The Dissertation Committee for Tal Tversky
certifies that this is the approved version of the following dissertation:

Motion Perception and the Scene Statistics of Motion

Committee:

Risto Miikkulainen, Supervisor

Wilson S. Geisler, Supervisor

Benjamin Kuipers

Bruce Porter

Peter Stone

Motion Perception and the Scene Statistics of Motion

by

Tal Tversky, B.S.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

May 2008

Dedicated to my parents, Amos and Barbara.

Acknowledgments

None of this work would have been possible without the guidance of my adviser, Bill Geisler. Bill is an outstanding mentor, in equal parts critical and kind. I cannot imagine a better role model for a scientist; he has unflinching integrity, unflagging diligence and a very big brain. Many thanks go to Risto Miikkulainen. Risto would always ask the hard question, challenge me to do more and convince me that it could be done. How many advisers will go surfing with you before a conference? Thanks to the rest of my committee, Bruce Porter, Ben Kuipers, and Peter Stone for reading this document and for their guidance during my graduate career.

I have to thank my lab companions here at the People's Republic of Psychophysics, Jeff Perry, George Najemnick, David Ing and Chris Bradley. Jeff is a perfect lab partner, always up for white-boarding equations, a dive into code to find a bug, a cup of coffee or P4 at Madam Mam's. There is never a dull moment when George is around, and the conversation always seems suddenly elevated to new philosophical levels.

Several of my fellow graduate students have left before me, but I owe them thanks for their help and conversations along the way. In particular Harold Chaput is the best of friends and I have missed his companionship these past few years. Thanks to Ken Stanley as well for his clarity and conversation. Many thanks to Michael Bogomolny. Bogo greeted me when I arrived in Austin, made me feel at home and then became like a brother to me here in Austin.

I have an unbelievably supportive group of friends here in Austin, without whom I would never have survived to finish this degree. Thanks to all of movie-group. In particular, thanks to Luisa, the best personal assistant a graduate student ever had. Thanks to big buddy Stu for his washers, poker and wisdom. Thanks to GC and Monika for their infinite hospitality and wonderful hottub. Shout-outs also to Leigh and David who were incredible mentors into parenthood and to Liz and Patrick who have been our partners in parenthood.

Thanks to my family. They are my constant support in too many ways to catalog. Thanks to my parents-in-law, David and G-ann. They are a tireless source of babysitting and family comfort that made this difficult project much easier. Thanks to my sister, Dona, and brother-in-law, Eran, who have basically served as my tertiary adviser on this whole dissertation. Thanks to my brother, Oren, who knows and understands me in ways only he could. Thanking my mother, Barbara, is an impossible task. She was a constant support and is likely the reason I came here in the first place. Her interest and excitement for life, art and ideas is inspirational. Thanks to my father, Amos, for teaching me fun math early in the evenings when I was a child, and for watching bad television with me late in the evenings when I was an adult.

Thanks most of all to my wife, Jenna. Without her constant support and love, this dissertation would not have been completed. While I slowly gave birth to this unwieldy project, she managed to birth two wonderful babies, start a successful career, and still managed to nurture our marriage as well. I have boundless respect for her intelligence, energy and compassion.

Tal Tversky

The University of Texas at Austin
May 2008

Motion Perception and the Scene Statistics of Motion

Publication No. _____

Tal Tversky, Ph.D.
The University of Texas at Austin, 2008

Supervisors: Risto Miikkulainen, Wilson S. Geisler

Motion coding in the brain undoubtedly reflects the statistics of retinal image motion occurring in the natural environment. Measuring the statistics of motion in natural scenes is an important tool for building our understanding of how the brain works. Unfortunately, there are statistics that are either impossible or prohibitively difficult to measure. For this reason, it is useful to measure scene statistics in artificial movies derived from simulated environments. This is a novel and important methodological approach that allows us to ask questions about optimal coding that are impossible otherwise. This dissertation describes a course of research that develops this research methodology, the simulated scene statistical approach.

This dissertation applied the artificial scene statistical approach to understanding the visual statistics of motion during navigation through forest environ-

ments. An environmental model of forest scenes was developed based on previously measured range and surface texture statistics. Spatiotemporal power spectra were measured in both simulated and natural scenes for the task of first person motion through a forest environment. These image statistics measurements helped validate the environmental model.

Next, the environmental model was used to simulate across-domain statistics to study the ideal aperture size of motion sensors. It was found that across a variety of different scene conditions, the optimal aperture size of motion sensors increases with the speed to which the sensor is tuned. This is an important constraint for understanding both how the brain encodes motion as well as for designing computer motion detectors.

This theoretical research inspired a psychophysical experiment estimating the receptive-field size of human foveal motion discrimination. It was found that for narrow-band stimuli the ideal aperture size increases with spatial frequency, but is unchanging with respect to velocity or temporal frequency.

This dissertation shows an approach to the study of vision that has applications in psychophysics, neuroscience and computer vision. The emphasis on accurate and validated environmental models for simulating scene statistics can help improve our understanding of the structure and function of the human visual system and also help us build more accurate and robust computer vision systems.

