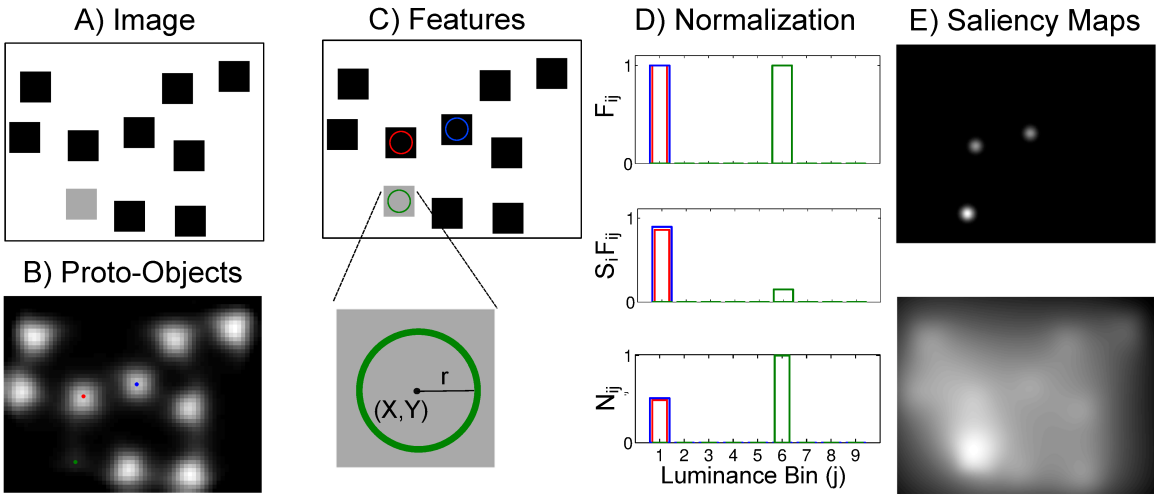

Figure 3.4: Simplified illustration of the POC model. (A) Example input image. (B) Map of the strengths of proto-objects with a radius of $r = 32$ pixels only. Three colored dots correspond to proto-objects used in the following panels. Note the weak response to the gray square, centered on green dot. (C) Illustration of the feature computation for the proto-object highlighted by the green dot in B. Histograms are formed over the pixels within the circle of radius r around the center (X, Y) of the proto-object. Note that all pixels are gray, which corresponds to the sixth luminance bin in panel D. (D) Top panel, feature representations for three proto-objects. Two are on black regions (red and blue dots in B), these proto-objects have identical colors (black) and their values in the histogram are identical, unity in the first histogram bin and zero in all others. The third is on the gray region of the image (Panel C and green dot in panel B), it has unity value in bin 6 and zero elsewhere. Middle panel, feature values multiplied by proto-object strength S. Bottom panel, strength after normalization. (E) Top panel, saliency output for the three proto-objects processed in D. Output for all other proto-objects omitted for clarity. Bottom panel, output for all proto-objects. Note the enhanced saliency of the gray square in D, E.

Likewise, all pairs of All-Red and Pink/Red images differed only in the color of exactly one square which was red, like all other squares, in the first case, and pink in the second.

We obtained the main result of this study by analyzing tap locations for the very first presentation of a simple scene (*i.e.* the second image the participants were presented with since the first was always a natural scene). We found that participants

tapped the singleton square (gray or pink) significantly more frequently than the corresponding square that shared its color with the other squares (black or red). This result holds both for a gray square among black squares (Figure 3.5A) and for a pink square among red squares (Figure 3.5C). In the former case, we observed 14 taps on the gray square taps *vs.* 6 taps on the black square, both out of 51 taps. A one-tailed Fisher's exact test gives $p = 0.039$. In the latter case, we observed 16 taps on the pink square *vs.* 8 on the red square, both out of 47 taps. A one-tailed Fisher's exact test gives $p = 0.048$.

This result confirms the intuition that a unique stimulus that is closer to the background color is more salient than a non-unique stimulus with higher contrast to the background.

3.3.2 All Presentations

Each image had a different singleton tap rate, defined as the fraction of times that participants tapped on the singleton square. We did not ensure that the same number of participants saw the same scene as their first, second, or n-th simple scene. Therefore, we can not quantify whether the location of an image in the image sequence influences tap rates. For the remainder of our analysis, we therefore aggregated data over all six presentations of simple scenes.

Results are highly significant for aggregated data. Including all six views by each participant, gray squares are tapped more than the black squares in corresponding

positions (90 taps on gray squares *vs.* 21 on black squares, out of 252 taps in both cases; a paired t-test gives $p < 10^{-14}$; Figure 3.5B) and the same holds for pink squares *vs.* corresponding red squares (101 *vs.* 59 out of 252; $p < 10^{-5}$; Figure 3.5D). Note that a paired t-test is appropriate in this case because the same participants saw paired Gray/Black and All-Black images on different presentations (and the same for Pink/Red and All-Red images). More detailed analysis shows that direct comparisons between Gray/Black and All-Black images were significant individually for each of the five pairs (see Figure 3.7), as well as for three out of the five pairs of Pink/Red and All-Red images (Figure 3.8). For the two images without significantly increased tap rates on the pink squares, the corresponding red square was the most tapped square on the All-Red image and in both images it was located close to the center of the scene. A ceiling effect, likely due to the geometrical arrangements of stimuli (center bias, see next paragraph), may thus be the reason why we did not find a significant effect in these cases.

As in previous studies (Buswell, 1935; Parkhurst et al., 2002; Tseng et al., 2009), we found a strong center bias in our results. Figure 3.5E shows the rates at which participants tapped on a singleton square as a function of the square's Euclidean distance from the center of the image. The lines in the figure are generated from a linear regression model where each type of stimulus and the distance from the center are used to predict the tap rate. Also shown are the tap rates and linear fits for the non-singletons in the All-Black and All-Red images. A significant effect of distance

from the center was found for each line (F-test, all $p < 10^{-5}$). The negative slope of all lines confirms the existence of a center bias in all conditions. Interaction terms between the distance from the center and the stimulus type were not found to be significant except in the case of the non-singletons.

By analyzing the intercepts of the fit lines (Figure 3.5F) we can roughly gauge the salience for the different image types independently of the center bias. By performing pairwise comparisons between the intercepts, we found that the Blue/Yellow intercept was significantly higher than the Gray/Black and the Pink/Red intercepts (F-test, all $p < 0.05$), and intercept of the fit line for non-singletons was lower than for any of the images containing singletons (all $p < 10^{-11}$). These results held after performing a False-Discovery Rate correction (Benjamini and Hochberg, 1995) to control for multiple comparisons. We also found that the singletons in Black/Gray and Blue/Yellow images are generally more salient than either the singletons in Gray/Black or singletons in Pink/Red images. These results agree qualitatively with previously found search asymmetry studies (Treisman and Gormican, 1988) since the Gray/Black singletons are less salient than the Black/Gray singletons, (Fig. 3.5E) while confirming that the singleton gray squares in the Gray/Black images can still be salient.

3.3.3 POC Model

The POC model is able to predict increased attention to unique objects in a scene. For each of the simple scene stimuli shown to participants, the POC model

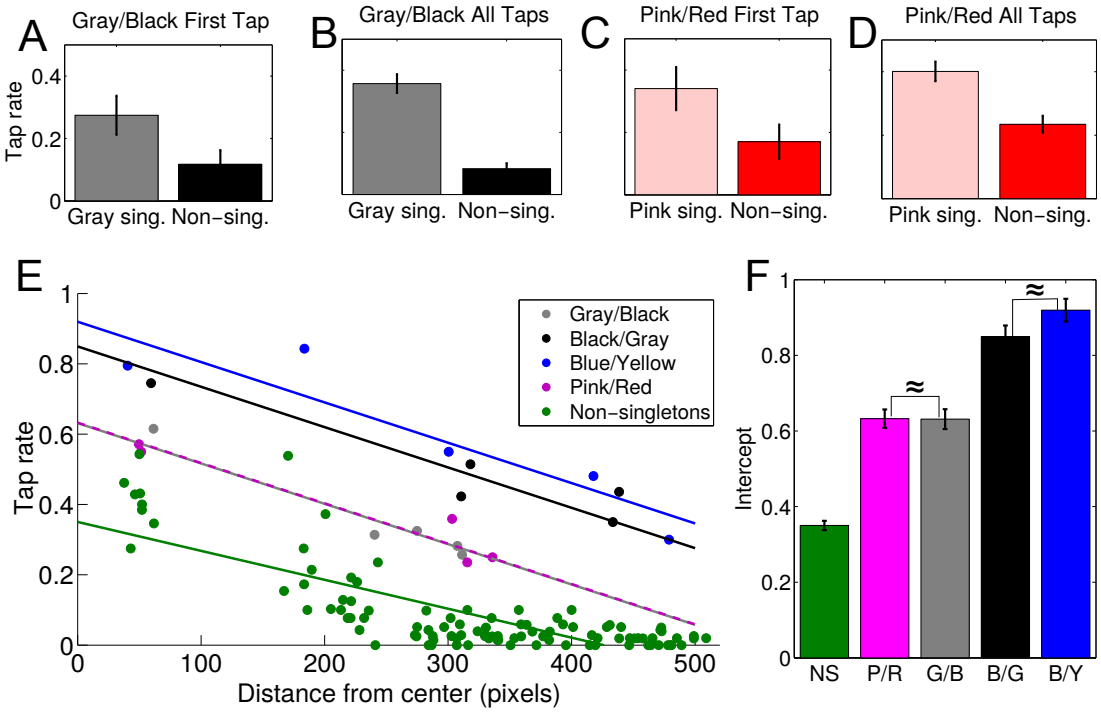

Figure 3.5: (A-D) Rates at which participants tapped on singleton (sing.) squares *vs.* non-singleton corresponding squares (Non-sing.) in a control image. Error bars represent standard error. (A) Gray/Black *vs.* All-Black comparison for each participant's first tap. (B) Gray/Black *vs.* All-Black comparison for all taps. (C) Pink/Red *vs.* All-Red comparison for each participant's first tap. (D) Pink/Red *vs.* All-Red comparison for all taps. (E) Rates at which participants tapped on the singleton squares (colored circles, see legend), and each of the various non-singleton squares in the All-Red and All-Black images (green circles). The horizontal axis is the Euclidean distance from the center of the image. Fit lines were generated for each singleton image type individually and for Non-singletons combined, colors same as for the corresponding circle symbols. (F) The vertical intercept of each fit line from (E) with standard error bars (G/B = Gray/Black, P/R = Pink/Red, B/G = Black/Gray, B/Y = Blue/Yellow, NS = Non-singletons). The symbol $\approx$ indicates that no pairwise difference was found ($p \geq 0.05$). All other intercept pairs differed significantly ($p < 0.05$).

was able to predict that the unique object was the most salient object in the image (see Figure 3.6, top row for an example). This held for all stimuli shown in Figures 3.7 and 3.8, even without accounting for center bias. Parkhurst et al. (2002) showed that even a very simple filter (convolution with a 2-D Gaussian) that emphasizes processing in the center of the visual field improves salience prediction (their Figure 9) but since the POC model already has perfect performance in this respect, it seems besides the point to add this modification. However, if in the future the model performs less than perfect on other data sets, it is highly likely that this simple change will improve its performance.

We also wondered whether the POC model is able to maintain the same level of performance on natural scenes as the Russell et al. (2014) model on which it is based. We tested the output of the POC model on images where we had previously recorded fixations to generate saliency maps, and used the Pearson correlation R between fixation maps and these saliency maps as our measure of accuracy (see Jeck et al. (2017) for details). Average R over the 100 images we tested was 0.484 for the POC model, actually slightly higher than the value of $R = 0.472$ for the Russell et al. (2014) model.

We also tested the POC model without inter-object competition, replacing P_i in equation 3.3 with S_i. This test was included to ensure that among the differences between the Russell et al. (2014) model and the POC model, the normalization process itself was not detrimental to performance on natural scenes. We found that the aver-

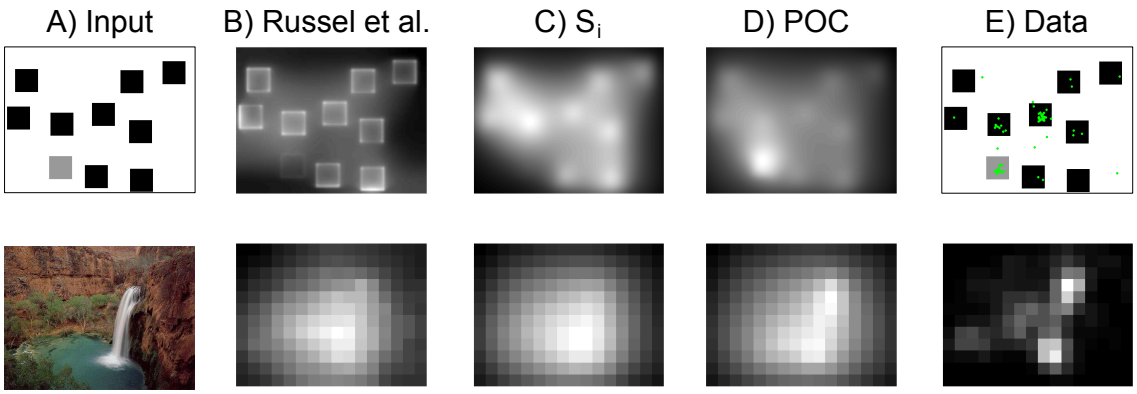

Figure 3.6: Model behavior on an example simple scene (top row) and natural scene (bottom). (A) Input image. (B) Output of the Russell et al. (2014) model. (C) Output of the model without normalization, using S_i instead of P_i in equation 3.3. (D) Output of the POC model. (E) Image with tap data (top) and fixation map (bottom) overlaid. For the natural scene saliency and fixation maps are downsampled to the 12×16 resolution used in Chapter 2.

age R for this modified model was 0.464, indicating that the normalization procedure does provide value on natural scenes as well as simple scenes.

3.4 Discussion

3.4.1 Simple Scenes and Models of Salience

Using the tapping paradigm we have shown that participants preferentially report that the first place they look is on a unique object. It was to be expected, and is predicted by quantitative models of salience computation, that this applies to unique objects that have high contrast relative to all other scene elements including the background. We show that this is the case for intensity contrast (black square

surrounded by white background and gray squares) and for color contrast (blue square surrounded by white background and yellow squares). In contrast, all computational models we tested predict the opposite outcome for a "faint" singleton object: As long as the singleton is well-isolated from other objects so that local center-surround differences incorporate substantial input from the background, models assign low salience to a singleton with lower contrast against the background than other objects in the image (gray or pink squares surrounded by white background and black/red squares). We show that humans select these singletons over otherwise identical stimuli that are not singletons. As long as the singleton, background, and other objects are sufficiently far apart from each other in color space, the singleton will be preferentially selected.

