Title
Spatiotemporal Signal Characteristics and Processing During Natural Vision

Author
DuTell, Vasha Guerin

Publication Date
2021

Peer reviewed|Thesis/dissertation

Spatiotemporal Signal Characteristics and Processing During Natural Vision

by

Vasha Guerin DuTell

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Vision Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Bruno A. Olshausen, Chair
Professor Martin S. Banks
Professor Kannan Ramchandran

Fall 2021

Spatiotemporal Signal Characteristics and Processing During Natural Vision

Abstract

Spatiotemporal Signal Characteristics and Processing During Natural Vision

by

Vasha Guerin DuTell

Doctor of Philosophy in Vision Science

University of California, Berkeley

Professor Bruno A. Olshausen, Chair

A current limitation in our understanding of the visual system is its function under natural viewing conditions, especially in the context of dynamic, human behavior. While there have been many advances in understanding the spatial response properties of visual neurons in relation to static stimuli such as natural images, understanding of the corresponding temporal properties has been limited by the lack of high-fidelity datasets that document the properties of the signal that reaches the human retina under natural conditions. In this thesis, I describe the design and construction of a mobile tracking device that can re-create the signal present on the retina of a human as they perform everyday tasks. This device is used to collect high fidelity video and tracking data from human subjects performing a set of tasks that sample the everyday human environment and behavioral repertoire.

This new dataset makes it possible to characterize the spatiotemporal statistics of natural time-varying signals as they occur on the retina. Here I examine the spatio-temporal power spectrum, which is of interest as a natural scene statistic in part because it is the Fourier transform of the autocorrelation function. In the absence of ego motion (movement of head and body), the spatiotemporal power spectrum of the dynamic environment has similar power-law structure to that previously reported for Hollywood movies. Head and eye motion modulate the spatiotemporal signal, boosting mid- and high-range temporal frequencies, such that the visual input on the retina is nearly whitened. This can be beneficial for reducing signal redundancy and maximizing the use of available bandwidth in the optic nerve.

The phase spectrum, which compliments the power spectrum, also carries relevant information about natural image statistics. Despite the strong perceptual signal carried by the classically defined global phase, I show that it has limited utility to differentiate natural images from noise. However, phase congruency, a locally-defined property of the phase, shows marked differences between the distributions of natural images and noise, as well as differences within separate categories of natural scenes.

Finally, I explore the relationship between natural signals and the human visual system by optimizing a neural network to carry the most amount of information through a bottleneck inspired by the human optic nerve, while limiting the energy utilized by neural spikes. I show that a previously proposed model exhibits computational instabilities that hinder the use of autodifferentiation software in training this model, and I offer methods of addressing them. I also show that this model can be reformulated with a restructuring of the network, from a single layer model to an autoencoder framework, avoiding computational instabilities altogether. I conclude with a summary of contributions, as well as a discussion of future areas of exploration.

I dedicate this work to the memory of

Esperanza Beverly Ortiz Dale

May we all be so lucky as to have someone in our life that brings us so much hope.

Contents

List of Figures

List of Tables

Acknowledgments

None of this work would have been possible without the support of those in my life who helped me with the journey leading up to graduate school. Espie and L.A. Dale and the rest of the Dale aunts, uncles, and cousins who supported and and pushed me in the right direction. Rob and Sharon Harrison, who encouraged me to pursue academics by attending college, and supported me through it and beyond. My mother Rosa DuTell, who taught me seize opportunities and push through the tough times, and my father Gary DuTell, who passed on his gusto for life to me. I am grateful to the many teachers, mentors, and collaborators that prepared me professionally and academically to pursue a PhD - Nancy Meyer, Dean Livelybrooks, Mike Wehr, Nima Dinyari, Stacey Wagner, Farrell Ford, Malcom Cook, Ariel Paulson, Kirsten Gotting, Summer Elasady, Ashley Woodfin, Richard Dannebaum, and Ron Yu.

My graduate student work benefited greatly from the faculty that taught, guided, and mentored me throughout my time as a graduate student. First and foremost is my mentor Bruno Olshuasen, who in guiding me through the PhD, has always had my best interests at heart, encouraging me at every step, always reminding me of the bigger picture, but also helping me with the details. I am incredibly grateful to have been his student. I am also grateful for the mentorship of Marty Banks, who invited me to collaborate in his lab, work from which became a major part of my thesis.

I also thank the faculty of the Redwood Center, Vision Science, and EECS departments at Berkeley, who challenged me and gave me helpful feedback on each of my many projects in graduate school, as well as countless additional support: Kannan Ramchandran, Fritz Sommer, Kris Bouchard, Gene Switkes, Jaijeet Roychowdhury, Avideh Zakhor, Marla Feller, Rowland Taylor, Karsten Gronert, Emily Cooper, Hani Farid, John Flannery, Dennis Levi, Austin Roorda, Will Tuten, Jorge Otero-Millan, Susana Cheung, and Chris Wildoset. Also my qualifying exam committee, who challenged me in knowledge of everything related to my project from nerual circuitry to information theory: Michael Silver, Teresa Putassery, Stan Klein, and Mike DeWeese.

I am also deeply grateful to my collaborators on the projects throughout my PhD. First and foremost is Agostino Gibaldi, who's creativity and engineering skills made building a head mounted eye tracker from scratch possible. The collaborators from my internship project, Rachel Brown, Pete Shirley, Bruce Walter, Morgan McGuire, Dave Luebke, and Ruth Rosenholtz. Work at C.Light with Dylan Paiton, Ethan Bensinger, Zach Helft, and Christy Sheehy. Alex Anderson for my rotation project on eye motion modeling. Steven Shepard, for work on sparse coding of optic flow. Sarah Marzen for work on continuous-time reservoir computing.

My work was also supported by fellow students, postdocs, and other members of the Redwood center, those who I collaborated with, and who mentored, taught, worked with, and encouraged me, and who fostered a welcoming, collaborative community: Chris Warner, Yubei Chen, Spencer Kent, Dylan Paiton, Ryan Zarcone, Neha Wadia, Shariq Mobin, Brian Cheung, Mayur Mudigonda, Jesse Livezey, Paxon Frady, Alex Anderson, Pratik Sachdeva,

Andy Vargas, Guy Isley, Mike Fang, Charles Frye, Sophia Sanborn, Charles Garfinkle, Chris Hillar, Connor Bybee, Eric Dodds, Eric Weiss, Gautam Agarwal, James Arnemann, Jasmine Collins, Kata Slama, Louis Kang, Mike Fang, Ryan Moughan, Saeed Saremi, Steven Shepard, Sylvia Madhow, Chris Kymn, Ji Huyn Bak, Christian Shewmake, Alex Belsten, Adrianne Zhong, Denis Kleyko, Ping-chen Huang, Steven Lee, Pentti Kanerva, and Jeff Teeters. Also the students my junior who helped me learn to mentor and teach: Christian Tomani, April Myers, Giulia Focarelli, and Diya Mecheri.

I am also grateful for the members of the Vision Science student and postdoc community whom I learned from, collaborated with, and shared friendship: Emilia Zin, Cécile Fortuny, Kelly Bryne, Avi Aizenman, Tere Cañas-Bajo, Liz Lawler, Sanam Mozfarri, Angie Godinez, Rachel Brown, Kavitha Ratnam, Ethan Bensinger, Christy Sheehy, Lauren Spano, Michele Winter, Bala Yellapragada, Norick Bowers, Billie Beckwith-Cohen, Adaeola Harewood, Elise Harb, Natalie Stepien-Bernabe, Nevin El-Nimiri, Sarah Kochik, Jazzie Junge, Zach Helft, Sahar Yousef, Zeynep Basgoze, Fabio Feroldi, Amanda McLaughlin, John Eric Vanston, Alyssa Braun, and members of the Banks lab, Steven Choliweck, Vivek Labhishetty, and Josselin Gautier.

I am incredibly grateful to my husband Dar, who encouraged me to apply to graduate school, and supported me both scientifically and personally throughout it, listening to every presentation and proofreading every paper. My sister Alee DuTell, for inspiring me with her strength and always just getting it. My mother-in-law Karen who has always shown up for me. Alice Dale, Frank Evans, and Nic Evans-Dale for helping me survive and thrive during a pandemic and internship. Michiko Curtis, for inspiring me with her fearlessness. My fur babies Adra and Tysa DuTahlen, for the unconditional love. My Oregon friends - Emily Griffith, Whitney Donileson, Ryan Moore, Carly Wright, and Elyse Martin who always make the time, inspire me with their own successes, and help me keep perspective.

I am also immensely grateful to those that carried me through the last 3 months of graduate school, who checked in on me and helped me keep going. Neha Wadia, for constant friendship, and always having the best advice. Ally Beohem and Yves Lamson for family dinners and friendship. Sanam Mozfarri, for all the runs, bike rides and walks. Angie Godinez for challenging me to see my inner strength. Liz Lawler for writing sessions together. Chrissy Bloome, for constantly encouraging me. Lisa Camasi for keeping me grounded, and feeding me delicious food. Rob and Sharon Harrison for positivity, encouragement, and love. Mike Fang and Mayur Mudigonda, for helping me with the final pushes to finish my thesis.

Finally, I am thankful for the sources of funding that supported both myself and my laboratory and computational equipment: The United States taxpayers through the Air Force Research Office and NDSEG fellowship, the Center for Innovation in Vision and Optics, and the UC Berkeley Student Technology Fund.

Chapter 1

Introduction

1.1 What are Natural Scene Statistics?

The study of 'natural scene statistics' refers to the description of both the regularities and modes of variability present in the natural world around us. Typically, as an area of study spanning Computer Science, Vision Science, and Neuroscience, these properties are studied in the context of images and/or videos that were recorded or otherwise collected from the natural world.

Interest in studying natural signals first emerged in Vision Science and Neuroscience as an alternative to the common practice of measuring neural and psychophyical responses to the commonly used artificial stimlui such as bars, gratings, and Gabors. While these artificial stimuli offer ease of experimental control and repeatability, because they are often relatively low-dimensional and simplistic as compared to complex natural signals, they are less likely to elicit the rich responses seen when a subject or organism interacts in its natural environment. Measured responses in this context of these simple, relatively impoverished stimuli, will likely cover a limited regime of responses, and will therefor limit our understanding of the system. By forgoing these manufactured inputs in favor of stimuli more reminiscent of the ecologically relevant environment of the organism, one can observe a more full range of behavior and neural responses, and better understand the utility of neural organization and response patterns.

When utilizing natural signals as stimuli, how can we address this loss of control and parameterization? Solving this issue is one of goals of natural scene statistics: to describe the statistical regularities and modes of variability within natural signals. Armed with descriptors for the differences between categories of natural signals, as well as the ability to modify natural signals, allowing them to deviate away from their natural properties in a controlled manner, one can study the relationship of the visual system to its input in a more fruitful manner.

1.2 Properties of Natural Scenes

The statistical regularities in natural scenes are thought to emerge from physical properties of the natural world. For example, the tendency of matter to combine together and form objects, when coupled with sufficient density of those objects in 3D space, form occlusions that emerge from layering basic shapes in depth; this is known as a 'dead leaves model' [66], and has been shown to reproduce some of the commonly studied statistical regularities in natural images such as the 'power law', explained in further detail below.

Another regularity present in natural images relates to the horizon. Due to the force of gravity pulling matter towards the ground in our 3D environment, there is often a horizontal boundary present (the horizon), especially in images where far viewpoints are present and the camera is pointed parallel to the ground (typical of human gaze positions). When quantities such as the 2D power spectrum are measured over groups of images, this causes a statistical bias towards higher power in vertical spatial frequencies [96, 98].

Beyond these logically explainable regularities of our human environment, the scene statistic known as 'critical scaling' appears to emerge from the fractal properties present in a whole host of natural phenomena [69], including images of nature, art [81], and even ecology [18]. In the context of images, critical scaling refers to the property that at any scale, from a photo of a small leaf, to a large aerial landscape, many statistical properties are conserved, when images are analyzed in aggregate [87]. This fractal nature has strong ties to metrics such as the slope of the 2D power spectrum[103, 11], which we will soon discuss in detail.

Finally, the vast majority of efforts in natural scene statistics have involved properties of the power spectrum. But as we will see in chapter 4, the counterpart of the power spectrum, that is the phase information present in images, appears to be more perceptually relevant than the power spectrum, containing edge information that forms objects and shapes. Some observed regularities in natural scenes are more closely tied to phase. As an example, the distribution of activations for a set of log-scaled Gabor filter banks, when convolved with natural images, takes on a kurtotic (sparse) distribution [39, 77]. We explore various phase-related statistics and their relationship to natural images in chapter 4.

1.3 Theoretical Neuroscience

The human visual system evolved in the natural world, processing the visual signal from the natural environment. This visual signal is what natural scene statistics aim to describe. It follows, then, that the visual system has likely adapted to the visual properties present in that natural environment, and that taking this ecologically-relevant visual environment into account lends insight when aiming to understand visual processing [44, 9]. One of the most early descriptions of this relationship was in 1981, with Laughlin's study of the natural contrast distribution and its match to the neural response properties of interneruons of the

fly compound eye [65]. Since then, in addition to further work on contrast [16], more complex examples parallels have been drawn between the visual system and its visual environment.

One framework for understanding that adaptation is in terms of efficiency, in that the visual system may be adapted to most efficiently encode the incoming visual signal, which can be quantitatively measured using tools from information theory such as entropy and mutual information [8] [4]. This 'efficient coding', or 'redundancy reduction' framework has informed many areas of investigation within Neuroscience, Vision Science, and Computer Science, both experimentally and theoretically.

From a theoretical neuroscience standpoint, utilizing natural image patches when training neuron models, while optimizing these models for efficiency, has been shown to reproduce a wide variety of neuron properties, including receptive field characteristics of visual neurons [77, 56]. In the experimental literature, countless experiments have documented the marked changes in the response properties of real visual neurons when stimulated with natural images and movies, as opposed to synthetic stimuli such as white noise, and there is evidence supporting the theory that the neural responses to natural stimuli are preferable by information theoretic measures [23].

Considering and utilizing naturalistic stimuli in both experimental and theoretical neuroscience pursuits however, introduces challenges that have deterred many in these communities from pursuing their use. Primarily, natural scenes are much more difficult to parameterize than more simple, synthetic stimuli such as Gabor wavelets. That is, that we lack the ability to describe, in a principled, neurally-relevant way, what varies between different instances of natural images and movies; this is due in large part to a lack of statistical descriptors sufficient to describe natural images and movies. This lack of sufficient parameterization presents challenges to creating controlled vision science and neuroscience experiments. Through improved descriptions of the statistical regularities and variability within natural signals, we can increase the ability of such experiments to utilize natural images and videos.

1.4 Datasets

One difficulty in studying natural scene statistics is the broad, unspecific definition of what constitutes a natural scene. For example, though the human visual system evolved in an environment without modern structures such as buildings and roads, the ecologically relevant environment for most modern humans contains man-made structures. This dilemma is reflected in the wide variety of databases available. Very early natural image databases such as the classic Van Hateren Natural Image Database [49], contain greyscale images of natural outdoor content such as wooded areas, ground and plant matter. Since Van Hateren's dataset, new natural scenes databases have emerged with higher-resolution color content, and sample a variety of environments including man-made content. Note that these datasets differ from popular databases such as ImageNet [25], as natural image databases aim to avoid biases such as being cropped to a single object, but aim to represent a snapshot in time of a human observer. In fact, some of these databases such as the Places dataset address the

broad definition of 'natural scene' by including a hierarchical categorization system to label image content both broadly such as man-made/nature and indoor/outdoor, and well as more specifically such as 'bedroom' and 'garden' [108]. Not only are these various image categories known to vary in their power spectra, as we will see in chapter 4 that these categories also show different phase profiles.

As interest has grown in extending beyond the spatial domain of images, and into the temporal domain with natural video databases, this ambiguity regarding what content is 'natural' continues to persist. Like with natural images, Van Hateren's natural video dataset, created by sampling video content from nature documentaries, remains a classic [50]. Since then, databases have much improved spatial and temporal resolution, as well as color channels[54, 93, 80]. For natural video, the large amount of time that many now spend on screens [17] arguably makes screen-based content an ecologically-relevant visual stimulus for modern humans. We know that statistical differences between screen-based and real-world visual environments exist [104], and that they may be relevant for many aspects of human perception [13, 91]. Many video databases such as YouTube-8m now include categorical descriptors [3].

When considering the relationship between dynamic natural scenes and the human visual system, it becomes important to account for all the motion types that are present on the signal that reaches the retina. Many databases include videos recorded from a 'first person view', commonly with a GoPro or other action camera strapped on a person's head, and often as they perform extreme-sport activities such as BMX, parkour, or skydiving [3]. While these videos introduce body and head motion, the datasets have strong biases towards action-filled moments, which are distinct from the motion profile seen by most humans in their day-to-day life. In addition, head-strapped cameras, in the absence of paired eye-tracking, cannot re-create the eye motion imparted on the signal by the constant movement of the human eyes. A variety of efforts have been made to collect such datasets, but as discussed in 2, most have been limited in either the types of tasks being performed, and/or in their spatial and temporal resolution.

1.5 Spatiotemporal Power Spectrum

One of the most basic and widely studied regularity of natural scenes is the shape of the spatial power spectrum, which is described by the '(inverse) power law', or 'one over f squared' relationship. This is described in detail in chapter 3. This regularity, first described by David Field in studies of natural images [39], describes the property that the 2D power spectrum of natural scenes, when averaged over many images, falls approximately inversely proportional to the square of the spatial frequency. As the square root of the power, the amplitude spectrum falls proportional to the spatial frequency. While this 1D property is typically measured by collapsing the 2D image space onto a single frequency plane, other investigations have studied the complex, 2D shape of the power spectrum, and found it to vary depending on the scene category [96] [98] [7]. Mathematically, the power spectrum can be derived from the auto-correlation function through the convolution theorem, and the

power law interpreted to describe the property of natural images that a given pixel has a high correlation with those closest to it, with this correlation decreasing at a given rate with increasing pixel distance.

While the spatial power spectrum of natural images is a critical regularity of our natural world, spatial regularities alone can only describe a single snapshot in time; these statistics ignore the temporal regularities present in the dynamic signals of a moving and changing environment the human visual system evolved in. Previous work measured the joint spatial-temporal power spectrum of Hollywood movies [29], showing an inseparable relationship between spatial and temporal frequencies and power. Like the spatial power spectrum, it is thought that the human visual system, thorough evolution, has adapted to these regularities in order to most efficiently process the incoming visual signal, including using spatial and temporal whitening [30, 86, 85], an information theoretic process that increases efficiency of information processing.

These power spectrum analyses on Hollywood movies however, were done in the 1990's where spatial resolution and framerate, as well as computational power were limited. And while statistics of Hollywood movies are well poised to inform video compression, their content strongly differs from the dynamic scenes experienced by the human visual system. Furthermore, these videos do not incorporate the complex ego-motion from the body, head, and eyes, that introduce additional motion into the signal processed by the human retina and brain.

In chapter 2, we discuss the design and build of a head mounted eye tracking system designed to capture, in high resolution, the dynamic visual input for a human observer, along with their accompanying body, head, and eye motion. This device is the first of its kind to capture such data in high-fidelity, incorporating specialized hardware for eye and body tracking, as well as color and depth of the scene. This allows for the high-fidelity reconstruction of the dynamic signal that reaches the retina. It is also designed as a mobile device, such that this data can be collected outside of the laboratory, and even outdoors, as a human subject navigates the natural environment sitting, standing, or walking.

Using this device, as discussed in chapter 3, we have collected a dataset of natural video and matched ego motion from human subjects performing a set of 15 everyday tasks such as reading a book, making a sandwich, and throwing a ball, as well as standing and walking in both indoor and outdoor environments.

Using this dataset, we then repeat the spatiotemporal power spectrum analysis of Dong and Atick [29], with our high-fidelity video dataset, and compare the power spectrum for videos with no ego motion, head and body motion only, and finally incorporating eye motion to reconstruct as closely as possible the visual signal that falls on the retina for the human subject. As we shall see in chapter 3, head and body motion strongly modulate the signal, boosting a portion of the temporal frequency spectrum. Eye motion further modulates the signal, resulting in a temporally whitened visual signal.

1.6 Phase

Despite the importance of the power spectrum in natural scene statistics, an arguably more perceptually relevant measure of natural images and video is in their phase spectrum. The phase spectrum is an additional quantity produced when the Fourier Transform method is used to calculate the power spectrum of a natural image or video [45], but is typically unused. While the power spectrum describes the relative amount of different sine/cosine waves (frequencies) present in a signal, the phase spectrum describes their relative positions. We refer to this phase spectrum as the 'global phase'.

While both power/amplitude and phase information are necessary for complete reconstruction of an image or video using the inverse Fourier Transform, phase information appears to carry more perceptually relevant information. That is, as shown in chapter 4, an image created from the amplitude spectrum of one image, and the phase spectrum of another, will resemble the structure of the phase's corresponding image, even maintaining recognizable objects. Given this, it stands to reason that regularities in the phase spectrum may have even more importance to understanding the relationship of the human visual system to the natural visual input than the amplitude/power spectrum.

1.7 Implications

In addition to the basic science questions addressing the relationship between the physics of our natural world, and the relationship of our human visual system to that world, the impact of an improved understanding of natural scene statistics has wide reaching potential in many areas of applied science and medicine. For example, one of the main impediments to using retinal implants to restore vision in the millions of visually impaired, is our limited understanding of the coding strategies utilized by the visual neurons these implants would replace, limiting our ability to properly relay visual signals to downstream brain regions. Despite advances in implantable hardware [20], only through better understanding of the spike-based communication visual neurons use, a communication strategy that evolved to process the natural visual world, will we be able to fully restore vision to these patients [36].

Another challenge is facing the increasing bandwidth requirements of streaming video content on the internet [38]. In developing regions of the world where internet access is slow and/or data-metered, huge bandwidth requirements make such content inaccessible [95]. Stored in raw formats, the size of video content quickly becomes unwieldy. The accessibility of video content at high resolutions such as 4K, and the standardization of 'lower' resolutions such as 1080p, has been enabled by compression algorithms such as AVC (H.264) [83], and more recently HEVC (H.265) [97]. These compression algorithms take advantage of basic statistical regularities such as difference coding [45] that are common in natural video in order to save streaming/storage space. More sophisticated descriptions of the statistical regularities present in the natural world will allow for improved compression algorithms, that are more closely adapted to the properties of their video signals, and can therefore encode

higher resolution video content in fewer bits. Given humans move their eyes when viewing such content, an understanding of the dynamics of the signal as seen by the visual system will be helpful in understanding what type of perceptually-based compression mechanisms may be effective. In addition, an understanding of how signal regularities vary between different types of content (nature, sports, cartoons, etc), would allow for adaptive compression algorithms which can tailor themselves to most efficiently encode the given type of video signal.

Also related to video coding, the world's increasing dependence on the internet as a source of information and news has brought about a crisis of misinformation, with ever-improving 'deep-fakes' and other artificially generated/manipulated videos that impersonate world leaders and claim to serve as 'proof' of events that never happened, quite literally threatening democracy [21]. One avenue of addressing these issues is through improved understanding of the regularities in natural videos, and better computational models of how the human visual system processes them. These will allow us to better compress video data, by identifying what visual information is not captured by the human visual system, and need not be transmitted, as well as what regularities are not obeyed in fraudulent videos that we as humans cannot perceive, improving verification of the validity of visual media[107].

Contributions

Here, I present a list of scientific contributions during my PhD. A portion of these works are presented in this thesis.

- DuTell, V., Gibaldi, A., Focarelli, G., Olshausen, B., and Banks, M. (2020). The Spatiotemporal Power Spectrum of Natural Human Vision. Journal of Vision, 20(11), 1661-1661 (Chapter X) (Abstract and Poster at Vision Science Society)
- DuTell, V., Gibaldi, A., Focarelli, G., Olshausen, B., & Banks, M. S. (2021, May). Integrating High Fidelity Eye, Head and World Tracking in a Wearable Device. In ACM Symposium on Eye Tracking Research and Applications (pp. 1-4).
- Gibaldi, A., DuTell, V., & Banks, M. S. (2021, May). Solving Parallax Error for 3D Eye Tracking. In ACM Symposium on Eye Tracking Research and Applications (pp. 1-4).
- Brown, R., DuTell, V., Walter, B., Rosenholtz, R., Shirley, P., McGuire, M., & Luebke, D. (2021). Efficient Dataflow Modeling of Peripheral Encoding in the Human Visual System. arXiv preprint arXiv:2107.11505.
- A. Vlasits; R. D. MORRIE; A. TRAN-VANMINH; A. BLECKERT; C. F. GAINER; V. DUTELL; D. A.DIGREGORIO; M. B. FELLER.Synaptic input distribution plays a role in the dendritic computation of motion direction in the retina. Society for Neuroscience, 2016

Chapter 2

Integrating High Fidelity Eye, Head and World Tracking in a Wearable Device

2.1 Abstract

We describe the design and performance of a high-fidelity wearable head-, body-, and eye-tracking system that offers significant improvement over previous such devices. This device's sensors include a binocular eye tracker, an RGB-D scene camera, a high frame-rate scene camera, and two visual odometry sensors, for a total of 10 cameras, which we synchronize and record from with a data rate of over 700MB/s. The sensors are operated by a mini-PC optimized for fast data collection, and powered by a small battery pack. The device records a subject's eye, head, and body positions, simultaneously with RGB and depth data from the subject's visual environment, measured with high spatial and temporal resolution. The headset weighs only 1.4kg, and the backpack with batteries 3.9kg. The device can be comfortably worn by the subject, allowing a high degree of mobility. Together, this system overcomes many limitations of previous such systems, allowing high-fidelity characterization of the dynamics of natural vision.

2.2 Introduction

The visual system evolved and developed in the natural environment, so obtaining a full understanding of its function requires studying how vision is engaged in everyday tasks. For this reason, there is a great need to expand vision science beyond the controlled laboratory setting and into the natural world. Data collected in such natural conditions provide crucial information about mechanisms underlying stereopsis [68, 94, 40, 41], eye movements [40] and their coordination with head movements [60, 51, 64], eye optics [43], and other motor behaviors [73, 14, 15]. To create a better account of natural sensory-motor relationships,

Figure 2.1: Left to right: Sample frames collected from Ximea camera, RealSense RGB stream, RealSense depth stream, and Pupil Labs binocular eye-tracking cameras. Images shown are frames as captured by each sensor, before post-processing.

data must be collected along with eye tracking, depth, and motion information when the subject is performing everyday tasks in real world. Furthermore, many applications, such as measurement of the power spectrum [31], require data to be recorded with high spatial and temporal resolution. Designing and building a device that fits these requirements presents many serious technical challenges. We first review previous work and then describe our device.

Early work in mobile eye tracking was restricted to the indoor laboratory environment: for instance using hard-wired acquisition computers and coil-based eye tracking [46]. Later work pioneered the collection of real-world scene and gaze-tracking data, adapting eye-tracking hardware designed for use in the laboratory into devices that allowed mobile recording outside the lab [55, 33, 68, 106, 94, 40]. Unfortunately, cameras in these devices had very limited spatial and temporal resolution, and heavy and bulky eye-tracking hardware limited subject mobility.

More recent efforts utilized compact hardware that is amenable to mobile data collection outside the lab; see [22] for a recent review. In particular, the introduction of lightweight, mobile-friendly eye trackers such as Pupil Labs tracker [57] and Tobii glasses [1], as well as lightweight sensors such as Intel RealSense devices [58], has led to more work in this area [73, 90, 92]. In addition, improved usability of collection software has allowed collection of hundreds of hours of data for many subjects [99, 90]. However, these datasets offer only low to medium temporal resolution and medium to high spatial resolution because of the limited capabilities of the scene cameras. An exception is the high-resolution data reported by [35]; but this is for subjects navigating virtual environments. Many of them also employ cameras with on-device H.264/H.265 encoding, which introduces compression artifacts into the data.

We present a solution to these issues with a wearable device optimized to obtain robust, high-fidelity, multi-modal data, while remaining lightweight and portable enough to enable data collection during everyday behavior in the natural environment. Our solution adapts consumer electronics and laboratory hardware to the needs of mobile, head-mounted tracking. The hardware is combined with custom software that enables accurate, high-resolution data acquisition and post-processing in a convenient interface.