Contents

List of Figures

Chapter 1

Introduction

How does the brain encode motion information? Motion is a rich source of visual information for perceptual tasks. Imagine the difference in visual experience between standing in a forest, or walking through a forest. Motion gives an enormous amount of additional information about the scene. Information about the observer's direction of motion, the shape and distance of individual leaves, branches and trees, and the separation of the trees and leaves from each other is all readily accessible from the first step. It is easy to see how motion provides useful information for many perceptual tasks including heading estimation, image segmentation, distance estimation and shape estimation. The heading estimation task is particularly relevant because, in general, most of the visual motion experienced by a human is due to self-motion. In addition, if an animal is able to estimate heading accurately, it is more likely to survive and reproduce.

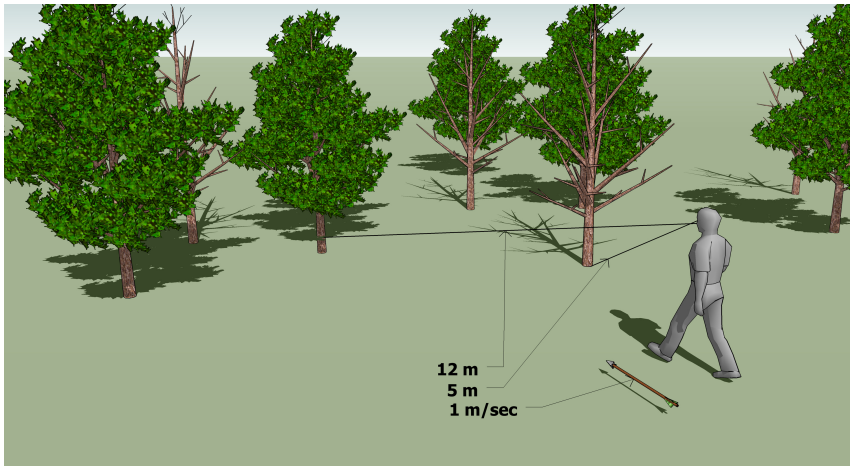
Due to the forces of evolution and natural selection, coding in the brain must reflect the properties of the environment in which an animal performs the tasks that are necessary to survive and reproduce. Therefore, to understand how the brain estimates heading, it is important to understand the statistics of the environment relevant to this task. This dissertation will characterize the statistics of motion that occur when navigating through natural environments, and explore the implications of those statistics for motion coding in the brain.

1.1 Motivation

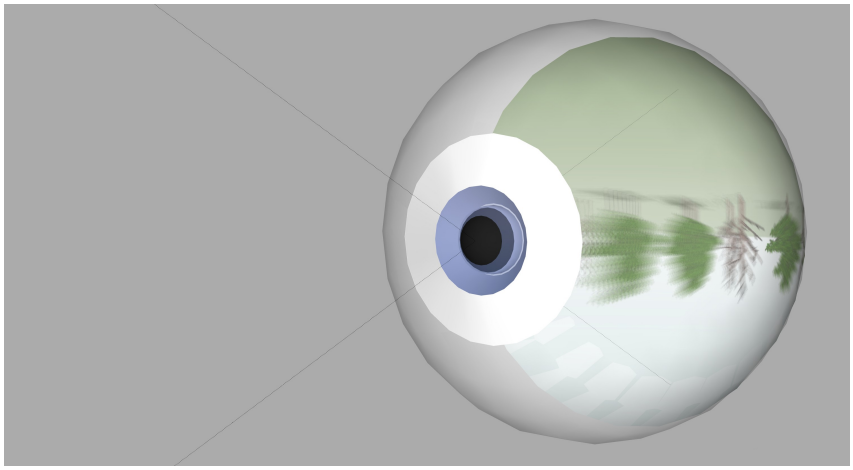
In order to characterize the information relevant to the heading estimation task, one needs to measure statistics simultaneously in two separate *domains*: the image plane, and the environment (Figure 1.1). All the visual information available to the brain for heading estimation is contained in the spatiotemporal image produced on the retina (or image plane) by self-motion through the three-dimensional environment (Figure 1.1*b*). For a heading estimator, this spatiotemporal image can be considered the input, but in order to design an optimal system, one also needs to know the environmental 'ground truth'. Specifically, for each location in the image plane, one needs to know the true three-dimensional velocity of the corresponding location in the environment (Figure 1.1*a*). Assuming that the only source of motion for this task is first-person motion, knowing the three-dimensional velocity of each location in the image plane is precisely equivalent to knowing the motion of the image plane through the environment and the range of each object in the environment. (In this context the word 'range' means the distance of the object from the focal point of the imaging lens.) Separate measurements of statistics from the image domain or the environmental domain are informative and relevant to understanding coding in the brain. However, simultaneous measurements in both domains allows us to tie the visual inputs with the ground-truth solution to the estimation task that the brain is performing, and thus allows us to compute the optimal way to perform the task.

Measuring and analyzing motion statistics in the image domain is a useful scientific endeavor by itself and there has been some research effort in this vein (Dong and Atick, 1995a; Dror et al., 2000; van Hateren and Ruderman, 1998). Most notably, Dong and Atick (1995a) analyzed the second order statistics of moving images from commercial movies as well as custom shot videos of their own. They found that their image statistics are well fit with a model that assumes a power law distribution of velocities in the world. They also showed that some of the temporal response properties observed in cells in the lateral geniculate nucleus (LGN) could be explained by the measured motion statistics (Dong and Atick, 1995b).