While previous research in the visual search and psychophysical literature have arrived at similar conclusions about salience (Bauer et al., 1996; Nothdurft, 2006; Treisman and Gormican, 1988), the participants involved in those studies were either explicitly informed about the nature of the images being presented, or they performed enough trials that they likely expected certain types of images. It is therefore not clear to what extent responses influenced by systematic top-down effects rather than controlled by the perceptual qualities of the visual stimuli. We therefore developed an experiment in which participants received minimal information on the stimuli. Our results show significantly increased salience of unique objects even for the very first time a participant sees a scene.

3.4.2 Top down Influences and Possible Experimental Confounds

Our experimental paradigm makes it feasible to measure behavioral responses from a large number of participants while minimizing expectations of stimulus contents they may have or develop during the experiment. Nevertheless, given the complexity of the human mind, it is extremely difficult to completely exclude top-down influences. In the following, we discuss potential remaining top-down effects that may have biased our results, from the more generic to the more specific.

Our experimental paradigm certainly does not remove all top-down influences on attention. Participants' behavior will naturally be affected by outdoor distractions, their internal state *etc.*. However, we suggest that removal of (explicitly or implicitly provided) information about the visual stimuli removes those top-down influences that are specific to the images they see, leaving only those of a generic nature that are independent of the images. Furthermore, in our experiment each image with a faint singleton square is paired with another one that is identical in all respects except that the singleton is replaced by a distractor square. There is no reason to assume that top-down influences due to generic distractions *etc.* are different between the two images of a pair. Therefore, differences in internal state of a general nature cannot explain our results.

A second, rather pessimistic interpretation of our results is that the participants

were priming themselves to look for unusual objects because they knew that they were participating in a psychology experiment, and that they responded in a way that they believed the scientist wanted them to respond ("demand bias;" Firestone and Scholl, 2016). It has, indeed, been found that participants in psychology experiments will modulate their responses based on what they believe the purpose of the experiment is. For example, Durgin et al. (2009) showed that participants will give a higher estimate of a slope to be scaled while wearing a heavy backpack if they infer that experimenters expect that the weight will influence their judgment, compared to a situation where they carry the same backpack but believe its weight is irrelevant for estimating the slant (because they are told that it contains measurement equipment). We acknowledge the possibility that subjects tap the unique object because they think the experimenter wants that response, though it does not seem probable. Our experiment was designed specifically to minimize this effect which, if present, should be much more prevalent in the cited previous behavioral studies of the salience of faint objects. In our experiment, a significant effect is observed when participants respond to the very first singleton image they ever see (after one other image showing an unrelated natural scene), with the response given within about one second. It seems highly unlikely that participants come to a conclusion of what the experimenter expects from them in literally a split second without any additional information but the image itself. In addition, the fact that our results in Chapter 2 (Jeck et al., 2017) showed significant correlations between behavior in this task and several measures

of saliency strongly suggests that taps do occur on salient stimuli. Furthermore, in informal debriefing of participants after the experiment, none of the participants asked if they were supposed to tap on the singletons. All this supports the interpretation that our results are not due to demand bias or similar effects.

Another criticism may be that the participants had enough time viewing the image to engage in a mixture of top-down and bottom-up processing. Under this view, the fact that the participants have a reaction time greater than a second is a serious design limitation. Rather than their attention being drawn immediately to the most salient stimulus and then reporting it, during that amount of time the participants may saccade to multiple locations and modify their choice of where they report they first look based on a higher-level interpretation of the scene. While it is true that the median reaction time of 1.3s would theoretically allow several fixations, this does not take into account the time needed for making a controlled hand-movement to a specific location in a task executed without any previous training, performed in a casual environment (standing on a walkway on campus), and without encouragement for a rapid response. We consider it likely that the vast majority of the 1.3s long period between the time the image was presented and the finger reached the tablet surface was devoted to motor planning and actual limb motion. We also note that this criticism would likewise apply to fixation data which is typically gathered over several seconds of free viewing (Borji and Itti, 2013).

Finally, we briefly address two concerns that are not related to top-down influ-

ences. The first is that, if the scene gist hypothesis (Review: Oliva, 2005) is true, the overall structure of the scene may begin to have an effect on attention immediately (within $\approx 100ms$). This is not a top-down effect by our definition since the gist is a property of the scene, rather than of the internal state of the observer. It would be extremely difficult to separate this effect from guidance of attention to salient stimuli in any experiment. In fact, if the gist of our singleton scenes can be described by "several similar objects and one dissimilar object on a homogeneous background" it would even conceptually be difficult to distinguish its effect from salience-driven guidance of attention to the singleton. Both explanations may simply be different descriptions of the same underlying process.

One possible methodological concern may be that participants may not follow our instructions to "tap the first place you look." Indeed, we do not control whether they do but we see this formulation rather as a non-technical way to instruct participants to indicate where they are attending than as an action that needs to be followed precisely. Our interest is to assess where attention is deployed, rather than finding a precise surrogate for eye movements. Pointing with a finger is a very natural and universal human behavior (Kita, 2003) which already appears during infancy, at about one year of age (Leavens et al., 2005; Tomasello et al., 2007). The purpose of finger pointing is typically to direct attention (either that of another person or occasionally of the pointing person him or herself) towards a specific part of the world. This behavior is thus a direct, voluntary expression of attentional selection.

While the term "overt attention" is traditionally used for eye movements, pointing can therefore be seen as another form of overt attention that makes the outcome of the agent's attentional selection process explicit. This is all we need to gauge deployment of attention in our scenes.

3.4.3 Object-based Models

Regardless of interpretation, a model that would capture the observed behavior must rely on a computation more advanced than spatially defined center-surround operations. A natural step in this direction is the formation of proto-objects by grouping together low level features of the same type. The Russell et al. (2014) model includes processing that computes where proto-objects are, computing strength of proto-object representations based on the strength and organization of edges in the image. However, the model is unable to capture the observed behavior (see Figure 3.2B) because it does not perform any comparison between proto-objects. Instead, the edges of the gray square with lower contrast result in a weaker proto-object representation. The same applies to the Walther and Koch (2006) model.

Models that do capture the observed behavior will need to operate on proto-objects or object representations. The Perazzi et al. (2012) model (which can capture the observed behavior after fine-tuning parameters, see Section 3.1.1) breaks the image into patches and looks for a unique, compactly distributed color. Comparing patches on a global scale when computing color uniqueness and color distribution allows this

(modified) model to generate the correct output. Proto-objects offer a representation that has more fidelity to the physical world, with distinct objects occupying consistent locations in visual space. Such representations would be more behaviorally useful than color patches, as predictable changes in the visual scene could be encoded for a proto-object but not for a color patch. Additionally, existing proto-object representations like Russell et al. (2014) separate out the background by assigning low S_i values to regions corresponding to the background. This reduces the number of regions that need comparing, as the Perazzi et al. (2012) model assigns many patches to the background.

3.4.4 What Are Unique/Rare Objects?

If, as we assumed based on our intuition and confirmed by our behavioral experiments, uniqueness contributes to the saliency of a (proto-)object, a definition or at least characterization of what *uniqueness* means would be useful. As Kadir and Brady (2001) observe, this is not a trivial question. Generalizing to the more general concept of *rarity,* they point out that what is considered rare depends not only on the contents of a scene but also on the method by which it is described. If feature descriptors are highly discriminatory, features of all (proto-)objects differ substantially and *everything is rare.* On the contrary, if descriptors are very general, all (proto-)objects have much in common and *nothing is rare.*

Once a descriptor is chosen, a simple Bayesian argument shows (Kadir and Brady,

2001) that those areas are of interest that have high entropy. It was shown empirically that, indeed, fixated (and therefore likely salient) image regions show increased entropy, contrast and decorrelation between neighboring pixels (McCamy et al., 2014; Parkhurst and Niebur, 2003, 2004; Reinagel and Zador, 1999). The reason is that their contents are not well explained by the applied feature descriptor, therefore these areas require more detailed scrutiny by more powerful mechanisms. The selection of areas that need to be processed in more detail is, of course, the fundamental function of selective attention. In the language of perception, these areas are *salient* which directs attention to them. A well-chosen feature descriptor explains large parts of the image satisfactorily, therefore this happens only rarely. This is the case even for very simple descriptors, *e.g.* the assumption of homogeneity in elementary features or combinations of them. For instance, the classical saliency map models (Itti and Koch, 2001; Itti et al., 1998; Koch and Ullman, 1985; Niebur and Koch, 1996) make the implicit assumption that color, intensity and orientation are homogeneous over space and that deviations from this assumption are marked as salient, to be selected by attention. This is easily generalized to time-varying input when change or motion are involved with the implicit assumption that energy in space-time is constant (Itti, 2005; Niebur and Koch, 1996; Parkhurst, 2002) and, again, deviations from this assumption are salient. The present work supports the idea that a simple descriptor with a prior of uniformity in pixels is not sufficient to explain behavior, since high-contrast would represent a strong deviation from the mean of the uniform distribution. However,

a descriptor that at first pass describes the numerous objects and background separately may label a unique low-contrast object as salient because it deviates from expectations.

3.5 Supplementary Information

3.5.1 Linear separability of the unique faint square in feature space

While Figure 3.1C shows that the Itti et al. (1998) model of visual salience does not produce a strong response for the unique gray square, we wondered whether *any* linear combination of the intensity features in that model could produce the observed behavior where the unique square is selected more than an otherwise identical black square at the same location. If a linear combination of the intensity center-surround features could produce the desired behavior, then there would exist some set of weights on the center-surround maps that enhance the gray square while suppressing the black squares as well as the background.

We therefore set up a set of linear constraints. If they are impossible to satisfy then the feature space is unable to replicate human behavior. The center surround features c generated by the original code from Itti et al. (1998) are six 48×64 pixel maps, one at every spatial scale in the Laplace pyramid of the intensity submodality.

For a given (x, y) location on the image, $c(x, y)$ is therefore a vector with six elements corresponding to each center surround feature value at (x, y). Our constraints are that a set of weights w must satisfy

$$w^T c(x, y) < K \tag{3.4}$$

for some constant K, when (x, y) is outside of the gray square. Additionally, for some point (x', y') inside the gray square,

$$w^T c(x', y') > K \tag{3.5}$$

must also be satisfied. For a given value of (x', y') this set of constraints can be reformulated as a linear programming problem which allows us to check the satisfiability of the constraints using off-the-shelf software (Matlab R2012b). We found that for the image in Figure 3.1A the constraints could not be satisfied for any given (x', y') value on the gray square.

3.5.2 Additional Figures

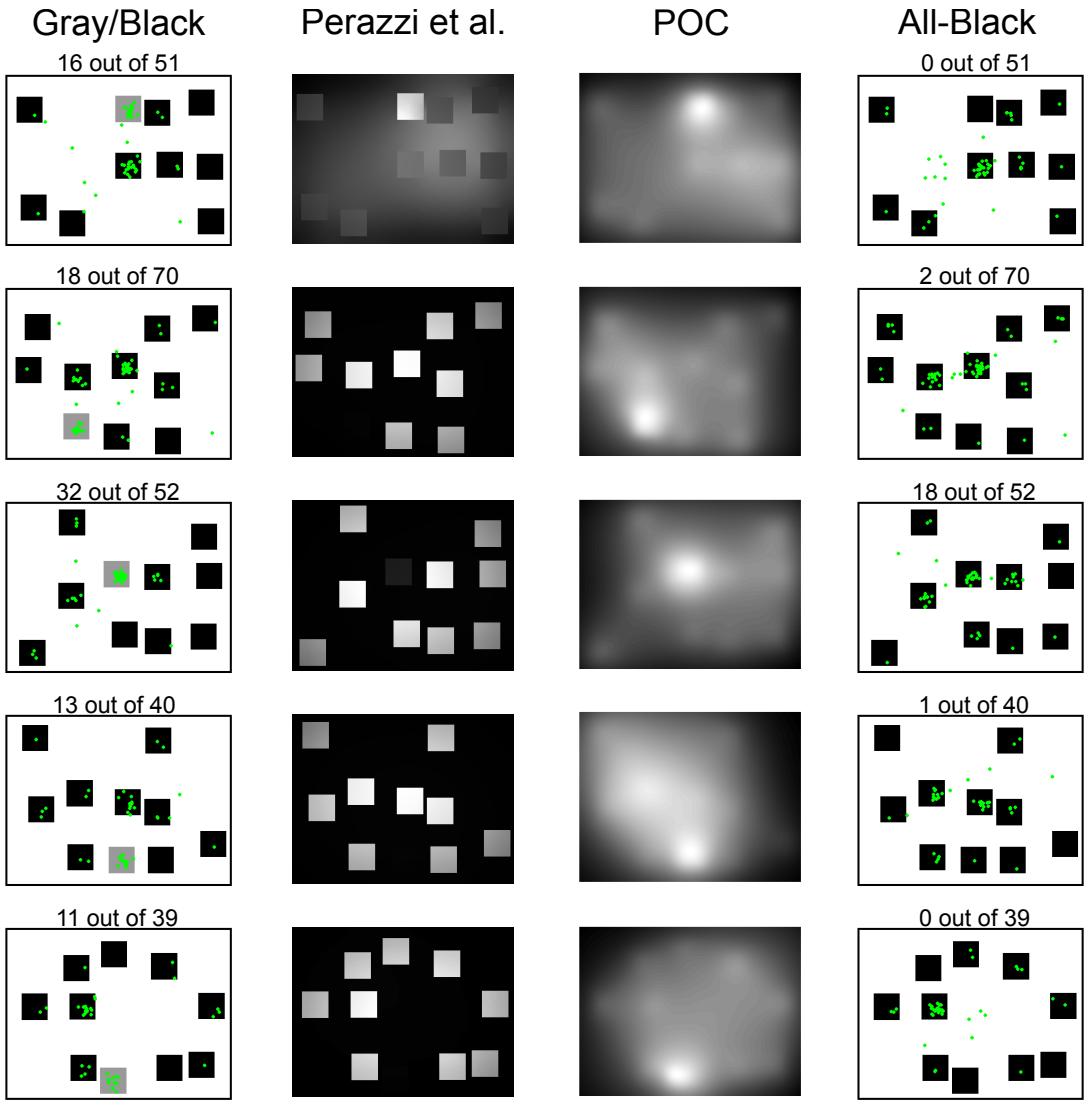

Figure 3.7: Model output and data collected on all Gray/Black images. First column, images with tap results overlaid (green dots). Text above each image indicates number of taps on the gray square out of the total number of participants who saw the image. Second column, output of the Perazzi et al. (2012) model. In four of the five cases, the unique square is not given high salience. Third column, output of the POC model. The unique square is always given high salience. Fourth column, corresponding All-Black images shown to participants with overlays of tap results (green dots). Text above each image indicates the number of taps on the corresponding black square. p values for direct comparisons between first and fourth column were computed using a paired t-test ($1.57 * 10^{-5}$, $7.99 * 10^{-5}$, $6.96 * 10^{-4}$, $2.10 * 10^{-4}$, and $4.22 * 10^{-4}$ respectively from top to bottom)