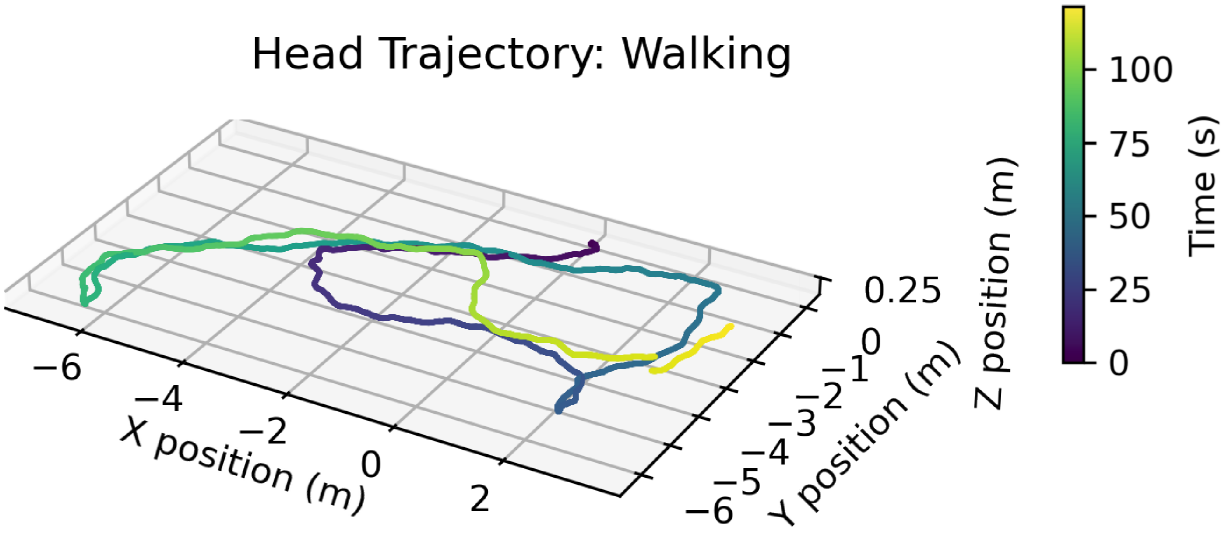


Figure 2.2: Example trajectory of head position as a person walks through an indoor environment. Color evolves over time from purple to yellow over 2 minutes. A RealSense T265 tracking sensor collects position data (such as this one) along with orientation and velocity data at 200Hz. Another tracker on the head provides the same odometry information for the head.

2.3 Hardware

Devices and Sensors To record information from the subject and scene, our device uses six sensors (Table 2.1). To capture high-fidelity video, we use a Ximea PCIE RGB camera with a global shutter running at 200Hz. We supplement the color video with corresponding depth information by including an Intel RealSense D435i, which records both depth and RGB video streams (Figure 2.9). Our device allows us to match the high-fidelity world-camera data to a lower-resolution depth signal. It also allows us, by coordinating with the eye tracker, to estimate the subject's fixation point in the three-dimensional scene. To track the eyes, we use the Pupil Labs binocular eye-tracker [57]. To track the subject's head and body motion, we use two Intel RealSense T265 tracking sensors [51] (Figure 2.2). One is mounted on the subject's back to measure body position and motion. The second is mounted on the head, attached rigidly to the headband, to measure head position and motion.

At full resolution and framerate, the total data flow produced is substantial at ∼700MB/s. The Ximea camera contributes more than 90% of this. The mini-PC, with 3TB on-board M.2 storage, allows just over an hour of recording time at the highest framerate.

Because our device pushes the framerate limits of the sensors, one challenge was minimizing dropped frames, especially in the visual sensors. The combination of image resolution and framerate settings reported in Table 2.1 maximizes spatial and temporal resolution without causing a significant number of dropped frames. With this configuration, frame loss is

Table 2.1: Device sensors and settings utilized by the system. These settings yield the best overall results for our experimental setup, but resolution and frame-rate settings for the RGB-D and eye tracking cameras can be easily modified in the GUI. The Ximea camera's spatial and temporal resolution is easily changed in a YAML file, and the field of view adjusted with a lens change.

Device	Resolution	Field of View	Model	Location	Data Format
High-Fidelity RGB Camera	2064 × 1544 @ 200Hz Global Shutter	61° × 46°	Ximea MX031CGSY-X2G2-FL	Head	8-bit CMYK Raw Binary
RGB-D Camera	640 × 480 @ 60Hz (color) 848 × 480 @ 90Hz (depth)	64° × 41° 86° × 57°	RealSense D435i	Head	MPEG-4 NumPy/PNG
Binocular Eye Tracker	192×192 @ 200Hz	37° × 37°	Pupil Labs	L/R Eye	MPEG-4
Odometry 1	200Hz	-	RealSense T265	Head	.pldata
Odometry 2	200Hz	-	RealSense T265	Body	.pldata

less than two frames over 2 minutes of data collection with the Ximea and RealSense RGB cameras. The depth stream typically varies in its effective framerate between 70–90Hz. We handle the frame drops that do occur with up-sampling during post-processing.

Device Ergonomics We had two key goals in designing the head-mounted part of the device (Fig. 2.3): (1) to be as lightweight and comfortable as possible, and (2) to be adjustable to accommodate each participant's head and face shape, and the task at hand. The headband is modified from a binocular indirect ophthalmoscope and adapted to hold the sensors. Custom components were designed in SolidWorks and 3D printed in PLA, making them robust yet lightweight. The three scene cameras (Ximea, RealSense D435i and T265) are mounted together on the same 3D-printed bracket. This is connected to the headband via three-point 3D-printed adjustable ball-and-socket joints and is secured by clamps. This arrangement enables adjustment of the pitch of the camera ensemble depending on the task. For tasks involving far viewing (e.g., outdoor walking), pitch can be adjusted upward to $\sim 0°$, and for tasks involving near viewing (e.g., cooking) pitch can be adjusted to $\sim 30°$ downward; mid-range tasks (e.g., seated chatting) are recorded using a mid-angle pitch. The Ximea camera's switch box is strapped to the back of the headband (Fig. 2.4). This switch box converts the PCIE connection from the computer to the ribbon-cable connection on the camera.

The two eye-tracking cameras are connected to the headband with custom designed and 3D-printed spherical joints (Fig. 2.3), which allow convenient, stable positioning of the

Figure 2.3: Subject wearing the device. The two scene cameras (Ximea and RGB-D) and tracker (IMU 1) are mounted together with a custom 3D-printed mount, adjustable in position with a 3-point ball-and-clamp adjustable mount. Custom ball-and-socket joints combined with set screws enable positioning of eye trackers below the eyes. A white ribbon cable connects the Ximea camera to the rear switch box.

Table 2.2: Details of operating computer used to control sensors and store acquired data.

Device	Model	Form Factor	Size	Notable Specs
Motherboard	Asus ROG Strix Z390-I	Mini ITX	-	Dual M.2, Wifi, PCIE
Hard Drives	2x Samsung 970 Evo	M.2	1TB, 2TB	Write 1.2 GB/sec
Memory	Crucial Ballistix Sport LT	DDR4 RAM	2x16GB	3200 MHz
CPU	Intel i7-8700	-	-	6 Cores, 65 Watts
Batteries	BPS Freedom CPAP	2x bricks (7.5"x5"x1")	2x 100Wh	12V/8A out

camera. We anticipated degradation of eye tracking in outdoor scenes due to intense scene illumination. To deal with this, a neutral-density filter can be placed in front of the lenses when recording outdoors [12]. The filter makes pupil detection more difficult, but this can be addressed when running pupil detection offline during post-processing.

The power and data cables connecting the sensors and computer are bound together into a clean band (Fig. 2.4); we loop this band behind the subject's back with slack in the loop. The binding and slack eliminates tangling while allowing the subject to move freely. We secure the body tracker on the back using a posture-correcting strap, which is underneath the backpack, but leaves the back tracker's cameras exposed. This avoids occluding the subject's and camera's views of the scene ahead, which would have occurred with front mounting.

The head mount weighs only 1.4kg. In the future we will investigate whether the device affects natural motion dynamics.

Operating Computer To collect data from all sensors simultaneously, we built a PC using consumer parts (Table 2.2). The high-speed camera requires a x8 PCIE port for which no laptop solutions were available, so we custom-built the computer. To minimize the form factor, we use a Mini ITX motherboard with 32GB of RAM, dual M.2 support, a PCIE port, and integrated WiFi. We use the Intel i7-8700 processor, which has sufficient computational power, yet maximizes battery life due to its low power consumption (65W). To maintain sufficient disk-write speed and avoid RAM overflow, we use M.2 SSDs—one with 1Tb and one with 2Tb—capable of writing at 1.2GB/s. We mounted a touchscreen inside the PC case for quick viewing and control of the computer while mobile. Power is provided by a pair of compact batteries designed to power CPAP machines. The batteries are connected in parallel and power both the computer's DC power supply and the PCIE camera's external power supply. We modified a standard mini ITX computer case with a custom 3D-printed enclosure. The enclosure covers the ports at the back of the computer case, exposing only

Figure 2.4: Rear view of a subject wearing the device with the backpack in mobile configuration. Computer and batteries are housed in the backpack. Ximea switch box is mounted with Velcro on the back of the head. Cords are bound in an adjustable loop enabling head mobility. Another motion tracker is mounted on back strap under backpack with the sensor positioned just above backpack.

Figure 2.5: Hardware control during data collection is performed in Pupil Lab's Pupil Capture with custom plugins running on the acquisition computer. The computer is controlled remotely with Chrome Remote Desktop over WiFi. Settings are adjusted and acquisition started and stopped by the experimenter using an iPad or laptop.

the ports for DC power, an external monitor, and Ethernet, leaving the band of sensor cables permanently connected. One CPU heatsink/fan is sufficient for cooling the computer. To cool the high-speed camera, we attached two 25mm fans to either side of the camera, powered by the camera switch box.

A video overview of the device hardware is available at: `https://www.youtube.com/playlist?list=PLEloutX3oXFbi2CoA3_koqFSwKpdxLliF`

2.4 Acquisition Software

Software Structure We wrote the device acquisition software in Python 3 [101] as plugins for Pupil Labs' Pupil Capture software [57] allowing for control of all the devices in a single graphical interface (Fig. 2.5). We use the RGB sensor on the Intel D435i as the world camera, and a plugin to the Pupil Capture software to save depth information as either raw *NumPy* values [48] or lossless PNG images with multiple images per file rather than the default lossy MPEG-4 encoding. Our software includes a plugin to align the RealSense depth

and RGB streams online. This online alignment reduces the highest achievable framerate, so we perform spatial alignment of frames during post-processing instead. We also wrote a plugin to view and record from the Ximea camera as well as load and apply camera settings from a YAML file. For the odometry sensors, we use the tracker code from [52], modified slightly to support recording from both tracking devices and the Intel RGB/depth device simultaneously.

During data collection, we use the Pupil Labs' Capture software, modified by our plugins, to observe and control the computer, switch between visual stream views, run eye-tracking calibration, adjust camera framerate and gain, and start and stop collection. When the subject's task does not involve locomotion, we control the computer and observe the video stream using an external monitor and Bluetooth keyboard and mouse with the computer placed on a table next to the subject. During tasks involving locomotion, we control the acquisition computer through Remote Desktop over WiFi with a laptop or iPad (Fig. 2.5). For eye tracking, we use the default Pupil Capture eye-camera recording software, which records infrared video of each eye at 200Hz. We turn off Pupil Capture's online pupil detection and accomplish detection offline with the Pupil Player software after data collection is completed. This reduces the computational load on the acquisition computer, and allows manual adjustment of the pupil-detection parameters, which in turn minimizes the number of frames with failed pupil detection.

To accommodate various lighting conditions, we include an analog (sensor) gain adjustment switch for the Ximea camera in our GUI, which can be used in combination with aperture adjustment for the varying light levels in indoor and outdoor data collection. This adjustment along with imaging a standard color checker chart [37] allows the experimenter to account for the system's luminance gain and perform color balancing.

We incorporate various software scripts related to eye-tracking calibration. The high-fidelity raw image data (particularly from the Ximea camera) is very storage intensive. To deal with this, we include a framerate adjustment switch for the Ximea camera in our GUI. The adjustment allows us to reduce framerate during calibration, which saves storage space significantly. We also use a custom Pupil Capture plugin to visualize a 9-point marker placed within the world camera's field of view (Figure 2.7) together with a custom 3D calibration routine adapted from [42].

High-Speed Acquisition The most significant design challenge for this system was acquiring and writing the high-speed RGB data from the Ximea camera, particularly in accommodating the high rate of data input (637MB/s for this sensor alone). To interface with and control the camera, configure settings, and collect data, we use Xiapi, Ximea's Python API. We utilize Python's threading and queue packages to create data-collection worker threads that continuously check for and collect images and their associated timestamps from the camera's buffer, placing them in FIFO queues. These queues are simultaneously checked by data-saving worker threads, which write queued frames and timestamps to disk. We save frames in the raw binary format from the camera (1000 images per file) for offline conversion

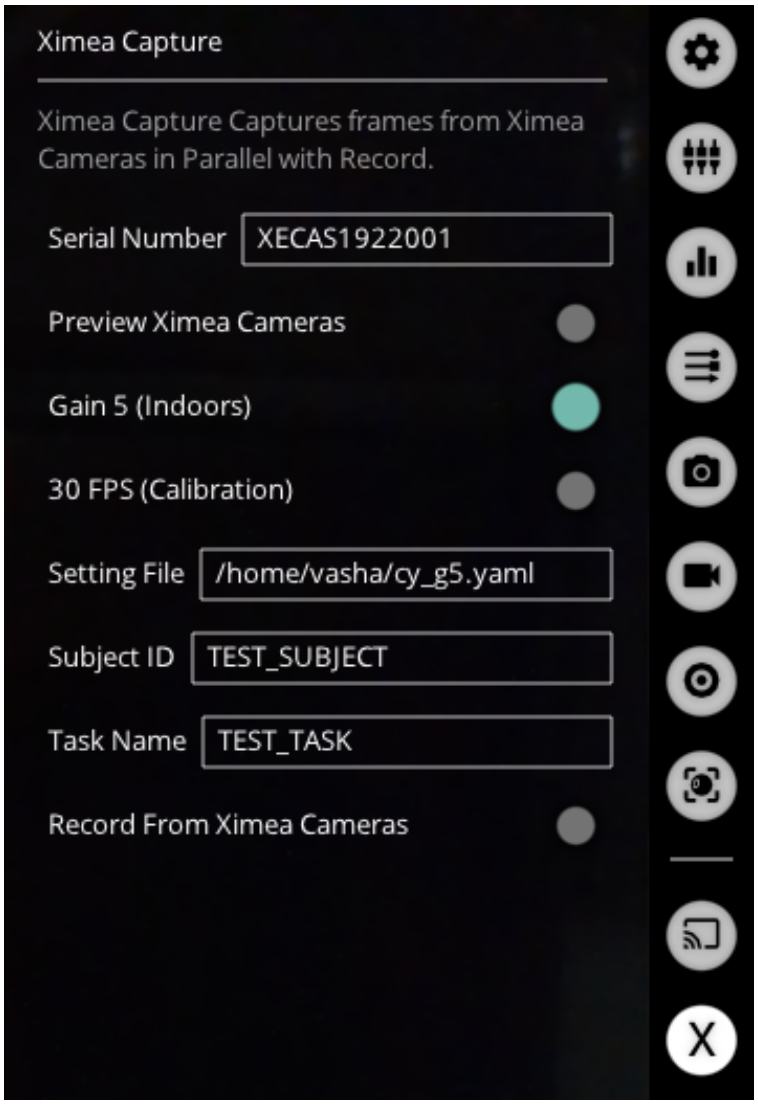


Figure 2.6: Custom plugin for recording from the high-speed Ximea camera has a GUI interface built as a plugin for the Pupil Capture software allowing control of camera settings and recording by the experimenter.

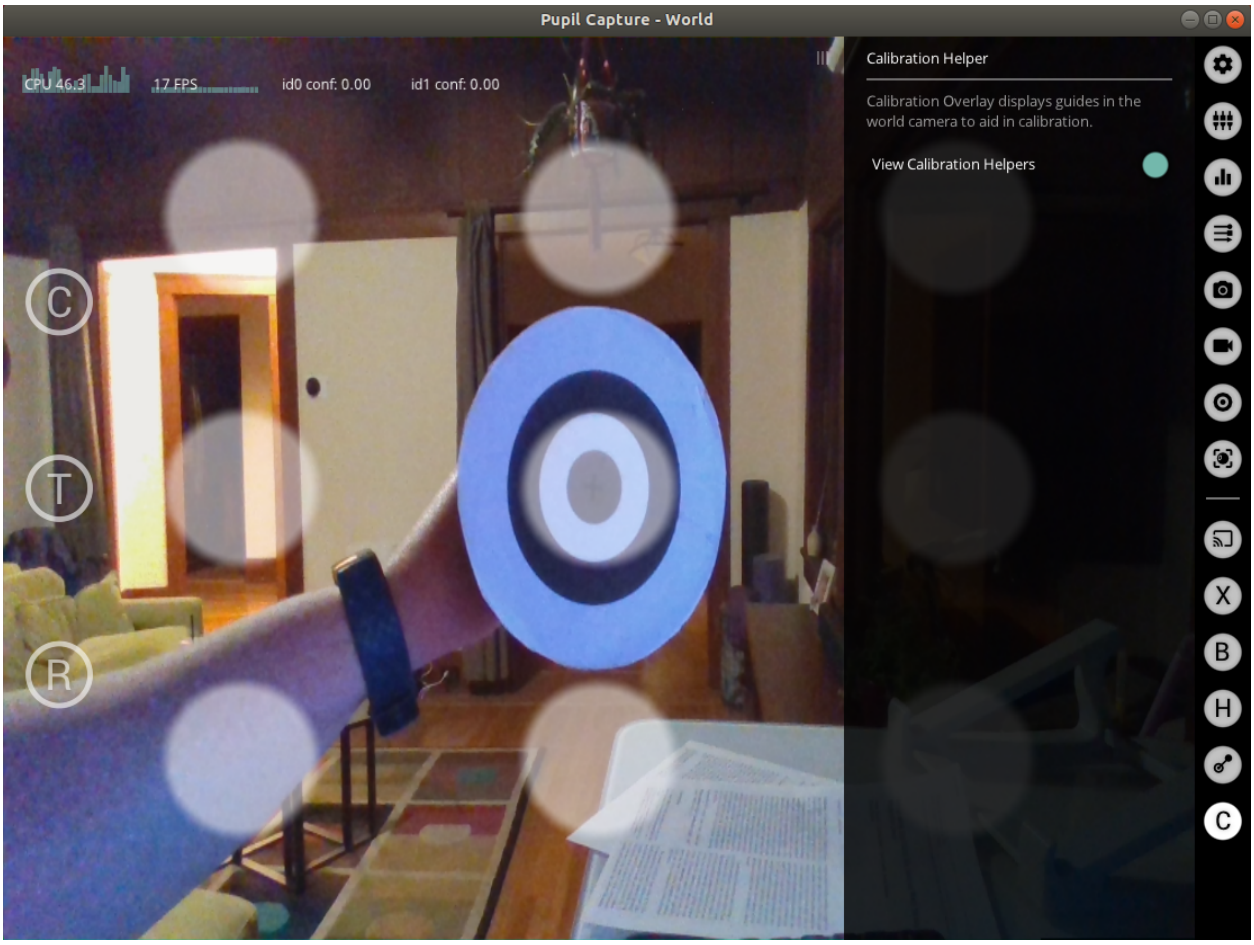


Figure 2.7: The custom Pupil Labs plugin toggles 9-point calibration positions overlaid on the video stream used for directing the subject to position a handheld calibration target for the calibration procedure.

to a standard image format. We tried other acquisition methods and they failed because either the camera's internal buffer was overwritten due to buffer overflow or because of a buildup of frames in the computer's RAM due to insufficient transfer of frames from RAM to disk. We use a similar multi-threaded queuing strategy for saving depth frames (also stored in raw format) to stabilize the effective framerate for the depth stream, and to avoid dropped frames.

The data collection software plugins are available at: `https://github.com/vdutell/hmet_aquisition`. The analysis software is available at: `https://github.com/vdutell/st-bravo_analysis`.

2.5 Post Processing

After each recording session, pupil detection is performed offline inside the Pupil Player software. Then, the data are transferred to a computational server via the exposed Ethernet port for post-processing. During this offline phase, timestamp synchronization and image registration are performed to align the streams spatially and temporally. This alignment allows the depth-map and gaze-position information to be overlaid in the high-framerate camera space. Finally, we perform the remainder of the eye-tracking analysis pipeline including calibration and gaze-point estimation.

Temporal Synchronization To temporally align data from the multiple streams, we first align the timestamps of all streams. Many multi-sensor devices address temporal synchronization issues with a synchronized triggering system so that timestamps are already aligned during data collection. This method is not supported in Pupil Labs, so to maximize each device's frame-rate, we instead allow individual sensors to 'free-run' at their specified framerates during data collection, and then synchronize their timestamps in post-processing. For the Ximea camera, we measure clock offsets between the sensor's internal clock and the computer's Unix timestamp at the beginning and end of recording. We then align the recorded timestamps to ensure there is minimal temporal drift between the two clocks during recording. For the other devices, Pupil Labs' software handles timestamp synchronization internally with Unix timestamps directly. We investigated the accuracy of the synchronization and found that the match between cameras is within one 200Hz frame (±5ms) (Figure 2.8) with typically fewer than one dropped frame over a 2 minutes of data collection. In post-processing, a ground-truth timeline at the desired frame-rate is generated, and frames from each stream resampled at their nearest matching timestamp. This addresses any dropped frames and allows for resampling lower framerate streams at higher frequencies as needed.

Spatial Registration For spatial registration of the images, a standard offline camera and stereo calibration is combined with depth-dependent alignment. This is done twice, once for an 'indoor' setting with an open aperture on the Ximea camera and once for an 'outdoor' setting with a smaller aperture. For each aperture setting, we first use a checkerboard grid to estimate the distortion matrix for the Ximea camera. The RealSense RGB distortion matrix is factory calibrated and the image is undistorted on the chip. Then, we use the same checkerboard grid to run a stereo calibration, fixing the distortion matrices and estimating the extrinsics matrix between the RealSense RGB and Ximea RGB streams. Because the rectification of the depth stream into the RGB frame of reference is depth dependent, we use the Pyrealsense2 align_to() method to rectify the depth stream to Ximea RGB space in two steps: 1) storing the frames in .bag file format and 2) reading in the .bag file for alignment. In the first step, we provide the RealSense camera's self-reported depth to RGB extrinsics to the alignment method, rectifying the depth frames into the RealSense RGB camera's frame of reference. This puts depth information into the RealSense RGB camera's frame of reference

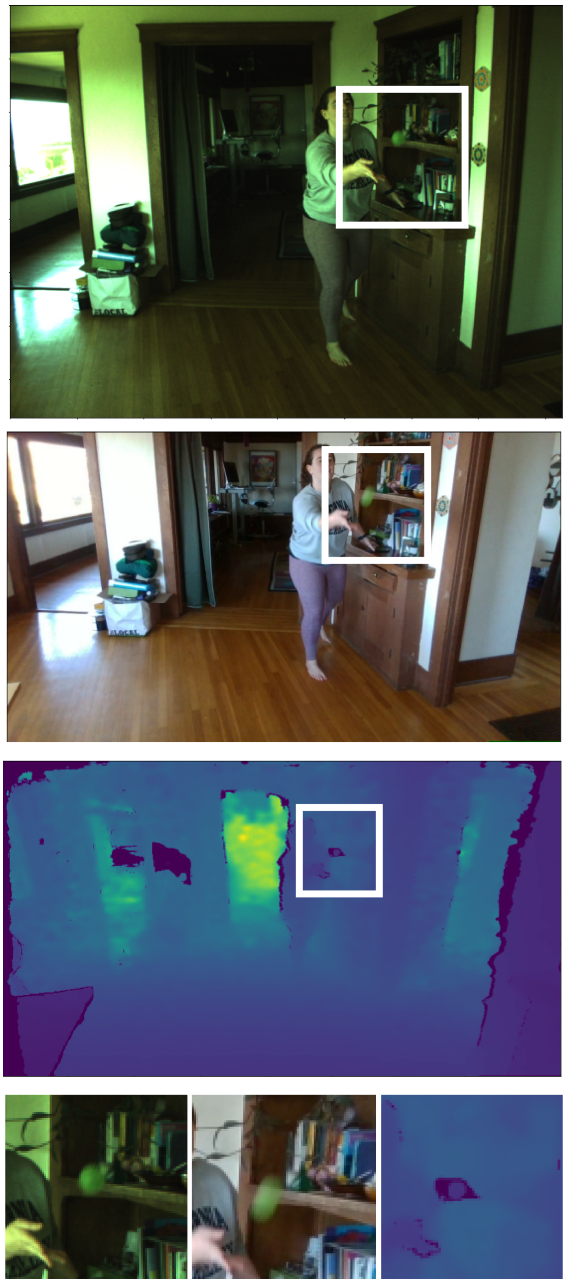


Figure 2.8: Visual streams are temporally synchronized to the framerate of the slowest visual stream (60Hz). Temporally synchronized frames from three visual streams (top to bottom): Ximea RGB stream, RealSense RGB stream, RealSense depth stream. White boxes indicate zoom-in on bottom panel, showing ball in same position at moment of release from hand during toss, which is evidence that timestamps are well matched. Note the greater motion blur of the ball in the RealSense RGB stream running at 60Hz (bottom middle panel) compared to the Ximea RGB stream running at 200Hz (bottom left panel).

for gaze localization. Next, we combine the RealSense to Ximea RGB extrinsics matrix (measured during the stereo calibration) with the RealSense camera's self-reported depth to RGB extrinsics matrix to create a depth-to-Ximea extrinsics matrix. Finally, we used this combined RealSense RGB to Ximea RGB extrinsics matrix in the .bag file alignment, rectifying these depth frames directly into the Ximea camera's frame of reference. Performing the alignment in one step with a combined extrinsics matrix avoids loss of image data due to the vertical field of view of the RealSense RGB being smaller than the depth and Ximea RGB streams. We perform all spatial registration offline after data collection. An example set of aligned frames is shown in Figure 2.9.

Spatial Accuracy There are various sources of error within the sensors, their synchronization, and eye tracking calibration which individually contribute to the overall spatial and temporal accuracy limits of the system. The largest source of spatial uncertainty in our system is in the eye tracking. We use a custom, depth-aware calibration and gaze localization method, which reduces estimated error to 0.25° in the best case, and 0.5-0.6° for an average subject, which is better than the <1° and the 1.5-2.5° accuracy reported for the Pupil Labs 2D and 3D gaze mapping methods, respectively [57]. We report the details of this custom method in previous work [42]. With the magnification factor of the lens used in our system, 0.25° corresponds to approximately 8 pixels - a wider angle lens would reduce this, and a temporal smoothing window can also be applied to the eye trace to reduce high-frequency jitter.

Temporal Accuracy The largest source of temporal uncertainty in our system is in the depth stream, which is frame rate limited by maximum sampling rate of the RealSense depth sensor at 90Hz. We up-sample the depth stream in post-processing, from the native 90Hz to the 200Hz sampling rate of the high fidelity cameras. Because both our gaze mapping and spatial re-projection methods are depth-dependent, this depth accuracy limitation propagates through our analysis and is a limiting factor of our system.

2.6 Discussion

To our knowledge, the apparatus and data collection and analysis methods are the first to enable high-fidelity, data-intensive, and synchronized multi-sensor signal capture in a mobile eye-tracking device. It enables a high-quality reconstruction of the natural visual input as experienced by the human eye as a subject goes about everyday activities. At the same time, it records the subject's body, head, and eye movements.