Clearly, measuring motion statistics in the image domain is a much simpler task than measuring the environmental sources of that motion, i.e., the statistics of



(a) Environmental Information



(b) Image Information

Figure 1.1: **Two domains for gathering motion statistics.** In order to completely characterize the information relevant to the heading estimation task, information needs to be gathered simultaneously in two domains. In the environmental domain (a) one needs to know the motion of the observer through the environment and the ranges of all the objects visible in the scene. Assuming that there are no moving objects in the scene, this knowledge is equivalent to knowing the three-dimensional motion of all the visible objects in the scene. In the image domain (b) one needs to measure the changing pattern of luminance and chromaticity projected onto the image plane as the observer moves through the scene. Separate measurements of statistics from the image domain or the environmental domain are informative and relevant to understanding coding in the brain. However, simultaneous measurements in both domains allow us to tie the visual inputs (b) with the ground-truth solutions (a) to the estimation task that the brain is performing and thus allows us to compute the optimal way to perform the task.

the structure of the environment and the statistics of the eye's motion through it. A number of studies provide some of the relevant environmental scene statistics by measuring ranges as a function of location in the image plane (Huang et al., 2000; Yang and Purves, 2003). These are useful studies, but only measure environmental statistics, not image statistics and environmental statistics. Potetz and Lee (2003) measure coregistered range and luminance images, but only in static scenes and only in the context of estimating depth from shading. None of these studies measures the complete information needed to characterize the statistics of heading estimation.

The majority of previous work on the statistics of motion has two major limitations. First, the statistical measurements are not done in a task-dependent manner. For instance, Dong and Atick used spatiotemporal images from movies and hand-held video cameras. The motion in their scenes derives from some unknown combination of object motion, camera motion and lighting changes. Instead, it would be preferable to collect spatiotemporal images in a manner that was known to be representative for a particular perceptual task. It might very well be the case that image motion statistics are quite different for the task of heading estimation than they are for tracking moving objects. One study that aimed at measuring image motion statistics for a specific task was that Boeddeker et al. (2005); they gathered image statistics by moving a camera on a robotic gantry through known insect flight paths. However, their work is specific to the motion and visual environment of insects and therefore is not relevant to human motion perception since the motion of insects and humans is quite different.

The second limitation of current motion statistics work is that statistics are typically measured just within the image domain. In order to understand a visual task, it is important to measure both the image motion and the environmental sources of that motion. This is what Geisler (2008) calls "across-domain" statistics and, in the jargon of machine learning, is the difference in the training set between supervised and unsupervised learning. Across-domain statistics are necessary to understand how an optimal visual system would solve the perceptual task and allows one to build hypotheses about the human visual system works.

Without ground-truth environmental information about the task being solved there is a limit to the theoretical work that can be done. For example, in order to draw conclusions about the behavior of early sensory neurons, Dong and Atick (1995b) relied on the "efficient coding hypothesis" (Attneave, 1954; Barlow, 1961).

This hypothesis assumes that the brain codes the information in the world so as to remove statistical redundancies in the sensory input. Given that the brain has a limited channel capacity (i.e. a limited number of noisy neurons) it is more efficient to remove redundancies in the perceptual inputs. Using this assumption is a reasonable approach and has yielded interesting results, but has limitations that across-domain statistics remove. In essence, lacking a well defined perceptual task, the efficient coding hypothesis assumes that the brain's perceptual task is to remove redundancies. It also assumes that all the input information is equally important. Without a perceptual task and ground-truth environmental information, it is impossible to know what aspects of the visual input are more important for the task.

In contrast, with simultaneously measured image and environmental information, one can ask theoretical questions about the specific perceptual task being solved. Within computational constraints, one can calculate how the optimal visual system ought to behave. In addition one can build hypothesis about how the human visual system behaves by comparing it to the optimum.

1.2 Approach

In order to measure across-domain statistics for the heading estimation task, one would need a calibrated camera that simultaneously shoots coregistered range and luminance images. One could then measure human walking paths through natural environments and their corresponding eye motions. The range/luminance camera could then be mounted on a robotic gantry that moves the camera plane through precisely the same path that the human retina took. Making the proper set of measurements is not beyond the realm of possibility; the procedure would be very similar to what Boeddeker et al. did, but using the coregistered range/image camera of Potetz and Lee and using humans walking and eye fixation paths instead of insect flight paths. However, making these measurements is clearly an expensive, difficult and time intensive task.

There is a simple alternative to direct measurement: simulation. With a ray tracer, one can simulate motion through an environment by building a model of that environment. In computer simulations, one can know with precision all the information relevant to the heading estimation task. There are two benefits to this approach. First and foremost, making the correct set of measurements is possible.

Second, one can quickly alter scene parameters and generate much larger statistical samples in simulation than would ever be possible by direct measurement. This is the approach taken in this dissertation.

Using simulated statistics as a theoretical tool for understanding the human visual system is a novel approach. This dissertation develops the simulated statistical methodology by following the scientific method from beginning to end: observations are made from simulated statistics, a hypothesis is developed, and an experiment is conducted to test the hypothesis. In so doing it serves as a blueprint for the simulated statistical methodology and an example of its scientific utility. In addition to having applications to the science of vision, the methodology also has application to engineering and computer vision.

When measuring scene statistics in simulation, there are some potential pitfalls. Simulated statistics are only as accurate as the model environment used in the simulation allows. If the environmental model is inaccurate or overly simplified, it may lead to misleading or inaccurate statistical measurements. For this reason, the environmental model used in this dissertation is based on the measured properties of range and surface texture in natural scenes. Additionally, the model is validated by comparing simulated image statistics with natural image statistics that were measured in an identical fashion.