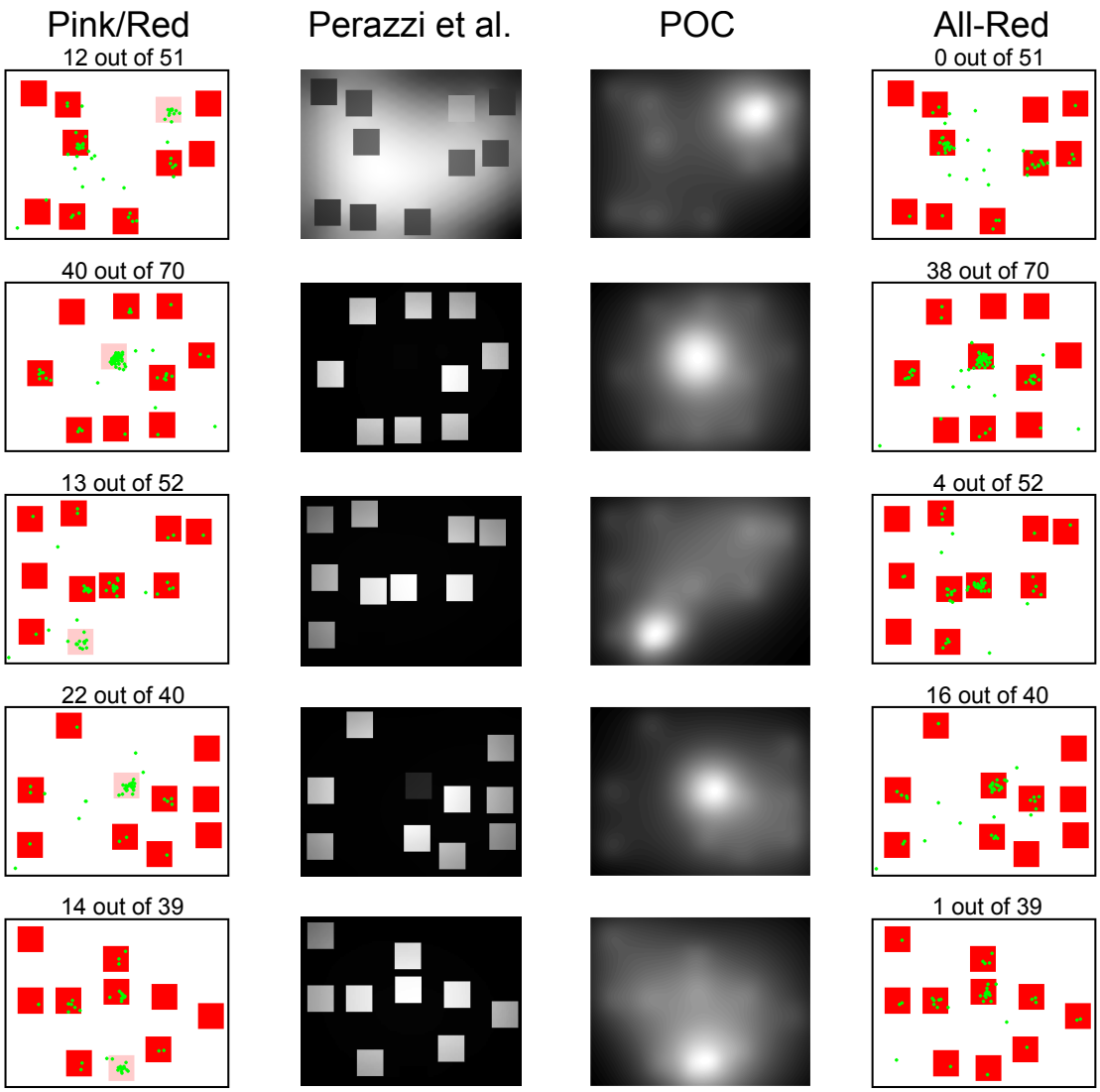

Figure 3.8: Model output and data collected on all Pink/Red images. First column, images with tap results overlaid (green dots). Text above each image indicates number of taps on the gray square out of the total number of participants who saw the image. Second column, output of the Perazzi et al. (2012) model. In four of the five cases, the unique square is not given high salience. Third column, output of the POC model. The unique square is always given high salience. Fourth column, corresponding All-Red images shown to participants with overlays of tap results (green dots). Text above each image indicates the number of taps on the corresponding black square. p values for direct comparisons between first and fourth column were computed using a paired t-test ($2.68 * 10^{-4}$, 0.686, 0.0111, 0.160, and $3.48 * 10^{-4}$ respectively from top to bottom)

Chapter 4

Closed Form Jitter Analysis of Neuronal Spike Trains

4.1 Introduction

Most neurons communicate by series of action potentials (spikes). It is known that in some cases the detailed time structure of these spike trains is used for information transmission while in others, only the overall number of spikes in some interval seems to be important but not their position in the interval. Examples of the first type are various kinds of "temporal coding" schemes proposed for different neural system and for different functional roles, *e.g.* refs Abeles (1991); Niebur et al. (1993); Riehle et al. (1997); Singer and Gray (1995); Softky (1995); Steinmetz et al. (2000), while the latter is the well-known rate-code mechanism, *e.g.* refs. Adrian

and Zotterman (1926); Shadlen and Newsome (1998). In the context of visual proto-objects, Martin and von der Heydt (2015) argued that the detection of a particular temporal code in extrastriate cortex (V2) supports the hypothesis that those neurons are receiving feedback from a proto-object representation. To distinguish between the two possibilities of rate and temporal codes, it is necessary to find whether reproducible correlations at the relevant time scale are present in neuronal data. One common way to approach this problem is to use auto- or cross-correlation functions as test statistics. Then one can (a) search for non-trivial structure in the function, like deviations from uniformity, and (b) detect whether there are differences between these functions in different experimental (*e.g.* behavioral) conditions.

The situation is complicated by the influence of rate variations on the raw correlations. To increase the signal-to-noise ratio, correlations are typically computed as averages over many trials. Changes in the behavioral state of an animal, *e.g.* due to onset of sensory stimuli or motor responses that occur always at the same time during a trial, typically result in changes in neural firing rates which are common to many neurons. While these are genuine correlations, they are unrelated to the neuronal coding question. Different techniques have been developed to remove them, *e.g.* subtraction of a "shuffle predictor" Perkel et al. (1967), the average of cross correlations between spike trains from permuted trials[1]. While this correction removes correlations that are time-locked to trial onset, it was later pointed out that peaks in the

[1]A "shift predictor" is very similar but the correlation function is computed from trials that immediately follow each other, rather than from randomly selected trials.

correlation function that may be taken as indicative of correlated firing (*e.g.* at zero lag) can also be caused by slow rate covariations Brody (1998, 1999). After finding a significant peak in the cross correlation function, this ambiguity can be addressed by analyzing the time scale at which the measured correlation arises.

It was pointed out more recently Amarasingham et al. (2012) that the null hypothesis of spike trains being independent in earlier work Perkel et al. (1967) is useless if their mean rates co-vary. Rate co-variation of means the spike trains are not independent which leads to its immediate rejection of the null without providing any further insight. Amarasingham and his colleagues instead proposed a more detailed null hypothesis, namely that within an interval of width Δ, the exact location of spikes does not matter. Then, under the null hypothesis, simulated spike trains can be generated by modifying the spike times of an original measured spike train within a range of Δ. The cross correlations obtained from these modified ("jittered") spike trains are then compared to those obtained from the original. If significant differences are found, the null hypothesis is rejected and it is likely that non-random correlations at time scales $\leq \Delta$ are present in the data. Additionally, this method gives rise to the computation of jitter-corrected cross correlograms, which have been used to compare changes in synchrony across experimental conditions (Hirabayashi et al., 2013a,b; Martin and von der Heydt, 2015; Smith et al., 2013). Because the method relies on repeated simulation of spike trains, it will be referred to as the *Monte Carlo jitter method* for the purposes of this thesis.

While the Monte Carlo jitter method (described fully in Section 4.2.1) is useful and easily generalized to complex statistical tests and hypotheses, its practical utility is limited by the inherent trade-off between accuracy and computation time in all Monte Carlo methods. As we will show in Section 4.3.2, the computation time may be prohibitively long, and even at this cost, it will only be a numerical approximation of the true solution. In the case where the test statistic is the cross-correllation value at a single lag, the p value can be computed exactly, as was shown by Harrison (2013). In the present study, we therefore explore the benefits of computing in closed form the distribution which is only approximated by the Monte Carlo simulations. Accordingly we refer to this method, described in Section 4.2.2, as the *closed form jitter method*. In addition to computing the p value for rejecting the null hypothesis exactly we show that the computation of the jitter-corrected cross correlogram follows readily from that derivation. The computational performance of the closed form jitter and Monte Carlo jitter methods are compared theoretically (as computational complexity) in Section 4.3.1 and practically (as computational time) in Section 4.3.2.

4.2 Materials and Methods

4.2.1 The Monte Carlo Jitter Method

Utilizing the Monte Carlo jitter method (Amarasingham et al., 2012), it is possible to determine whether correlations arise from fine temporal structure or larger scale

T	Length of binned spike trains
X, Y	Two spike trains
τ	Correlation Lag
$C(\tau)$	Correlation of X and Y
Δ	Jitter interval width
i	Index over Monte Carlo simulated signals
j	Index over jitter intervals
$N(X, j)$	Number of spikes in X in interval j
X_i^{MC}	ith Monte Carlo simulated signal
N_{MC}	Number of Monte Carlo simulated signals
C_i^{MC}	Correlation of X_i^{MC} and Y
R_τ	Number of cases where $C_i^{\mathrm{MC}}(\tau) > C(\tau)$
p_τ	p Value for correlation at lag τ
$\mathrm{JCCG}(\tau)$	Jitter-corrected Cross Correlogram
$P_\tau(C^{\mathrm{MC}})$	Monte Carlo estimate of the distribution of correlations at lag τ
C_j^{int}	Number of coincidences in the jth interval
$P_\tau(C(\tau))$	True distribution of correlations at lag τ
$\tau_{\max}$	Maximum τ value processed
$N_{\max}$	Maximum value of $N(X, j)$ or $N(Y, j)$
$C_{\max}$	Maximum possible number of coincidences

Table 4.1: Glossary. Variables are listed in the order in which they are introduced.

variations, sometimes referred to as rate covariations. This determination is made by comparing a test statistic (in this case cross correlations) of an original pair of spike trains against those computed from a set of jittered spike trains as described below. The jitter method, like cross correlation, operates on binned spike trains which we take as binary signals with values 0 and 1 and integer arguments 0 to $T-1$, where T is the number of bins in the spike train. The binary assumption implies that the bin size is small enough (typically 1ms or so) such that two spikes cannot be recorded in a single time bin. A sufficiently small bin size can always be chosen since there are limits on the minimal inter-spike interval time due to the refractory period of the neurons in question.

Let $X(t)$ and $Y(t)$ be two such binned spike trains. The processing then consists of the following steps:

1. Compute the cross correlation $C(\tau)$ between the original X and Y,

$$C(\tau) = \sum_t X(t-\tau)Y(t)$$

   where the sum runs from 0 to $T-1$ and X is assumed to be 0 if its argument is outside that range.

2. Subdivide one of the signals, say X, into intervals of width Δ.

3. Count the number of spikes in each interval of X. In interval j this is,

$$N(X, j) = \sum_{k=j\Delta}^{(j+1)\Delta-1} X(k) \tag{4.1}$$

4. For X generate N_{MC} Monte Carlo simulated signals $\{X_i^{\mathrm{MC}}\}$, in which the spike counts for each interval are the same as in the corresponding interval in X, such that $N(X_i^{\mathrm{MC}}, j) = N(X, j)$ for all i, j. However, now spike times within the interval are all equally likely. Spike times should be sampled without replacement to ensure that the spike count stays constant without putting multiple spikes in a single bin.

5. Compute the cross correlation $C_i^{\mathrm{MC}}(\tau)$ for lag time τ between each X_i^{MC} and the second spike train Y to get an estimate of the distribution $P_\tau(C^{\mathrm{MC}})$ of cross correlation values for each time lag τ.

$$C_i^{\mathrm{MC}}(\tau) = \sum_t X_i^{\mathrm{MC}}(t - \tau)Y(t)$$

6. Let R_τ be the number of simulations where $C_i^{\mathrm{MC}}(\tau) \geq C(\tau)$. Then the p value for a given lag τ is computed as

$$p_\tau = \frac{R_\tau + 1}{N_{\mathrm{MC}} + 1}$$

7. If desired, a jitter-corrected cross correlogram (JCCG), defined using the expectation operator $E[\cdot]$, can be computed as

$$\mathrm{JCCG}(\tau) = \mathrm{C}(\tau) - \mathrm{E}[\mathrm{C}^{\mathrm{MC}}(\tau)] \approx \mathrm{C}(\tau) - \frac{1}{\mathrm{N}_{\mathrm{MC}}} \sum_{\mathrm{i}} \mathrm{C}_{\mathrm{i}}^{\mathrm{MC}}(\tau) \tag{4.2}$$

where the approximation approaches equality with $N_{MC} \to \infty$.

Jittering the spikes within an interval of size Δ destroys all correlations at time scales within this interval. The cross correlations computed from the jittered spike trains therefore are not correlated on time scales Δ or smaller, and $P_\tau(C^{\mathrm{MC}})$ is the distribution of correlations at time lag τ obtained under the null hypothesis that correlations at time scales $\leq \Delta$ are indistinguishable from random correlations. If the measured cross correlation $C(\tau)$ is significantly outside this distribution, we have to reject the null hypothesis and we conclude that nonrandom correlations at lag τ with time scales $\leq \Delta$ are found in the observed spike trains. If, on the other hand, the observed correlation is consistent with what is seen in the distribution of jittered spike trains, then we cannot reject the null hypothesis. This means we cannot exclude that the observed synchrony is caused by correlations on time scales *outside* the jittered range, in other words that the observed correlation at lag time τ is caused by rate variations on time scales greater than Δ.

In practice, $X(t)$ and $Y(t)$ do not have to be gathered from a continuous block

of time. In the case of multiple trials of the same experimental condition, it may be useful to concatenate the recorded spike trains (possibly after removing sections of them, like those recorded during stimulus onsets). In doing so, a period of no spiking of width $\tau_{\max}$ (the largest correlation lag of interest) should be added between the trials so that correlations between trials don't affect the outcome of the jitter procedure.

As mentioned in Section 4.1, the practical utility of the Monte Carlo method is limited by the trade-off between accuracy and computation time inherent in all Monte Carlo algorithms. Furthermore, in practice a single set of Monte Carlo simulations is often generated for many hypothesis tests (*i.e.* tests at multiple lags), introducing potential dependencies between the different tests when they should be treated independently [2]. In order to avoid both of these issues, the probability distribution $P_\tau(C^{\mathrm{MC}})$ can be computed exactly and independently for each time lag as described in the following.