A limitation of the device is the high framerate camera (particularly in its switch-box/PCIE system) because the camera adds weight and bulk to the head, which may restrict subject movement during data collection. In other words, the weight and bulk may affect the statistics of the measured body, head, and eye motion. When we designed the device, this camera was the best available option for high-speed collection without introducing artifacts

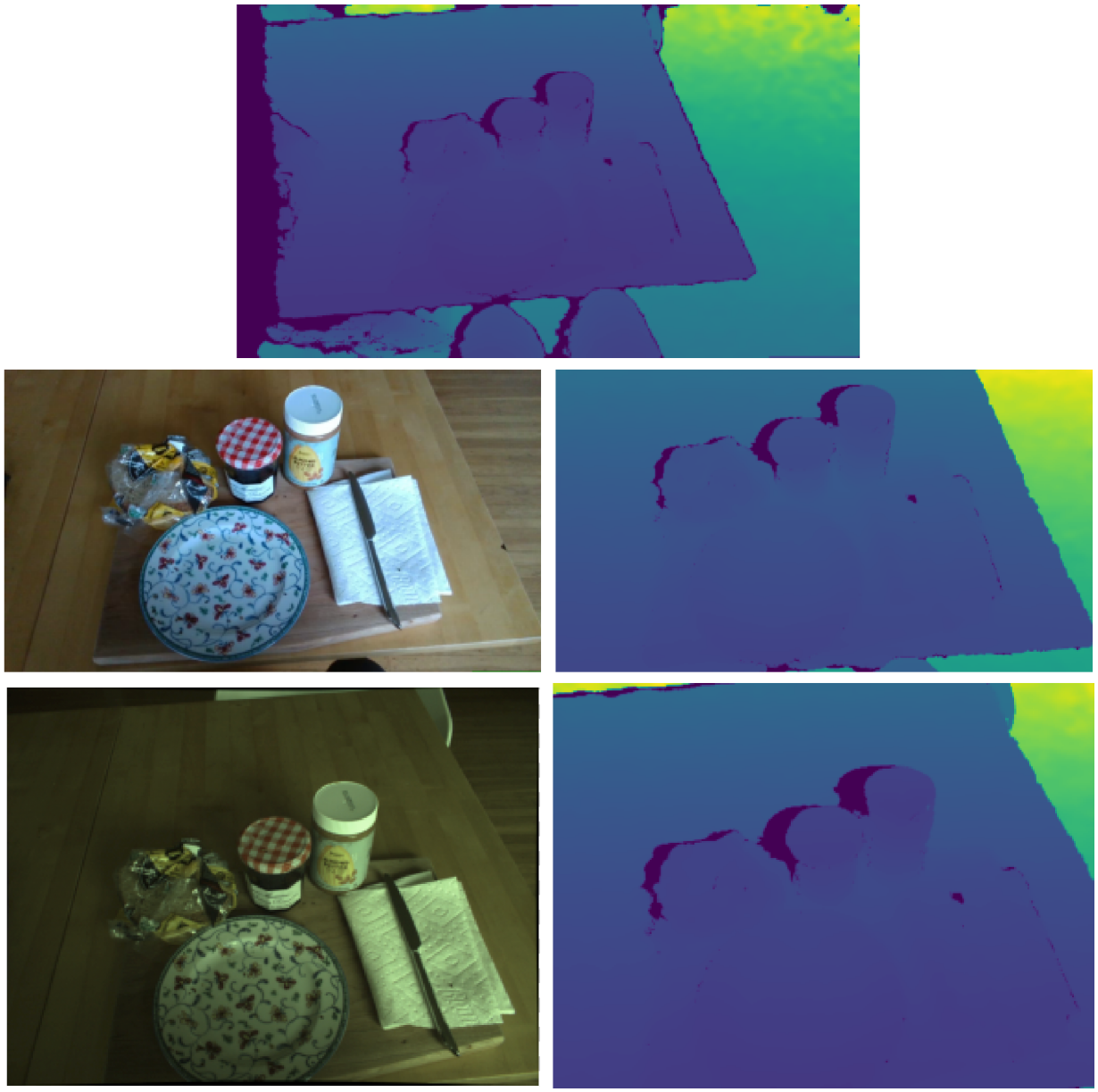


Figure 2.9: Visual streams are spatially aligned through registration with extrinsics matrices. Top: Original depth frame as provided by RealSense camera before spatial alignment. Middle: RealSense RGB frame reference on left and aligned depth on right. Bottom: Ximea RGB frame reference on left and aligned depth on right.

due to on-camera compression. Since then, new cameras have been released that connect via other methods including USB-C and they allow high-fidelity and lighter weight.

Data collected with the device will reveal the complex spatiotemporal patterns of light that strike the retina during everyday life. Quantifying the statistics of these patterns will be important for gaining a better understanding of the human visual and motor systems and how they have adapted to the natural environment. The data collected with this device will be useful to a number of scientific and technical communities including vision science, experimental psychology, neuroscience, bioengineering, computer science, and display technology.

Chapter 3

The Spatiotemporal Power Spectrum of Natural Human Vision

Abstract

When engaging in natural tasks, the human visual system processes a highly dynamic visual data stream. The retina, performing the very first steps in this processing, is thought to be adapted to take advantage of low-level signal regularities, such as the autocorrelation function or power spectrum, to produce a more efficient encoding of the visual signal. Previous work examined the joint spatiotemporal power spectrum of handheld camera videos and Hollywood movies, showing that power falls as an inverse power-law function of spatial and temporal frequency, with an inseparable relationship. However, these data are far from an accurate characterization of a day in the life of the retina due to body, head, and eye motion, which overlay additional diverse types of complex motion to the incoming signal, modifying the overall statistics. In addition, the distribution of natural tasks will influence the statistics of this signal. Here, we aim to characterize these statistics of natural vision using a custom device that consists of a head-mounted eye tracker coupled with high frame-rate world cameras and orientation sensors. Using video data captured from this setup, we analyze the joint spatiotemporal power spectrum for three conditions: 1) a static camera viewing a natural task being performed (environmental motion only) 2) a head mounted camera worn by a subject engaged in a natural task (head and body motion added) 3) videos simulating the dynamic retinal image, created by overlaying the subject's eye motions on the head-mounted camera video stream (eye motion added). This allows for analysis of the signal properties imparted by each of these motion types individually and an account of the final signal that reaches the retina, which incorporates all motion types. Results suggest that compared to a static camera, body and head motion have the effect of boosting high temporal frequencies. Furthermore, eye motion enhances this effect, particularly for mid to high spatial frequencies, causing this portion to deviate from the power-law and become nearly flat. These data will be helpful in developing efficient coding models relevant to

natural vision.

3.1 Introduction

The power spectrum of natural images, averaged over many images, is known to follow a power-law [39]. This relationship can be described with the equation:

$$P(f) \approx \frac{\beta}{f^{\alpha}}$$

where f is the spatial frequency, β is a scaling factor that varies with the image's contrast, α defines the slope of the falloff, and $P(f)$ is the power as a function of frequency f. The value of α obtained from a function fit to a single individual image or image patch is quite variable. Averaged over a large number of natural images/patches, however, α is generally found to have a value of ~ 2 for the power spectrum ~ 1 for the amplitude spectrum [39].

This power-law property of the power spectrum of natural scenes results in the amount of power being the same, regardless of spatial scale (zooming in or out of an image); this is known as scale invariance, or self-similarity, and is thought to emerge from the structure of object sizes in the natural environment [87]. However, missing from this purely spatial perspective is an account of the dynamic properties of natural scenes, which contain motion that contributes to a corresponding temporal frequency spectrum. These temporal frequencies interact with the spatial frequency spectrum in a non-trivial way.

Previous literature on spatiotemporal natural scene statistics focused on analyzing widely available video content, typically from sources such as nature documentaries. A seminal study analyzing the spatiotemporal power spectrum of Hollywood and handheld camcorder movies reported that movies depicting naturalistic scenes have a joint spatiotemporal power spectrum that follows a power-law both in space and in time [30] The results of their analysis is shown in figure 3.1. These power spectrum curves show a complex relationship between spatial and temporal frequencies that cannot be described by a simple separable function of spatial frequency f and temporal frequency w. Non-separability means that the spatiotemporal power spectrum $G(f, w)$ cannot be factored into a product of two functions $F(f)$ and $W(w)$, which each depend only on the spatial and temporal frequencies, respectively.

$$G(f, w) \neq F(f)W(w)$$

Visually, the inseparability of these curves is evident by the variable slopes for the different dotted lines in figure 3.1, indicating a spatial frequency falloff that is dependent on temporal frequency and a temporal frequency falloff that is dependent on spatial frequency. A separable function would have independent spatial and temporal frequency falloffs, resulting in parallel lines bounded by lines of the same slope both above and below the curves. Separability can be evaluated mathematically using Principal Component Analysis (PCA), to determine the amount of variability in the data explained by a product of two 1D functions [32].

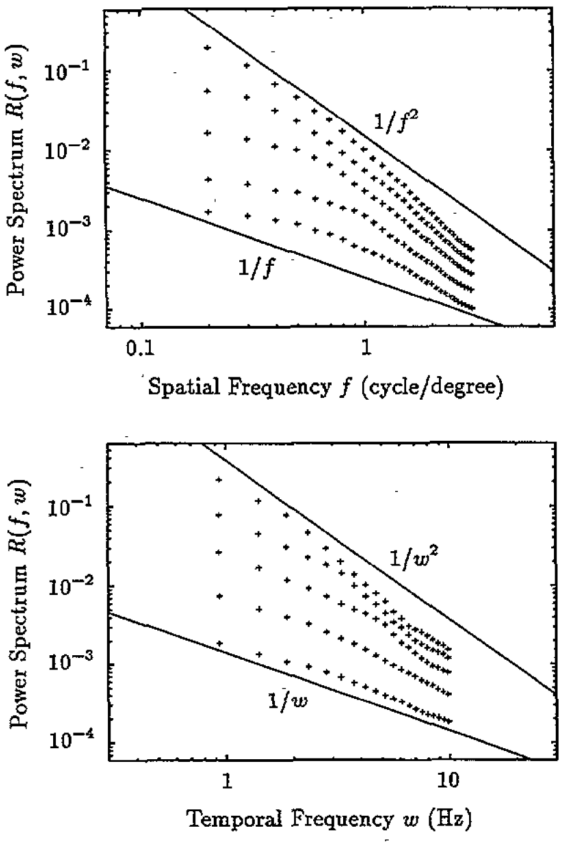


Figure 3.1: Spatiotemporal power spectrum of naturalistic videos follow a power law. Figure From [30]. Sampled temporal frequency lines plotted along the spatial frequency axis, with the lowest temporal frequencies bounded below by $\frac{1}{f}$, and the highest temporal frequencies bounded by $\frac{1}{f^2}$ above (upper). Sampled spatial frequency lines plotted along the temporal frequency axis, with the lowest spatial frequencies bounded below by $\frac{1}{w}$, and the highest spatial frequencies bounded by $\frac{1}{w^2}$ above (lower).

These power spectra reflect the statics of Hollywood and camcorder movies, with tracking, panning, zooming, and other camera motion. However, these motion types are not representative of the signal that reaches the human retina, which is modulated by ego motions such as head and body movements, as well as the complex motions of the eye.

Understanding the shape of the spatiotemporal power spectrum has strong implications for understanding the coding properties of the human visual system. In the efficient coding framework [9], the visual system is thought to be adapted through evolution to process visual information optimally, and importantly, to have adopted coding strategies that are optimal given the statistical properties of the incoming visual signal. For example, in the spatial domain, it is thought that the center-surround receptive fields of retinal ganglion cells are matched to the $\frac{1}{f^2}$ power spectrum of natural images, and act to spatially whiten the retinal signal, an effect that is beneficial from an information theoretic standpoint. However, in the temporal domain, understanding this relationship has been limited due to a lack of high fidelity data documenting all the complex motions present in the dynamic retinal signal. Collecting such data and accounting for the joint spatiotemporal aspects of the incoming signal's power spectrum is necessary to extend our understanding of the complex relationship between the human visual system and its environment.

3.2 Methods

We collected data from a set of three human subjects, each performing the same set of 14 everyday tasks while wearing a custom-designed tracking device that collected visual and depth data from the environment, the subject's eye position, and motion of the subject's body and head. We also collected data with the tracking device mounted on a mannequin head while the mannequin 'viewed' the task being performed from an approximately first-person perspective.

Subjects

Three subjects (two females and one male, ages 26-33, all self-reported emmetropic) participated in the experiment. The subject protocol was approved by the Institutional Review Board at the University of California, Berkeley. All subjects gave informed consent before starting the experiment.

Apparatus

Subjects wore a custom mobile head-mounted eye tracking device, which is described in detail in Chapter 2, and consisted of a helmet with world cameras, head, and eye trackers, a back strap with a body tracker, and a backpack which held an acquisition computer and batteries. The helmet held a high fidelity RGB camera with an adjustable lens, a lower fidelity RGB camera, a depth sensor, a motion, velocity, and acceleration tracker, and a pair

of binocular eye trackers. The cameras were attached to the helmet using custom-designed 3D printed parts that allowed adjustment of the camera positions. The back strap was fit snugly to the body and held a second odometry tracker. The mesh backpack held a custom-built data acquisition computer and a pair of batteries. The backpack was worn only during mobile tasks and sat on the table next to the subject during seated tasks.

The adjustable lens was used in two different modes, open and pinhole aperture, which were used during indoors and outdoor tasks, respectively. The lens' focus was adjusted once before collecting data from the first subject, and its position was locked in with a set screw throughout the entire experiment, unaltered through the collection of data for all subjects. The focus and aperture settings were set with the following protocol: First, the aperture was fully opened to create the smallest depth of field possible, and the focus adjusted to 0.75 meters, fixing this as the lens focal position. Next, we closed the aperture to as close to pinhole as possible, validating that there was sufficient light for an indoor task with the indoor gain setting of 5dB, then opened the aperture to the position that avoided a washed out scene for the outdoor tasks, with the outdoor gain setting of 0dB. Both of these were tested at the experimental exposure time of 4.75ms, and chosen as the two aperture settings for indoor and outdoor tasks. Finally, we returned to the open aperture, validating that objects were in focus at both at near and far distances, testing this with an increased exposure time to increase the overall light levels. Finally, we noted the open and closed aperture positions and the focal position and secured them with a set screw.

Acquisition Software

The data acquisition software was written as individual plugins to Pupil Lab's Pupil Capture software, such that camera streams could be easily viewed, settings could be easily adjusted, and acquisition of all cameras could be both temporally synchronized, and recording easily started and stopped. The plugins included a plugin for Ximea camera acquisition, a plugin for saving depth in raw NumPy format, a plugin for acquiring data from the head and body trackers (adapted from [51]), and for acquiring and target positioning for the 9-point calibration. These plugins are available online at `https://github.com/vdutell/hmet_aquisition`, and are described in further detail in Chapter 2.

Tasks

While wearing the tracking device, each subject was recorded while performing a set of 14 tasks, each for 2 minutes. These tasks listed in Table 3.1, were inspired by the American Time Use Survey [47, 63], as well as reports indicating the increased use of screen time in recent years[17], and selected to represent as much as feasible the common of everyday tasks of an average person living in modern society.

In addition to data recorded from the human subjects, we also recorded 2 minutes each of the 14 tasks with the recording device mounted on a mannequin head attached to a tripod. The mannequin head was positioned as close as possible to where the subject's head would

Table 3.1: Set of 14 everyday tasks surveying a 'day in the life' of an average person. Walking tasks were paired with standing tasks in the same location for comparison to environmental-only recordings from the mannequin. Tasks varied in location (indoor vs outdoor), mobility (seated, standing, walking), and viewing distances. Tasks were included that engaged various aspects of vision such as reading, passive viewing, sensory-motor engagement, smooth pursuit, and complex navigation such as stairs. For the 'watch movie' task, subjects watched a clip from Indiana Jones, which was studied in [29].

Task	Location	Mobility	Viewing Distance	Notable Aspects
Read Book	Indoor	Seated	Near	Reading
Use Cellphone	Indoor	Seated	Near	Reading,Active,Screen
Use Computer	Indoor	Seated	Near/Mid-Range	Active,Screen
Make Sandwich	Indoor	Seated	Near/Mid-Range	Sensory-Motor
Watch Movie	Indoor	Seated	Near/Mid-Range	Passive,Screen
Chat with Person	Indoor	Seated	Mid-Range	Human,Eye-Contact
Fold Laundry	Indoor	Seated	Near/Mid-Range	Sensory-Motor
Stand in House	Indoor	Mobile	Mid-Range/Far	Carpentered
Walk in House	Indoor	Mobile	Mid-Range/Far	Carpentered,Navigation
Play Catch	Indoor	Mobile	Near/Mid-Range/Far	Sensorimotor,Pursuit
Stand on Patio	Outdoor	Mobile	Mid-Range/Far	Nature
Walk on Patio	Outdoor	Mobile	Mid-Range/Far	Nature,Navigation
Stand on Road	Outdoor	Mobile	Far	Urban
Walk on Road	Outdoor	Mobile	Far	Urban,Navigation

have been performing the task themselves. The mannequin then observed the task being performed from an egocentric position but absent from ego-motion.

Calibration

We utilize a custom calibration routine described in [42], which consists of a modified 9-point calibration routine and a custom 3D-printed handheld calibration target with a fixation-cross-shaped hole in the middle and a flashlight in the back. For each eye, at each of the 9 positions, the subject would keep the target covered with their thumb, then align the target with the gaze point of one eye, in the manner of a gun-sight, until they could see the flashlight light coming through the hole in the front of the target. This ensured that the normal vector to the plane of the target was aligned with the eye position. The subject would then uncover the target to allow automatic target detection in both depth and RGB streams. This method created a 3D point cloud of eye and gaze positions, which allowed accurate binocular gaze positioning in 3D space during data collection.

Each full calibration routine included the above described 9 point calibration taken once in each eye, a binocular validation with the target placed at ~ 8 random positions within the field of view, and measurement of the primary position (eye position with subject focused at infinity). This full routine was done once every 3-4 tasks, or if the headset was re-positioned on the head. In addition, a single point target validation was taken at a 2-3 meter distance before and after each task, allowing the identification and correction of any drift in the calibrated gaze positions during data collection.

Motion Categories

Video cubes of 512x512 pixels by 400 frames (2 seconds) were sampled from the recorded data in each of three conditions

- Static: Starting at a random location within the static camera (tripod + mannequin) movie, and remaining spatially localized in the frame as the "movie" progressed.
- Head-Body: Starting at a random location within the head-mounted camera movie, but remaining spatially localized in the frame as the move progressed.
- Retinal: Starting at a measured spatial-temporal eye position location within the head mounted camera movie frame, moving spatially matching measured eye positions as the movie progressed.

Fourier Analysis

For each movie chunk, we averaged over 3 color channels to get a greyscale movie, applied a raised cosine window, subtracted the chunk mean, then calculated the 3D Fourier transform. We then squared this value to calculate the power spectrum, and took the rotational average over the two spatial dimensions. We repeated this method for 250 movies per subject/task combination, and calculated the mean over all subjects, and over all tasks. These mean power spectra are reported.

3.3 Results

We report here the results taken from a subset of the data. Shown here are the results from a single subject performing a subset of six tasks (read book, use phone, use computer, watch movie, chat, and make sandwich).

Mean Power Spectra - All Tasks and Subjects

We first report the grant means of the spatiotemporal power spectrum for one subject performing the six seated tasks.

When considering environmental motion only, without any modulation by ego motion, we find that the spatiotemporal power spectrum follows the general inseparable shape previously reported by Dong and Atick [29] (Figure 3.2). The highest power is found in the lowest spatiotemporal frequencies, and the lowest power in the joint high spatial, high temporal frequency regime. In the spatial frequency domain, the lowest temporal frequencies have the highest power and fall the most sharply with increasing spatial frequency, bounded by approximately $\frac{1}{f^2}$. The highest temporal frequencies, by contrast, fall shallowly with increasing spatial frequency, bounded by approximately $\frac{1}{f}$. This leads to an overall triangular shape in the joint 2D frequency space (Figure 3.2, top).

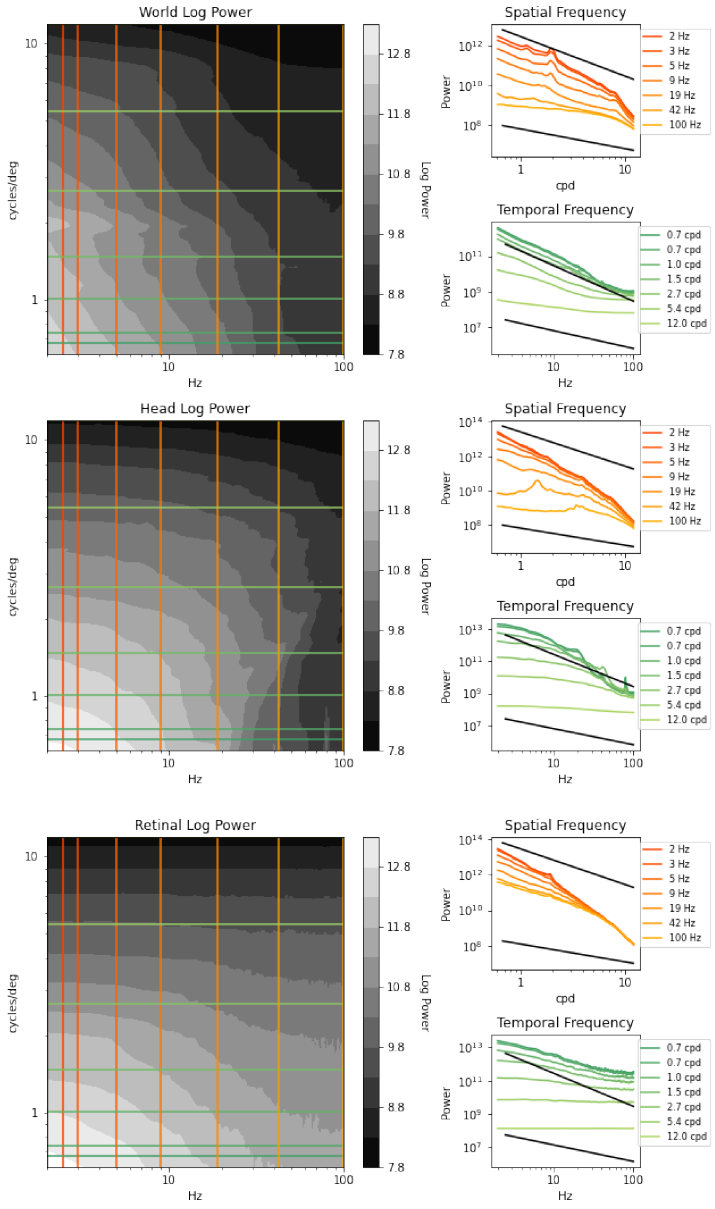


Figure 3.2: Mean Spatiotemporal Power Spectra over all subjects and tasks for three conditions: World (top): video data taken from tripod/mannequin with environmental motion only. Head (middle): video data taken directly from the head-mounted camera with environmental, body, and head motion. Retinal (bottom): video data taken from the head-mounted camera with eye motion overlaid; this includes environmental, body, head, and eye motion. There is a progressive change in the shape of the power spectrum as additional motion types are included. Head and body motion boosts mid-range temporal frequencies, while retinal motion boosts high temporal frequencies, resulting in a partial temporal whitening of the signal that reaches the retina.

When head and body motion is introduced, adding ego motion on top of the environmental motion signal, this modulation strongly affects the resulting spatiotemporal power spectrum. The effect in the spatial frequency domain is subtle but is more pronounced in the temporal and joint spatiotemporal frequency space. The effect of head and body motion boosts the power for mid-range temporal frequencies, particularly those in the low to mid-spatial frequency range. This dramatically reduces the slope of the temporal frequency falloff in the mid-range (Figure 3.2, middle). This mid-range boosting strongly changes the shape of the joint 2D space, resulting in a box-like shape. As we will show in section 3.3, the mid-range spatiotemporal frequencies boosted are consistent with the corresponding velocities of body and head motion.

When eye motion is introduced, overlaid on top of the environmental, body, and head motion, approximately recreating the retinal signal, the spatiotemporal power spectrum is further changed. Again, this boosting effect is most strongly seen in the temporal frequency domain but is most present in the high temporal frequencies, especially those in the low to the mid-range spatial frequency range (Figure 3.2, bottom). This further reduces the slope of the falloff with increasing temporal frequency, reducing the slope to zero for the high-range spatial frequencies, and reducing the slope significantly even for the lowest spatial frequencies. This is seen again in the shape of the joint 2D space, where mid and high spatial frequencies with the lowest power no longer depend on temporal frequency. As we will show in Figure 3.3, the high-range temporal frequencies boosted are consistent with the corresponding velocities of eye motion.

To separate the effect of head/body motion and eye movements on these power spectra, we calculated the difference between the mean power spectra shown above (Figure 3.3). The effect of head and body motion is simply calculated by subtracting the 'World' power spectrum from the 'Head' power spectrum. We see that the effect of head and body motion is a dampening of low temporal frequencies at all spatial frequency ranges and a boosting along with a subset of mid-range spatial and temporal frequencies. The effect of eye motion has a similar dampening effect; however, the boosting imposed by eye motion lies in the low to mid spatial frequency range, at high temporal velocities only. We will see in 3.3 that these profiles are more easily interpreted in terms of their corresponding velocities.

Velocities

Each point in the joint spatiotemporal-frequency plane corresponds to a specific velocity, allowing for the translation of spatiotemporal power to power of velocities. Velocity is calculated by dividing temporal frequency in units of $\frac{cycles}{second}$ by spatial frequency in units of $\frac{cycles}{degree}$. This yields a speed in units of $\frac{degrees}{second}$. Figure 3.4 shows the corresponding velocities for the range of spatiotemporal values shown in Figure 3.2. When plotted in log-log space, areas of equivelocity appear as parallel right-diagonal lines (slope of 1). By comparing this velocity plot to 3.3, we see that the effect of head and body motion is an increase in mid-range velocities on the order of $10\frac{deg}{sec}$. Eye motion, however, boosts power for higher velocities on

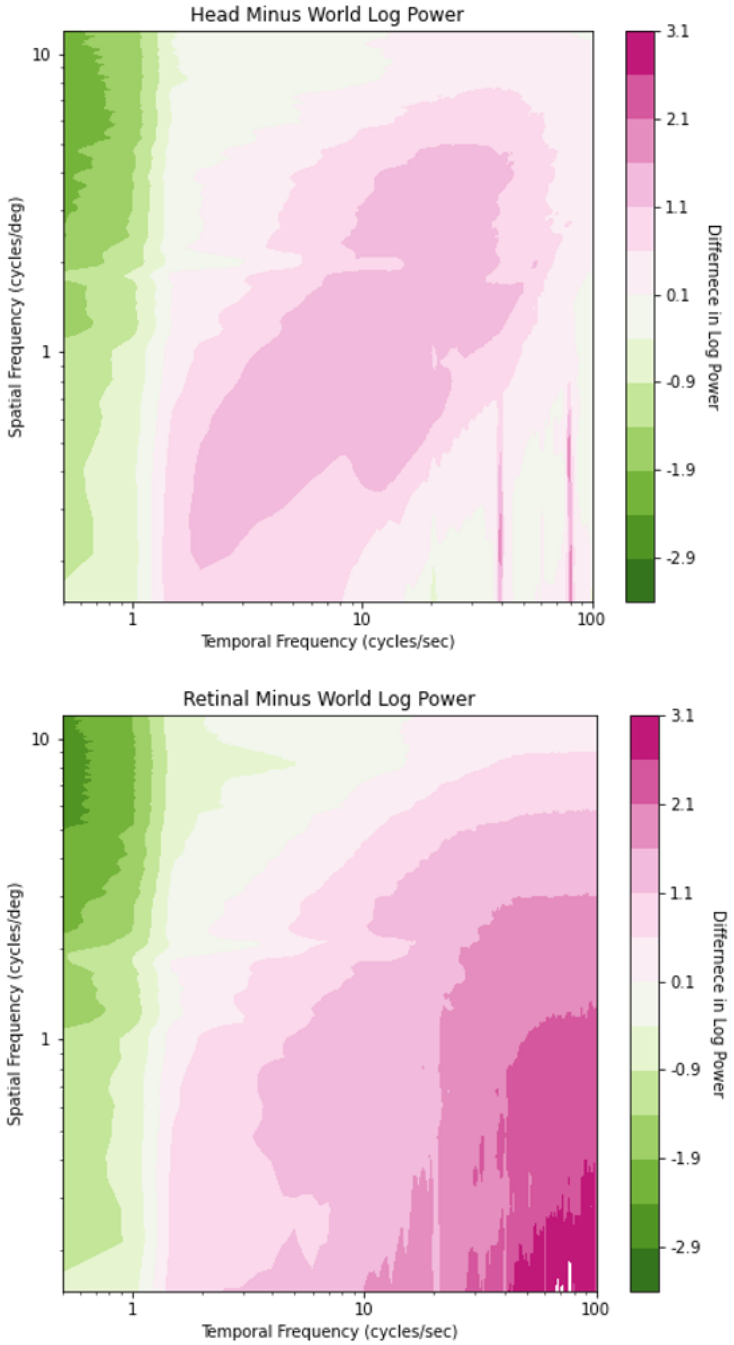


Figure 3.3: Effect of Body and Head Motion (top) and Eye Motion (bottom) as calculated by differences between respective power spectra. Pink indicates a boost in power. Green indicates power is dampened.