The simulated scene statistics approach has broad applicability to perception research and cognitive neuroscience as well as to computer vision and robotics. The work described here focuses on motion statistics for heading perception and helps further our understanding of the human visual system, but the results have applications for designing computer vision systems. More generally, the methodology has applicability to many visual perception problems such as shape from shading, segmentation based on motion and segmentation based on color.

1.3 Outline of Dissertation and Results

In this dissertation, research projects were conducted, each study using the previous one as its foundation. First, the environmental model was validated by measuring natural image statistics and comparing them to the simulated statistics of the environmental model. In the next research project, across-domain simulated statistics were gathered using the validated environmental model and an optimal visual sys-

tem's behavior was modeled, allowing for the development of a hypothesis about the human visual system. Finally, a psychophysical experiment was conducted to compare human performance with the hypothesis.

The next chapter, **Chapter 2**, describes the research foundations of this work. Because of the interdisciplinary nature of this work, three different research areas are reviewed: the natural scene statistics of motion, simulated environments and psychophysical measurements of receptive-field sizes. Finally, the literature related to the environmental model is also reviewed.

Chapter 3 describes the first of the three research projects in this dissertation, the validation of the environmental model. Image domain measurements were made both in simulated and in natural forest scenes in the context of a specific perceptual task, heading estimation. Image sequences were gathered in natural scenes using a calibrated camera mounted on a custom built sliding rail. Additionally, artificial image sequences were generated using a ray tracer with a model of forest scenes based on measured natural scene statistics.

These measurements revealed that the power spectra of the natural and artificial movies resemble each other qualitatively. The measured power spectra were not well fit by the current dominant model (Dong and Atick, 1995a). There are two main assumptions that the Dong and Atick model make that are broken by the measured motion sequences. First, there is a large amount of non-translational motion in the scenes and, second, the distribution of velocities in the scenes does not follow a power-law. These results imply that it is important to sample natural scene statistics in a task-specific manner. Also, the results suggest that using artificially generated scenes statistics can be a valuable supplement to natural scene measurements. The results pave the way for the work in the next chapter, by validating the environmental model.

In **Chapter 4** across-domain statistics were measured to investigate a specific question about motion receptors used in heading estimation. What is the ideal aperture size for local motion detectors? It was found that over a large range of scenes the optimal aperture size increases monotonically with the speed of the motion being detected. Additionally, over a range of speeds, the ideal aperture size increases linearly with the log of the speed. This result leads to a hypothesis about the visual system: receptive-field sizes of motion detectors should increase with the speed to which they are tuned.

Chapter 5 describes a psychophysical experiment testing the receptive-field size hypothesis. Human foveal receptive-field sizes were estimated at different speeds for a motion discrimination task. Narrow-band stimuli at different spatial and temporal frequencies were used to distinguish between an effect of speed and the effects of spatial and temporal frequency. It was found that psychophysical receptive-field sizes change with spatial frequency, but not with temporal frequency implying that there is no specific dependency on speed. This raises more questions, both theoretical and experimental, for future work. The first and most obvious question is whether there is receptive-field size dependency on speed at non-foveal locations.

Chapter 6 concludes the dissertation by reviewing and discussing the results, and outlining some future work and discussing the broader implications of the research.

Appendix A describes the math behind the modeling of the Fourier power spectra from Chapter 3.

Appendix B describes the ideal observer model used in Chapter 5.

Appendix C contains the complete set of plots of the power spectra from the artificial and simulated scenes described in Chapter 3.

Chapter 2

Foundations

This chapter describes the research foundations of this dissertation. First, the natural scene statistical approach is reviewed, with an emphasis on work related to motion. Second, previous work related to simulated scene statistics is discussed. Next, the natural scene measurements that serve as the foundation for the environmental model used in this dissertation are described. The chapter concludes with a review of the psychophysical measurements of receptive-field sizes.

2.1 Natural Scene Statistics of Motion

The motivation for research in natural scene statistics derives from the fact that the process of natural selection has shaped the human visual system to effectively perform visual tasks in natural environments. Therefore, much can be learned about the design of the human visual system by studying the properties of natural environments and tasks.

The roots of this approach are related to Gibson's (1979) "ecological" approach to vision and perception. Gibson emphasized the environment and the context of visual perception and discouraged the view of vision as abstract computation. Contemporary research in natural scene statistics is ecological in the sense that the environmental context is of key importance, but differs from Gibson's approach in that visual computation is not ignored. The focus of this review is on work relating to motion in natural scenes, but for good reviews of the natural scene statistics approach as a whole, see Geisler (2008) and Simoncelli and Olshausen (2001).

There has been a wide variety of work done measuring the natural scene statistics of motion. Most significantly, Dong and Atick (1995a) have started the task of measuring and analyzing the statistics of time-varying images. Their dataset includes both segments from commercial movies and videos that they have filmed. In analyzing this library, they first computed the pairwise probability distribution of light intensities at two different locations, varying in space and time. Not surprisingly, they found that image intensities tend to change slowly in both space and time, i.e. it is likely that two pixels close in space and time have the same or similar light intensities.

Second, Dong and Atick took randomly sampled space-time patches from their library, computed the Fourier transform and analyzed the power spectrum. They found that if one knows the average static spatial spectrum of the environment and one also knows the probability of velocities in the environment then one can compute with some accuracy the spatiotemporal power spectrum. In other words, the measured statistics can be explained in large part by a simple environmental model where the motion in the images is caused either by camera motion or by objects moving relative to the camera. Their model makes several seemingly reasonable assumptions:

1. All light intensity changes are due to object motion in the world. In this case observer motion can be considered a special case of object motion where all objects have the same trajectory.