4.2.2 Closed Form Computation

4.2.2.1 Probability Distribution For One Interval

First, let us consider a single interval consisting of Δ time bins. For example, if spike times have been binned to 2 ms, for an interval of width 20 ms we have $\Delta = 10$.

[2]The procedure of generating one set of spike trains for multiple lags is appropriate inappropriate only if each lag is being tested independently. If the test statistic is the sum of $C(\tau)$ over a range of lag values, a single set of simulated spike trains is appropriate.f

Since time has been discretized, it is still possible to discuss this unitless value as a length of time, a time scale, or a interval width for a given bin size. As before, we assume that the sequence is binary, so each bin has either zero or one spike. This is true even when the spike times are jittered because spike times are sampled without replacement. For this single interval, the probability of a given number of coincidences occurring is determined by three values: Δ, $N(X, j)$ the number of spikes in interval j of spike train X, and $N(Y, j)$ the number of spikes in interval j of spike train Y.

As a first step, we count the number of perfect coincidences, in which one spike occurs in both X and Y within the same time bin, meaning $\tau = 0$. Using the standard notation of $\binom{a}{b}$ for the combinatorial operation (a choose b), we find that there are $\binom{\Delta}{N(X,j)}$ ways to distribute $N(X, j)$ spikes in Δ available bins. The number of empty (spike-less) bins in spike train Y is $[\Delta - N(Y, j)]$. The number of ways to distribute $N(X, j)$ spikes such that each of them falls into one of these empty bins is $\binom{\Delta - N(Y,j)}{N(X,j)}$. These are all possible cases in which a coincidence is avoided. The probability that zero coincidences occur in the j-th interval is therefore

$$P(C_j^{\text{int}} = 0 | \Delta, N(X, j), N(Y, j)) = \frac{\binom{\Delta - N(Y,j)}{N(X,j)}}{\binom{\Delta}{N(X,j)}} \tag{4.3}$$

where C_j^{int} is the number of coincidences in this interval.

We can generalize equation 4.3 to a non-zero number c of coincidences by breaking the numerator up into the number of ways that c spikes can coincide with the spikes

in Y, and $N(X,j) - c$ spikes coincide with the gaps (or non-spikes) in Y. We can thus compute a probability distribution for each interval j,

$$P(C_j^{\text{int}} = c|\Delta, N(X,j), N(Y,j)) = \frac{\binom{\Delta - N(Y,j)}{N(X,j)-c}\binom{N(Y,j)}{c}}{\binom{\Delta}{N(X,j)}} \tag{4.4}$$

where we follow the customary convention of setting the value of a "choose" operation to zero if either of its arguments is negative, or if its upper argument is less than the lower. If this happens in the numerator of eq. 4.4, the probability on the left hand side becomes zero. Of course, the denominator is always positive since $N(X,j) \leq \Delta$. This is a hypergeometric distribution.

Equation 4.4 is easily generalized to nonzero values of τ by applying the analysis leading to it to a shifted version of Y. For the computation of $N(Y,j)$, this implies adding τ to the summation limits in eq. 4.1. As with other correlation algorithms, the boundaries of finite spike trains (beginning and end) result in fewer intervals to analyze as τ gets further away from zero. Thus generalizing eq. 4.4 to non-zero τ, we denote the resulting number of coincidences as $C_j^{\text{int}}(\tau)$ and the associated probability distributions as P_τ^{int}.

4.2.2.2 Jitter-Corrected Cross Correlation

Once we have the analytical probability distribution for the correlations, we can obtain all relevant quantities to characterize the pairwise correlations between two

spike trains. It is straightforward to compute the commonly used jitter-corrected cross correlogram (e.g., Hirabayashi et al., 2013a,b; Martin and von der Heydt, 2015; Smith et al., 2013) which shows the correlation function after all correlations on time scales longer than Δ have been removed. It is defined in analogy to equation 4.2 where the expectation value of the stochastic solution, $E[C^{\mathrm{MC}}(\tau)]$, was used. We can replace this approximation by the exact solution $E[C^{\mathrm{int}}(\tau)]$. Furthermore, by the null hypothesis each interval is conditionally independent based on the spike counts. Therefore, the JCCG can be computed without approximation by

$$\mathrm{JCCG}(\tau) = \mathrm{C}(\tau) - \mathrm{E}\left[\sum_{\mathrm{j}} \mathrm{C}_{\mathrm{j}}^{\mathrm{int}}(\tau)\right] = \mathrm{C}(\tau) - \sum_{\mathrm{j}} \mathrm{E}\left[\mathrm{C}_{\mathrm{j}}^{\mathrm{int}}(\tau)\right] \tag{4.5}$$

which, as should be remembered, is computed for a specific jitter interval width Δ. The expectation on the right can either be calculated for each window as $N(x, j) \times N(y, j)/\Delta$.

The jitter-corrected cross correlogram is used, for instance, when the scientific question of interest is whether there are significant changes in synchrony between conditions, rather than a test of the presence or absence of synchrony. It is then used as part of a bootstrap statistical test in which the observed pairwise correlation is compared with the distribution obtained from eq. 4.5.

4.2.2.3 Probability Distribution For Spike Train

One can also obtain the probability distribution for the entire signal $P_\tau(C(\tau))$ as the convolution of the individual probability distributions for all intervals, P_τ^{int}. This is identical to computing $P_\tau(C^{\text{MC}})$ from Section 4.2.1 with an infinite number of Monte Carlo simulations for each value of τ. One can then evaluate how likely it is that the observed cross correlation $C(\tau)$ is explained by this probability distribution. The likelihood p that this is the case is obtained as the integral of the probability density function exceeding $C(\tau)$, as in

$$p_\tau = \sum_{c=C(\tau)}^{\infty} P_\tau(c) \tag{4.6}$$

4.3 Results

4.3.1 Computational Complexity

In many situations, the statistical distributions underlying the phenomena under study are complicated or unknown and performing Monte Carlo simulations are the only way to make progress, even though it may be costly and it introduces additional randomness in the processing. In the case considered here (binary spike trains, null hypothesis of uniform spike time distribution in fixed interval, cross correlation test statistic), the distribution $P_\tau(C)$ can be computed directly, using the closed form jitter

method described above, without the need for repeated simulations. This section will compare the computational complexity of using the Monte Carlo jitter method against the direct computation using the closed form jitter method.

4.3.1.1 Monte Carlo Method

In the Monte Carlo algorithm, the data generation step requires a permutation of Δ data points for each interval and simulation. Since a single permutation operation has a computational complexity $O(\Delta)$, and Δ times the number of intervals is the length of the signal T, generating the set of Monte Carlo simulations $\{X_i^{\mathrm{MC}}\}$ is $O(N_{\mathrm{MC}} \times T)$. The complexity of cross correlation or convolution of two signals with lengths T is $O(T \times \log T)$, assuming an FFT-based method (Cooley and Tukey, 1965) is used. So computing the full Monte Carlo probability distribution for all values of τ is $O(N_{\mathrm{MC}} \times T \times \log T)$. In many cases, not all values of τ are needed. If the correlation is computed only for the subset of delay values from 0 to $\tau_{\max}$, the complexity for the Monte Carlo jitter method is

$$O(N_{\mathrm{MC}} \times T \times \tau_{\max})$$

Computing the jitter-corrected cross correlogram by this method only requires one additional sum, with complexity $O(N_{\mathrm{MC}} \times \tau_{\max})$ so the total complexity remains unchanged.

4.3.1.2 Closed Form Probability Distribution

To compute the exact probability distribution with the closed form jitter method, note that the values of the distribution can be precomputed based on the maximum values of $N(X, j)$ and $N(Y, j)$ over all j; call this maximum $N_{\max}$. Also, n choose k operations can be as fast as $O(\min(k, n-k))$ (Manolopoulos, 2002). Therefore, a three dimensional table of all possible values of $P(C_j^{\text{int}}|N(X, j), N(Y, j))$ can be precomputed and then looked up for each interval. Generating this table requires up to $N_{\max}$ different values of C, $N_{\max}$ values of $N(X, j)$, and $N_{\max}$ values of $N(Y, j)$. Computing each value requires three n choose k operations, which are on the order of $O(N_{\max})$, so the total computation of the probability table is $O({N_{\max}}^4)$. While the exponent is high, the expression does not have any dependence on the length of the signal and, furthermore, $N_{\max} \leq \Delta$ is a small number in essentially all cases of interest. In practice, for analyzing neurophysiological data it is rare that a time resolution finer than 1 ms is needed, or controlling for cross correlations at time scales larger than approximately 100 ms (*i.e.* $\Delta \approx 100$). The full lookup table is therefore maximally a $100 \times 100 \times 100$ matrix, which requires negligible resources to compute and store.

To compute the combined probability distribution $P_\tau(C)$ over all intervals, all interval probability distributions $P_\tau(C_j^{\text{int}})$ must be convolved, and the computational complexity of the problem is dominated by these convolution operations. As will be shown, we can improve performance by taking advantage of the structure of the

problem at hand, since many of the convolution operations are identical. As a result, the convolutions can be grouped together based on $N(X, j)$ and $N(Y, j)$ and quickly combined so that $O(T)$ convolutions will turn into $O({N_{\max}}^2)$ convolutions. This can be done by the following procedure:

1. Take the Fast Fourier Transform (FFT) of $P_\tau(C_j^{\text{int}}|N(X, j), N(Y, j))$ for each encountered value of $N(X, j)$ and $N(Y, j)$.

2. Raise each complex frequency spectrum value to a power equal to the number of times that the $(N(X, j), N(Y, j))$ pair appears.

3. Multiply these frequency spectra.

4. Take the inverse FFT of the result to get the final probability distribution $P_\tau(C)$ and compute p_τ as in equation 4.6.

5. Repeat steps 2 through 4 for each value of τ to be tested.

The FFT operations in step 1 must be zero-padded up to the maximum number of coincident spikes $C_{\max}$ to account for the highest possible number of synchronous spikes in the combined probability distribution. Therefore the FFT operation in step 1 is $O(C_{\max} \times \log(C_{\max}))$. In step 2, exponentiation is $O(1)$. However there are $O(T \times {N_{\max}}^2)$ exponents to be taken, repeated $\tau_{\max}$ times in step 5. Step 3 requires $O(T \times {N_{\max}}^2)$ multiplications, again repeated $\tau_{\max}$ times. In step 4, the length of the spectral signal (to be inverted by FFT) is $C_{\max}$, so the operation is

$O(C_{\max} \times \log(C_{\max}))$ repeated $\tau_{\max}$ times. When combining these steps, note that $C_{\max}$ is proportional to T. However $C_{\max}$ will be used when relevant because it captures the frequency dependence of the computation time. Therefore the total computational complexity is

$$O(C_{\max} \times \log(C_{\max}) \times \tau_{\max})$$

.

Note that because the zero-frequency component of a probability distribution is always exactly unity, the inverse FFT computation will have accuracy limited by the precision of the numerical system. In practice this implies that p values less than 10^{-13} will not be estimated accurately.

4.3.1.3 Jitter-Corrected Cross Correlation

Both the complexity analysis and the actual computation of the jitter-corrected cross correlogram is much simpler than that of the probability distribution. We generate a lookup table of possible $E\left[C^{\text{int}}\right]$ values and, from equation 4.5, the jitter-corrected correlogram can be computed at a speed of

$$O(T \times \tau_{\max})$$

.

4.3.2 Computational Execution Time

For practical applications, consumption of resources is an important limitation for any computational method. For the size of problems encountered in typical neurophysiological experiments, the only limiting resource is execution time. To compare the performance of the Monte Carlo jitter and the closed form method, the two algorithms were run side by side in the MATLAB environment (Mathworks, Natick MA). Synthetic spike trains were generated that varied in both frequency of spiking (5 to 500 Hz) and length (1 to 91 seconds). For each (time, frequency) condition, 50 spike trains were generated, binned to 1 ms, and the average processing time was computed. Processing was performed with $\tau_{\max} = 100ms$ and $\Delta = 20$. All computations were performed on an Intel i7 920 processor with 12 GB of RAM running Linux Ubuntu 12.04.

For the Monte-Carlo method, N_{MC} was set to 1000. This selection of N_{MC} is unrealistically low for two reasons. First, it can at best result in a Bonferroni corrected p value of 0.201 due to the 201 p values being tested in the range of $-\tau_{\max}$ to $\tau_{\max}$. As the execution for $N_{\mathrm{MC}} = 1000$ already takes 5.7 days to run, increasing N_{MC} is impractical. Second, only a single set of Monte-Carlo trials were generated for all lag values computed, inducing potential correlations between the p values. These correlations should decrease as more trials are generated. Therefore results are extrapolated to $N_{\mathrm{MC}} = 20,000$ (resulting in a minimum $p \approx 0.01$) under the assumption that the processing for 20 times as many simulations would take 20 times as long. Though the

bonferroni correction used here is conservative, it is less conservative (by an order of magnitude) than simulating a whole new set of spike trains for each p value as would be required to entirely eliminate any correlations between the p values.

For the closed form jitter method, all lookup tables were computed *de novo* for each spike train. This is a conservative approach (favoring the Monte Carlo technique) since performance of the closed form jitter method could be improved by computing the tables only once and using them for all spike trains. This is certainly advised in a "production environment."

The results of this simulation, shown in Figure 4.1, illustrate a number of features about the speed of the two algorithms. Plotted is the performance gain, defined as the ratio of the computation time between the Monte Carlo jitter method and the closed form jitter method. The first observation is that the closed form method is substantially faster than the Monte Carlo method in all cases considered. Second, while the performance gain depends only weakly on spike train length, it does decrease with increasing firing rate. This is because the computation time of the closed form jitter method increases with firing rate. In practice, however, it is rare to observe firing at sustained frequencies exceeding 100 Hz in physiological recordings. In the physiological range, the closed form jitter algorithm is faster by a factor of approximately 180 to 7200.

Harrison (2013) uses importance sampling to accelerate the Monte Carlo hypothesis testing process which requires drawing fewer samples. In that work the number

of samples needed, even for a low Bonferroni corrected p value, is reduced to 100. However, each sample is reported to take 18 times as long to generate and process as before, effectively resulting in a simulation about 11 times faster than the Monte Carlo simulation with $N_{\mathrm{MC}} = 20,000$. Therefore, under physiological conditions the closed form computation has an expected speed-up of 16 to 650 times compared to the importance sampling method. It should be noted that in cases where even lower p values are needed because of multiple hypothesis constraints, importance sampling will provide larger gains in estimating very small p values. In these cases, increasing the p value requirements has no effect on the computation speed of the closed form method, so the closed form method can be expected to be faster in all cases.