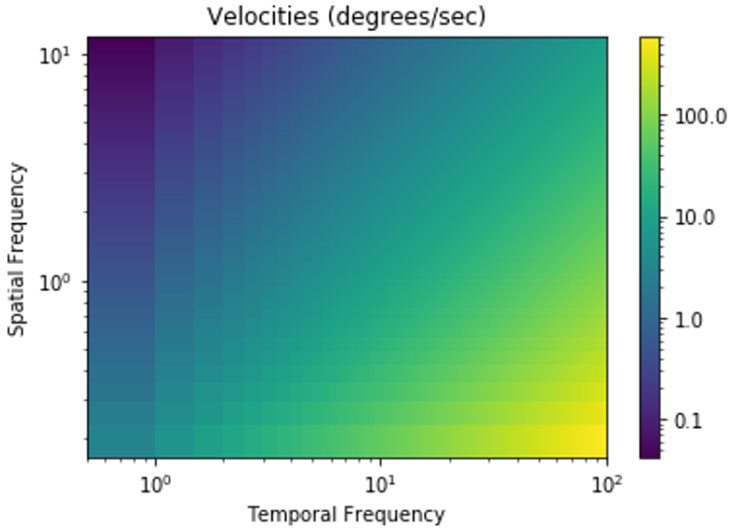


Figure 3.4: Right-diagonal lines in the spatiotemporal frequency plane correspond to lines of equal velocity.

the order of $100\frac{deg}{sec}$. These results are consistent with the biomechanics of body/head and eye motion and previously reported velocity distributions [33].

Separability Analysis - Approximation by 1D Functions

Separability, in the context of the spatiotemporal power spectrum, is the degree to which the 2D power can be described by a product of the separate spatial and temporal frequency spectra. Qualitatively, we find distinct separability in the power spectra of all three conditions by noting the temporal-frequency dependent difference in the slope of power falloff in spatial frequency spectrum, and the spatial-frequency dependent slope in the temporal frequency spectrum. A completely non-separable spectrum reported by Dong & Atick [30], would result in the power falloff in spatial frequency to be independent of temporal frequency and vice versa, resulting in spatially offset parallel falloff lines. We instead see notable differences in slope, and qualitatively, the spatiotemporal power spectrum appears separable. Moreover, although the introduction of ego motion dramatically alters the overall shape of the power spectrum, we find that this does not alter the separability property, at least by visual inspection.

To investigate these results quantitatively, we followed the Singular Value Decomposition (SVD) method [32], which in decomposing the 2D power spectrum into it's singular values and vectors, measures the degree to which this 2D function can be described by the product of two 1D functions (one function of spatial frequency, one function of temporal frequency). The degree of separability is quantified by the index of separability α, which is the ratio of the first singular value to the sum of all singular values. Thus, a high α indicates high

separability in that the product of two 1-D functions can explain a majority of the variance in the data. By this quantitative metric, we find very high α values of 0.9973, 0.9949, and 0.9994 for the environmental, head/body, and retinal conditions, respectively, which is in agreement with the high (> 0.99) α values previously reported for natural video sequences [32].

3.4 Discussion

High Fidelity Spatiotemporal Power Spectrum

Surveying the visual input from various everyday human environments, in the absence of ego-motion, we find that the environment and task has little effect on the qualitative shape of the spatiotemporal power spectrum. Even when compared with the highly motion-biased dataset studied by Dong & Atick [29], the overall spatiotemporal power spectrum retains the same general properties. There is an inseparable relationship between spatial and temporal frequencies, with the falloff in both spatial and temporal frequency bounded by $\frac{1}{f}$ below, and $\frac{1}{f^2}$ above, where high temporal frequencies falloff faster in space, and high spatial frequencies falloff faster in time.

This relationship is scale-invariant within the measured range, extending into spatial and temporal frequencies much higher than previously investigated. This is not surprising in the spatial domain, as scale invariance in natural images has been widely described [88]. Given the relatively few investigations into temporal natural scene statistics as compared to spatial. However, the extension of this relationship into higher temporal frequencies is not obvious. Furthermore, the highest temporal frequencies investigated here (100Hz), though beyond the classic flicker fusion limit of 35-60 Hz [53, 34, 70], are within the range of both perceptually visible frequencies, which have been detected at 500Hz [24], as well as the high temporal frequency range known to stimulate and entrain visual neurons [105]. In the spatial frequency domain, the maximum detectable frequency is $\sim$ 60 cycles per degree, which is also beyond the measurement capabilities of our system. Future studies may wish to investigate the range of even higher spatiotemporal frequency stimuli beyond our current system's capabilities, but that is also relevant to the human visual system sensitivity range.

Separability

Our results for separability are somewhat conflicting, in that visual inspection of the power spectra show an inseparable relationship, while a quantitative analysis indicates a high separability metric (~ 0.999). Given the clear inseparability indicated in the visualizations (Fig 3.2), as well as the previous work by Dong & Atick [30] outlining this visually defined separability metric, we interpret this mismatch in favor of the visual inseparability result and conclude that for some reason, the separability metric does not seem to reflect the separability seen in the data. We note that for our environmental motion condition, our value for α is

in agreement numerically with the results from Eckert et al. [32] in which Hollywood video data was analyzed. In addition, we run separability analysis on the log-normalized power spectrum, as in Dong & Atick, as this reveals its scale-invariant structure (the same reason for visualizing it in log space); when this step is omitted, α is closer to 0.990. We suspect that given the power-law structure of the data, either the SVD method may be over-fitting to a subset of data points, or that a seemingly large value of ~ 0.999 is not high enough to indicate separability in our case. These issues deserve further investigation.

In everyday task environments, we find that the spatiotemporal power spectrum of the environment itself is qualitatively inseparable. The shape of falloff of power in spatial frequency is temporal frequency-dependent, and vice versa. Such a quantitative analysis is performed simply by observing the differing slopes of the sampled power spectra in Figure 3.2. These are the same qualitative results as reported previously for Hollywood movies [29]. Furthermore, this inseparable quality is unchanged by adding ego motions of the body, head, and eye. However, when speaking quantitatively, numerical analysis points towards a more separable relationship, which is also in agreement with previous studies on natural video [32]. We also find that ego motion does not affect this quantitative metric of separability.

Incorporating the modulation of ego motion on the environmental signal, we find that this separability metric is not majorly altered despite the substantial changes imparted on the power spectrum plot. This is the case for both qualitative as well as quantitative measures of separability. It stands to reason that the visual system would not need to make spatial and temporal frequencies separable, as there is little evidence for a system that factorizes spatial and temporal signals. Instead, the signal is split downstream into the magnocellular/parvocellular/koniocellular pathways, which each code a portion of the join spatiotemporal frequency spectrum in parallel [67].

Temporal Whitening

We find a strong temporal boosting effect caused by the combination of head/body and eye motion. When analyzing these portions of the spatiotemporal power spectrum in terms of their corresponding velocities, this makes sense given the velocity distributions of head and eye motion. However, in the context of the vestibulo–ocular reflex (VOR), a visual motor system that aids in gaze stabilization, one might have expected more corresponding eye motion that would cancel out head and body motion in the service of gaze stabilization, rather than independent boosting. Given the evidence for varied VOR strength depending on locomotor task [27, 26], investigation of the whitening effects for seated, mobile, and standing tasks should be investigated individually.

This boosting effect has computational implications, particularly in the context of whitening, a strategy that is computationally beneficial by decorrelating or removing the redundancy in a signal [9]. Spatial-domain whitening, for example, is often thought to occur in the retina as an effect of the center-surround structure of retinal ganglion cells [5], and is often modeled as a necessary pre-processing step in V1 models [77, 28]. Temporal whitening, studied on a scale smaller than discernible by our methods, is thought to be a feature of fix-

ational eye motion including microsaccades and fixational drift [89]. While we find evidence of whitening by larger, saccade-level eye motion, we do not believe that mirosaccade-scale temporal whitening is mutually exclusive of larger-scale temporal whitening. This makes sense given the saccidic main sequence [6], in that the categorization of saccades and microsaccades is somewhat arbitrary, and the two are well described as portions of a continuous spectrum.

Our temporal whitening results are somewhat in contrast to previous theoretical work arguing for the role of the downstream Lateral Geniculate Nucleus (LGN) in temporal whitening of the visual signal [30]. While this theory points to the lagged and non-lagged cells of the LGN as the stage at which temporal whitening occurs, our evidence points to whitening of the signal before even reaching the retina. As LGN whitening is expected to occur on the same scale as the eye motion we record, it seems counter-intuitive that the visual system would perform temporal whitening/decorrelation twice, as this would over-whiten, and reintroduce correlations, removing any computational benefit. This apparent conflict can perhaps be resolved in noting that our results show a trend towards a whitened signal rather than a complete decorrelation. Especially for lower spatial frequencies, the signal reaching the retina is only partially temporally whitened. It is possible that these two systems work in concert, and given the complex dynamics of both eye motion and the LGN, that the LGN completes the partial whitening imparted by eye motion.

Saccadic Supression

Finally, the power spectra analyzed here are measured from two-second time windows drawn uniformly from the two-minute tasks. However, it is known that the statistics of eye motion are not stationary over time, varying between periods of saccades and fixations, as well other eye motions, including vergence and smooth pursuit, depending on the task. Saccades are of particular interest due to the effect of saccadic suppression, the effect of visual information being suppressed during saccades [72]. This is particularly relevant in the context of whitening, as it is unclear what portion of the whitening seen here is a result of saccadic eye motion, which is less likely to be perceptually relevant. Future work will calculate the power spectrum for saccadic and inter-saccadic intervals separately to determine the degree to which this whitening occurs during perceptually relevant timepoints.

Chapter 4

Phase Analysis of Natural Images

4.1 Introduction

Toward the goal of understanding the properties present in natural visual signals, several lines of research have identified regularities in the statistics of natural images. One of the most well-known of these is the 'power law', or 1/f relationship, which describes the amplitude of the frequencies present in natural images; specifically, the amplitude of a given frequency in any given natural image is found to fall off as the inverse of the frequency [39]. This 1/f spectrum has implications in visual neuroscience, having been shown to be matched to the band pass filter properties of the early stages of the human visual system, where this system seemingly works to remove redundancy in the incoming visual signal [5].

The power law alone however, is only a simple description of one property of natural images, and without additional parameters, is insufficient to describe natural images completely. For example, one can generate 1/f noise images (Figure 4.1) for which their power spectra follow the power law, but are distinct from a true natural image. Furthermore, statistical descriptors are needed to characterize what commonalities are shared among natural images but are not present in other power law signals such as noise images.

In addition to the commonalities among all natural images, additional lines of research have sought to understand what statistical differences exist that separate subcategories of natural images. Such analyses can be used to quickly identify, for example, the location of a scene as indoor/outdoor or natural/man-made. Though there are differences in the orientation-averaged power spectrum for different scene categories, they are relatively subtle. However, when considering the full 2D power spectrum, which allows the differentiation of slope at different orientations, the slope of the power in the vertical, horizontal, and oblique directions is quite variable between images of natural and man-made scenes and objects [98].

These noted statistical regularities of natural images are derived from one aspect of the natural image, its amplitude spectrum, as calculated from the discrete Fourier transform (DFT). However, all of these analyses ignore the other quantity produced by the DFT, an image's phase spectrum. We know that the phase component of a Fourier transformed

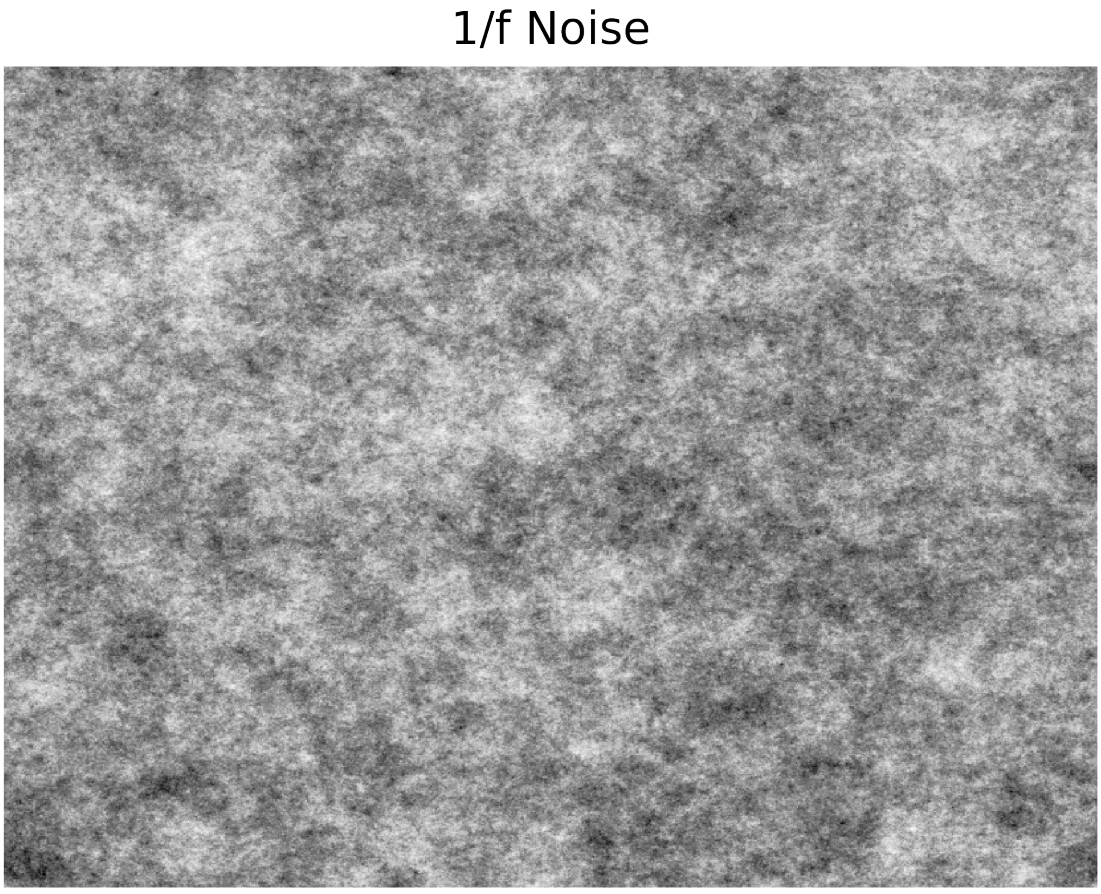
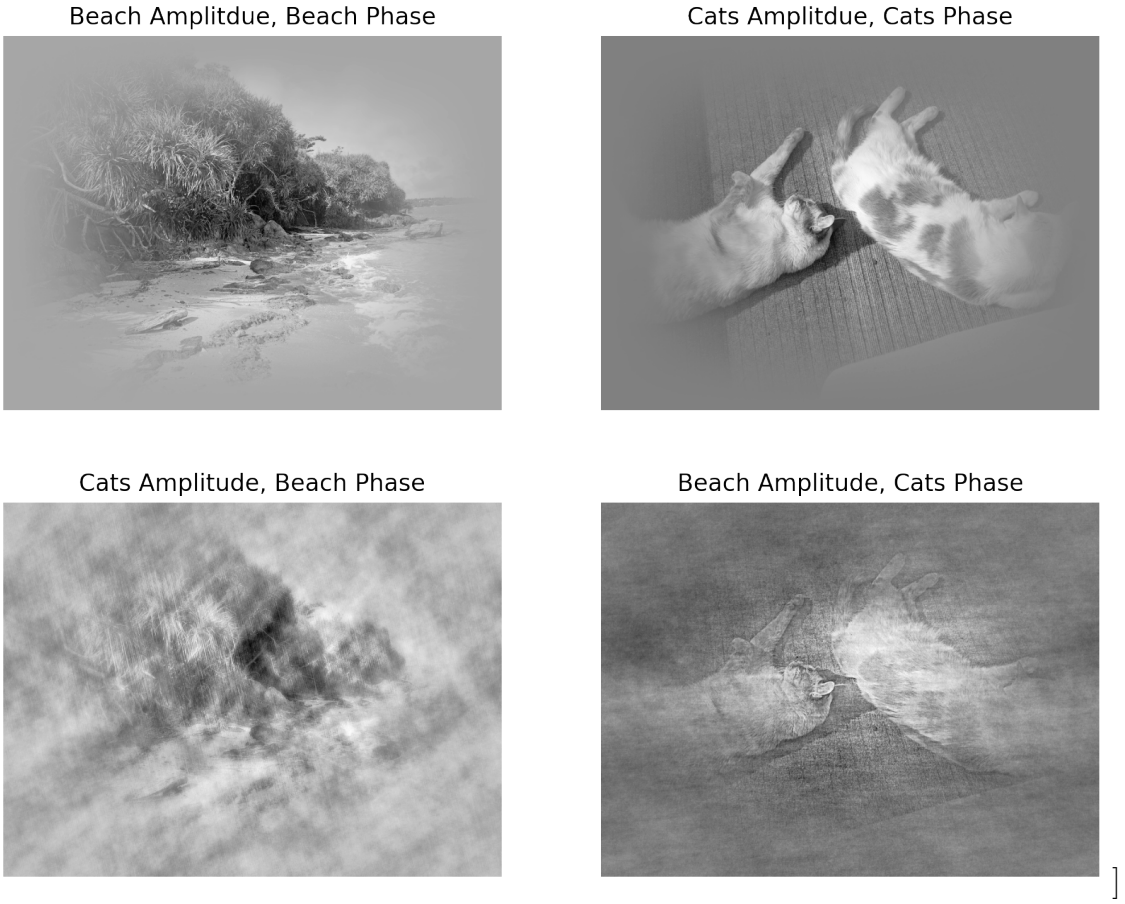


Figure 4.1: 1/f Noise has a power spectrum with a distribution characteristic of natural images. The phase spectrum typically takes a random uniform distribution. While this 1/f noise image has no discernible structure, it is visually distinct from white noise, which has a power spectrum distribution that is uniform.



]

Figure 4.2: Top: Original Images with cosine window applied, reconstructed from original phase and amplitude spectrum information. Bottom: Images with phase information swapped. Image with phase information from cat image but amplitude spectrum from beach image retains the cat appearance, while the image with cat amplitude spectrum and beach phase spectrum appears more like the beach image.

image contains more perceptually relevant information than the amplitude. For example, combining the phase of one image with the amplitude of another image, and taking the inverse Fourier transform results in a reconstructed image that looks much closer to the image that contributed the phase information, never the image that contributed the amplitude (Figure 4.2). In addition, image reconstruction given only the phase spectrum is readily implemented, whereas reconstruction given only the amplitude spectrum is much more challenging. The phase angle contains more perceptually relevant information about the image than the amplitude. It stands to reason then that by analyzing only the power spectrum, the natural scene statistics, the field is missing out on a large part of an image's perceptually-relevant information.

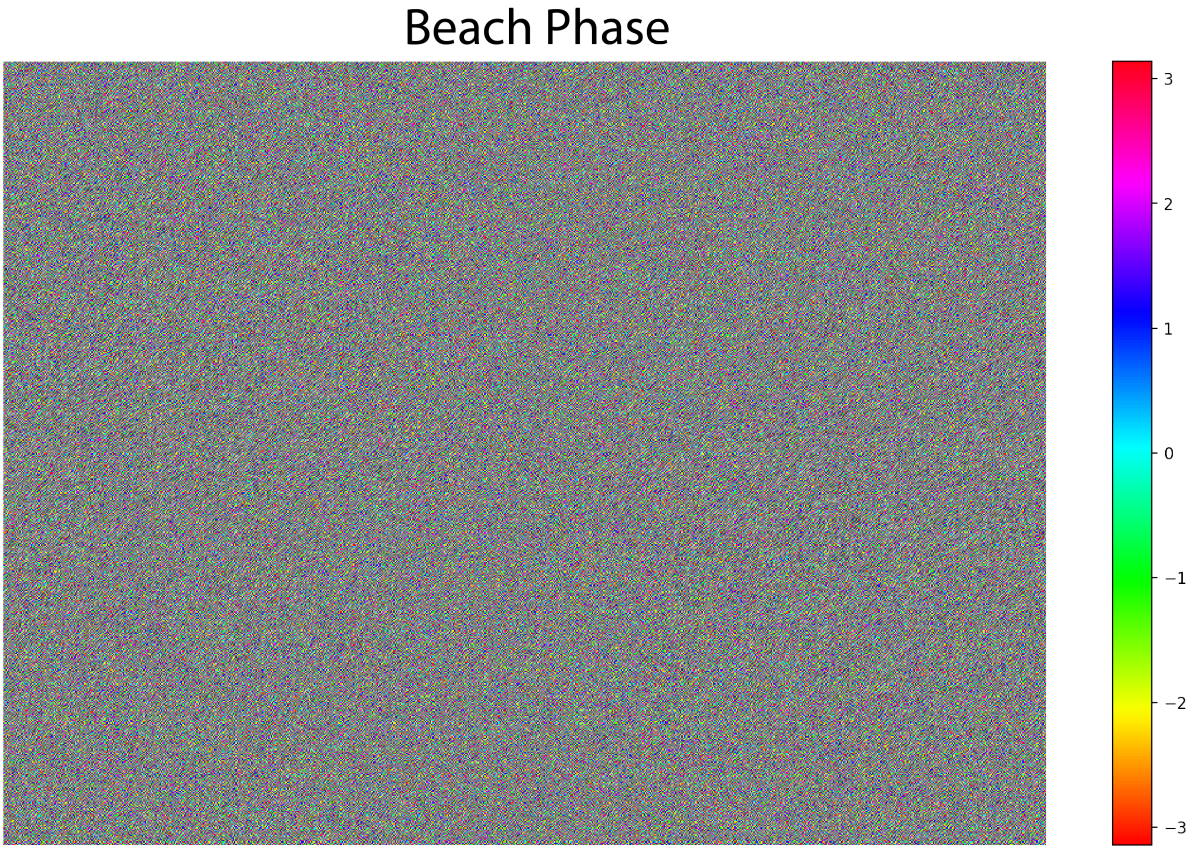


Figure 4.3: Phase angles for the beach image, taken from the Fourier Transform, with locations intact. Though the shape retains the original image's dimensions, each pixel corresponds to a given horizontal and vertical spatial frequency pair.

4.2 Spatial Structure in Global Phase

We first attempt to derive information from the global phase of natural images, analyzing these images to determine the feasibility of a global phase analysis method. Such a statistic would be useful to combine with the power law of the amplitude spectrum and used to synthesize images that more closely resemble true natural images than 1/f noise.

As a first step to understanding the phase component of natural images, we view the raw phase values derived from the Fourier Transform, keeping their relative locations intact. An example of this is shown for the beach image in Fig 4.3, with individual pixels in the phase image corresponding to a pair of vertical and horizontal spatial frequencies, rather than spatial location. Thus, despite the strong perceptual signal contained in the phase spectrum as seen in Figure 4.2, it is difficult to discern any spatial structure at all, let alone any structure corresponding to the image.

In the cat image, however (Figure 4.5, left), there is some structure visible in a small

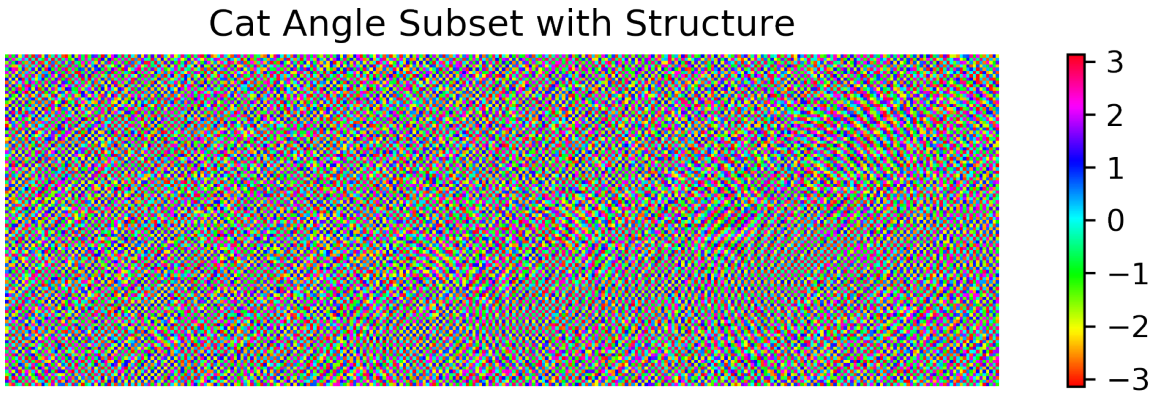


Figure 4.4: Subset of phase angle from cat photo shows areas of localized structure within the phase angle 'image'.

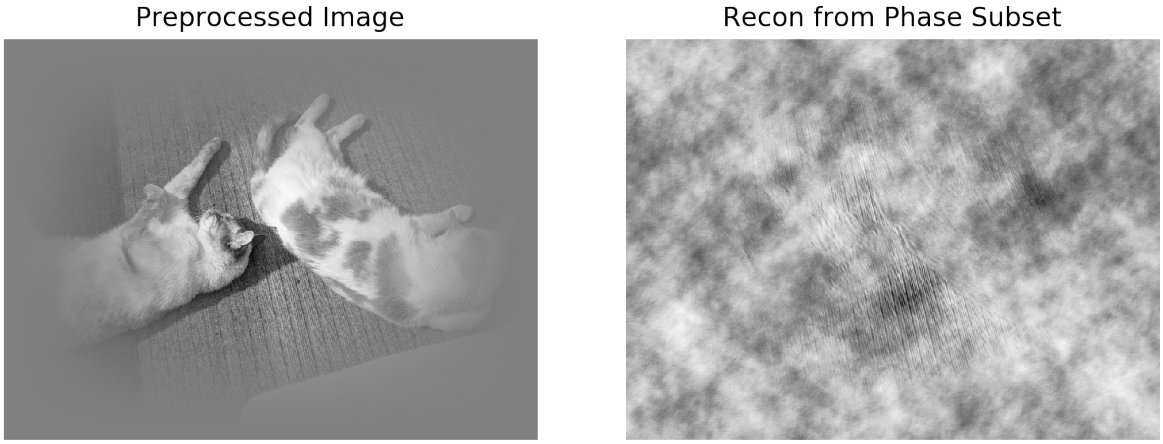


Figure 4.5: Image and Reconstruction from Spatially Structured Phase

section of the phase image (Fig 4.4). In the phase plot, there is visible striped structure along the diagonal. This area corresponds to locations of increased amplitude. Some other analyzed images contain small amounts of visible structure, but he cat image contains the most visible spatial organization in its phase plot.

In order to determine what properties of the image cause this structure to appear, a uniformly random phase spectrum is augmented by manually inserting only this spatially structured portion of the image's phase angle into the random spectrum. This is then combined with the cat image's amplitude spectrum and reconstructed using the inverse Fourier transform. Interestingly, the reconstructed image contains only the texture of the carpet (Figure 4.5). The repetitive striped pattern of the carpet can be described well by sine

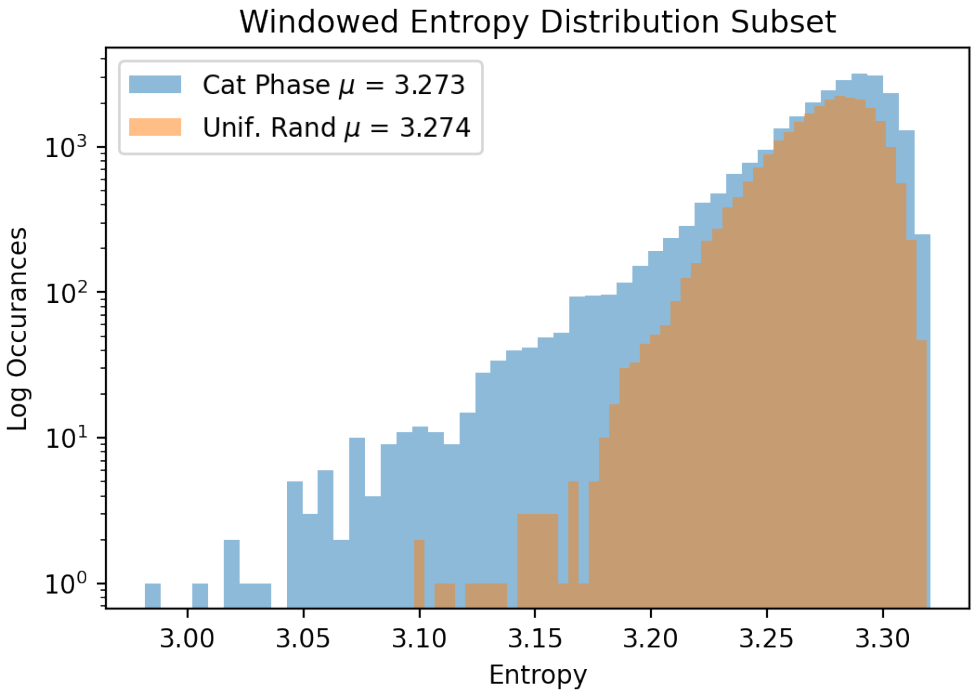


Figure 4.6: Windowed Entropy Distribution

waves and is therefore sparse in the Fourier domain. This explains the amplitude spectrum's peak and perhaps correlations between neighboring phase angle values, leading to interesting structured patterns in the global phase image.