2. All motion in the world is approximately translational relative to the image plane. Examples of motion that break this assumption include sheer, expansion and contraction. However, in a sufficiently small spatial area and a sufficiently small period of time, expansion and contraction would also be approximately translational.

3. The distribution of motion velocities in the image plane is rotationally invariant.

4. The static power spectrum has a consistent average that is independent of the motion of the scene.

5. The velocities of the moving objects in the world are drawn from a power-law distribution.

The assumptions used in Dong and Atick's Fourier model are not just reasonable, but are also common. For example, in a piece of theoretical work that precedes Dong and Atick's, van Hateren (1993) used an almost identical model with a similar set of assumptions. Van Hateren did not make any physical measurements, but used the set of assumptions and the fact that static images have a $\frac{1}{f^2}$ power spectrum to predict the dynamic power spectra of natural scenes. The properties of these power spectra were then used to explain the spatiotemporal contrast sensitivity function and explain a wide variety of psychophysical results.

Imposing the translational motion assumption is useful from a theoretical point of view because it greatly simplifies the mathematical modeling of Fourier power spectra (Appendix A). In addition, these assumptions are, on their surface, perfectly reasonable if the empirical data support them. However, in the next chapter, Dong and Atick's model will be compared with measured scene statistics and it is found that the model does not fit the motion statistics of first-person motion well.

In a paper published concurrently with their Fourier analysis, Dong and Atick (1995b) used their environmental measurements in conjunction with the efficient coding hypothesis to model the temporal tuning properties of cells in the lateral geniculate nucleus (LGN). Their model offers an ecological explanation for the temporal tuning properties of LGN cells and also for the existence of the two classes of temporal tuning in LGN cells, lagged and non-lagged.

In contrast to the work performed in this dissertation, the movies captured by Dong and Atick were not collected in the context of any particular task nor is there any information about the ground-truth sources of the motion in the image. In addition, the computational analysis performed is not in the context of any specific visual task, but instead uses the efficient coding hypothesis to optimize receptive field properties. Obviously, only image information is captured and the ground-truth motion in the scene is unknown.

Van Hateren and Ruderman (1998) also created a database of movies. Their movies were digitized from broadcast television. They sampled over half a million 12x12x12 space-time patches from their video sequences and applied independent-components analysis (ICA) to this collection of blocks. In this context the derived independent components (IC's) can be considered a coding scheme for natural

movies, where a movie can be encoded as follows:

$$I(x, y, t) = \sum_i a_i C_i(x, y, t),$$

where $I(x, y, t)$ describes the movie intensity at each space-time location and a_i is the amplitude of independent component C_i. The independent component filters (ICF's) are used to extract amplitudes, a_i, for the IC's for a particular movie.

Under the efficient coding hypothesis, these filters serve as a model of how the visual system ought to encode information. The explicit assumption made by van Hateren and Ruderman is that the neurons in the primary visual cortex create a basis of independent vectors that represent the environmental image information. The properties of the ICF's derived from this analysis qualitatively match the receptive-field properties of simple cells in the primary visual cortex. The derived space-time filters are similar to Gabors, moving sinusoids with a Gaussian intensity envelope. Like V1 cells, the filters are local in space, and the filters are tuned to a range of velocities and spatial frequencies. One significant aspect in which the derived filters differ from visual cells is the fact that the filters are localized in time. This is an artifact that relates to the fact that the temporal dimension in ICA is finite.

Because ICA creates vectors that are not orthogonal, the ICF's used to extract amplitudes have higher spatial and temporal frequencies than their corresponding IC's. Van Hateren and Ruderman suggest that these differences between the filters and what they represent can explain a visual illusion where space-time Gabors appear as if they are moving in space even though their spatial envelope remains fixed.

In a related piece of work Olshausen (2001), also created a set of filters based on van Hateren's natural image sequences that are, in addition to being independent, also sparse. This means that the response characteristics of the filters, match those of V1 neurons, where only a few filters are activated at a time in short bursts.

In contrast to the Dong and Atick and van Hateren movie databases where the data were collected in an ad hoc manner, Boeddeker et al. (2005) gathered image statistics by moving a camera on a robotic gantry through known insect flight paths. They used a pair of precisely located high speed digital cameras to record movies of blowflies navigating their natural environment. The movies were then processed to extract the precise flightpath of the blowfly through the scene. Then,

using a robotic gantry, a panoramic video camera was moved through the flightpath of the blowfly, recording the visual input experienced by the blowfly. Later, in a laboratory, this recorded stimulus was played back to blowfly visual neurons involved in motion processing while the neuronal responses were recorded. This complicated methodology made it possible to make realistic recordings in a laboratory of neuronal responses that would occur during natural behavior in natural environments. They used this paradigm to validate a model of the motion sensitive horizontal-system (HS) neurons. They also showed that the HS neurons responded to both rotation and translation information supporting the hypothesis that these neurons are involved in depth computations as well as self-motion.

Like the other research described here, Boeddeker et al.'s work still lacks ground-truth information about the true projected motion on the image plane and is specific to the motion and visual environment of insects. There is no previous work that has measured image statistics in the context of human locomotion through the environment. This dissertation measures image statistics in the context of first person motion through a forest environment.