Another improvement mentioned in Section 4.2.2.2 is the ability of the closed form jitter method to compute the jitter corrected correlogram very rapidly, without computing the null hypothesis distribution of correlation values. To show the magnitude of the improvement, the simulation was repeated with only the mean of the null hypothesis distribution computed under the closed form jitter method since this is all that is needed for the corrected correlation function, eq. 4.5. We also restricted firing frequencies to the range 5–200 Hz. Figure 4.2 shows the ratio of the time it takes to compute equation 4.2 *vs.* equation 4.5. In these cases, the closed form jitter calculation is substantially faster (480x–13,000x), with increasing benefits for increasing spike train lengths. As discussed previously, the spike train length is typically not that of individual trials but of the concatenation of many trials.

4.4 Discussion

The importance, or absence of it, of precise timing of neural spikes has been discussed for the last half-century. Several techniques have been developed to characterize neuronal responses at fine time scales and it is clear that statistical methods have to be developed with much care to avoid wrong conclusions (e.g. Gawne and Richmond, 1993; Roy et al., 2000). One important difficulty is that firing rates can co-vary in the neurons under study. It is well-known that such co-variations are observable in quantities like pairwise cross correlation functions but they are typically considered as irrelevant from the point of view of neuronal coding or of determining the connectivity in the underlying circuitry. For instance, the onset of a stimulus will typically generate a temporary increase in firing rates in sensory cortex but the resulting increase in cross correlation is usually not considered of importance for neural coding (for an exception see **?**, who showed that spike timing relative to onset-related population activity is informative). One common way to subtract such stimulus-locked effects is by subtracting a "shuffle predictor" (Perkel et al., 1967), obtained by computing cross correlations between spike trains from permuted trials. It has been pointed out repeatedly (see references in the Introduction) that this does not eliminate spurious correlations, including close to $\tau = 0$ (synchrony).

Brody (1998, 1999) and Amarasingham et al. (2012) proved that adopting the null hypothesis of independent neurons can not solve the problem. Observation of such correlations is, indeed, evidence against the null hypothesis of independence between

the two observed spike trains. Rejection of this null hypothesis can, however, occur either because spikes in the two spike trains are correlated "one-by-one" (synchrony), or because of slow firing rate covariations common to both spike trains. The fact that this null hypothesis can be rejected does not tell us *why* it is rejected. If the question is whether synchrony exists at less than a given time scale (only), this is the wrong null hypothesis. Instead, the time scale needs to be specified explicitly. The null hypothesis chosen by Amarasingham et al. (2012) is that changes of spike times within a time interval of size Δ have no effect on the computed statistic, in this case the correlation function. It is this null hypothesis that is tested by computer simulation in the Amarasingham et al. (2012) study and analytically in this report.

A key element of the methods discussed here is that the jitter intervals are defined without reference to the original spike trains. This ensures that if the null hypothesis is true, there is no way to distinguish the original spike trains from the Monte Carlo simulated spike trains. This characteristic (called exchangeability) ensures that the obtained p values are from a well formulated hypothesis test. If, on the other hand, the resampling method was changed so that each spike was jittered about it's original spike time, then even under the null hypothesis the original spike train would stand out from the rest because all of its spikes would be at the center of the jitter intervals. Therefore the resulting test would not be a proper statistical test and should be avoided (Amarasingham et al., 2012).

We have discussed two ways one can choose to characterize the correlations be-

tween two spike trains. One is a strict hypothesis testing approach. A null hypothesis is formulated, namely that the observed correlations are indistinguishable from correlations between spike trains whose spikes have been distributed randomly within intervals of length Δ, without changing the number of spikes in each interval. By comparing the observed correlation with those in the distribution generated under the null hypothesis, it is then decided for a given α whether the null hypothesis can be rejected.

The alternative is to compute the time-resolved correlation function and "correct" for the correlations as observed under the null hypothesis, by subtracting the expectation value of the latter. This is the more commonly chosen approach, perhaps because the time-resolved correlation function is both intuitive and familiar. The distribution of JCCG values can be compared between experimental conditions (indicating a change in 'excess synchrony') using a bootstrap test to test for significance. Also, its shape (*e.g.* the location of peaks) may provide insight that goes beyond the yes-no answer whether the null hypothesis can be rejected or not.

In the Amarasingham et al. (2012) study, the Monte Carlo procedure is further developed to account for more potential causes of fine timing effects besides synchrony such as ramping spike rates within an interval or inter-spike interval distribution effects. These methods are straightforward and statistically well-defined. Like any Monte Carlo method, however, they only generate an approximation to the underlying distribution whose quality depends on the number of surrogate spike trains. In

practice what is more problematic is that the method can be computationally very costly. For instance, as discussed in section 4.3.2, our example problem using the simplest of the null hypotheses discussed (50 spike trains of a few seconds long each, mean rates between 1 and 100 Hz, maximal time lag of 100 ms, $\alpha = 0.01$ with Bonferroni correction applied) would have required a simulation several *months* long on a reasonably fast machine. We therefore only simulated $N_{MC} = 1000$ trials and extrapolated to the execution time needed for $N_{MC} = 20,000$ but even that abbreviated Monte Carlo run took nearly six days. Some progress can be made by using much faster machines or many machines (the problem parallelizes easily) but execution time is clearly a problem.

In contrast, the closed form jitter methods this report focuses on are exact, rather than approximate. More important for practical applications may be that they are extremely efficient, with a speed-up of at least two orders of magnitude for the hypothesis testing approach, and four orders of magnitude for the full correlation functions. Even over importance sampling methods (Harrison, 2013), they have been shown to provide a substantial increase in speed. For the hypothesis testing examples used in our study (whose scope is quite comparable to that of typical neurophysiological experiments, assuming a proper Bonferroni correction is applied), computation time is reduced from more than 100 days under the original Monte Carlo method to about one night. Computational time required for the full correlation function is reduced from over 100 days to a few minutes. An increase in performance on this scale is more

than merely a quantitative improvement. For instance, it is essentially impossible to explore variations in the analyses (like the influence of the jitter time scale Δ) if each computational run takes a few months, but it is easy to do if it takes minutes.

So far we were only concerned with correlations between two spike trains. Modern recording techniques are already increasing the number of simultaneously recorded spike trains to tens or hundreds. Unfortunately, the closed-form jitter method is limited in the ability to analyze large ensembles. This is because the correlation functions of some pairs in an ensemble will restrict the possible correlation values of other pairs. For example, if there are three neurons X, Y, and Z, and the pairs XY and YZ have perfect correlation, then the pair XZ must also have perfect correlation. A Monte-Carlo jitter analysis that jitters an entire ensemble of neurons and then performs a hypothesis test on the ensemble can be performed relatively simply, but no such closed-form method exists yet. In order to avoid the constraints of the type described above, only $N-1$ pairs of neurons can be analyzed with closed form methods when N neurons are recorded.

Additionally, the nature of the exact solutions provides an opportunity for further exact analysis. Having a closed form solution allows questions about the effects of spike sorting errors, the value of Δ, or the structure of $JCCG(\tau)$ to be addressed rigorously and more precisely than is possible with any numerical method.

In conclusion, we study a statistical framework for quantifying correlations between spike trains at given time scales. It can be applied both for hypothesis testing

and for correcting observed correlation functions for correlations at these time scales. Results are exact, and both computational complexity and computational time for realistic examples are several orders of magnitude lower than related approaches based on Monte Carlo simulations.

Matlab code is available at `https://github.com/dannyjeck/closed-form-jitter`

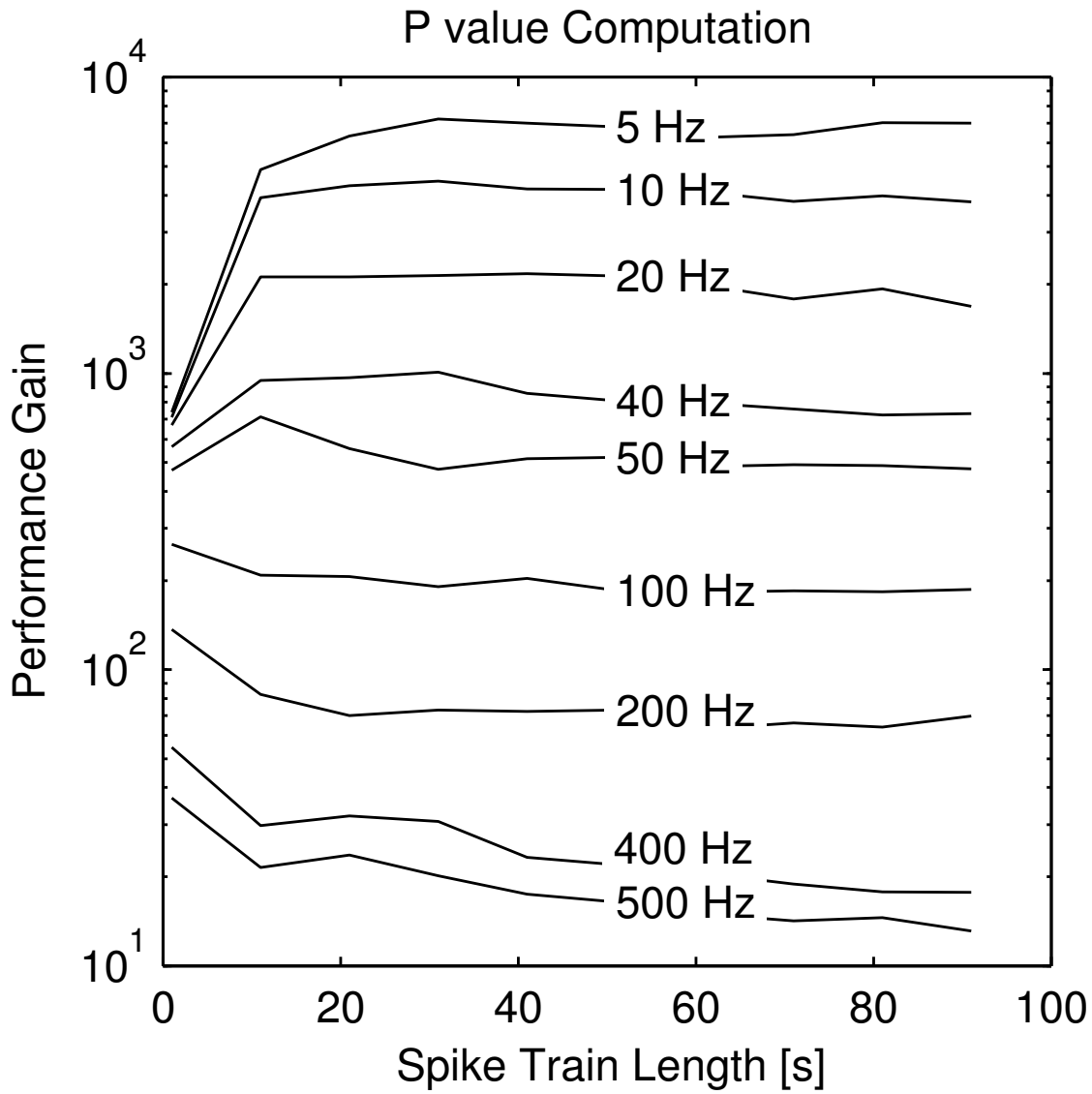

Figure 4.1: Performance gain in implementing the closed form jitter method for p value computations. Gain is defined as the ratio in computation time between the Monte Carlo Jitter method and the closed form jitter method. Processing parameters used are $\tau_{\max} = 100ms$, $\Delta = 20$, and $N_{\mathrm{MC}} = 20,000$ (see text for details).

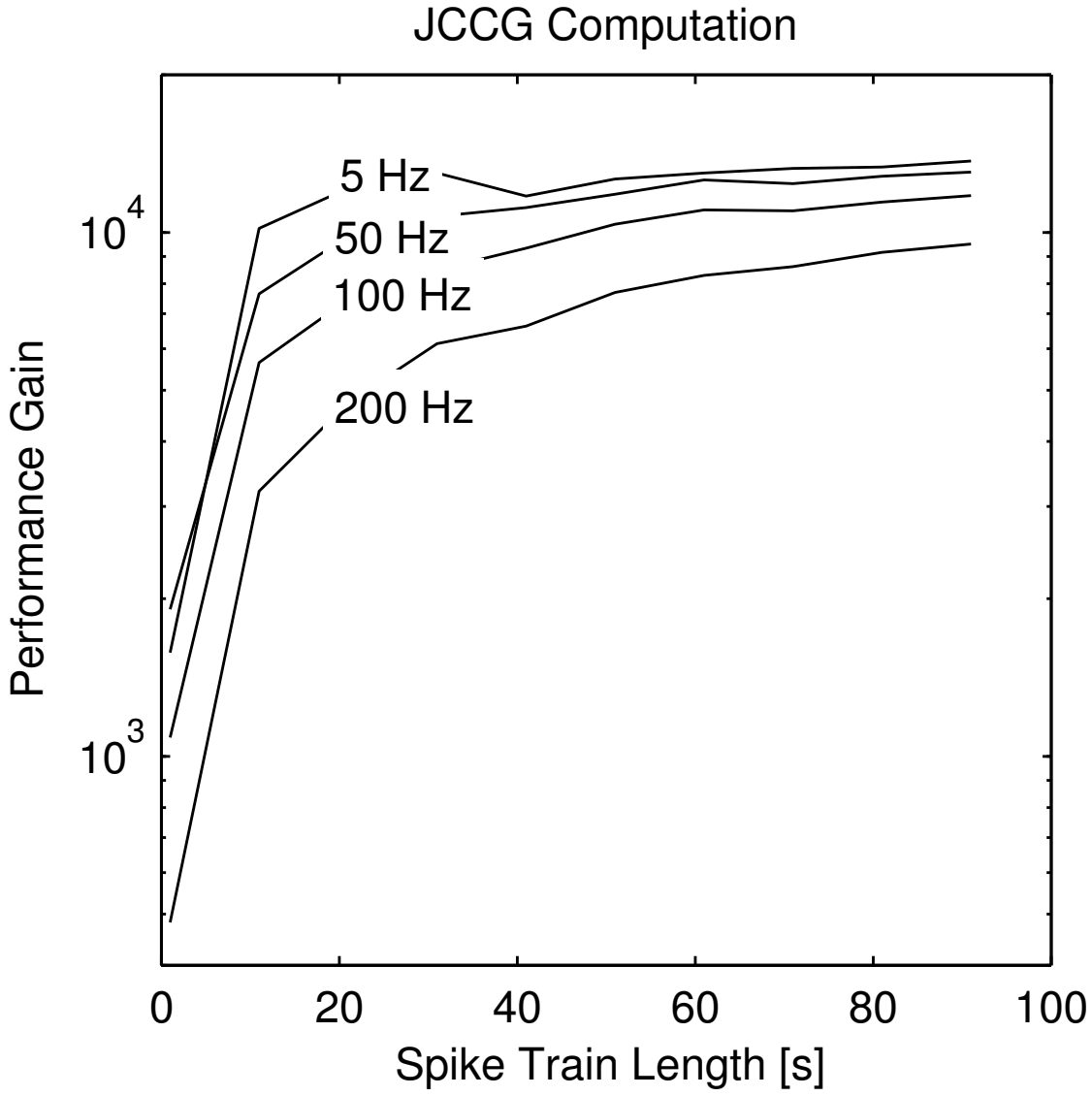

Figure 4.2: Performance gain in implementing the closed form jitter method for Jitter Corrected Correlogram computations. Gain is defined as the ratio in computation time between the Monte Carlo Jitter method and the closed form jitter method. Processing parameters used are $\tau_{\max} = 100ms$, $\Delta = 20$, and $N_{\mathrm{MC}} = 20,000$ (see text for details). The lowered performance gain at signal length of 1 second is due to the overhead of computing the probability table *de novo* for each spike train.