Total & Local Entropy of Global Phase

To quantify these interesting structured patterns in the phase spectra, and potentially utilize this quality for image categorization, entropy is used as a measure of structure in the phase angle image. Global values for entropy in the phase angle are conserved between images and vary only slightly between true phase distributions and uniform noise, (difference of 0.0001 for values of around 3.32), (Figure 4.6). As a comparison, values for entropy of the images themselves are around 3.07, and vary by around 0.02 between images. Thus, entropy measurements of the global phase do not appear to differentiate natural images from noise.

Because the structure is localized to one spatial area of the phase angle plot, entropy is measured locally, using a 30x30 pixel window convolved with the phase angle image to look for spatially-localized variations in entropy, which is expected to correspond to the apparent spatial structure in the image. Indeed, the structures corresponds to spatially localized areas of decreased entropy. Binning these values to evaluate their distributions reveals that the phase angle values have a very similar distribution compared to a uniform random phase spectrum. While the true phase angle values tend towards lower entropy, their means are

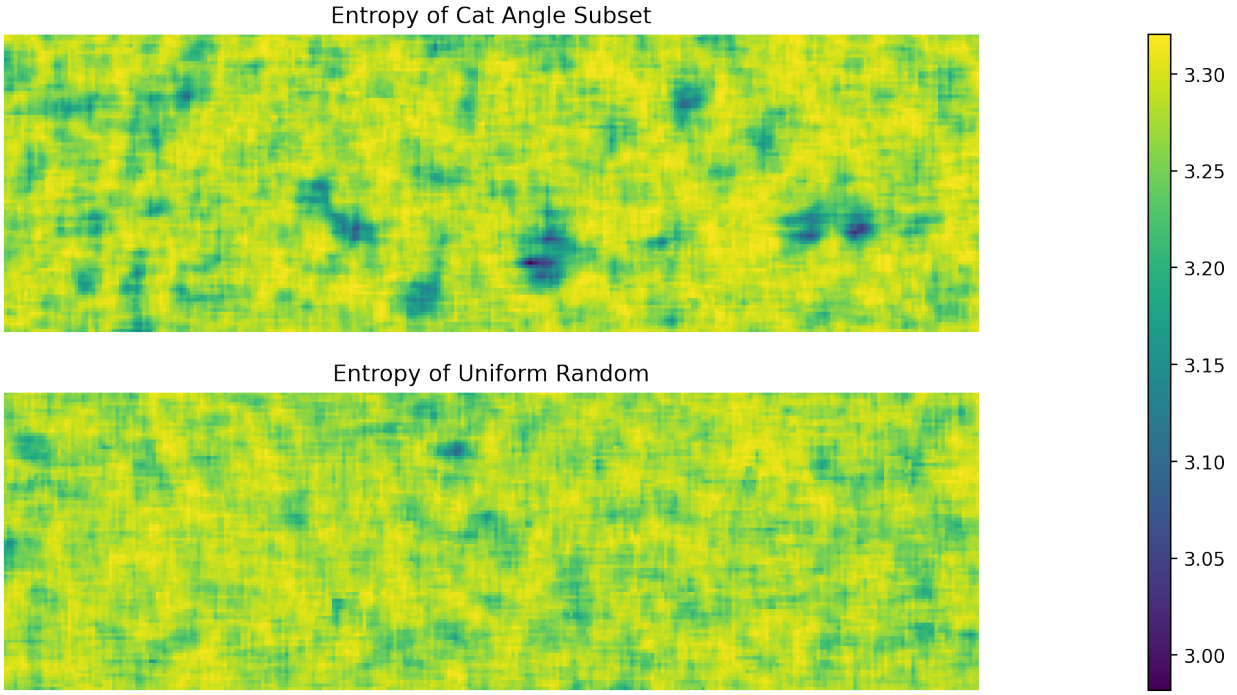


Figure 4.7: Spatial Distribution of Entropy of subset of phase angle for cat image (top), and for phase angle of uniform noise image of matched size (bottom). Natural image shows regions of low entropy, leading to a smaller entropy value overall.

very similar.

To determine if the spatial areas of low entropy are significantly different, the same analysis is performed using a smaller window (5x5 pixels) convolved with the subset of the phase angle which contained the pinwheels. Again, localized areas of lower entropy appear near the center of the pinwheels. Binning these entropy values and comparing their distribution to the same windowed analysis over a uniform random phase spectrum reveals that the distributions have very similar means. However, the true phase values have a large distribution of values at very low entropy, as expected, where the uniform random distribution does not (Figure 4.7).

Windowed entropy measured across the spatial distribution of phase angles does appear to separate some natural images from noise. However, this is only for a subset of images and appears to be those with particular spatial structure, such as textures with repeating patterns.

Large Scale Image Analysis of Global Phase

Given the promising differences between a single image and random distributions, the analysis is now expanded to a larger dataset of greyscale natural images and noise images. We

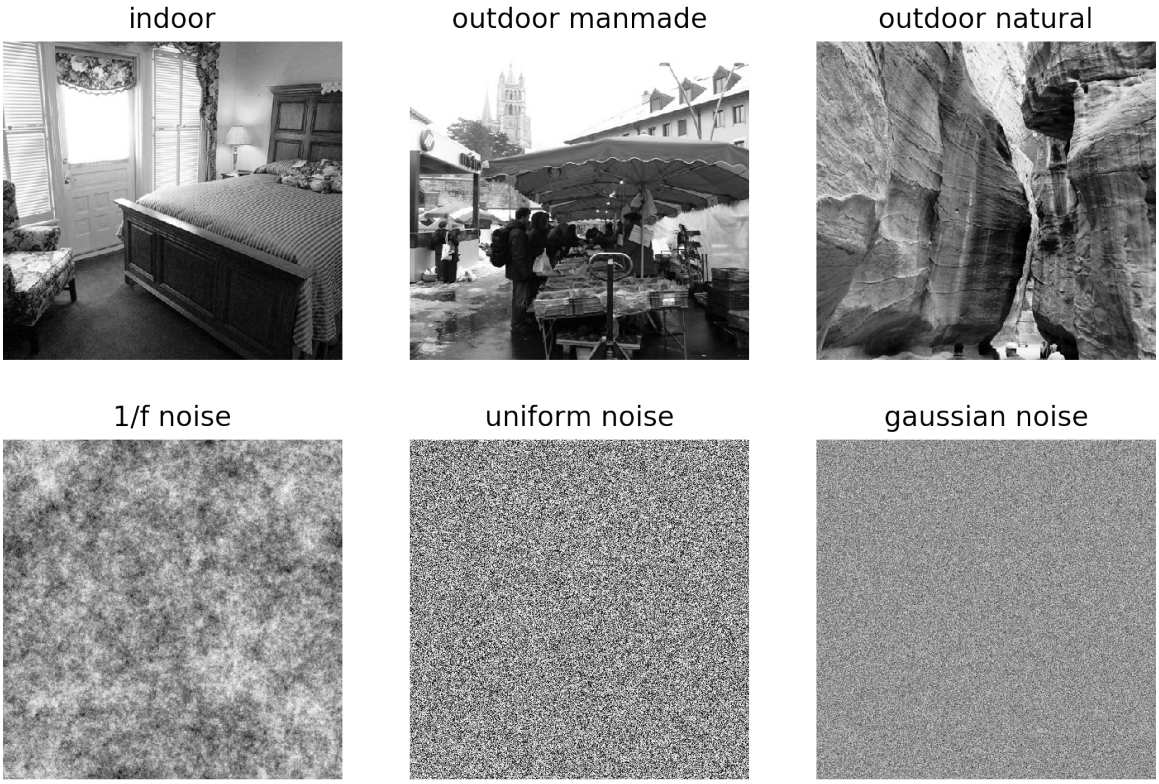


Figure 4.8: Image Categories

use the Places dataset [108]; the categories included are: indoor images, outdoor:man-made images, outdoor:natural images, 1/f distributed noise, uniformly distributed noise, and Gaussian distributed noise, with 1000 images per category. This allows us to examine the statistics that vary between images and distributions and compare within natural images to determine if any measured statistics varied between natural and less-natural images. Figure 4.8 gives examples of the image and noise categories tested.

Employing this dataset, global entropy analysis of the raw images and the global phase spectrum is measured. For the entropy of the images, the mean global entropy for all image categories is around 3.1-3.2. Uniform and 1/f noise share these values, and Gaussian noise is slightly higher at 3.3. In terms of the entropy of the global phase spectrum, all image categories and noises have similar distributions and mean entropy values of 3.32, with the exception, surprisingly, of 1/f noise, which had a slightly lower entropy of 3.10.

Overall, some specific images show differences in the spatial structure of their global phase angle spectrum, as measured by windowed entropy. However, this appears to be restricted to specific images which have repeated texture patterns. It does not appear that windowed entropy alone can differentiate natural images from noise images, or differentiate categories of natural images. It may be possible, given 1/f noise images have lower entropy values,

that such an entropy analysis could be used in conjunction with the amplitude spectrum to differentiate natural images from noise, but this is left to future work.

4.3 Phase Congruency & Energy

Background

While an image's global phase spectrum itself can difficult to interpret due to its diffuse nature, an alternative and more tractable quantity for analysis is the instantaneous phases of constituent frequencies at each point in the image. In particular, we can measure the amount of instantaneous phase alignment between all the frequencies at a given point. This ratio of alignment, or the phase congruency (PC) of an image, has been described previously [75], [62]. Phase congruency has shown to be useful as an edge detector, as areas in the image where many Fourier components are all aligning spatially are in 1D, likely to be edges, and in 2D, likely to be corners [62]. While PC has certain advantages as compared to other edge detectors such as the Canny, its relatively large computational requirements have hindered its wide adoption as an edge detector. Rather than an edge detector, we propose phase congruency as a tool to study the statistical regularities of natural images and variation between subcategories of natural images.

Phase congruency (PC) is a quantity defined for every sample of a discrete signal, which measures the ratio of Fourier components for which their instantaneous cosine phases are aligned (regardless of phase value). In one dimension, it is defined as:

$$PC(x) = \max_{\bar{\phi} \in [0,2\pi]} \frac{\sum_n A_n \cos(\phi_n(x) - \bar{\phi})}{\sum_n A_n}$$

Where ϕ_n is the phase of frequency n, and $\bar{\phi}$ is the mean phase over all frequencies at point x.

However, for computational ease, we calculate phase congruency equivalently by first calculating the local energy E, then normalizing the local energy to a number between 0 and 1. Thus, we define local energy E as:

$$E(x) = \sqrt{f^2(x) + f_H^2(x)}$$

where f is the Fourier transform of signal x, and f_H is the Hilbert transform of f, which rotates the signal f by $\frac{\pi}{2}$. This is derived from the definition of the analytical signal. We then normalize our local energy $E(x)$ by the sum of the n Fourier amplitudes A_n.

$$PC(x) = \frac{E(x)}{\sum A_n(x)}$$

Figure 4.9 shows an example of this analysis applied to a 3-level discrete step function (top). The local energy spectrum (middle) of this step function yields a function that peaks

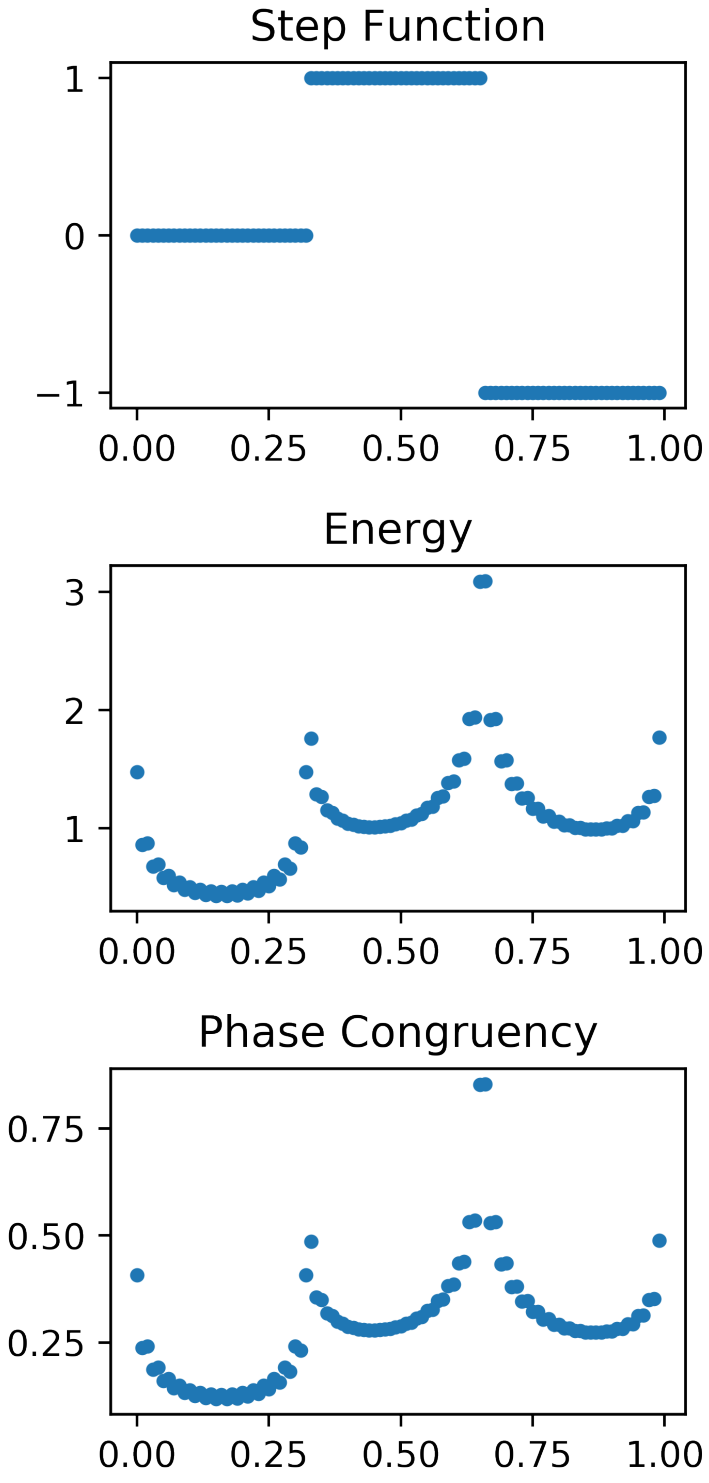


Figure 4.9: PC & Energy of a 1D Step Function

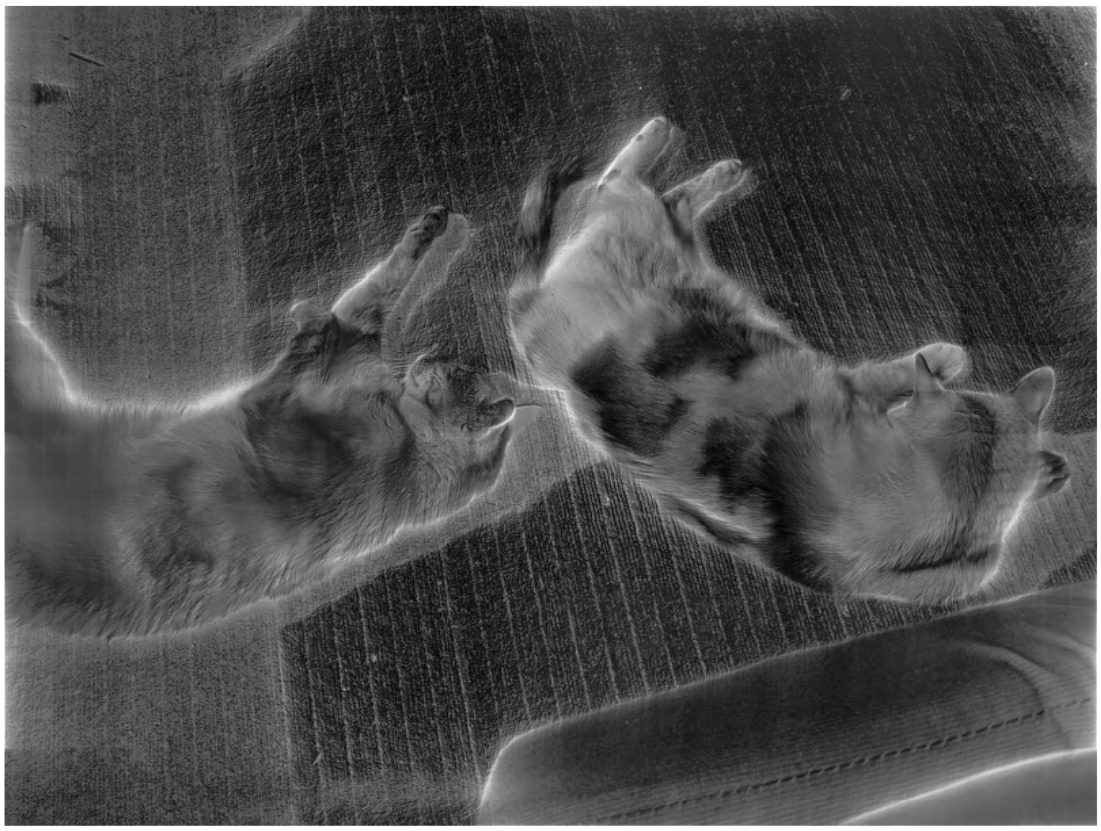


Figure 4.10: 2D PC of Cat Image. Note how phase congruency highlights edges and boundaries, by selecting for pixels for which phases are in alignment.

at the highest discontinuity, at the fall from 1 to -1, with smaller peaks at the step from 0 to 1, and at the start and end of the function. Note, however, that the values are unnormalized. When we divide by the sum of the Fourier amplitudes to compute the phase congruency (bottom), we have normalized values that sit between 0 and 1 but retain the same shape, with the same maximum and minimum values.

To expand the phase congruency analysis to images, we calculate the local energy independently in each x and y direction of the image using a 1D Fourier transform applied in parallel along one axis of the image. To independently analyze the PC for the horizontal and vertical directions, we divide each directional energy by the mean corresponding directional Fourier amplitude. Finally, we sum the two directional energies together at each pixel, and divide by the mean amplitude of 2D Fourier transform. This yields the 2D PC for the image (one PC value per pixel).

Other interpretations of 2D PC use azimuthal sums to calculate the 2D PC, measuring

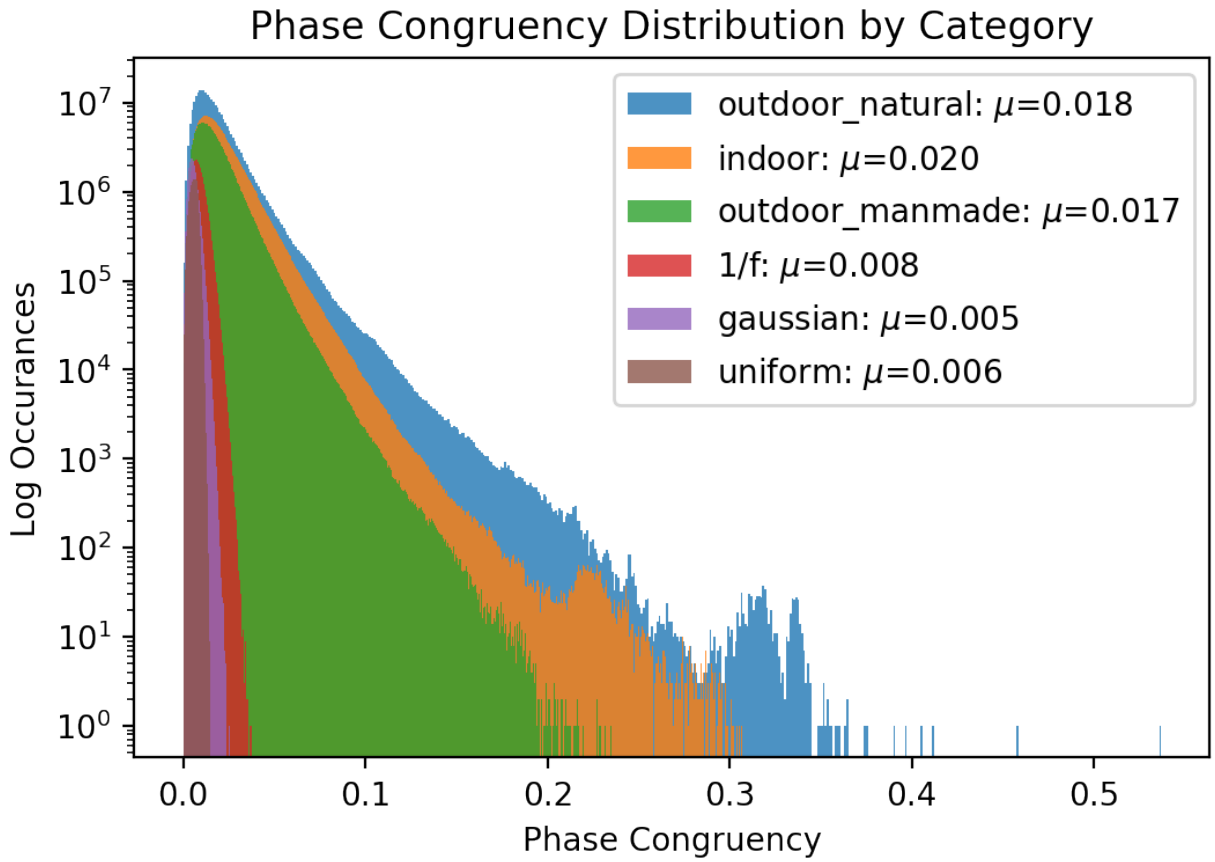


Figure 4.11: PC Distributions

phase alignment from all phases rather than just the x and y directions. These implementations have better localization and avoid the streak-like artifacts that appear in our two directional implementations (Figure 4.10). Avoiding these artifacts is crucial in certain contexts, such as using PC as an edge detector. In our case, however, for the computational feasibility of calculating the 2D PC over many images, we use the sum of the x and y directional PC values as the 2D PC.

Distributions of PC for Natural Images and Noise

To explore the feasibility of phase congruency to differentiate natural images from noise, as well as between different image categories, we apply the phase congruency calculation to the images in the categories from the Places dataset [108], shown in Figure 4.8. We calculate both the 1D (axial) and 2D PC at each pixel image and the mean axial and mean 2D PC value overall pixels for each image.

As an initial comparison of the PC for the different categories, their distributions are

visualized in Fig 4.11. When comparing the 2D PC values among the various image categories, despite sampling the PC for 1000 images per category, the mean values for PC are not significantly different among the three image categories. However, there does appear to be some separation, with indoor images having a slightly higher mean value for overall PC. The difference between distributions, however, is not large enough to classify one image based on its PC distribution. Results are similar for image-wise mean PC value distributions.

However, when comparing the distribution of all images as compared to all noise types, all the noise distributions have significantly lower values, even the 1/f distribution, which matches the natural images in the power spectrum. This can be explained in that images contain structure and edges, which contribute to phase congruency, while noise distributions have fewer sharp edges. Again, results are similar for image-wise mean PC value distributions. Mean PC values for natural images are above 0.1, whereas mean PC values for noise distributions were below 0.1. This result points to a baseline of phase congruency values as an invariant property of natural images.

Axial Phase Congruency

A characteristic difference between indoor/outdoor and natural/man-made images is in the variation of the amplitude falloff between the horizontal, vertical, and oblique axes [98]. Therefore, calculating horizontal and vertical phase congruency separately, variability may be present in the axial phase congruency for different image categories. The value of mean horizontal PC to mean vertical PC for each image and fit a line to these points is shown in Figure 4.12. The 1/f distribution has a very strong fit to the 1:1 line; this should be expected because there is no directional tendency in the noise distributions. However, PC values are so low for Gaussian and uniform noise they could not be fit and are not shown. The next closest fit to 1:1 is the indoor image category, with a slope of 0.54. Both outdoor image categories have poor fits, and their distributions look similar (this similarity could be due to poor labeling - see methods). Of note in this distribution is the strong tendency of all images, especially both outdoor categories, to fall far above the 1:1 line, indicating more PC as measured along the vertical direction (horizontal lines) than as measured in the horizontal direction (vertical lines). In addition, the images with the overall largest mean PC values are for indoor images.

4.4 Methods

Global Phase Analysis

All analyses were performed in Python3 [102] inside the JupyterLab environment [59]. For the global phase analysis of the individual images, we preprocessed 6 hand-picked 620x826 pixel color images of different types: two cats sunbathing on a textured carpet, a remote tropical beach, a running trail in the woods, an aerial photograph of a city, a portrait of a

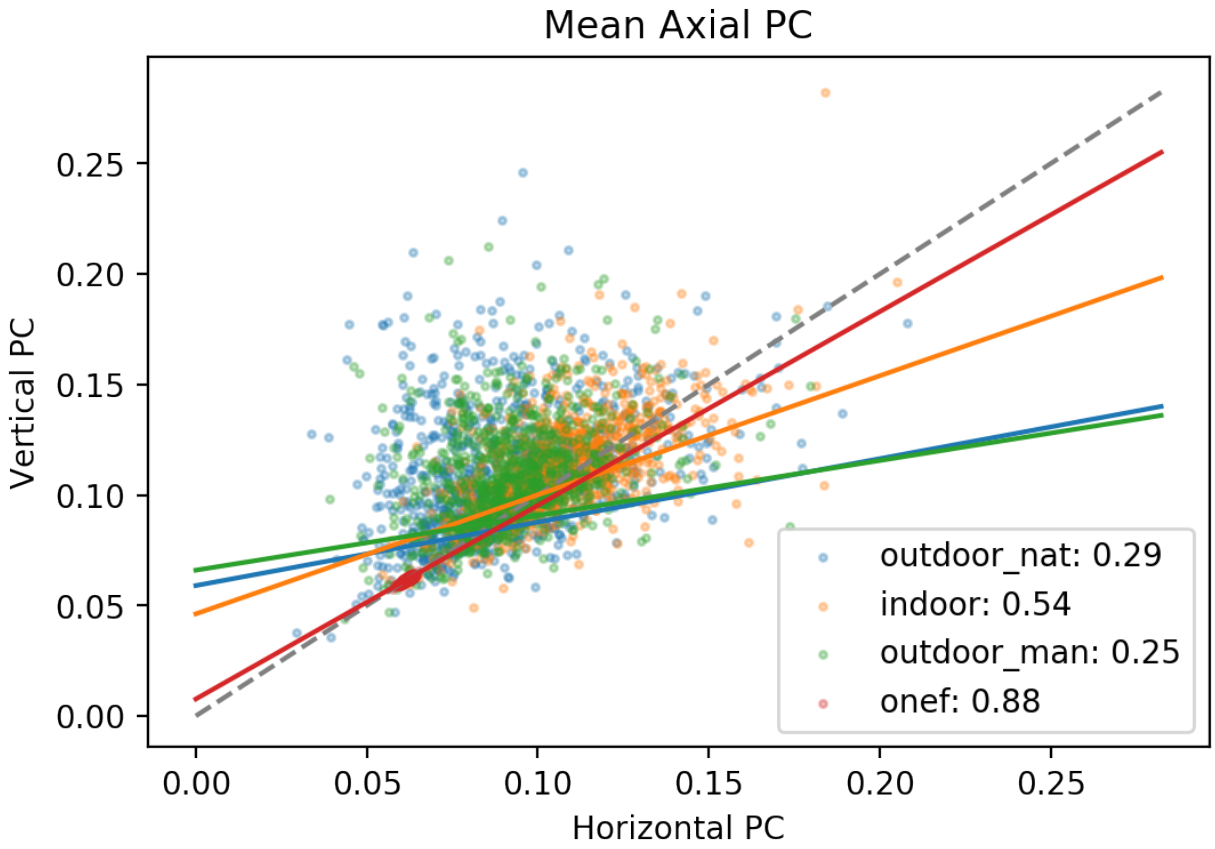


Figure 4.12: Axial PC Ratios

face, and a riverboat in front of a city skyline. All were captured on a Pixel2 smartphone and resized using Google Photos. We preprocessed these images by averaging the three color channels to produce a greyscale image. We then subtracted the mean of each image to center its pixel values about zero (removing the DC component), and divided by the mean value, resulting in the image being normalized between -1 and 1. Before taking the Discrete Fourier Transform (DFT), a cosine window was applied to the image, removing low frequency phase artifacts caused by abrupt image edges during the DFT calculation. Finally, the DFT was calculated, extracting the amplitude and real part of the phase spectra.