2.2 Simulated Statistics

To overcome the difficulties of simultaneously gathering image motion statistics and information about the ground-truth sources of that motion, environmental statistics can be simulated. Several studies have used the Brown range image database of Huang et al. to create a set of optic flow statistics for computer or human vision research (Calow et al., 2005; Calow and Lappe, 2007; Roth and Black, 2007). Other researchers have used computer graphics techniques to create simulated movies for use as ground-truth test sets.

2.2.1 Using Range to Simulate Flow

Calow et al. tested the performance of a biologically inspired technique for averaging motion information across the visual field. In human vision, motion sensitive neurons in the middle temporal (MT) area of the visual cortex have receptive-field sizes that vary with eccentricity. Small receptive-fields are found at the fovea and larger receptive-fields at more eccentric locations. Calow et al. applied this principle to optic flow estimation to see if it would improve estimates of heading. They

built three different optic flow algorithms each of which fed into the same heading estimator. One had no filtering, a second had uniform filtering, and a third used the biologically inspired "space-variant filtering", where local flow estimates were averaged over larger areas of the image at larger eccentricities.

The algorithms were tested on two motion sequences shot from a moving car. Since the ground-truth motion from these movies was unknown, Calow et al. could only compare the variability in the estimated heading. They found that the two filtering algorithms gave much less noisy results than the unfiltered. In order to create a database of motion sequences with ground-truth information, Calow at al. used the Brown range image database. Given a range image from this database and an instantaneous camera velocity, one can easily compute the instantaneous optic flow (local image velocity) at each location in the range image. In lieu of measured image information, however, Calow et al. created image information by simply treating the range value as a gray scale luminance value. In order to create space-time movies of motion, they made new image frames by interpolating the initial image according to the three-dimensional motion of the image plane. Using these image sequences as a test bed, Calow et al. found that their space-variant filtering algorithm gave more accurate results than both the unfiltered and the constant filtered algorithms.

In another study focused on human vision instead of computer vision, Calow and Lappe (2007) used the Brown range image database to perform a statistical analysis of the retinal optic flow signal experienced by humans during walking. In order to model realistic motion through the environment, Calow and Lappe combined the information from the range image with a heading direction, an eye fixation and a model of head bobbing and swaying during walking. Heading directions were chosen randomly from a range image by picking uniformly from locations without obstacles. Eye fixation locations were chosen by measuring human eye fixations of the static range images. A model of human vertical and horizontal head motion during walking was developed by measuring a subject's head motion during walking using a positional tracking system. Head motion was sampled from this model and added to the heading and eye tracking information to create a large sample of realistic retinal optic flow information.

Calow and Lappe then performed a statistical analysis on different aspects of their simulated retinal optic flow signal. They found that there is only a small

statistical dependence between the speed and direction of optic flow. They found a strong correlation between retinal speed and the depth structure of the scene. This result makes intuitive sense since the closer objects are the faster they appear to move. The retinal flow direction is correlated with the direction of gaze and direction of ego-motion. In addition, Calow and Lappe found that different areas of the visual field provide more or less information about particular flow dimensions. Finally, their study also included an analysis of the difference in the motion signal when eye tracking movements were included or not.

In another computer vision study, Roth and Black also used the Brown range image database to create a database of optic flow fields. Instead of using range images to create a ground-truth *test* bed, Roth and Black used their optic flow database as a *training* set for an optic flow estimation algorithm. To create the optic flow data, the range images from the Brown database were combined with a database of camera motions. The camera motions were extracted from a database of hand-held or car-mounted video sequences. Camera motions were combined with three dimensional scene models extracted from the range images. The resultant flow fields were used to train an estimate of priors on optic flow using a Markov random field model. They found that the motion in these scenes was often slow and smooth, but occasionally fast and discontinuous. The prior was incorporated into an optic flow estimation algorithm and was found to improve the performance of the algorithm.

Combining ego-motion information with range data is a promising approach, but still lacks a few key pieces of information. Foremost, there is no way to connect the ground-truth flow with image information. In other words, the range images only have information about the distance of the objects in the scene, and are completely lacking in information about the surface characteristics of the scene. Using the range data as a luminance image, like Calow and Lappe, allows one to create interpolated images that resemble natural images, but there is no guarantee that the surface and texture characteristics of the world are captured in these interpolated images. Also, occlusions in the world cause gaps in the range information available in the database. Although instantaneous velocities are accurate, it is impossible to know how these velocities or the interpolated images change over time. All of these difficulties are resolved by the approach taken in this dissertation, creating simulated image sequences in a ray tracer.

2.2.2 Simulated Image Sequences

Computer vision researchers have used computer rendering techniques (ray tracers or OpenGL^{TM}) to simulate movies for testing and refining computer vision algorithms. For example, McCane et al. (2001) used a ray tracer to create benchmarks for comparing and testing optic flow algorithms. They created both a synthetic test set and test set of real image sequences with ground-truth information. The synthetic test-set was generated using a custom ray tracer that recorded positional information in addition to generating image sequences. Image sequences in the test-set were categorized by the complexity of objects in the scene and by the complexity of object and camera motion in the image sequence. The ground-truth information in the real image sequences was generated by using polyhedral objects of known dimensions and a calibrated camera moving along a known trajectory. They analyzed the performance of seven different optic flow algorithms on both their synthetic and real image test beds. They found that the order of the performance results were consistent for the synthetic and real image sequences and concluded that this serves to validate their synthetic test set.