Chapter 5

Neuronal Common Input Strength is Unidentifiable from average firing rates and synchrony

5.1 Introduction

One of the fundamental questions of neuroscience is the nature of neuronal codes. It is well-established Adrian and Zotterman (1926) that in some cases information is represented in the mean firing rates of neurons, averaged over a suitable interval, typically some fraction of a second. There is also a rich literature suggesting that in many other cases, relationships between spikes at a much finer scale are employed for neural coding, at resolutions of milliseconds Abeles (1991); Riehle et al. (1997); Singer

and Gray (1995); Steinmetz et al. (2000) or even less Rokem et al. (2006). Whether fine timing relationships are available for use at least potentially can be determined by observing whether reproducible correlations at the relevant time scales are present.

Independent of any functional role of temporal correlations between spikes, it has been proposed that the relative timing of spikes from different neurons can provide information about the architecture of the circuit they are part of. For instance, if a direct connection exists between two neurons, a correlation between their spikes can be expected with a non-zero time lag which is given by the sum of the propagation delays along axon and dendrite and across the synapse, all of which are strictly positive. If, on the other hand, two neurons receive synapses from another neuron (common input), a peak in the cross correlation function is expected whose time lag is determined by the difference in time at which the common input reaches these neurons; this time difference may be zero. Indirect (multi synaptic) connections can at least in principle be identified, too, from the correlation between spikes in simultaneously recorded neuronal responses.

The first issue to address when analyzing two spike trains is whether the correlations are significant given the firing rates of the neurons. As noted, firing rates are determined by averaging over some period of time, which introduces a time scale (to be distinguished from time lag) to the analysis. One rigorous method to determine the presence of a significant correlation at a particular time scale is the jitter and spike resampling algorithm Amarasingham et al. (2012). However, finding sig-

nificant correlations does not necessarily prove that the correlations are part of the neuronal code for any particular variable. To show that correlations are involved in coding neuronal contents, spike trains must be recorded under two experimental conditions that differ with respect to these contents, and correlation should be found to change significantly between them. For example, to understand coding of attentional modulation, in one condition the subject should be paying attention to a stimulus represented by the recorded neural population, and in another condition attention should be elsewhere Steinmetz et al. (2000).

Determining whether correlation has changed significantly between conditions inherently implies an underlying model. Previous work has been based on highly simplified assumptions, *e.g.* that the rate of detecting synchrony does not vary due to a change in firing rate alone Steinmetz et al. (2000), or that correlations are created when otherwise independent spike trains have synchronous spikes added Amarasingham et al. (2012); Martin and von der Heydt (2015). These models can then rely on simple metrics (the rate of detected synchrony, and the correlation above chance, respectively) to determine whether changes in the spike trains are solely due to changes in firing rates or not. However, for any but the most basic models, determining a change in synchrony independently of a change in firing rate is at best highly complex, and at worst, impossible. Here we will illustrate this point using a very simple network consisting of two leaky integrate and fire (LIF) neurons receiving both independent and common input (Figure 5.1). We will define an increase in synchrony independent

of firing rate as occurring when the fraction of the input arriving from the common input increases. We show that identifying changes in common input current from average firing rates and average jitter-corrected correlation is impossible.

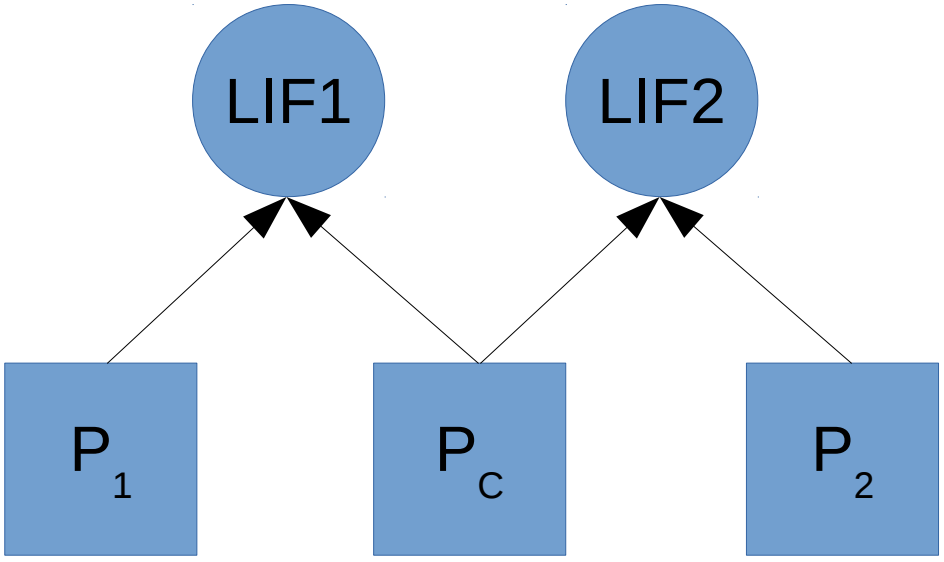

Figure 5.1: Network Structure. Two LIF neurons (LIF1 and LIF2) receive input with Poisson statistics that is the sum of independent (P_1, P_2) and common (P_C) spike trains.

5.2 Methods

The network we use, shown in Figure 5.1, is among the simplest possible that still allow the study of synchrony. It consists of two LIF neurons (LIF1 and LIF2) that receive excitatory spike train inputs modeled as Poisson processes. Each LIF

neuron has one input that only it receives (P_1 and P_2 respectively), and one that is common to both (P_C). They are parameterized by their rates λ_1, λ_2, and λ_C respectively. The inputs are then multiplied by their synaptic weights w_1, w_2, and w_C respectively; for simplicity, the synaptic weights of P_C to both LIF neurons are identical. As a simple model of synaptic dynamics, inputs to each neuron are filtered by applying exponential decay with time constant τ_e, the same for all synapses, and then summed, shown in eq. 5.1 below. To perform simulations of this system, we have to assume numerical values for all parameters. We chose all parameters within physiological ranges; for the synaptic time constant, we use $\tau_e = 2\ ms$. Our results do not, however, depend on the details of these choices. Note that there no interactions between neurons, to keep the network as simple as possible.

The membrane voltage of each LIF neurons is then

$$\frac{dV_i}{dt} = \frac{-V_i}{\tau_m} + I_i(t) + I_C(t) \tag{5.1}$$

where V_i is the membrane voltage of the i-th LIF neuron, τ_m is the membrane time constant (chosen as 20 ms), $I_i(t)$ is the input current (exponentially filtered stochastic Poisson process) from the independent input P_i, and $I_C(t)$ is the input current[1] from the common process P_C. If a neuron's voltage exceeded a threshold of 1, a spike was recorded for that neuron and the voltage of that neuron was reset to 0. This choice

[1]Technically, $I_i(t)$ and $I_C(t)$ are not currents but changes in voltage caused by synaptic current inflows but they are commonly referred to as input currents.

of threshold value allows easy interpretation of the synaptic weights as the fraction of an input required to cause a spike to occur. Simulated trials of the network consisted of numerical integration of eq. 5.1 for 1 second (forward Euler, time step 0.1 ms). Synaptic weights for each input as well as Poisson input rates were varied between simulated trials.

We limit the size of the explored parameter space by restricting expected total input current to each LIF neuron ($w_1\lambda_1 + w_C\lambda_C$, and $w_2\lambda_2 + w_C\lambda_C$) to values such that the neurons fire in a range around 20-30 Hz. Thus, given values for w_1, w_2, and w_C, as well as for the fraction of current coming from the common input $R = w_C\lambda_C/(w_1\lambda_1 + w_C\lambda_C) = w_C\lambda_C/(w_2\lambda_2 + w_C\lambda_C)$, all six input parameters can be determined.

For each input parameter set, the simulation was run $10,000$ times and a sample of the LIF neuron firing rates and the jitter-corrected correlogram (JCCG) value at zero lag was determined for each run (see Amarasingham et al. (2012) for jitter correction). These initial samples were used to generate a distribution of firing rates and JCCG values given the input parameter set. The simulation was repeated and the distribution was updated until the Kullback-Leibler divergence between updates dropped below a threshold of 10^{-6}.

As previous work (*e.g.* ref Steinmetz et al. (2000)) has focused on changes in the means of synchrony and firing rates between experimental conditions, our further analysis will focus on these values, which are found by taking the expected value of the

output distribution for a given parameter set. It should be noted that if the network model was an accurate representation of real neurons (*i.e.* if the model captured all aspects of the neuron that influence the spike counts and JCCG) then it may be possible to use the full distribution of the output measurements to estimate all of the model parameters. However, in designing our model, we deliberately ignored a number of factors that could affect the outputs, including different temporal (non-Poisson) structure of the inputs, slow variations in firing rate, and time-dependent synaptic weights. Future work will focus on how these changes affect the output distribution of a given parameter set, and whether the distribution of measured output values can practically be utilized to estimate network structure.

5.3 Results

The input parameters w_1, w_2, w_C, and R form a four-dimensional space. The simulation can be thought of as a map from this space to a three-dimensional space defined by the three output statistics: the expected values of the two firing rates and of the JCCG at zero lag, or synchrony. Given that the input parameter space is of a higher dimension than the output space, the inverse problem of finding the inputs from the outputs is likely impossible. However, we are only interested in whether the fraction of common input R can be determined from the test statistics. A necessary condition for achieving our goal is that there exists some reverse mapping of the 3-D

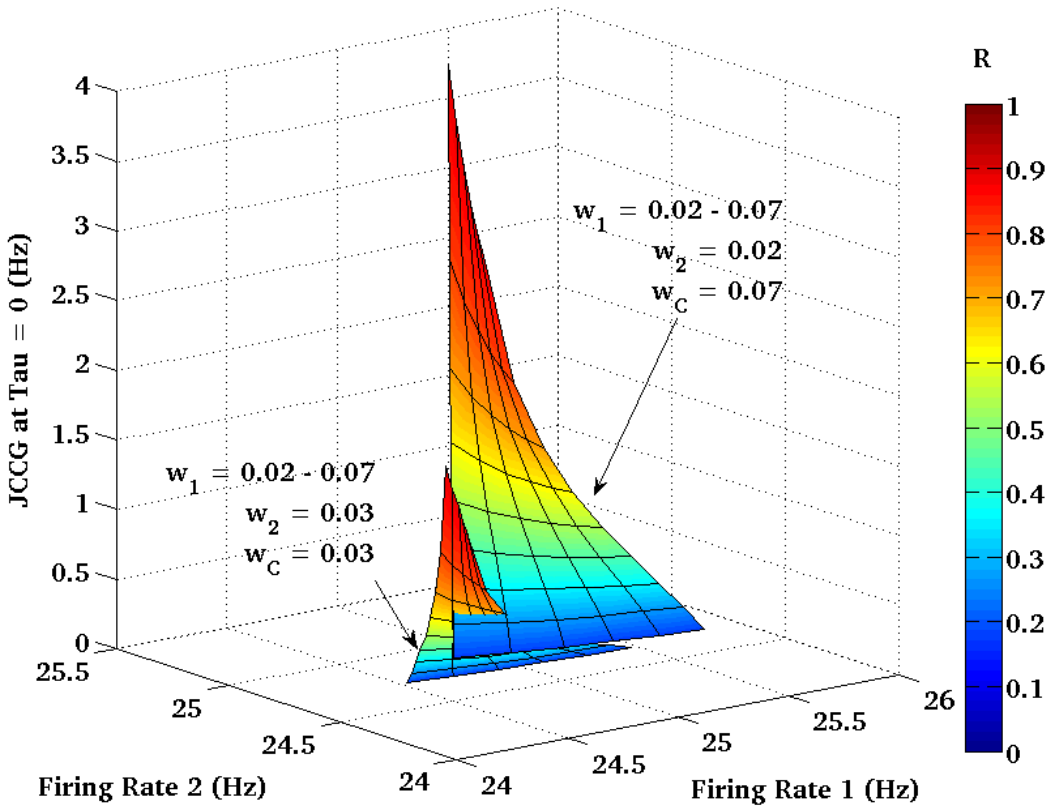

Figure 5.2: Example results. Two slices through the input parameter space are shown to intersect in output statistic space. Color indicates the value of R

output space back onto R. Note that the rest of the input parameters need not be determined.

Figure 5.2 shows how two sets of input parameters are mapped onto the output statistic space. Two of the four parameters are kept fixed on each of the surfaces, ($w_2 = 0.02, w_c = 0.07$ on one and $w_2 = w_c = 0.03$ on the other) and the other two were varied continually, w_1 in the range $0.02 - 0.07$ and R from 0.18 to 0.91. These ranges were chosen to ensure that output firing rates were in the target range, 20-30 Hz, and JCCG values were within a reasonable measurement range. The projections of the two-dimensional (w_1, R) space into the three-dimensional space spanned by the two firing rates and synchrony are shown as the colored surfaces where the coloring represents the value of R along each surface. The figure shows that the mappings intersect, and notably they do so in such a way that there is a color discontinuity on the line of intersection. As color represents the value of R, no reverse mapping can be obtained since multiple values of R are mapped on the same points in the space of measurements. It follows that for this example, R is not identifiable from the mean firing rates and the mean level of synchrony.