Local Phase Analysis

For the local phase (PC analysis) and the large-scale global phase calculations, images from the Places365 dataset [108] were used, which classifies images into the 'indoor', 'outdoor:man-made', and 'outdoor:natural' categories. Images used were only those classified as falling into only one of the these three categories, throwing out all others. Images were cropped to a uniform 512x512 size, averaged over the three color channels to get greyscale images, and

normalized to lie between 0 and 1. For the PC analysis only, no cosine window was applied, as this completely changed the shape of the local energy distribution. Then, 1000 images from each category were randomly sampled for the analysis. While overall, the categorization appeared mostly consistent, especially for outdoor vs. indoor, several images labeled outdoor:natural were noted to contain significant man-made structures. This may have reduced any difference noted between the two outdoor categories. Finally, phase congruency was calculated as described in the Phase Congruency & Energy section.

4.5 Discussion & Future Work

Global Phase

Global phase entropy does not appear to be a defining factor of natural images. In fact, with the textured carpet as an example, it appears that only particular man-made objects contain sparse representations and lower entropy in the Fourier domain, whereas completely natural (nature) images do not. Even in the specific subgroup of man-made images containing such structure, the change in entropy is minimal in the specific area of the phase plot where this structure is visible. While entropy of an image itself appears to be a somewhat helpful statistic in image space, the entropy of the phase does not appear to be. We suspect this is because the information is too spatially distributed. Future work could explore and verify this for the tested images' phase entropy measures.

Local Phase (PC)

The most notable result in this analysis is the difference in PC distribution values between images and noise. This difference in distribution points to an invariant image property. A potential extension is to sample from this distribution to synthesize new images that follow the same distribution. Images generated using this method may look more like natural images than 1/f noise, as they will contain edges. However, it is unclear if such an image synthesized from a natural image-like PC distribution will follow a 1/f amplitude spectrum or if this is an additional constraint that must be explicitly modeled. It is also unclear whether these generated images contain actual object-like structure with continuous edges or spatially distributed phase coherence.

A prerequisite to synthesizing such images is to parameterize the distribution of phase values such that this distribution can be sampled from. It is clear that the distribution of PC values in natural images is not Gaussian (Figure 4.11) - an exponential distribution may be more appropriate and should be explored.

Relatedly, a potentially fruitful line of exploration is an analysis of the spatial distribution of the PC values of images. Like the images themselves, PC values likely have a strong correlation with nearby pixels, but these correlations may be dependent on whether or not the PC value is high or not. For example, if a PC value is high, this is likely to be on an edge,

and that in one direction, the PC values at 2-neighbor pixels will also be high, and in the other direction, the 2-neighbor PC values will be low, or anti-correlated. By performing axial PCs individually, one may be able to predict which neighbors (vertical or horizontal) will be correlated and anti-correlated. By contrast, if the PC value is low at the reference pixel, nearby neighbors are also likely to be low on average and, therefore, positively correlated.

Another distribution that could be analyzed for comparison using these methods is the dead leaves model [66], which incorporates borders and occlusion with scale-invariance. We expect this model to more closely follow the PC distribution of real images rather than noise. This is because images generated using this method contain edges and borders that we expect to contribute larger PC values.

Chapter 5

Retinal Models of Natural Image Processing

5.1 Introduction

Among the computational tasks performed by the retina is the compression of visual information from the 6-7 million cone photoreceptors that detect incoming light, to the 1.5 million retinal ganglion cell fibers that convey this information to the brain, is one of the most important. We can think of the coding scheme used by the retina as an optimal adaptation that through evolution, is optimized to perform this compression, within a set of relevant biological constraints. In this context, the constraints for the case of spatial information we consider here are:

- The coding strategy must transmit relevant information in the visual signal as faithfully as possible to the brain for further processing (maximize information transmitted).
- The coding strategy must allow for an optic nerve small enough to allow for physical movement of the eyeball within the skull (limited number of neurons).
- The coding scheme must be robust to the limited precision of neural spikes (limited bandwidth; modeled as robustness to noise).
- The coding scheme must limit the number of spikes in order to minimize energy expenditure (limited number of spikes).

Through neural network modeling, we can study the relationship between these constraints and the optimal coding strategies to satisfy them by training networks with similar physical structure and loss functions reflective of these constraints. By adjusting the relative importance of these different constraints for the model through the choice of cost function, we can explore the landscape of these coding strategies. Finally, we can compare the properties of these coding strategies to the properties of the coding strategy seen in real biological retinas, which has been optimized through evolution rather than gradient descent.

In Karklin & Simoncelli, 2011 [56], the authors model this retinal compression task by training a single layer, linear/nonlinear neuron model on natural images to produce an output that conveys as much information as possible about the image. The model contains few constraints, all of them well biologically motivated. One main finding of this work is that when given the proper constraints, the model learns a spatial weight function with with properties that are strikingly similar to receptive fields of retinal neurons. As a first step towards extending this model into the temporal domain, we present work in reproducing this result, using Tensorflow [2], a popular software package for training neural networks.

Software packages such as TensorFlow and PyTorch [78] have gained popularity in recent years, in major part due to their ease of use, and specifically their use of automatic differentiation (autodiff, also sometimes referred to as autograd), in calculating the loss surface for backpropagation. While the automation and abstraction of these calculations lowers the barrier to entry, this opaque-ness can make debugging difficult.

This chapter describes an investigation into the mathematical underpinnings of training this model, with the aim of reproducing the same results. In particular, we discuss issues in the calculation and backpropagation of the error signal in the mutual information portion of the loss function. First, I describe and implement a method of circumventing some of these issues by reformulating the problem as an autoencoder; this effectively avoids the numerical precision issues by using reconstruction as a proxy for mutual information. I then dive back into the original issue in the mutual information calculation, and discuss numerical precision issues in this calculation that lead to difficulties in reproducing this result. Specifically, in calculating the inverse, and later the determinant of a singular matrix as a part of the mutual information calculation. In addition, I show that a Tensorflow implementation of the Moore-Penrose pseudoinverse would also work to estimate the true inverse.

Unlike other chapters in this thesis that highlight the final product and results of completed research, this chapter intentionally steps through the debugging process for a neural network model, navigating the steps, pitfalls, and solutions along the way to the desired final product. My hope is that this walk-through of the process, emphasizing the importance of simplifying the problem, and taking the time to understand the mathematical fundamentals that are abstracted away by neural network software, may be helpful to future students facing similar issues.

5.2 Model Description

In the Karklin & Simoncelli model [56], the early visual system is modeled using a set of 100 linear/nonlinear neurons to encode 16x16 pixel natural image patches, compressing them from a 256-dimensional space to a 100-dimensional space, in a manner reminiscent of the retina. In the model (Figure 5.1), natural image patches x are combined with static Gaussian noise n_x. These are fed to linear weights w which are trained, yielding linear output y_i. This in turn is fed to nonlinear functions f (f_i in Figure 5.1), which are learned in addition to the weights. These nonlinear functions are parameterized as mixtures of 500 Gaussians, so

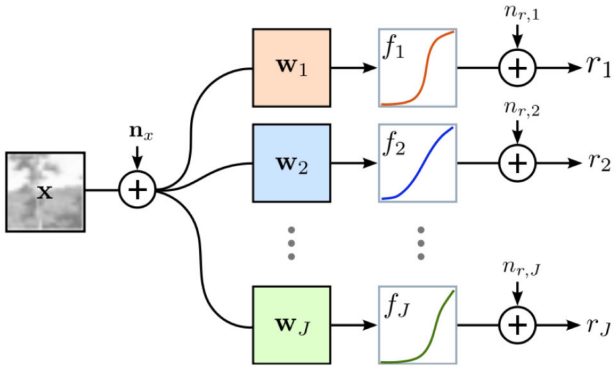


Figure 5.1: Linear-Nonlinear model from [56].

the shape of the resulting activation curve can be learned. Finally, output noise n_r is added to the output of f, and r measured. The response $\mathbf{r}^i$ of each neuron i is calculated as:

$$y_i = \mathbf{w}_i^T(\mathbf{x} + \mathbf{n}_x)$$
$$\mathbf{r}^i = f_i(y_i) + n_r$$

The nonlinear responses f_i are approximated via a first-order Taylor approximation by $\mathbf{G}^i$, a diagonal matrix where each diagonal entry is the derivative of the response of each nonlinear function at the output y from the filters $\mathbf{W}$, $f_i(y)$. In addition, a vector of learned offsets f_o are added to the output.

$$\mathbf{r}^i \approx \mathbf{G}^i\mathbf{W}^T(\mathbf{x}^i + \mathbf{n}_x^i) + \mathbf{n}_r^i + \mathbf{f}_o^i$$

The model is trained using gradient descent to maximize an objective function which maximizes mutual information between the image and the response $I(X;R)$. Mutual information is defined as the difference between the entropy of the input X and the conditional entropy of the input given the response R:

$$I(X;R) = H(X) - H(X|R)$$

The model also works to maintain a low overall spiking rate of the output $< r_j >$. These two quantities are weighted together using the constant λ as a trade-off parameter, with λ adjusted to achieve an average response value r of one per neuron per image. This gives the following cost function:

$$F_{cost} = -I(X;R) + \sum_j \lambda_j \langle r_j \rangle$$

$$F_{cost} = -H(X) + H(X|R) + \sum_j \lambda_j \langle r_j \rangle$$

Because the global input entropy $H(X)$ does not depend on the model, mutual information is maximized by minimizing the conditional entropy $H(X|R)$. The cost function becomes:

$$F_{cost} = H(X|R) + \sum_j \lambda_j \langle r_j \rangle$$

The conditional entropy $H(X|R)$ can be approximated to first order using the determinant of the covariance matrix $\mathbf{C}_{x|r}$ (see paper for a more detailed derivation).

$$H(X|R) \approx E\left[\frac{1}{2}\ln\left(2\pi e \det(\mathbf{C}^i_{x|r})\right)\right]$$

The posterior $\mathbf{C}^i_{x|r}$ is in turn calculated by the covariance matrix $\mathbf{C}_{r|x}$ of the prior, $p(r|x)$, multiplied by the weights $\mathbf{W}$ and the diagonal matrix of the nonlinear responses $\mathbf{G}$, and added to the inverse of the global covariance matrix of the input images $\mathbf{C}_x$.

$$\mathbf{C}^i_{x|r} = (\mathbf{C}_x^{-1} + \mathbf{W}\mathbf{G}^i(\mathbf{C}^i_{r|x})^{-1}\mathbf{G}^i\mathbf{W}^T)^{-1}$$

This is in turn calculated by combining the global covariance of the images $\mathbf{C}_x$, the weights $\mathbf{W}$, the slope of the activation function for a given image $\mathbf{G}$, as well as the co variances of the input noise $\mathbf{C}_{nx}$ and of the response noise $\mathbf{C}_{nr}$.

$$\mathbf{C}^i_{r|x} = \mathbf{G}^i\mathbf{W}^T\mathbf{C}_{n_x}\mathbf{W}\mathbf{G}^i + \mathbf{C}_{n_r}$$

This leads to a total cost function to be minimized:

$$F_{cost} = E\left[\frac{1}{2}\ln\left(2\pi e \det(\mathbf{C}^i_{x|r})\right)\right] + \sum_j \lambda_j \langle r_j \rangle$$

Where:

$$\mathbf{C}^i_{x|r} = (\mathbf{C}_x^{-1} + \mathbf{W}\mathbf{G}^i(\mathbf{C}^i_{r|x})^{-1}\mathbf{G}^i\mathbf{W}^T)^{i^{-1}}$$

$$\mathbf{C}^i_{r|x} = \mathbf{G}^i\mathbf{W}^T\mathbf{C}_{n_x}\mathbf{W}\mathbf{G}^i + \mathbf{C}_{n_r}$$

This training can be done using typical neural network training methods. An image x is fed into the network, resulting in a response vector r and slope of activation G. The cost function F_{cost} is then evaluated based on these values.

By training this network on many batches of natural image patches, and adding the correct amounts of noise, the authors obtained weight vectors that share many similarities to neurons in the retina. First, each of the 100 individual weight vectors (corresponding to the filter of one hidden layer unit or 'neuron' in the input image space) is spatially localized in one small area of the image (Figure 5.2). Second, the weight vectors self separate into two populations, one selective for light patches with dark annuli surrounding them (ON-center), and the other population selective for dark patches with light annuli surrounding them (OFF-center). Finally, all the weight vectors in a given population, put together, tile

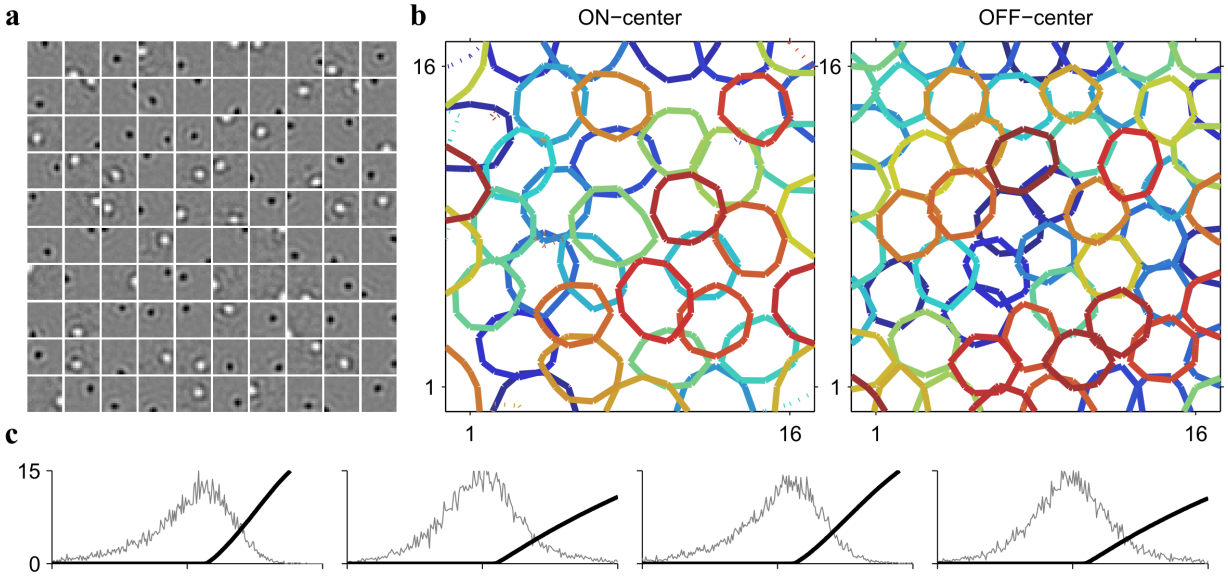


Figure 5.2: Expected Results - Two neural populations [56].

the entire space of the input image, causing the network to code both the 'on' and 'off' signal in parallel.

5.3 An Autoencoder Solution

The goal here is to reproduce these properties of the weight vectors in this model. However, as we shall see in the next section, there are numerical instabilities inherent to the mutual information calculation in the loss function. Before diving into solving this issue however, we will see that this problem can be avoided. Because these instabilities come about as part of the mutual information calculation, a reformulation of the problem allows us to avoid this calculation altogether. This is achieved by formulating the problem as a 2-layer autoencoder, using image reconstruction as a proxy for mutual information.

This auto-encoder network (Figure 5.3) adds a simple nonlinear layer to the output of the encoder, re-expanding the network from 100 nodes, back to the original 256, the dimension needed for a reconstruction of the original image patch. The loss function for this model then, is simply the reconstruction error between the original image path and this reconstruction. Here, we use the L2 norm.

This autoencoder model is trained with the following cost (loss) function:

$$F_{cost} = ||X_{recon} - X_{orig}||_2 + \lambda_{weight}\langle w_j \rangle + \lambda_{spike}\langle r_j \rangle$$

The first term,

$$||X_{recon} - X_{orig}||_2$$

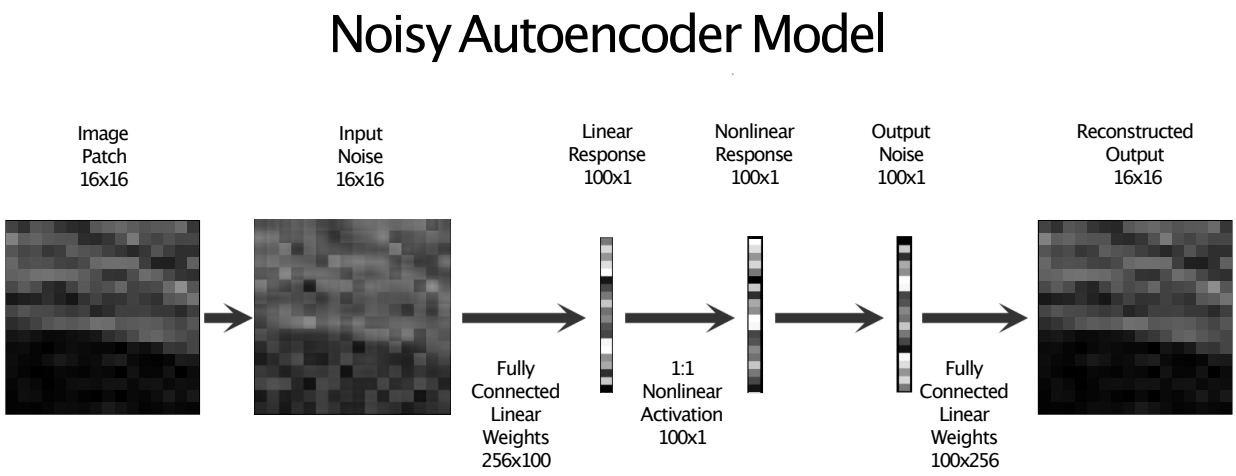


Figure 5.3: Simple Single Layer Autoencoder with added noise allows a loss function based on reconstruction of the image patch, rather than based on mutual information between the image patch and the neural representation. This is achieved by simply adding a set of a fully connected linear weights which learn to reconstruct the 16x16 image patch (shown). Optionally, a non-linear response can be added onto the output to increase the power of the reconstruction network.

signifies the L2 norm on the reconstruction of the image patch. The second term, $\lambda_{weight}\langle w_j \rangle$ is a soft weight constraint, to keep the weight vectors from imploding/exploding. The final term, $\lambda_{spike}\langle r_j \rangle$, is the activation constraint, encouraging the solution to have a low number of spikes (energy constraint).

This model can be trained using Tensorflow, which uses autodiff to calculate the derivative of this cost function with respect to the weights $\frac{dF_{cost}}{dW}$, and takes a small step in the direction of minimizing the cost function by slightly changing the weight values accordingly using backpropagation. This entire process is then repeated for a new batch of images, for the desired number of iterations.

5.4 Revisiting Mutual Information: Instabilities

While the autoencoder solution avoids numerical stability issues, it also introduces it's own additional issues and biases. First, it is at best debatable that the brain is actually attempting to reconstruct the image signal from the retina pixel-for-pixel. Even if this is the case, or that reconstruction is an appropriate stand-in for mutual information, an L2 loss is not necessarily appropriate for measuring the quality of such a reconstruction. Furthermore, the brain is not limited to a simple, single linear layer to reconstruct the input image patch, as in this model, though from an efficient coding perspective, a simple reconstruction would likely be preferable. While the addition of a pointwise non-linearity to the linear output

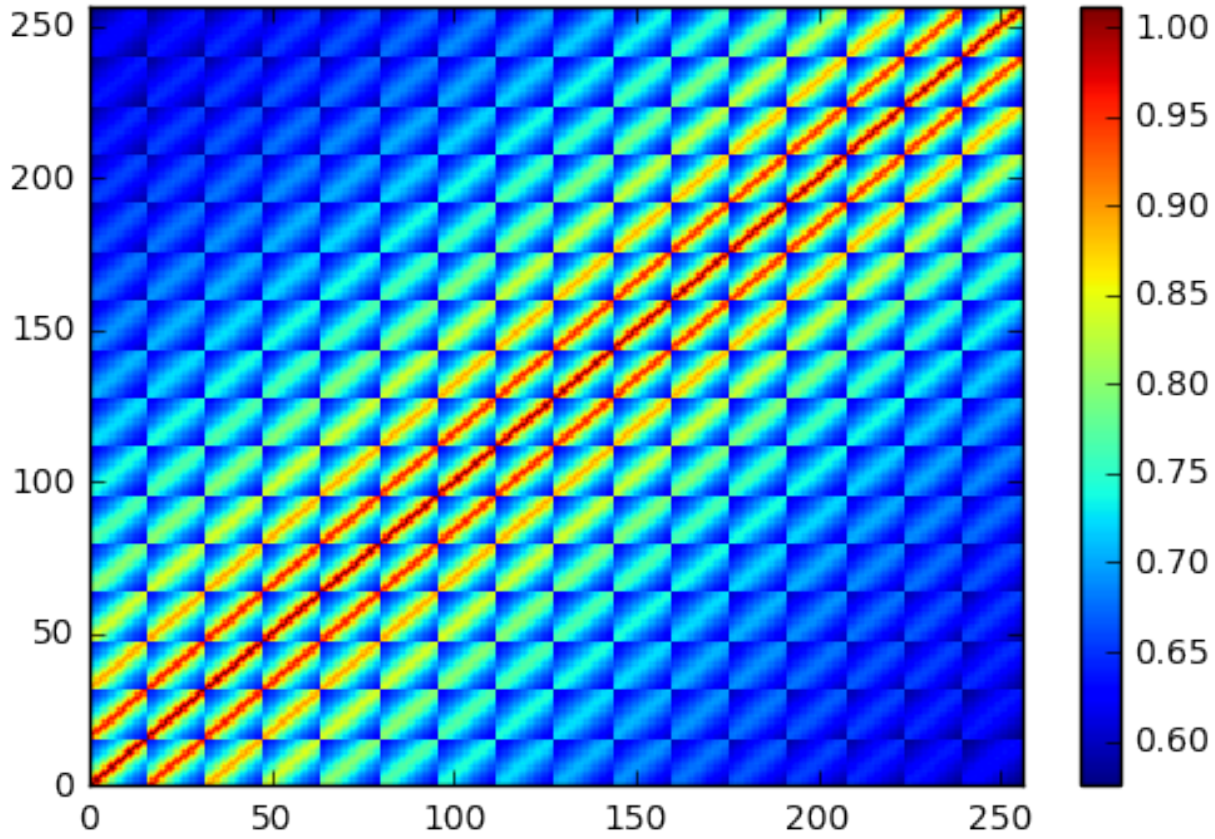


Figure 5.4: $\mathbf{C}_x$, Covariance of Input Images

of the reconstruction weights is possible, one must take care as the approximately normally distributed ground truth image patch with values in the range [-3,3] has negative pixel values that cannot be reconstructed with the output of a standard Rectified Linear Unit (ReLu), or Sigmoidal function (Sigmoid), which have positive-only outputs. It is worthwhile then, to revisit the mutual information formulation, and diagnose the issues, so they can be addressed in the context of the original model.

When first implementing this model in Tensorflow, one runs into the following error: "Cannot compute inverse, determinant is zero to working precision." Upon further investigation, this attempt to invert a matrix with zero determinant happens in the inversion of the $\mathbf{C}_{x|r}$ term, for which the determinant is calculated in the equation for $H(X|R)$.

Mathematically speaking, for a square matrix which $\mathbf{C}_{x|r}$ is, there are only a few reasons for a zero determinant. Either an entire row is zero, two rows or columns are equal, or a row or column is a multiple of another. Computationally speaking however, there is another potential cause of a determinant evaluating at zero. If the values in the matrix are small enough such that that when multiplied together (as is done in the calculation of the determinant), they become evaluated at zero by the computer. Given the small numbers in

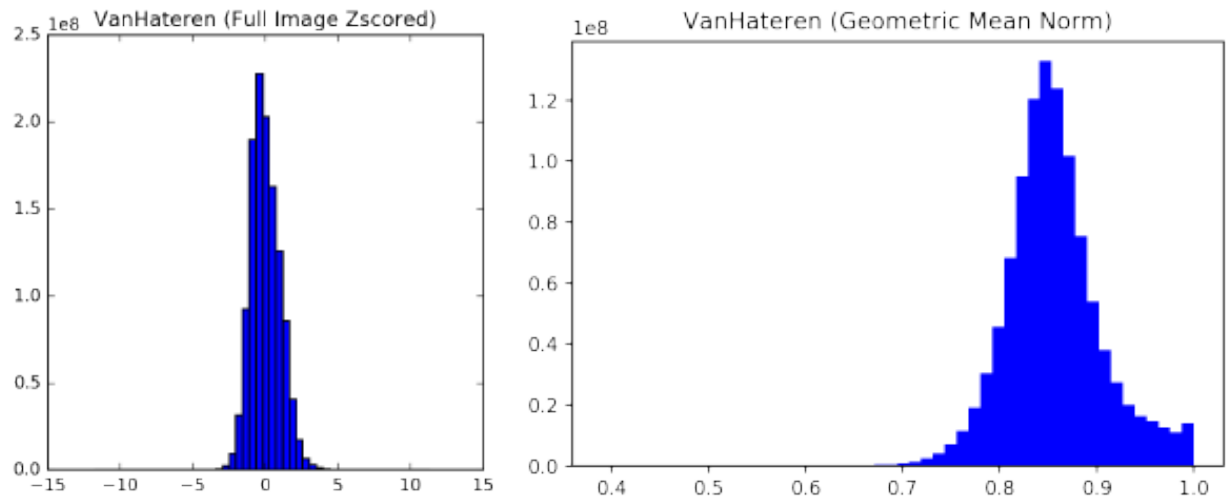


Figure 5.5: Image Pixel Distributions

these matrices, and the large 126x126 size of $\mathbf{C}_{x|r}$, the last of these is the likely culprit.

Looking at the calculation of $\mathbf{C}_{x|r}$ itself, we see that the root of this problem may lie in $\mathbf{C}_x$, the covariance matrix of the input images having too small values, though $\mathbf{C}_x$ appears as expected for vectorized image patches (Figure 5.4). Specifically, because the original 16x16 image patches are vectorized to 256, the covariance matrix appears as 16x16 blocks of the 256 pixel strips correlated highly within themselves and the patches surrounding them along the larger diagonal, and with distant strips less correlated to each other.

5.5 Understanding a Single Example

The approach to solving the problem is to simplify it as much as possible. We take the calculations out of the difficult to debug Tensorflow graph framework, implementing it directly in numerical python (NumPy) [48]. This allows us to use print statements withing the calculation to easily query various steps within the calculation, as well as keep calculated values in the global scope for later accessing. Another simplification is to remove the batch dimension, reducing all the dimensions by one, and ensuring there were no mistakes caused by batch application of matrix operations. This allows us to first work through the simpler process of calculating the mutual information for one image patch by itself, and then simply put that in a loop to sample multiple images.