Langer and Mann (2003) used synthetic image sequences to study a particular type of environment that causes problems for classical optic flow algorithms. They considered the case of an observer moving through a cluttered three dimensional environment with objects at many depths. They called this motion field "optical snow" since it is analogous to the experience of seeing snow fall. Movement through a dense forest scene also falls into this category of motion field. Since traditional optic flow algorithms rely on smooth changes in motion, they perform poorly in this type of environment where there are many large discontinuities in the motion field. Langer and Mann developed an algorithm that uses motion parallax information, changes in motion information across parts of the scene, to estimate the camera motion. They tested their algorithm on both synthetic cluttered scenes as well as real image sequences of camera motion through forest environments.

The approach taken by Langer and Mann resembles the approach of this dissertation in two key points. They are interested in how environmental properties relate to vision, and they used simulated scenes. In contrast to the simulated scene statistics approach, they did not base their environmental model on any scene measurements, nor did they test the validity of their environmental model. Additionally,

their focus was on improving the performance of specific algorithms in computer vision. The focus of this dissertation is on modeling the optimal design of some aspect of any visual system in the environment. These optimality constraints can be as informative to computer vision research as they are to human vision research.

Computer vision researchers regularly generate artificial scenes and use ground-truth information from their simulations to test their algorithms. Human vision researchers measure image statistics in the world to try to understand the design of the human visual system. But combining these two approaches, studying vision by simulating across-domain statistics has not yet been done and opens up avenues of theoretical research that are not possible with only within-domain statistics. Across-domain statistics, can help not only in creating more accurate computer vision systems, but also can help in attaining a deeper understanding of the human visual system.

2.3 Environmental Model

In order to build an accurate simulation of the world, the environmental model used in this dissertation is based on the previously measured range and surface properties of natural environments.

Using a laser range-finder with a rotating mirror, Huang et al. (2000) collected 54 panoramic range images from forest environments. They found that the range statistics of forest scenes can be modeled well with a world consisting of a flat ground plane populated with a Poisson distribution of cylinders (i.e. simulated approximations to trees). For the top half of the forest scenes, which were dominated by the trees rather than the ground plane, they modeled the probability distribution of ranges as an exponential: $f(r) \propto \lambda L e^{-\lambda L r}$, where r is the range, λ is the density of the cylinders and L is the width of the cylinders. They also examined the derivative and the bivariate statistics of range and found that they were consistent with a world of piecewise smooth regions. In addition they found that the statistics of range images were scale invariant. In comparison to image statistics, range seemed to be a better cue for deriving the object structure of the world than optical measurements like luminance or color.

In a seminal piece of research in natural scene statistics Field (1987) measured the power spectrum of static images of natural environments and examined their

relationship to a model of human visual cortical cells. He took six pictures of natural scenes using a film camera and then digitized them. He calculated the Fourier power spectrum of these images averaged across all orientations. He found that the power spectrum of these images consistently falls as $\frac{1}{f^2}$, where f is the spatial frequency. Note that a $\frac{1}{f^2}$ curve in the power spectrum corresponds to a $\frac{1}{f}$ curve in the amplitude spectrum. Field points out that, although it is surprising how consistent the power spectrum behavior is across images, it is what one would expect if the distribution of frequencies in the environment were scale invariant. Field then showed how a bank of Gabor filters can serve as an efficient code for representing images with this power spectrum. The key observation was that having filters with a constant octave bandwidth, like those observed in V1, will lead to equal energy in each spatial frequency channel when exposed to natural images.

The physical structure of the environmental model used in this dissertation is based on Huang et al.'s range model. The forest scenes were rendered as a flat ground plane with a random distribution of cylinders (trees). Notice that doubling the density of trees will shift the probability distribution of ranges, but doubling the density of trees and halving the width of the trees will not change the distribution of ranges. This manipulation will, however, increase the frequency of object boundaries visible in the scene.

The surface texture of all the objects (cylinders and planes) in the environmental model were covered with a surface texture whose amplitude spectrum falls as $\frac{1}{f}$. In some conditions, in order to vary the spatial frequency content of the simulated surfaces, the texture used had an amplitude spectrum falling as $\frac{1}{f^{1.5}}$.

2.4 Psychophysics

In the third research project in this dissertation, described in Chapter 5, a psychophysical experiment is described testing the hypothesis that receptive-field size of motion units increases with increasing speed. The experiment is designed to psychophysically measure the receptive-field size of motion discrimination units in the human visual system. This section will describe the previous psychophysical work leading up to this experiment.

There is a large body of research attempting to psychophysically estimate the receptive field properties of visual neurons. A selected reading list is cited here,

and the papers specifically dealing with how velocity and receptive field size interact are described in detail below (Anderson and Burr, 1987, 1989, 1991; Anderson et al., 1991; Burr et al., 2006; Fredericksen et al., 1994, 1997; Georgeson and Scott-Samuel, 2000; Spillmann et al., 1987; Tadin and Lappin, 2005; Tadin et al., 2003; van de Grind et al., 1986, 1983; van Doorn and Koenderink, 1984; Watson et al., 1983; Watson and Turano, 1995).