Unfortunately, this is not an unusual case. Figure 5.3 shows two views of the output statistic space for a variety of sets of input parameters. In Figure 5.3(a) each surface is generated by varying w_1 and R with a fixed value of w_C and w_2, as in Figure 5.2. The different surfaces correspond to different values of w_C and w_2 which were varied independently in steps of 0.01 from 0.02 to 0.07 (inclusive). The surfaces

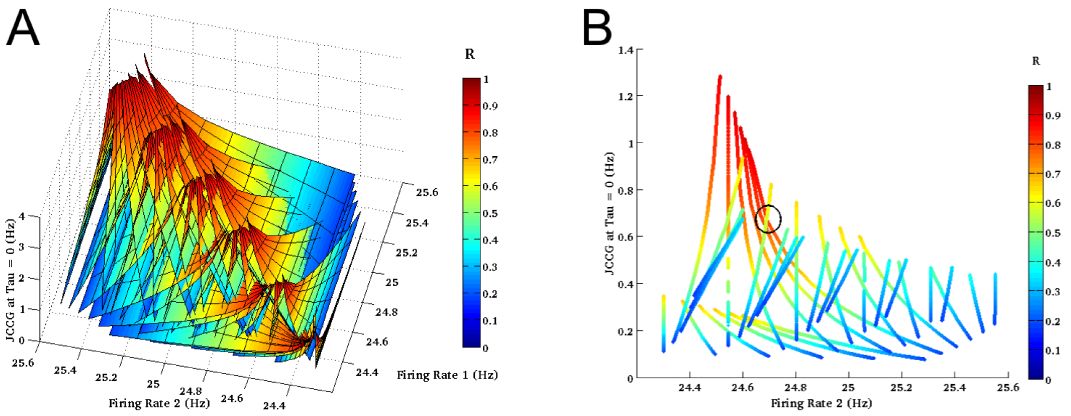

Figure 5.3: Maps of parameters onto output space for additional parameter sets. (a) Each surface is generated by varying w_1 from 0.02 to 0.07 and R from 0.18 to 0.91. Different surfaces are generated by stepping w_C and w_2 from 0.02 to 0.07. R value replicated in color. Existence of intersections between surfaces of different colors indicates that R is not identifiable in the output space. (b) Slice through the surfaces in (a) where the firing fate of LIF1 is 24.6 Hz. Black circle indicates one region where R is not identifiable.

overlap in a number of locations which makes it difficult to see the structure of intersections between surfaces. We therefore show, in Figure 5.3(b), a slice of the surfaces along the plane where the firing rate of LIF1 is 24.6 Hz, thereby reducing the surfaces to lines. Locations where the lines intersect indicate that multiple input parameters map onto the same output values. These intersections often do not correspond to the same value of R (*e.g.* the inside the black circle). This indicates that R is not identifiable using these metrics.

5.4 Conclusion

We have shown that the identification of changes in synchrony independent of firing rate is ambiguous. Even in an extremely simplified case where it is known that each neuron only has two inputs of known structure and that there are no interactions between neurons, observation of mean firing rates and synchrony (JCCG at zero lag) does not determine the level of input from a common source. While this shows that the fraction of common input current is unidentifiable, it may yet be possible to determine some other synchrony transfer function that can be identified from these spike train statistics. However, the "network" that we studied here is extremely simple compared to almost any biological system. Whether it is possible to find a set of functions that makes relevant system parameters identifiable is unknown.

Appendix A

Appendix

A1 Natural Scenes Tapping Experiment

Below is additional information referred to in Chapter 2.

A1.1 Demographics

Detailed demographics are shown in Figure A1. Participants were passers-by on the Johns Hopkins University campus. No deliberate selection criterion was applied, except for (possibly unconscious) perceptions of approachability and whether the individuals seemed in too much haste to be likely willing to participate in the experiment. *Post-hoc* we noticed that gender groups were generally balanced, with the exception of the 23-30 age group in which female participants dominated for unknown reasons.

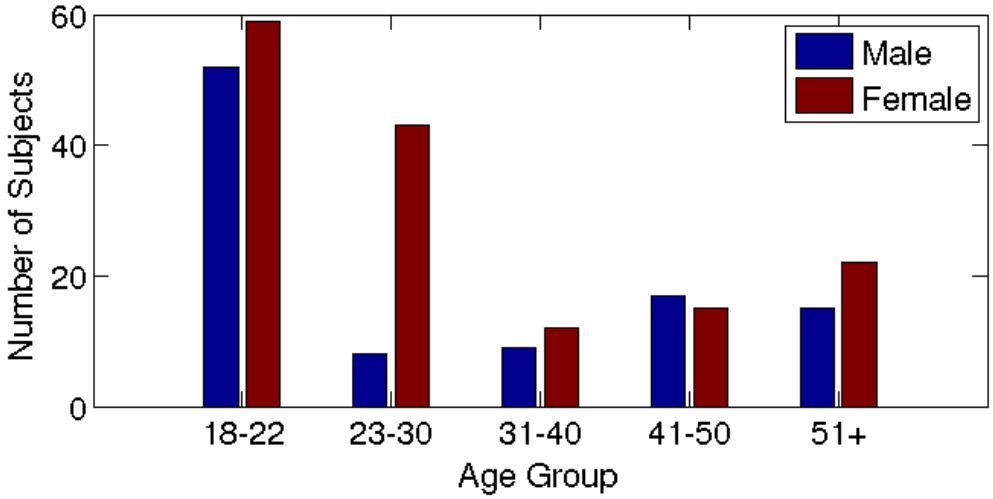

Figure A1: Demographics of the 252 participants.

A1.2 Error modes

The experiment had a number of error modes, as follows.

1. Two taps by one participant were lost when transmitting data from the tablet to the server.

2. Some participants would tap the black square on the right side of the initialization screen but the tablet registered the tap as being on the status bar (not visible to the participant) and did not process it within the experiment. This caused some confusion for some early participants before an image was presented, but no data was lost. Later participants were told to only use the square on the left, away from the status bar.

3. Some participants would accidentally tap either the test image or the initialization screen twice in rapid succession. 15 taps were recorded to take place

within 400 milliseconds of another tap.

4. There was some variability between the loading times of images. Some seemed to consistently load more slowly than others. As mentioned in the main text, we did not analyze reaction times in detail for this reason.

5. One participant seemed to understand the instructions when starting the experiment, but this became doubtful while she performed the experiment. She tapped in a tight group on the right side of the screen. The possible reason was that she was English-challenged, something that was not apparent while she was recruited and instructed.

6. Some participants seemed to consistently take a very long time to complete the task.

Since all these error modes resulted in a very small number of possibly problematic taps, no exclusion criteria were defined before analyzing the data, none was excluded. All participants and taps are included in the analysis of the paper barring the two taps that were not recorded.

A1.3 Statistical validation

To validate our statistical approach we will first repeat our tests using a standard bootstrap technique, and then introduce the motivation and validation of the technique used in the main text.

A canonical bootstrap technique (Efron, 1982) draws samples with replacement from some empirical distribution to generate new samples. This is the way we generate the surrogate maps under the sample error hypothesis. A standard way to gather p-values is to generate surrogate samples under a null hypothesis and compare a measured value to those samples. Consider as an example the sample error hypothesis that $R(\hat{F}, \hat{T})$ is a sample from $R(\hat{F}, \tilde{F}^T)$. Let N be the number of samples from $R(\hat{F}, \tilde{F}^T)$ that are drawn, and n be the number of those samples that satisfy

$$R(\hat{F}, \hat{T}) \geq R(\hat{F}, \tilde{F}^T)$$

We can then generate a valid p-value as

$$p = \frac{n+1}{N+1} \tag{A.1}$$

Here, the $+1$ in the numerator and denominator arise because when hypothesis testing we assume the null is true, and therefore the measured value of $R(\hat{F}, \hat{T})$ is also part of the null hypothesis.

We computed p-values using equation A.1 on the data shown in Figure 3D using 1000 samples drawn from the sample error hypothesis. The measured value of $R(\hat{F}, \hat{T})$ did not exceed any of the 1000 surrogate correlation values. We repeated this analysis for each of the sample error hypotheses shown in Figure 3 and obtained the same result. All p-values are therefore equal to $1/1001$. This includes the case of $R(S, \hat{T})$,

which had a p-value above 0.05 in the main text.

A hypothesis test is considered valid if, when the null hypothesis is true, the rate of getting a p-value below a threshold α is less than or equal to α (Casella and Berger, 2002). This is true if the distribution of p-values under the null hypothesis is uniform, or if the left side of the distribution is lower than a uniform distribution (in which case it is also called a conservative test). To further validate the simple bootstrap test from equation A.1, we generated 1000 p-values when $R(\hat{F}, \hat{T})$ is replaced with a sample from $R(\hat{F}, \tilde{F}^T)$ (*i.e.* assuming that the sample error hypothesis is true) to show that the distribution of p-values is uniform. This is, indeed, the case, as shown in Figure A2A.

While these results confirm the validity of our hypothesis test with the chosen $\alpha = 0.05$, we were curious how confident we can be that our results hold for stricter choices of α. We could choose to generate more samples from the sample error hypothesis, however these are computationally expensive and unreasonably large numbers of samples would be needed to obtain the low p-values we measure. An alternative approach is to use a closed-form approximation of the distribution of interest and then compute the p-values using that approximate distribution. Because the correlations we test are all averages over many images, we chose a Gaussian approximation. The associated hypothesis test is therefore a two-sample Z-test. In order to validate the approximation we must ensure that p-values generated under the null hypothesis are valid. To do so we repeat the processing used to generate Figure A2A, but now we

compute the p-values using the Z-test. The positive slope of the resulting distribution (shown in Figure A2B) indicates that the test is valid, and indeed conservative, with (much) fewer than 50 of the 1000 p-values below the threshold of 0.05 that would be expected under a uniform distribution.

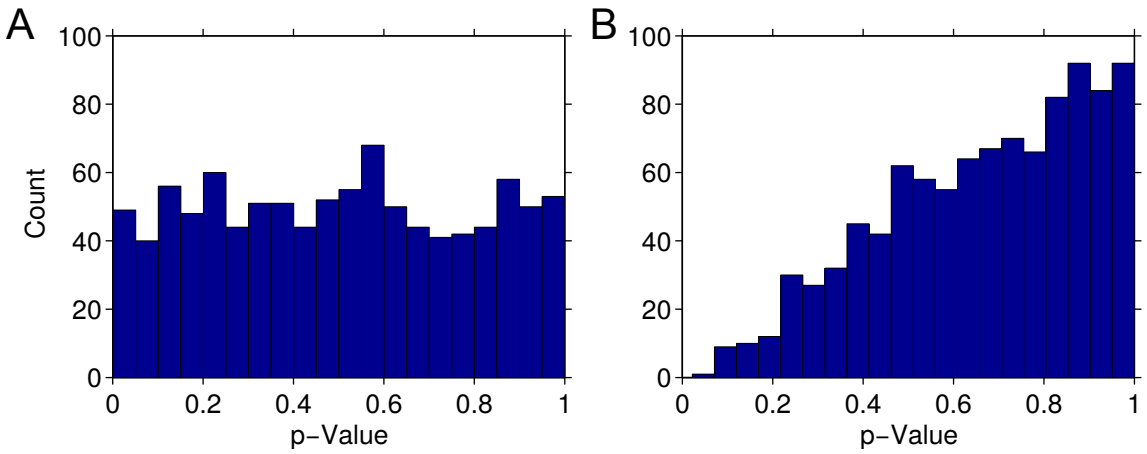

Figure A2: Histograms of 1000 p-values under the null hypothesis (A) under the empirical p-value from equation A.1, and (B) under the Gaussian assumption from the main text.

Bibliography

M. Abeles. *Corticonics – Neural circuits of the cerebral cortex*. Cambridge University Press, 1991.

E. D. Adrian and Y. Zotterman. The impulses produced by sensory nerve endings. part 2. the response of a single end organ. *J. Physiol.*, 61:151–171, 1926.

A. Amarasingham, M. T. Harrison, N. G. Hatsopoulos, and S. Geman. Conditional modeling and the jitter method of spike resampling. *Journal of Neurophysiology*, 107(2):517–531, 2012. PMC3349623.

B.A. Anderson, P.A. Laurent, and S Yantis. Value-driven attentional capture. *Proc. Nat. Acad. Sci., USA*, 2011.

D. Ardila, S. Mihalas, R. von der Heydt, and E. Niebur. Medial axis generation in a model of perceptual organization. In *IEEE CISS-2012 46th Annual Conference on Information Sciences and Systems*, pages 1–4, Princeton University, NJ, 2012. IEEE.

BIBLIOGRAPHY

W. F. Bacon and H. E. Egeth. Overriding stimulus-driven attentional capture. *Perception & Psychophysics*, 55:485–496, 1994.

D. Baldauf and H. Deubel. Visual attention during the preparation of bimanual movements. *Vision Research*, 2008.

B. Bauer, P. Jolicoeur, and C. B. Cowan. Visual search for colour targets that are or are not linearly separable from distractors. *Vision Research*, 36(10):1439–1466, May 1996. ISSN 00426989. doi: 10.1016/0042-6989(95)00207-3. URL `http://linkinghub.elsevier.com/retrieve/pii/0042698995002073`.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B*, 1995. URL `http://www.jstor.org/stable/2346101`.

A. Borji and L. Itti. State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013.

A. Borji and L. Itti. Cat2000: A large scale fixation dataset for boosting saliency research. *CVPR 2015 workshop on "Future of Datasets"*, 2015. arXiv preprint arXiv:1505.03581.

A. Borji, D. N. Sihite, and L. Itti. Objects do not predict fixations better than early saliency: A re-analysis of Einhäuser et al.'s data. *Journal of vision*, 13(10):18, 2013.

BIBLIOGRAPHY

A. S. Bregman. *Auditory Scene Analysis: The Perceptual Organisation of Sound.* MIT Press, Cambrigde, MA, 1990.

C. D. Brody. Slow covariations in neuronal resting potentials can lead to artefactually fast cross-correlations in their spike trains. *J. Neurophysiology*, 80(6):3345–51, December 1998.

Carlos D. Brody. Correlations without synchrony. *Neural Comput*, 11(7):1527–1535, Oct 1999.

G. T. Buswell. *How people look at pictures.* University of Chicago Press Chicago, 1935.

G. Casella and R. L. Berger. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

J. W. Cooley and J. W. Tukey. An algorithm for the machine calculation of complex fourier series. *Mathematics of computation*, 19(90):297–301, 1965.

E. Craft, H. Schütze, E. Niebur, and R. von der Heydt. A neural model of figure-ground organization. *Journal of Neurophysiology*, 97(6):4310–26, 2007. PMID: 17442769.

M. DeAngelus and J. B. Pelz. Top-down control of eye movements: Yarbus revisited. *Visual Cognition*, 17(6-7):790–811, August 2009. ISSN 1350-6285. doi:

10.1080/13506280902793843. URL http://www.tandfonline.com/doi/abs/10.1080/13506280902793843.

H. Deubel and W. X. Schneider. Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision research*, 36(12):1827–1837, 1996.

H. Deubel and W. X. Schneider. Delayed saccades, but not delayed manual aiming movements, require visual attention shifts. *Annals of the New York Academy of Sciences*, 2003.

Y. Dong, R. von der Heydt, and E. Niebur. Synchrony and the binding problem in macaque visual cortex. In *Society for Neuroscience Annual Meeting*, page 734.8. Society for Neuroscience, Washington, DC, 2006.

F. H. Durgin, J. A. Baird, M. Greenburg, R. Russell, K. Shaughnessy, and S. Waymouth. Who is being deceived? The experimental demands of wearing a backpack. *Psychonomic Bulletin & Review*, 16(5):964–969, 2009.