To address this problem, we run these calculations outside of the Tensorflow graph, and also with a new dataset (described below). This immediately gets past the error of the determinant of $\mathbf{C}_{x|r}$ term being zero, but causes an overflow error. The source of this overflow error is identified by deconstructing the $C_{x|r}$ term. There are some values in the C_x matrix that are very small. In the process of doing the full $C_{x|r}$ calculation, C_x ends up being inverted a total of 3 times, and this is causing values in $C_{x|r}$ in turn to be relatively

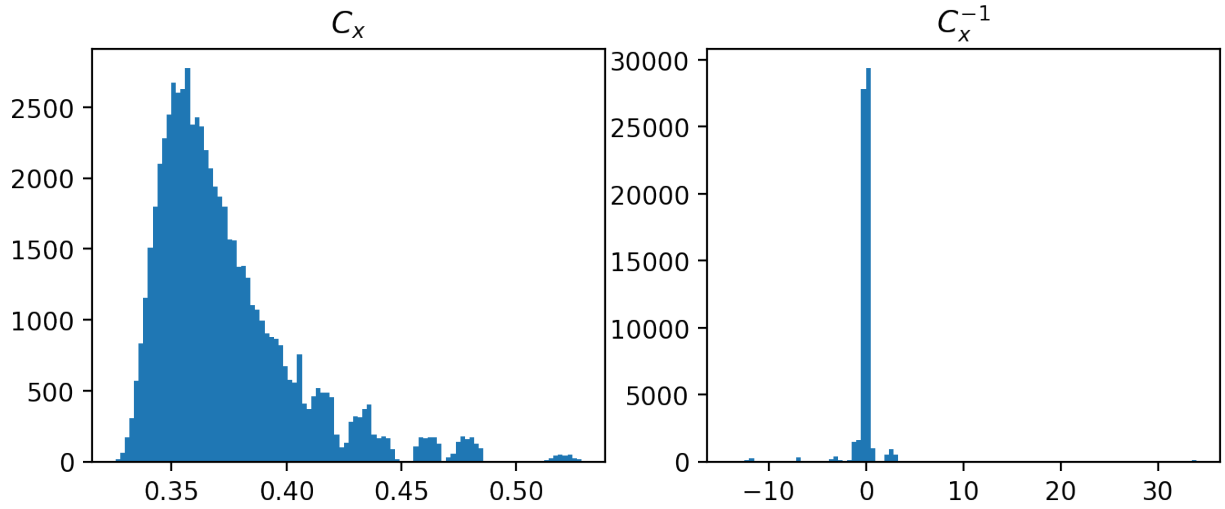


Figure 5.6: Inverting Matrix C_x with small values

large. When the determinant of $C_{x|r}$ is computed - which multiplies many of these large values together - there is a numerical overflow.

The reason for this discrepancy is uncovered by turning to the distribution of the image set. The original image set's pixel values is log-normed and then z-scored during preprocessing, giving a Gaussian distribution with mean zero. An alternative normalization technique is normalizing by the geometric mean. This yields a similarly shaped distribution, but with all values between zero and 1 (Figure 5.5). This change barely alters the shape of the distribution of C_x and $C_{x|r}$, but changes their ranges dramatically (Figure 5.6), and the overflow error in the determinant persists. This distribution changes slightly with the exact image set chosen, but the normalization has a huge effect.

This relatively small change in input resulting in a huge numerical change at the output is reminiscent some N-R solvers. All the matrix inversions cause the results to implode/explode, finally with the determinant causing either numerical overflow or underflow, with underflow coming in the form of a matrix being singular to working precision. However, this instability in the calculation can be pinned down to the magnitude of the covariance matrix being either below or above 1. To address this, we normalize the distribution, calculating the correlation matrix instead of the covariance matrix, putting all values in the range [0,1]. It seems plausible then that the original paper may calculating the covariance matrix rather than the correlation matrix.

5.6 Matrix Identity for Increased Numerical Stability

While image normalization stabilizes the value of C_x^{-1}, fixing the overflow error, the determinant calculation remains unstable, as some of the values in C_x^{-1} end up being fairly large,

on the order of 10. This in turn causes the determinant of C_x^{-1} to not be finite. To handle this problem, we turn to the matrix identity [74]:

$$\ln(\det(a)) = tr(\ln(a))$$

With regards to numerical stability, calculating the trace of a matrix is preferable to the determinant. Using this identity to replace the determinant calculation for the trace, we can modify the cost function as follows to give more numerical robustness:

$$-H(X|R) = -E[\frac{1}{2}\ln 2\pi e \det(\mathbf{C}^i_{x|r})]$$

$$-H(X|R) = -E[\frac{1}{2}(\ln(2\pi e) + \ln(\det(\mathbf{C}^i_{x|r})))]$$

$$-H(X|R) = -E[\frac{1}{2}(\ln(2\pi e) + tr(\ln(\mathbf{C}^i_{x|r}))]$$

$$-H(X|R) = -E[\frac{1}{2}\ln(2\pi e) + \frac{1}{2}tr(\ln((\mathbf{C}^i_{x|r})))]$$

Because we care about the ***gradient*** of $-H(X|R)$ rather than the value itself, the constant $\frac{1}{2}ln(2\pi e)$ term vanishes when taking the gradient of both sides:

$$-\nabla H(X|R) = -\nabla E[\frac{1}{2}tr(\ln(\mathbf{C}^i_{x|r}))]$$

Giving the following cost function:

$$F_{cost} = tr(\ln(\mathbf{C}^i_{x|r}))$$

5.7 Transitioning to Tensorflow

Using this trick finally gets past the non-finite determinant issue. The true test however, is implementing the equation within a graph, and testing if the equation truly calculates mutual information. If so, a neural network that uses this equation as it's loss function should become trained. That is, it's cost over time should go down, and it's weights should change. Ideally, the network's weights should reproduce the results from Karklin & Simconcelli, 2011.

Towards this end, we convert the single iteration NumPy code into a Tensorflow graph, and implement batch training, using the modifications described in earlier sections. Implementing batching was a bit tricky, as matrix multiplication in Tensorflow does not yet support broadcasting. To solve this, a vectorized batch matrix multiply function, while possible, is extremely slow, with 10 iterations taking on the order of minutes, even on a GPU. To speed things up, we implement batched matrix multiplication using Einstein summations, which are significantly faster, and utilizing them makes training time practical. They

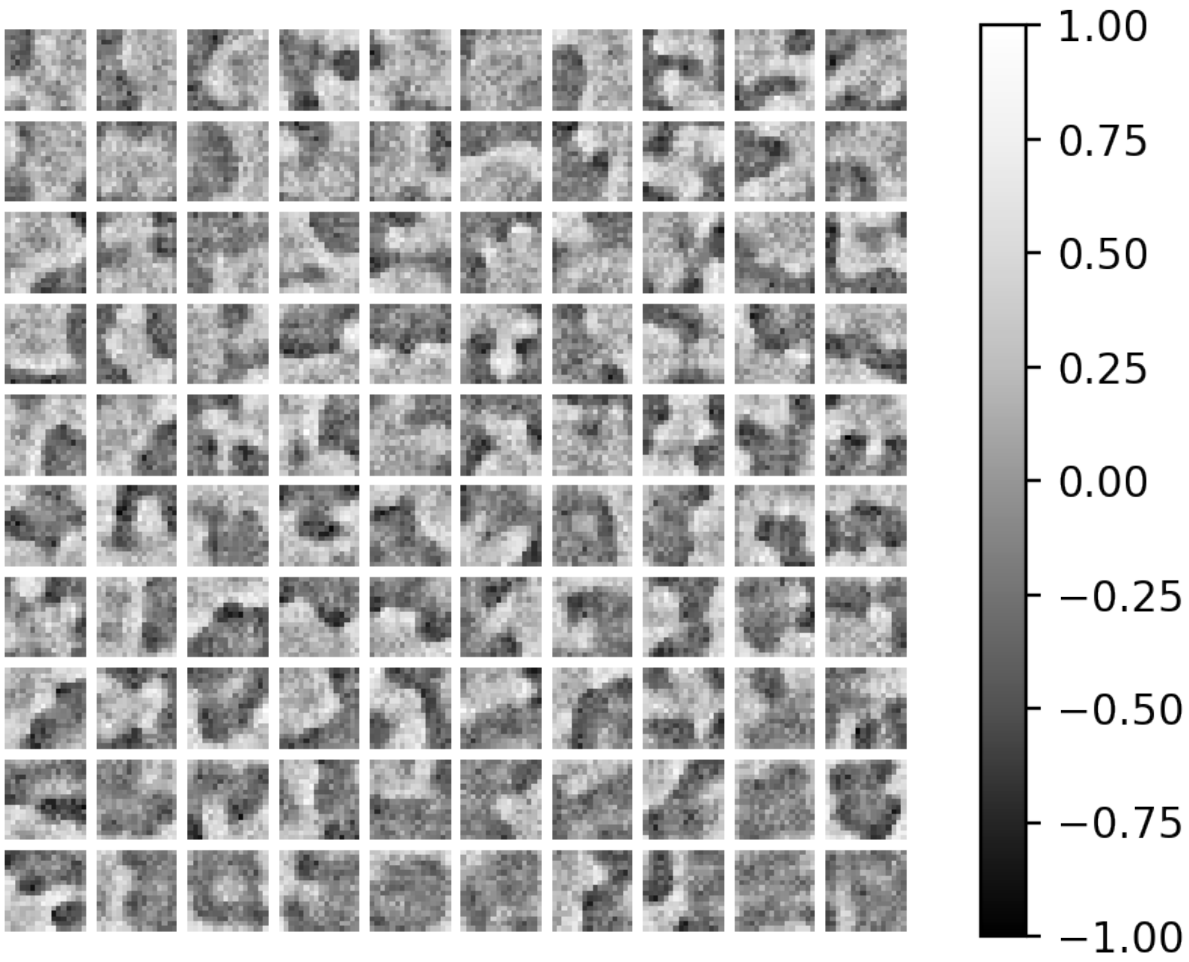


Figure 5.7: Initial Structure in Weights

also result in much cleaner code overall; multiple batched matrix multiplications which included transposes becoming brief one-liners. Take for example the calculation of C_{rx}, which becomes:

```
crx = tf.einsum('ijk,lk,lm,mn,ino->ijo' G, w, cnx, w, G) + cnr
```

5.8 Weights Becoming NaN

With these changes we can begin training the mutual information model. Structure begins to emerge in the weight matrix (Figure 5.7). This is a strong sign that the cost function is indeed meaningful, and the network is learning a set of weights that will extract information from the natural images to minimize the cost (maximize mutual information). This is good

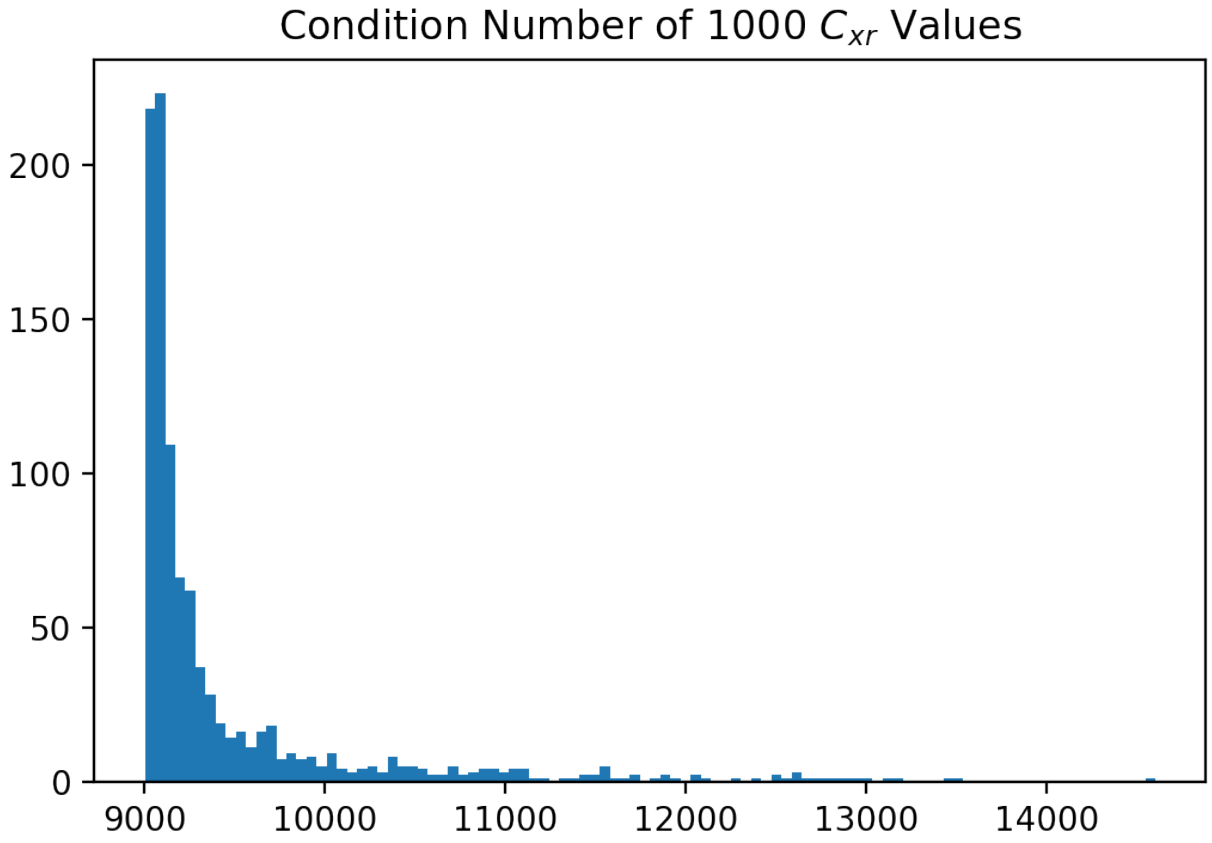


Figure 5.8: Poorly Conditioned Matrices

news, however we soon run into a different problem. After just over 300 iterations each of 1000 batched training examples, the values in the weight matrix become NAN. The number of successful iterations before this error happened depends on the chosen learning rate, but a learning rate small enough to avoid this error is not sufficient to train the network.

5.9 Poorly Conditioned Cost Function

In neural network training, this problem of weights becoming NAN is often times due to a poorly conditioned matrix. If such a matrix is present, a small learning rate r will be appropriate for the parts of the matrix with small eigenvalues, but a large learning rate is necessary to train the parts of the matrix with large eigenvalues. This comes from the backpropagation calculation on a poorly conditioned matrix M, calculated as $\frac{\delta Cost}{\delta M}$. Large learning rates will quickly cause the parts with smaller eigenvalues to become NAN, and smaller learning rates will never be able to train the larger parts of the network.

To determine if this may be the problem, we compute the condition number of matrix

C_{xr}, the matrix used to calculate the cost function. As this is not implemented in Tensorflow, we do this manually, calculating the ratio of the maximum to the minimum eigenvalues of C_{xr}. We use SVD for each individual instance of C_{xr} This greatly slows computation - as SVD operation on 1000 256x256 matrices for each of 300 runs is computationally expensive. Eventually, we find that indeed, the condition numbers are all poor, all above 9000 (Figure 5.8). Importantly, this distribution has a very long tail on the high side, meaning some condition numbers were above 14,000. It makes sense that once in awhile, an extremely high condition number would cause a divide by zero and a resulting NAN, that would propagate through to the weights.

Plotting the evolution of the condition numbers over training however, while one might expect that over training, the condition number increases over time until the matrix is so poorly conditioned that the weights became NAN. Instead, what we see is that the average condition number actually decreases over time, from an average well above 9000, then lowers over training until it bottoms out at around 9000. It seems then, some other constraint on the network is to blame, forcing the condition number to be above 9000, but not lower. And without the ability to go lower, the network cannot train any further.

5.10 ReLu & a Weight Constraint

One important simplification in this implementation as compared to the original paper is to use a predetermined off-the-shelf point-wise nonlinear function, rather than the reported method of learning a generalized nonlinearity as estimated by a sum of Gaussians. Up until now we have used a sigmoidal function. We note that the nonlinearities learned by the network appear to be less sigmoid looking, and more like a rectified linear unit (ReLu) (Figure 5.2. It stands to reason that we may want to use a ReLu function instead of the sigmoid, in order to better reproduce these results.

Applying this weight regularization as a soft constraint, we get the following cost function:

$$F_{cost} = tr(ln((\mathbf{C}^i_{x|r}))) + \lambda_{weight}\langle w_j \rangle + \lambda_{spike}\langle r_j \rangle$$

5.11 Moore-Penrose Pseudoinverse of C_{xr}

Finally, making this change we began to see localized, discrete filters that separated into on and off-like weight vectors, and appear to somewhat tile the input image space (Figure 5.9). Indeed, the network is learning something like mutual information, and we are close to reproducing the paper's results. However, the filters are not of similar sizes, and more importantly, do not show the center-surround shape we would have expected. Furthermore, these results are always be followed by a NAN error, as before.

Though it is future work to determine why the ReLu works so much better than the sigmoid, it is possible that the sigmoid may cause the network to be over constrained in

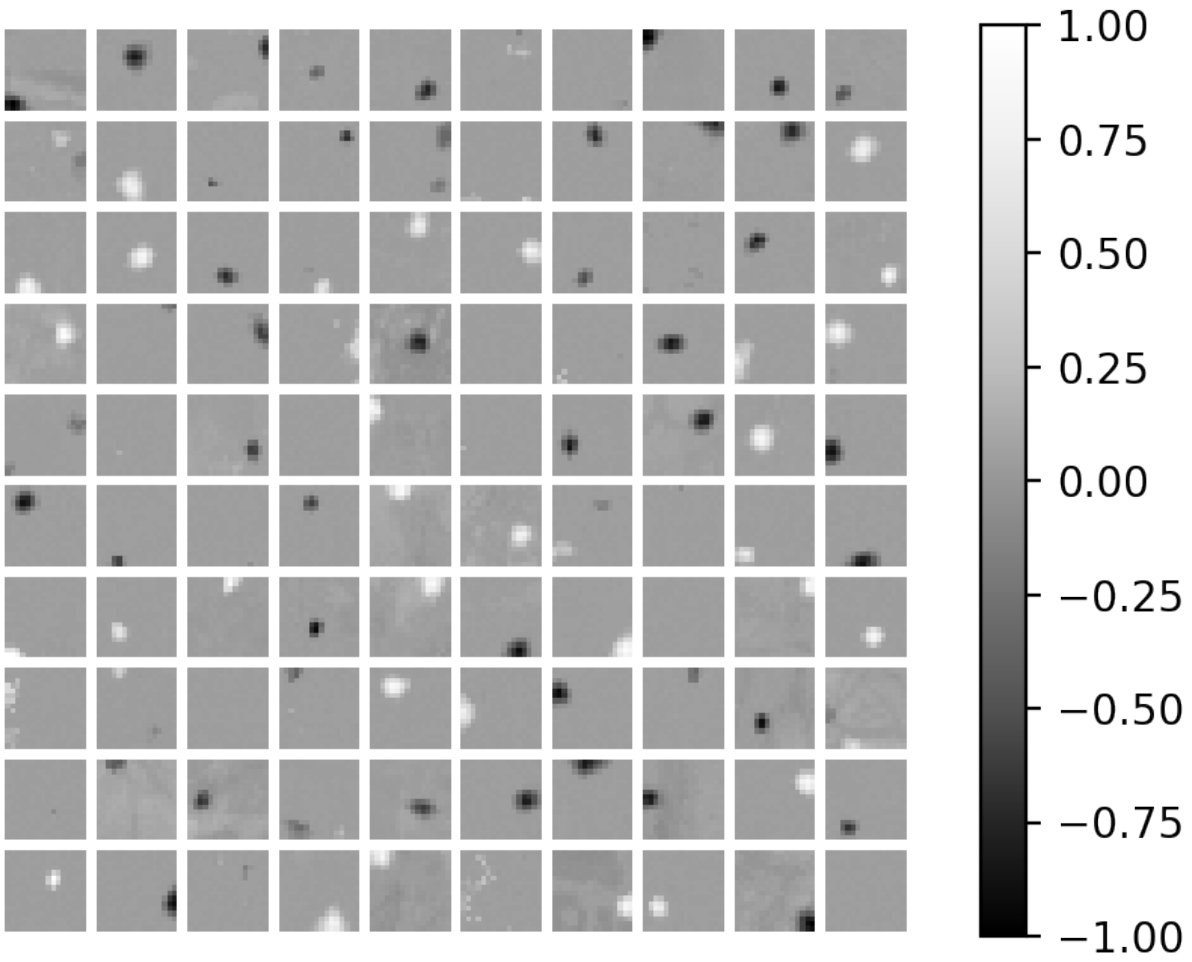


Figure 5.9: Localized, on and off Filters

some way, as the activation output of a neuron from a sigmoid function by definition cannot go above 1. The activation constraint λ encourages activations to be on average, one per neuron per image. With a sigmoid, this would mean all neurons are always maximally activated for all images, which is not an information dense code. The ReLu however, is unbounded in the maximum value of activation it can have, which would give the network enough freedom to learn an information dense code while maintaining the activation average of 1 per neuron per image.

Also of note is the improvement of the condition number of C_{rx}. While it still follows the pattern of improving over training, the value now starts and ends at a much more reasonable number, on the order of 300-500 (Figure 5.10).

There is an additional weight constraint implemented in the paper that we so far have not included in our model. Specifically, each of the 100 weight vectors $\vec{w_i}$ of W was con-

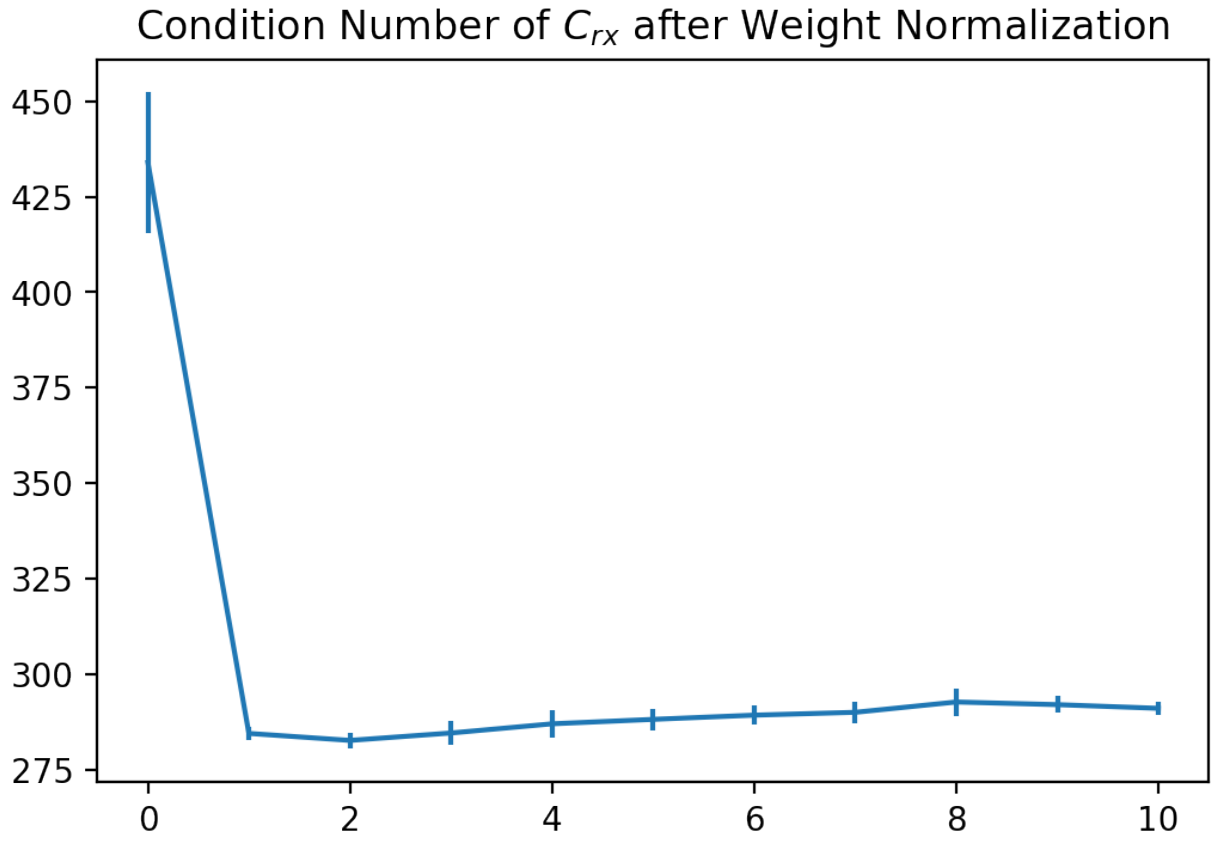


Figure 5.10: Condition Number Improves with Weight Constraint

strained to have $\|\vec{w}_i\| = 1$. This serves as a form of regularization that could keep values from running away. Indeed, adding this normalization ended the NAN error. In our implementation however, this constraint causes weight vectors to lose their spatial localization seen before; they appear more like the original distributed filters. Despite various hyperparameter adjustments (changing input/output noise, activation constraints, L1 vs L2 norm, etc), localized weights do not seem to emerge. Exploring this is left to future work.

Once the network is training and learning the mutual information calculation using the ReLu network, the final step is to implement the Moore-Penrose pseudoinverse [10], and try using it in place of the matrix inverse for the C_{xr} inverse calculation. If the inverse calculation is the same, it should give the same result. We implement it using the SVD method [79], which goes as follows for calculating A^+:

1. Calculate SVD on A: $U\Sigma V^* = A$
2. Calculate Σ^+ is the inverse of each diagonal value, except zeros, which remain zero.
3. Calculate A^+: $V\Sigma^+ V$

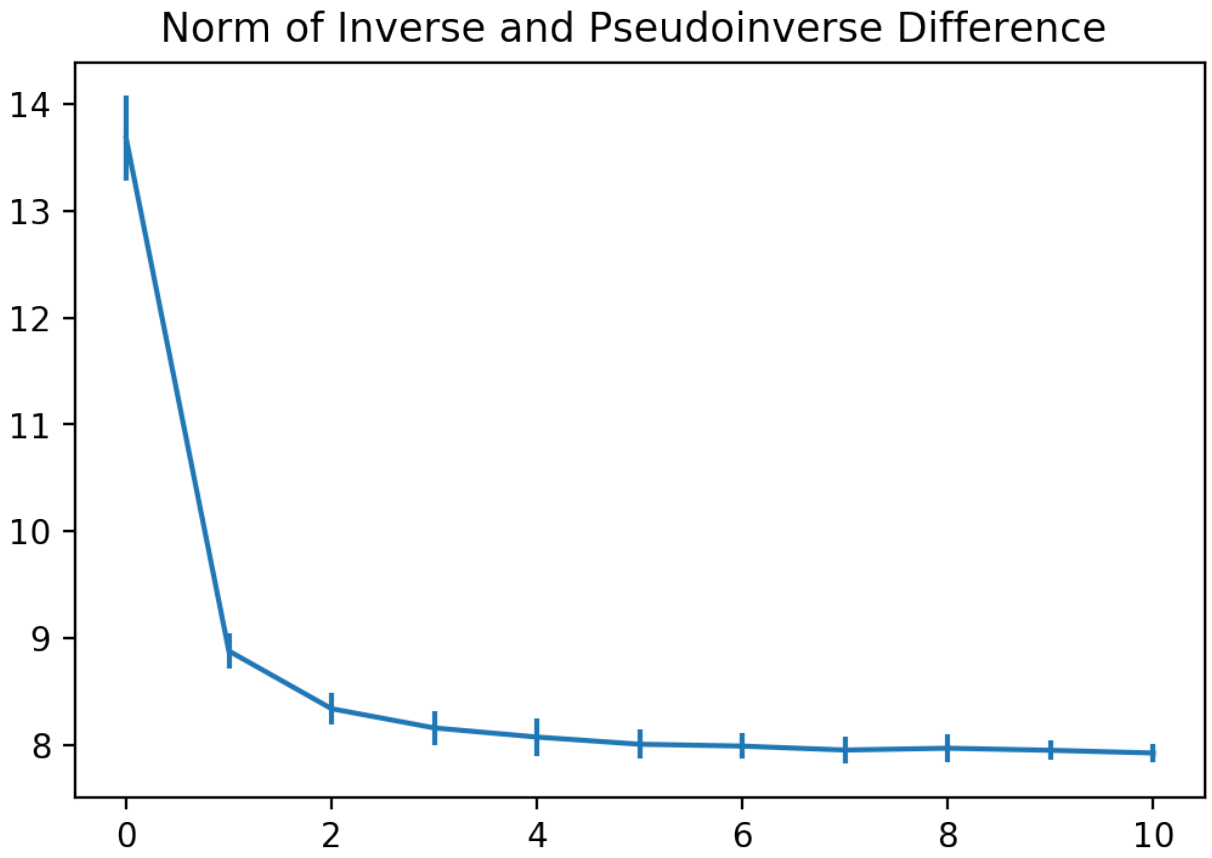


Figure 5.11: Difference between C_{xr}^{+} and C_{xr}^{-1} Decreases

We use this method to calculate C_{xr}^{-1} within the Tensorflow graph, and use it's value as the input to the cost function. However, this causes another major roadblock. The Tensorflow gradient descent optimizer cannot propagate the gradient through the SVD calculation. As of 2015, this functionality has been requested but has not yet been implemented. This problem does not arise when calculating SVD to investigate the condition number, as this value was simply reported, and not used for the optimization.

Another option is to use an alternative method for the pseudoinverse calculation, such as the Ben-Israel & Cohen iterative method. Rank decomposition and QR methods however, would likely run into the same problem of gradient descent not being defined over them.

In order to confirm that the calculated pseudoinverse is correct, we use the true inverse for the training, and report back the Frobenious norm for the difference between the two values. This allows us to do the SVD calculation without putting it in the cost function. If the calculations are equivalent, we expect this value to be low. Indeed, they are fairly small considering the size of C_{xr} and condition number, around 8 by the end of training (Figure 5.10). Interestingly, the norm decreases over training, and that the decrease followed the

same pattern as the condition number (Figure 5.10). This decrease in condition number may cause the decrease in the distance between the two methods. This would make sense if C_{xr} has some very small values, as these are the ones that will be inverted and become very large in the true inverse, when the true inverse would have just reported zero.

5.12 Conclusions

We step through the process of reproducing the result from [56], debugging implementation issues, particularly those related to computational instabilities which are exposed by autodifferentiation software in Tensorflow. Future work is in the direction of reproducing this work. One option is to take the derivative of the cost function by hand, and train the network by applying this gradient manually at each training step. This would avoid the instabilities cause with auto-differentiation software such as Tensorflow. Another area of exploration is in additional constraints on aspects of the model such as the weight matrix, that may alleviate the computational instibilities.

5.13 Acknowledgements

I firstly thank Dr. Bruno Olshausen for introducing me to this paper, proposing the Magno/Parvo project that prompted the investigation into reproducing this paper's results, and for guiding me along the way. I also thank Dr. Shariq Mobin for his write-up of the derivations for this paper that he wrote for journal club; this framework helped greatly in my understanding of the mathematics of this paper. I also thank Dr. Jaijeet Roychowdhury for giving me the mathematical background to more deeply understand the issues in training this model during his Numerical Simulation and Modeling course, as well as his open-mindedness in allowing me to investigate this as a final project during his course, despite his distaste for neural networks.

Chapter 6

Conclusion

6.1 Summary of Contributions

In this thesis, I have described a set of explorations surrounding natural scene statistics. This includes the development of custom hardware to collect a dataset representative of the signal that reaches the human retina. I have also described the analysis of this and other natural scenes datasets in terms of space, time, power spectrum and phase, with the ultimate goal of understanding the relationship of these statistical properties to the coding properties of the visual system.

In chapter 2, a novel, high-fidelity world, body, head, and eye tracking device is used to create a dataset of the visual signal experienced by humans as they go about their everyday life. This device can record data both inside and outside of the laboratory, in both seated and mobile tasks. It is designed with ergonomics in mind, and is relatively lightweight and unobtrusive for the subject, ensuring the recorded motion is as natural as possible. It also records this data at spatial and temporal frequencies higher than have been measured previously in such a device, and incorporates depth information for the scene in the same frame of reference as the high-resolution camera.

In chapter 3, this tracking device is used to collect a dataset of 15 everyday tasks, each performed by 3 human subjects. We analyze this data in terms of the spatiotemoral power spectrum, showing that the characteristics of the spectrum previously reported for environmental motion only are conserved at the higher spatial and temporal frequencies that were previously unable to be measured. In addition, three conditions are compared, each progressively increasing the amount of motion present on each scene from environmental only, then including head and body motion, and finally including eye motion to re-create as closely as possible the signal present on the subject's retina. I show that head/body motion modulates the power spectrum of the incoming spatiotemporal signal, boosting mid range temporal frequencies. I also show that eye motion also modulates this signal, further boosting the mid and high range portions of the temporal frequency spectrum. These modification to the signal can be thought of as a temporal whitening, a property which has favorable properties

from an information theoretic standpoint.

In chapter 4, considering the larger perceptual relevance of the phase spectrum as compared to the power spectrum, I identify statistical regularities present in the phase spectrum of natural images. First, I analyze structure in the phase spectrum of natural images, using a windowed entropy analysis to identify substructures in this spectrum that correspond to perceptually relevant information including regions with repeated textures. Then, contrasting this global phase measurement, we turn to a more locally defined property of phase, known as 'phase congruency' [61], a quantity that can be used to identify image boundaries such as edges and corners. We show that phase congruency values in natural images are much larger in natural images than for various categories of noise images, and that the distribution is slightly different even among different categories of natural images (indoor/outdoor/man-made/natural). As phase congruency can be defined separately for different orientations, we also explore directional biases between these different categories. While this difference may not be large enough to allow automatic categorization, such prior knowledge about the underlying structure of natural images could be used to inform image and video compression algorithms, even allowing for adaptive compression algorithms based on content type.

In chapter 5, we utilize the statistics of natural signals to model the human visual system using a single layer neural network. We show many of the difficulties in training such a model, addressing computational instabilities that hinder the training of this model using auto-differentiation software. In addition to modifying the loss function and training protocol to address these issues, we also demonstrate a method of reformulating the model as an autoencoder, to avoid these computational instabilities.

6.2 Future Work

Mobile Tracking Device and Data Set

There are areas for improvement and continuation of data collection with our head mounted eye tracking setup. Firstly, in future iterations of the tracking device, there are improvements possible such as less bulky cameras, a second camera to collect high-fidelity binocular data, as well as a laptop acquisition computer for improved battery life and ease of data collection. With respect to the type of data collected, while we collected data using this device for a specific set of 15 tasks, this device could also be used to study human perception in a wide range of additional settings. To improve our dataset, a wider variety of outdoor and mobile tasks could be included, including tasks performed in more nature-heavy environments. Furthermore, this device could be used for completely separate studies, for example to study specific sensory-motor tasks like grasping, or eye movements in a variety of fast-paced tasks, and could be used to compare clinical and non-clinical populations. In more applied settings, data recorded from our device could help inform more specialized tasks such as driving or specific job-related tasks, especially those in which fast motion is present and task-relevant. Additionally, the use of object recognition, pose recognition, and image segmentation soft-

ware [82, 84, 71] would allow for high-level analysis of the video data, and could even be performed real-time in some cases.

Natural Tasks Data Set Analysis

With the dataset we have collected, there are countless potential analyses possible. Within experimental neuroscience, this dataset could be used as the visual stimulus while performing any variety of cellular recording techniques (intracellular, extracellular, calcium imaging, etc). The most obvious recordings would be from the postmortem human retina, to study the response of neurons in the human retina to this natural stimulus, but other recording situations could be easily imagined.

From a natural scene statistics point of view, as we have seen with spatial phase, the power spectrum analysis described above only just scratches the surface. An analysis of temporal phase congruency, for example, is an obvious extension. Given the temporal natural of this data, the statistics of optic flow on the retina would be an interesting, and neurally-relevant direction [100], as this has been shown to be important for tasks such as navigation [73]. In addition, while the first step of our Fourier analysis collapses the three color channels into a single greyscale channel, we pair our RGB recording with images of a known color calibration target for each lighting condition, allowing any of these natural scene analyses to incorporate color. This is particularly relevant in the magno/parvo/konio system of the retina and LGN, which pairs spatiotemporal signal properties with color opponent channels in parallel pathways.

As we used a lens subtending approximately 60° horizontally, there is an opportunity with our dataset to study the statistics of not only foveal vision as we have with our spatiotemporal Fourier analysis, but also how the visual signal may deviate statistically as one moves into the mid-periphery. In this situation, it becomes even more important to model the defocus blur for peripherally-viewed objects that are in a different depth plane than the gaze point, an additional area of future work. In all of these contexts, our dataset allows for the comparison of these statistics for the separate conditions of environmental-only motion, environmental with body and head motion included, and finally for the complete recreation of the retinal signal, with eye motion layered on top of the other three.

Finally, in addition to the recorded visual data, the data from the motion trackers that record the head and body position and motion during the tasks is an additional untapped area of exploration from this dataset. Even on its own, the variability of motion profiles between tasks as well as between individual subjects will be an interesting dataset to explore. Furthermore, when combined with the visual portion of the dataset, it could be used to compare to the visual motion seen in the recordings, as well as to tease apart the separate effects of head versus body motion, which cannot be separated out from the visual data alone.

Theoretical Neuroscience

Extending classical natural scene statistics approaches, we can also explore the natural visual signal by studying the decomposition of the recorded video, as well as the statistical properties of the convolution of this video data with various classes of filters. For example, PCA and ICA [50], though likely much simpler than the true decomposition performed by the visual system, can likely shed light on the major modes of variation in these dynamic signals, and can also be used to better understand the effect of various motion types on that variability. Such methods can be applied to full frames of the video, chunks of the video, or even optic flow computed from the motion vectors between two subsequent frames. Increasing in complexity, pyramidal decomposisitions such as the spatiotemporal steerable pyramid, are more neurally plausible, and may begin to shed light on the relationship between the natural retinal signal and the representation in the early layers of visual area V1. Learned filter representations such as in spatiotemporal sparse coding [76] may show interesting differences in the learned basis functions when ego-motion is included. And as we have seen in Chapter 5, this type of video data can be of use to understand the principles underlying earlier stages of visual processing including the retina. Manifold learning [19] is also a promising direction in this area, which could benefit from this high-fidelity dataset of the retinal signal.

In summary, we have designed and built a new data high-fidelity mobile recording device, and used it to create a novel dataset for use by the Vision Science, Neuroscience, and Computational Vision communities. We then use this data to show the effects of environmental and ego motion on the dynamic visual signal processed by the human retina, and statistical properties of this signal that have implications for the coding properties of the human visual system.

Bibliography

[1] Tobii Pro AB. *Tobii Pro Lab User's Manual*. Version 1.145. Danderyd, Stockholm, 2014. URL: `http://www.tobiipro.com/`.

[2] Martín Abadi et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. 2015. URL: `https://www.tensorflow.org/`.

[3] Sami Abu-El-Haija et al. "Youtube-8m: A large-scale video classification benchmark". In: *arXiv preprint arXiv:1609.08675* (2016).

[4] Joseph J Atick. "Could information theory provide an ecological theory of sensory processing?" In: *Network: Computation in neural systems* 3.2 (1992), pp. 213–251.

[5] Joseph J Atick and A Norman Redlich. "What does the retina know about natural scenes?" In: *Neural computation* 4.2 (1992), pp. 196–210.

[6] A Terry Bahill, Michael R Clark, and Lawrence Stark. "The main sequence, a tool for studying human eye movements". In: *Mathematical biosciences* 24.3-4 (1975), pp. 191–204.

[7] Rosario M Balboa and Norberto M Grzywacz. "Power spectra and distribution of contrasts of natural images from different habitats". In: *Vision research* 43.24 (2003), pp. 2527–2537.

[8] Horace B Barlow. "Unsupervised learning". In: *Neural computation* 1.3 (1989), pp. 295–311.

[9] Horace B Barlow et al. "Possible principles underlying the transformation of sensory messages". In: *Sensory communication* 1.01 (1961).

[10] Adi Ben-Israel and Thomas NE Greville. *Generalized inverses: theory and applications*. Vol. 15. Springer Science & Business Media, 2003.

[11] Alexander J Bies et al. "Relationship between fractal dimension and spectral scaling decay rate in computer-generated fractals". In: *Symmetry* 8.7 (2016), p. 66.

[12] Kamran Binaee et al. "Pupil Tracking Under Direct Sunlight". In: *ACM Symposium on Eye Tracking Research and Applications*. 2021, pp. 1–4.

[13] Markus Bindemann. "Scene and screen center bias early eye movements in scene viewing". In: *Vision research* 50.23 (2010), pp. 2577–2587.

[14] Kathryn Bonnen et al. "A role for stereopsis in walking over complex terrains". In: *Journal of Vision* 19.10 (2019), 178b–178b.

[15] Kathryn Bonnen et al. "Binocular vision and the control of foot placement during walking in natural terrain". In: *Scientific Reports* (2021).

[16] Nuala Brady and David J Field. "Local contrast in natural images: normalisation and coding efficiency". In: *Perception* 29.9 (2000), pp. 1041–1055.

[17] Miriam Brinberg et al. "The idiosyncrasies of everyday digital lives: Using the Human Screenome Project to study user behavior on smartphones". In: *Computers in Human Behavior* 114 (2021), p. 106570.

[18] James H Brown et al. "The fractal nature of nature: power laws, ecological complexity and biodiversity". In: *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 357.1421 (2002), pp. 619–626.

[19] Yubei Chen, Dylan M Paiton, and Bruno A Olshausen. "The sparse manifold transform". In: *arXiv preprint arXiv:1806.08887* (2018).

[20] Naïg Aurelia Ludmilla Chenais, Marta Jole Ildelfonsa Airaghi Leccardi, and Diego Ghezzi. "Naturalistic spatiotemporal modulation of epiretinal stimulation increases the response persistence of retinal ganglion cell". In: *Journal of Neural Engineering* 18.1 (2021), p. 016016.

[21] Bobby Chesney and Danielle Citron. "Deep fakes: A looming challenge for privacy, democracy, and national security". In: *Calif. L. Rev.* 107 (2019), p. 1753.

[22] Matteo Cognolato, Manfredo Atzori, and Henning Müller. "Head-mounted eye gaze tracking devices: An overview of modern devices and recent advances". In: *Journal of Rehabilitation and Assistive Technologies Engineering* 5 (2018), p. 2055668318773991.

[23] Yang Dan, Joseph J Atick, and R Clay Reid. "Efficient coding of natural scenes in the lateral geniculate nucleus: experimental test of a computational theory". In: *Journal of neuroscience* 16.10 (1996), pp. 3351–3362.

[24] James Davis, Yi-Hsuan Hsieh, and Hung-Chi Lee. "Humans perceive flicker artifacts at 500 Hz". In: *Scientific reports* 5.1 (2015), pp. 1–4.

[25] Jia Deng et al. "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[26] H Dietrich and M Wuehr. "Strategies for gaze stabilization critically depend on locomotor speed". In: *Neuroscience* 408 (2019), pp. 418–429.

[27] Haike Dietrich and Max Wuehr. "Selective suppression of the vestibulo-ocular reflex during human locomotion". In: *Journal of neurology* 266.1 (2019), pp. 101–107.

[28] Eric McVoy Dodds, Jesse Alexander Livezey, and Michael Robert DeWeese. "Spatial whitening in the retina may be necessary for v1 to learn a sparse representation of natural scenes". In: *BioRxiv* (2019), p. 776799.

[29] Dawei W Dong and Joseph J Atick. "Statistics of natural time-varying images". In: *Network: Computation in Neural Systems* 6.3 (1995), pp. 345–358.

[30] Dawei W Dong and Joseph J Atick. "Statistics of Natural Time-Varying Images". In: *Network: Computation in Neural Systems* 6.3 (1995), pp. 345–358. ISSN: 0954-898X. DOI: 10.1088/0954-898X/6/3/003. arXiv: arXiv:1011.1669v3.

[31] Vasha DuTell et al. "The Spatiotemporal Power Spectrum of Natural Human Vision". In: *Journal of Vision* 20.11 (2020), pp. 1661–1661.

[32] Michael P Eckert, Gershon Buchsbaum, and Andrew B Watson. "Separability of spatiotemporal spectra of image sequences". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 14.12 (1992), pp. 1210–1213.

[33] Wolfgang Einhäuser et al. "Human eye-head co-ordination in natural exploration". In: *Network: Computation in Neural Systems* 18.3 (2007), pp. 267–297.

[34] Auria Eisen-Enosh et al. "Evaluation of critical flicker-fusion frequency measurement methods for the investigation of visual temporal resolution". In: *Scientific reports* 7.1 (2017), pp. 1–9.

[35] Kara J Emery et al. "OpenNEEDS: A Dataset of Gaze, Head, Hand, and Scene Signals During Exploration in Open-Ended VR Environments". In: *ACM Symposium on Eye Tracking Research and Applications*. 2021, pp. 1–7.

[36] Cordelia Erickson-Davis and Helma Korzybska. "What do blind people "see" with retinal prostheses? Observations and qualitative reports of epiretinal implant users". In: *Plos one* 16.2 (2021), e0229189.

[37] Andreas Ernst et al. "Check my chart: A robust color chart tracker for colorimetric camera calibration". In: *Proceedings of the 6th International Conference on Computer Vision/Computer Graphics Collaboration Techniques and Applications*. 2013, pp. 1–8.

[38] Anja Feldmann et al. "Implications of the COVID-19 Pandemic on the Internet Traffic". In: *Broadband Coverage in Germany; 15th ITG-Symposium*. VDE. 2021, pp. 1–5.

[39] David J Field. "Relations between the statistics of natural images and the response properties of cortical cells". In: *Josa a* 4.12 (1987), pp. 2379–2394.

[40] Agostino Gibaldi and Martin S Banks. "Binocular eye movements are adapted to the natural environment". In: *Journal of Neuroscience* 39.15 (2019), pp. 2877–2888.

[41] Agostino Gibaldi and Martin S Banks. "Crossed–uncrossed projections from primate retina are adapted to disparities of natural scenes". In: *Proceedings of the National Academy of Sciences* 118.7 (2021), e2015651118.

[42] Agostino Gibaldi, Vasha DuTell, and Martin S Banks. "Solving Parallax Error for 3D Eye Tracking". In: *ACM Symposium on Eye Tracking Research and Applications*. 2021, pp. 1–4.

[43] Agostino Gibaldi et al. "The blur horopter: Retinal conjugate surface in binocular viewing". In: *Journal of Vision* 21.8 (2021), pp. –.

[44] James J Gibson. *The ecological approach to visual perception: classic edition*. Psychology Press, 1979.

[45] Rafael C Gonzalez, Richard E Woods, et al. *Digital image processing*. Prentice hall Upper Saddle River, NJ, 2002.

[46] Gerard E Grossman et al. "Frequency and velocity of rotational head perturbations during locomotion". In: *Experimental brain research* 70.3 (1988), pp. 470–476.

[47] Daniel S Hamermesh, Harley Frazis, and Jay Stewart. "Data watch: The American time use survey". In: *Journal of Economic Perspectives* 19.1 (2005), pp. 221–232.

[48] Charles R. Harris et al. "Array programming with NumPy". In: *Nature* 585.7825 (Sept. 2020), pp. 357–362. DOI: 10.1038/s41586-020-2649-2. URL: https://doi.org/10.1038/s41586-020-2649-2.

[49] J. H. van Hateren and A. van der Schaaf. "Independent Component Filters of Natural Images Compared with Simple Cells in Primary Visual Cortex". In: *Proceedings: Biological Sciences* 265.1394 (1998), pp. 359–366.

[50] J Hans van Hateren and Dan L Ruderman. "Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex". In: *Proceedings of the Royal Society of London. Series B: Biological Sciences* 265.1412 (1998), pp. 2315–2320.

[51] Peter Hausamann, Christian Sinnott, and Paul R MacNeilage. "Positional head-eye tracking outside the lab: an open-source solution". In: *ACM Symposium on Eye Tracking Research and Applications*. 2020, pp. 1–5.

[52] Peter Hausamann et al. "Evaluation of the Intel RealSense T265 for tracking natural human head motion". In: *Scientific Reports* 11.1 (2021), pp. 1–12.

[53] Selig Hecht and Simon Shlaer. "Intermittent stimulation by light V. The relation between intensity and critical frequency for different parts of the spectrum". In: *Journal of General Physiology* 19.6 (1936), pp. 965–977.

[54] Vlad Hosu et al. "The Konstanz natural video database (KoNViD-1k)". In: *2017 Ninth international conference on quality of multimedia experience (QoMEX)*. IEEE. 2017, pp. 1–6.

[55] Takao Imai et al. "Interaction of the body, head, and eyes during walking and turning". In: *Experimental brain research* 136.1 (2001), pp. 1–18.

[56] Y. Karklin and E. P. Simoncelli. "Efficient coding of natural images with a population of noisy Linear-Nonlinear neurons". In: *Advances in Neural Information Processing Systems (NIPS)* (2011), pp. 1–9.

[57] Moritz Kassner, William Patera, and Andreas Bulling. "Pupil: An Open Source Platform for Pervasive Eye Tracking and Mobile Gaze-based Interaction". In: *Adjunct Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. UbiComp '14 Adjunct. New York, NY, USA: ACM, 2014, pp. 1151–1160. ISBN: 978-1-4503-3047-3. DOI: 10.1145/2638728.2641695. URL: http://doi.acm.org/10.1145/2638728.2641695.

[58] Leonid Keselman et al. "Intel realsense stereoscopic depth cameras". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2017, pp. 1–10.

[59] Thomas Kluyver et al. "Jupyter Notebooks – a publishing format for reproducible computational workflows". In: *Positioning and Power in Academic Publishing: Players, Agents and Agendas*. Ed. by F. Loizides and B. Schmidt. IOS Press. 2016, pp. 87 –90.

[60] Rakshit Kothari et al. "Gaze-in-wild: A dataset for studying eye and head coordination in everyday activities". In: *Scientific Reports* 10.1 (2020), pp. 1–18.

[61] Peter Kovesi. "Phase congruency: A low-level image invariant". In: *Psychological research* 64.2 (2000), pp. 136–148.

[62] Peter Kovesi. "Phase congruency detects corners and edges". In: *The australian pattern recognition society conference: DICTA*. Vol. 2003. 2003.

[63] United State Bureau of Labor Statistics.

[64] Michael F Land and Mary Hayhoe. "In what ways do eye movements contribute to everyday activities?" In: *Vision Research* 41.25-26 (2001), pp. 3559–3565.

[65] Simon Laughlin. "A simple coding procedure enhances a neuron's information capacity". In: *Zeitschrift für Naturforschung c* 36.9-10 (1981), pp. 910–912.

[66] Ann B Lee, David Mumford, and Jinggang Huang. "Occlusion models for natural images: A statistical study of a scale-invariant dead leaves model". In: *International Journal of Computer Vision* 41.1 (2001), pp. 35–59.

[67] Barry B Lee. "Sensitivity to chromatic and luminance contrast and its neuronal substrates". In: *Current Opinion in Behavioral Sciences* 30 (2019), pp. 156–162.

[68] Yang Liu, Alan Bovik, and Lawrence Cormack. "Relationship between the Helmholtz shear of vertical meridians and disparity statistics in natural scenes". In: *Journal of Vision* 8.6 (2008), pp. 846–846.

[69] Benoit B Mandelbrot and Benoit B Mandelbrot. *The fractal geometry of nature*. Vol. 1. WH freeman New York, 1982.

[70] Natalia D Mankowska et al. "Critical Flicker Fusion Frequency: A Narrative Review". In: *Medicina* 57.10 (2021), p. 1096.

[71] Alexander Mathis et al. "DeepLabCut: markerless pose estimation of user-defined body parts with deep learning". In: *Nature neuroscience* 21.9 (2018), pp. 1281–1289.

[72] Ethel Matin. "Saccadic suppression: a review and an analysis." In: *Psychological bulletin* 81.12 (1974), p. 899.

[73] Jonathan Samir Matthis, Jacob L Yates, and Mary M Hayhoe. "Gaze and the control of foot placement when walking in natural terrain". In: *Current Biology* 28.8 (2018), pp. 1224–1233.

[74] A Morozov and Sh Shakirov. "Analogue of the identity Log Det= Trace Log for resultants". In: *Journal of Geometry and Physics* 61.3 (2011), pp. 708–726.

[75] M Concetta Morrone and Robyn A Owens. "Feature detection from local energy". In: *Pattern recognition letters* 6.5 (1987), pp. 303–313.

[76] Bruno A Olshausen. "Learning sparse, overcomplete representations of time-varying natural images". In: *Proceedings 2003 International Conference on Image Processing (Cat. No. 03CH37429)*. Vol. 1. IEEE. 2003, pp. I–41.

[77] Bruno A Olshausen and David J Field. "Sparse coding with an overcomplete basis set: A strategy employed by V1?" In: *Vision research* 37.23 (1997), pp. 3311–3325.

[78] Adam Paszke et al. "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach et al. Curran Associates, Inc., 2019, pp. 8024–8035. URL: `http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf`.

[79] KB Peterson and MS Pederson. "The Matrix Cookbook, E-manual". In: *Technical University of Denmark* (2006).

[80] Margaret H Pinson. "The consumer digital video library [best of the web]". In: *IEEE Signal Processing Magazine* 30.4 (2013), pp. 172–174.

[81] Christoph Redies, Jens Hasenstein, Joachim Denzler, et al. "Fractal-like image statistics in visual art: similarity to natural scenes". In: *Spatial vision* 21.1-2 (2008), pp. 137–148.

[82] Joseph Redmon and Ali Farhadi. "Yolov3: An incremental improvement". In: *arXiv preprint arXiv:1804.02767* (2018).

[83] Iain E Richardson. *H. 264 and MPEG-4 video compression: video coding for next-generation multimedia*. John Wiley & Sons, 2004.

[84] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation". In: *International Conference on Medical image computing and computer-assisted intervention*. Springer. 2015, pp. 234–241.

[85] Michele Rucci, Ehud Ahissar, and David Burr. "Temporal coding of visual space". In: *Trends in cognitive sciences* 22.10 (2018), pp. 883–895.

[86] Michele Rucci et al. "Miniature eye movements enhance fine spatial detail." In: *Nature* 447.7146 (2007), pp. 851–854. ISSN: 0028-0836. DOI: `10.1038/nature05866`.

[87] Daniel L Ruderman. "Origins of scaling in natural images". In: *Vision research* 37.23 (1997), pp. 3385–3398.

[88] Daniel L Ruderman and William Bialek. "Statistics of natural images: Scaling in the woods". In: *Physical review letters* 73.6 (1994), p. 814.

[89] Irina Yonit Segal et al. "Decorrelation of retinal response to natural scenes by fixational eye movements". In: *Proceedings of the National Academy of Sciences* 112.10 (2015), pp. 3110–3115.

[90] Bharath Shankar et al. "Ergonomic Design Development of the Visual Experience Database Headset". In: *ACM Symposium on Eye Tracking Research and Applications*. 2021, pp. 1–4.

[91] Amy L Sheppard and James S Wolffsohn. "Digital eye strain: prevalence, measurement and amelioration". In: *BMJ open ophthalmology* 3.1 (2018), e000146.

[92] Markus D Solbach and John K Tsotsos. "Tracking Active Observers in 3D Visuo-Cognitive Tasks". In: *ACM Symposium on Eye Tracking Research and Applications*. 2021, pp. 1–3.

[93] Li Song et al. "The SJTU 4K video sequence dataset". In: *2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX)*. IEEE. 2013, pp. 34–35.

[94] William W Sprague et al. "Stereopsis is adaptive for the natural environment". In: *Science Advances* 1.4 (2015), e1400254.

[95] Lakshminarayan Subramanian et al. "Rethinking wireless for the developing world". In: *IRVINE IS BURNING* (2006), p. 43.

[96] Eugene Switkes, Melanie J Mayer, and Jeffrey A Sloan. "Spatial frequency analysis of the visual environment: Anisotropy and the carpentered environment hypothesis". In: *Vision research* 18.10 (1978), pp. 1393–1399.

[97] Vivienne Sze, Madhukar Budagavi, and Gary J Sullivan. "High efficiency video coding (HEVC)". In: *Integrated circuit and systems, algorithms and architectures*. Vol. 39. Springer, 2014, p. 40.

[98] Antonio Torralba and Aude Oliva. "Statistics of natural image categories". In: *Network: computation in neural systems* 14.3 (2003), pp. 391–412.

[99] Matteo Valsecchi et al. "Pedestrians egocentric vision: individual and collective analysis". In: *ACM Symposium on Eye Tracking Research and Applications*. 2020, pp. 1–5.

[100] JH Van Hateren et al. "Function and coding in the blowfly H1 neuron during naturalistic optic flow". In: *Journal of Neuroscience* 25.17 (2005), pp. 4343–4352.

[101] Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009. ISBN: 1441412697.

[102] Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009. ISBN: 1441412697.

[103] Richard F Voss. "Characterization and measurement of random fractals". In: *Physica Scripta* 1986.T13 (1986), p. 27.

[104] Andrew B Watson, Albert J Ahumada, and Joyce E Farrell. "Window of visibility: a psychophysical theory of fidelity in time-sampled visual motion displays". In: *JOSA A* 3.3 (1986), pp. 300–307.

[105] Patrick E Williams et al. "Entrainment to video displays in primary visual cortex of macaque and humans". In: *Journal of Neuroscience* 24.38 (2004), pp. 8278–8288.

[106] Kentaro Yamada et al. "Can saliency map models predict human egocentric visual attention?" In: *Asian Conference on Computer Vision*. Springer. 2010, pp. 420–429.

[107] Peipeng Yu et al. "A Survey on Deepfake Video Detection". In: *IET Biometrics* (2021).

[108] Bolei Zhou et al. "Places: A 10 million image database for scene recognition". In: *IEEE transactions on pattern analysis and machine intelligence* 40.6 (2017), pp. 1452–1464.