B. Efron. *The jackknife, the bootstrap and other resampling plans*, volume 38. SIAM, 1982.

H. Egeth, J. Jonides, and S Wall. Parallel processing of multielement displays. *Cognitive Psychology*, 3(4):674–698, 1972.

BIBLIOGRAPHY

H. E. Egeth, R. A. Virzi, and H. Garbart. Searching for conjunctively defined targets. *J. Experimental Psychology*, 10(1):32–39, 1984.

R. Egly, J. Driver, and R. Rafal. Shifting visual attention between objects and locations: evidence for normal and parietal lesion subjects. *Journal of Experimental Psychology: General*, 123:161–77, 1994.

W. Einhäuser, M. Spain, and P. Perona. Objects predict fixations better than early saliency. *J. Vision*, 8(14):1–26, 2008.

C. Firestone and B. J. Scholl. "Please tap the shape, anywhere you like": Shape skeletons in human vision revealed by an exceedingly simple measure. *Psychological science*, 25(2):377–86, February 2014. ISSN 1467-9280. doi: 10.1177/0956797613507584. URL `http://www.ncbi.nlm.nih.gov/pubmed/24406395`.

C. Firestone and B. J. Scholl. Cognition does not affect perception: Evaluating the evidence for" top-down" effects. *Behavioral and brain sciences*, 39, 2016.

J. D. Fisk and M. A. Goodale. The organization of eye and limb movements during unrestricted reaching to targets in contralateral and ipsilateral visual space. *Experimental Brain Research*, 1985.

T. J. Gawne and B. J. Richmond. How independent are the messages carried by adjacent inferior temporal cortical neurons? *J. Neurosci.*, in press, 1993.

BIBLIOGRAPHY

M. T. Harrison. Accelerated spike resampling for accurate multiple testing controls. *Neural Computation*, 25(2):418–449, 2013.

T. Hirabayashi, D. Takeuchi, K. Tamura, and Y. Miyashita. Functional microcircuit recruited during retrieval of object association memory in monkey perirhinal cortex. *Neuron*, 77(1):192–203, January 2013a. ISSN 1097-4199. doi: 10.1016/j.neuron.2012.10.031. URL `http://www.ncbi.nlm.nih.gov/pubmed/23312526`.

T. Hirabayashi, D. Takeuchi, K. Tamura, and Y. Miyashita. Microcircuits for hierarchical elaboration of object coding across primate temporal areas. *Science (New York, N.Y.)*, 341(6142):191–5, July 2013b. ISSN 1095-9203. doi: 10.1126/science.1236927. URL `http://www.ncbi.nlm.nih.gov/pubmed/23846902`.

B. Hu, R. von der Heydt, and E. Niebur. A neural model for perceptual organization of 3d surfaces. In *IEEE CISS-2015 49th Annual Conference on Information Sciences and Systems*, pages 1–6, Baltimore, MD, 2015. IEEE Information Theory Society.

B. Hu, R. Kane-Jackson, and E. Niebur. A proto-object based saliency model in three-dimensional space. *Vision Research*, 119:42–49, 2016.

B. Hu, R. von der Heydt, and Niebur E. Proto-object based contour detection and figure-ground segmentation. In *CoSyNe*, 2017.

D.H. Hubel and T.N. Wiesel. Receptive fields and functional architecture of monkey striate cortex. *J. Physiol.*, 195:215–243, 1968.

BIBLIOGRAPHY

N. A Ibraheem, M. M. Hasan, R. Z. Khan, and P. K. Mishra. Understanding color models: a review. *ARPN Journal of science and technology*, 2(3):265–275, 2012.

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

L. Itti and C. Koch. Computational modelling of visual attention. *Nature Neuroscience*, 2:194–203, 2001.

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

W. James. *The Principles of Psychology*. Henry Holt, New York, 1890.

D. Jeck and E. Niebur. Neuronal common input strength is unidentifiable from average firing rates and synchrony. In *49th Annual Conference on Information Sciences and Systems IEEE-CISS-2015*. IEEE Press, 2015a.

D. Jeck and E. Niebur. Closed form jitter methods for neuronal spike train analysis. In *49th Annual Conference on Information Sciences and Systems IEEE-CISS-2015*. IEEE Press, 2015b.

D. Jeck, M. Qin, H. Egeth, and E. Niebur. Attentive pointing in natural scenes correlates with other measures of attention. *Vision Research*, 135, 2017.

BIBLIOGRAPHY

D. Jeck, M. Qin, H. Egeth, and E. Niebur. Unique objects have enhanced salience even when low-contrast. *In Preparation*, 2018.

D. Jonikaitis and H. Deubel. Independent allocation of attention to eye and hand targets in coordinated eye-hand movements. *Psychological science : a journal of the American Psychological Society / APS*, 2011.

T. Judd, F. Durand, and A. Torralba. A benchmark of computational models of saliency to predict human fixations. In *MIT Technical Report*, 2012.

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

R. Kimchi, M. Behrmann, and C. R. Olson. *Perceptual organization in vision: Behavioral and neural perspectives*. Psychology Press, 2003.

R. Kimchi, Y. Yeshurun, and A. Cohen-Savransky. Automatic, stimulus-driven attentional capture by objecthood. *Psychon Bull Rev*, 14(1):166–172, Feb 2007.

S. Kita. *Pointing: Where language, culture, and cognition meet*. Psychology Press, 2003.

C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiol.*, 4:219–227, 1985.

K. Koch, J. McLean, R. Segev, M. A. Freed, M. J. Berry, V. Balasubramanian, and P. Sterling. How Much the Eye Tells the Brain. 2006.

BIBLIOGRAPHY

M. Kümmerer, T. S. A. Wallis, and M. Bethge. Deepgaze II: Reading fixations from deep features trained on object recognition. *arXiv:1610.01563 [cs.CV]*, 2016.

D. A. Leavens, W. D. Hopkins, and K. A. Bard. Understanding the point of chimpanzee pointing epigenesis and ecological validity. *Current Directions in Psychological Science*, 14(4):185–189, 2005.

Y. Manolopoulos. Binomial coefficient computation: recursion or iteration? *ACM SIGCSE Bulletin*, 34(4):65–67, 2002.

A. Martin and R. von der Heydt. Spike synchrony reveals emergence of proto-objects in visual cortex. *The Journal of Neuroscience*, 35(17):6860–6870, 2015.

C. Masciocchi, S. Mihalas, D. Parkhurst, and E. Niebur. Everyone knows what is interesting: Salient locations which should be fixated. *Journal of Vision*, 9(11): 1–22, October 2009.

M. B. McCamy, J. Otero-Millan, L. L. Di Stasi, S. L. Macknik, and S. Martinez-Conde. Highly informative natural scene regions increase microsaccade production during visual scanning. *Journal of neuroscience*, 34(8):2956–2966, 2014.

S. Mihalas, Y. Dong, R. von der Heydt, and E. Niebur. Mechanisms of perceptual organization provide auto-zoom and auto-localization for attention to objects. *Proceedings of the National Academy of Sciences*, 108(18):7583–8, 2011. PMC3088583.

BIBLIOGRAPHY

C. Moore, S. Yantis, and B. Vaughan. Object-based visual selection: evidence from perceptual completion. *Psychological Science*, 9:104–10, 1998.

S. F. W. Neggers and H. Bekkering. Ocular gaze is anchored to the target of an ongoing pointing movement. *Journal of Neurophysiology*, 2000.

E. Niebur and C. Koch. Control of selective visual attention: Modeling the "where" pathway. In D. S Touretzky, M. C. Mozer, and M. E. Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 8, pages 802–808. MIT Press, Cambridge, MA, 1996.

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

E. Niebur, L. Itti, and C. Koch. Controlling the Focus of Visual Selective Attention. In J. L. Van Hemmen, J. D. Cowan, and E. Domany, editors, *Models of Neural Networks IV: Early Vision and Attention*, pages 247–276. Springer Verlag, New York, 2002.

H. C. Nothdurft. Salience from feature contrast: additivity across dimensions. *Vision Research*, 40(10-12):1183–1201, 2000. ISSN 00426989. doi: 10.1016/S0042-6989(00)00031-6. URL `http://linkinghub.elsevier.com/retrieve/pii/S0042698900000316`.

H. C. Nothdurft. Salience-controlled visual search: Are the brightest and the

least bright targets found by different processes? *Visual Cognition*, 13(6):700–732, April 2006. ISSN 1350-6285. doi: 10.1080/13506280544000237. URL http://www.tandfonline.com/doi/abs/10.1080/13506280544000237.

A. Oliva. Gist of the scene. In *Neurobiology of attention*, pages 251–257. 2005. URL http://cvcl.mit.edu/papers/oliva04.pdf.

D. Parkhurst. *Selective attention in natural vision: using computational models to quantify stimulus driven attentional allocation.* PhD thesis, Johns Hopkins University, April 2002.

D. Parkhurst and E. Niebur. Scene content selected by active vision. *Spatial Vision*, 16(2):125–54, 2003.

D. Parkhurst and E. Niebur. Texture contrast attracts overt visual attention in natural scenes. *European J. Neuroscience*, 19(3):783–789, 2004.

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

A. P. Pentland. Perceptual organization and the representation of natural form. *Artificial Intelligence*, 28(293-331), 1986.

F. Perazzi, P. Krahenbuhl, Y. Pritch, and A. Hornung. Saliency filters: Contrast based filtering for salient region detection. In *2012 IEEE Conference on Computer*

Vision and Pattern Recognition, pages 733–740. IEEE, June 2012. ISBN 978-1-4673-1228-8. doi: 10.1109/CVPR.2012.6247743. URL `http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6247743`.

D. H. Perkel, G. L. Gerstein, and G. P. Moore. Neuronal spike trains and stochastic point processes. II: Simultaneous spike trains. *Biophys. J.*, 7:419–440, 1967.

S. Ramenahalli, S. Mihalas, and E. Niebur. Local spectral anisotropy is a valid cue for figure-ground organization in natural scenes. *Vision Research*, 103:116–126, Oct 2014. doi: 10.1016/j.visres.2014.08.012. URL `http://dx.doi.org/10.1016/j.visres.2014.08.012`. NIHMSID 631573.

P. Reinagel and A. M. Zador. Natural scene statistics at the center of gaze. *Network: Computation in Neural Systems*, 10(1-10):4, 1999.

R. A. Rensink. The dynamic representation of scenes. *Visual Cognition*, 7(1/2/3): 17–42, 2000.

A. Riehle, S. Grün, M. Diesmann, and A. Aertsen. Spike synchronization and rate modulation differentially involved in motor cortical function. *Science*, 278:1950–1953, December 1997.

A. Rokem, S. Watzl, T. Gollisch, M. Stemmler, A. V. M Herz, and I. Samengo. Spike-timing precision underlies the coding efficiency of auditory receptor neurons. *Journal of Neurophysiology*, 95(4):2541–2552, 2006.

BIBLIOGRAPHY

A. Roy, P. N. Steinmetz, and E. Niebur. Rate limitations of unitary event analysis. *Neural Computation*, 12(9):2063–2082, 2000.

A. F. Russell, S Mihalas, R. von der Heydt, E. Niebur, and R. Etienne-Cummings. A model of proto-object based saliency. *Vision Research*, 94:1–15, 2014.

B. J. Scholl. Objects and attention: the state of the art. *Cognition*, 80(1-2):1–46, 2001.

M. N. Shadlen and W. T. Newsome. The variable discharge of cortical neurons: implications for connectivity, computation, and information coding. *J. Neurosci.*, 18:3870–3896, 1998.

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556v6*, 2015. URL `https://arxiv.org/abs/1409.1556`.

W. Singer and C. M. Gray. Visual feature integration and the temporal correlation hypothesis. *Annu. Rev. Neurosci.*, 18:555–586, 1995.

M. A. Smith, X. Jia, A. Zandvakili, and A. Kohn. Laminar dependence of neuronal correlations in visual cortex. *Journal of neurophysiology*, 109(4):940–7, March 2013. ISSN 1522-1598. doi: 10.1152/jn.00846.2012. URL `http://www.ncbi.nlm.nih.gov/pubmed/23197461`.

BIBLIOGRAPHY

W. Softky. Simple codes versus efficient codes. *Current Opinion in Neurobiology*, 5: 239–247, 1995.

C. Spearman. The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, 15(1):72–101, 1904. ISSN 0002-9556. doi: 10.2307/1412159. URL `http://www.jstor.org/stable/1412159`.

P. N. Steinmetz, A. Roy, P. Fitzgerald, S. S. Hsiao, K. O. Johnson, and E. Niebur. Attention modulates synchronized neuronal firing in primate somatosensory cortex. *Nature*, 404:187–190, 2000.

T. Sugihara, F. T. Qiu, and R. von der Heydt. Figure-ground organization and attention modulation in neurons of monkey area v2. *J. Vision*, 4, 2004.

M. Tomasello and M. Carpenter. Shared intentionality. *Developmental science*, 10 (1):121–125, 2007.

M. Tomasello, M. Carpenter, and U. Liszkowski. A new look at infant pointing. *Child development*, 78(3):705–722, 2007.

A. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12:97–136, 1980. PMID: 7351125.

A. Treisman and S. Gormican. Feature analysis in early vision: Evidence from search asymetries. *Psychological Review*, 95:15–48, 1988.

BIBLIOGRAPHY

P. H. Tseng, R. Carmi, I. G. M. Cameron, D. P. Munoz, and L. Itti. Quantifying center bias of observers in free viewing of dynamic natural scenes. *J. Vis.*, 9(7): 1–16, 7 2009. ISSN 1534-7362. URL `http://journalofvision.org/9/7/4/`.

D. Walther and C. Koch. Modeling attention to salient proto-objects. *Neural Networks*, 19:1395–1407, Nov 2006.

J. M. Wolfe. Guided search 2.0 – a revised model of visual search. *Psychonomics Bulletin & Review*, 1(2):202–238, 1994.

J.M. Wolfe. Guided search 4.0. *Integrated models of cognitive systems*, pages 99–119, 2007.

J.M. Wolfe, K.R. Cave, and S.L. Franzel. Guided search: an alternative to the feature integration model for visual search. *J. Exp. Psychology*, 15:419–433, 1989.

A.L. Yarbus. *Eye Movements and Vision.* Plenum Press, New York, 1967.

H. Zhou, H. S. Friedman, and R. von der Heydt. Coding of border ownership in monkey visual cortex. *J. Neurosci.*, 20(17):6594–6611, 2000. PMID: 10964965.

Vita

Daniel M. Jeck received the B.S. degree in Biomedical Engineering from North Carolina State University in 2009 and enrolled in the Biomedical Engineering Ph.D. program at Johns Hopkins University in 2012. His research focuses on perceptual organization of visual scenes and the relation of that organization to visual salience and neurophsyiological recordings.

Starting spring 2018 he will work at Matroid, Inc. as a software engineer specializing in machine learning algorithms.