In an aptly titled paper, "What does the eye see best?", Watson et al. (1983) measured human performance to find the stimulus that was most efficiently detected. They suggested that this stimulus describes the receptive-field of the most efficient human contrast detector. They used an ideal-observer model that includes a model of quantum noise to calculate the contrast energy of their stimuli. When the threshold contrast energy of the stimulus was maximized, the human was performing at maximum efficiency. For their stimuli, Watson et al. used drifting Gabor patches and varied their properties along five dimensions: the spatial frequency, the temporal frequency, the width, the height and the duration. All the dimensions except width and height were manipulated independently while keeping the other dimensions constant. The width and height were changed in conjunction, keeping the other dimensions constant. The optimal Gabor stimulus was found to be 7 cycles/degree at 4 Hertz with a width and height of 3 cycles (0.5 degrees) and a duration of 160 ms.

This ideal observer methodology from Watson et al. is the one employed in Chapter 5 of this dissertation. Instead of searching for the single best stimulus, the experiment in this dissertation found the optimal stimulus width as a function of the spatial and temporal frequency of the stimulus.

Van de Grind et al. (1983) were interested in seeing how motion detection performance varies as a function of eccentricity across the visual field. In a large parametric experiment, they measured threshold signal to noise contrast for motion detection at a range of eccentricities, stimulus widths and velocities. Their stimulus was an array of moving random pixels embedded in random pixel noise and was therefore broadband. They argued against the claim that motion detection performance in the eccentric visual field differs qualitatively from foveal performance. They found that all performance differences could be explained by cortical magnification, i.e. the scaling of receptive field sizes with eccentricity.

Van de Grind et al. had no explicit measurement of efficiency or optimal

velocity as a function of stimulus width. However, they modeled the threshold signal to noise ratio as a function of velocity, stimulus width, and eccentricity. Their model includes a high-velocity cutoff as a function of width. They found that this critical velocity is proportional to stimulus width, independent of eccentricity. This is evidence supporting the hypothesis that the receptive field size of motion neurons increases with velocity.

Van Doorn and Koenderink (1984) performed a similar experiment measuring threshold signal to noise ratio for foveal moving dot patterns. They found that for stimulus widths below some critical width, motion is impossible to detect. In addition they find that this critical width linearly increases with velocity. In a follow-up project, van de Grind et al. (1986) used the same methodology, but extended the work to eccentric visual locations in addition to the fovea. They replicated the finding that critical velocity increases linearly with stimulus width and also found that this is the case independent of eccentricity.

Burr et al. (2006) were interested in studying the spatial resolution for perceiving motion defined contours. They created a stimulus of alternating bars of horizontally moving dots, where each successive layer moved in the opposite direction. They varied the width of the bars and measured threshold width for performing two different tasks. One task was to localize the motion boundary relative to a marker on the screen and the second task was to discriminate the coherent motion from a control stimulus with incoherent motion. Since both experiments measured a critical width at which motion detection is possible, it can be argued that they were measuring a quantity that is proportional to the receptive field size of motion detectors. Burr et al. also varied the spatial frequency content of their stimuli starting with broadband stimuli and filtering the stimuli to narrower bands. They found that for both tasks the threshold stimulus width increased with increasing speed.

Despite this large body of research, there are some gaps. All the previous experiments that explicitly measure psychophysical receptive-fields size as a function of velocity did so with broadband stimuli. Also, none of the studies measuring receptive-field size as a function of velocity used the more reliable ideal observer methodology. Chapter 5 of this dissertation describes a psychophysical experiment that measures receptive field size for narrow-band stimuli as a function of both spatial and temporal frequency and uses the ideal observer methodology.

2.5 Conclusion

To summarize, research in several different areas has been reviewed and in each area, this dissertation builds on previous work. In the area of natural image statistics, no previous researchers have captured image sequences of natural environments in the context of a specific task. In Chapter 3, image sequences are captured in the context of first-person motion through a forest. In the area of across-domain measurements, computer scientists have created simulated motion sequences for use as training or testing data for computer vision algorithms. Also, some human vision researchers have directly used range data to simulate some of the relevant environmental statistics for vision during locomotion. However, combining the two approaches and simulating image sequences to try and understand how the environment and task influence an ideal visual system has not yet been done. Simulations of this nature are described in Chapter 4. In psychophysics, no one has estimated the psychophysical receptive-field size of motion sensors as a function of speed using narrow-band stimuli, as is described in Chapter 5.

Chapter 3

Image Statistics

This chapter addresses three broad questions:

1. What do the natural scene statistics of motion through forest scenes look like?
2. How do the statistics of the environmental model compare to real-world statistics?
3. How do the simulated and natural statistics compare to the current dominant model?

To answer these questions, natural movies of navigation through a forest environment were captured using a calibrated digital camera and a custom built telescoping rail. Simulated movies were generated using the environmental model and a ray tracer. Spatiotemporal Fourier power spectra were computed from the natural and simulated movies and compared. Additionally, both the natural and simulated power spectra were compared with with the current dominant model of motion statistics (Dong and Atick, 1995a). Note that in this chapter, only image domain statistics are measured and compared. Simulated scenes are the primary focus of this dissertation precisely because of the difficultly of measuring across-domain statistics in real scenes. This chapter serves to validate the environmental model in the image domain before it is used to gather across-domain information.

3.1 Methods

In order to compare real and simulated scenes, image motion sequences were captured in natural environments and generated in simulated ones. Fourier power spectra of these movies were computed and used for comparison. This section describes the methods for creating natural and simulated movies and for computing their power spectra.

3.1.1 Natural Scene Measurements

Using a custom built telescoping rail and a calibrated digital camera, sequences of images were captured that simulate observer motion through natural forest environments. Images were captured using a digital camera that is calibrated to give trichromatic pixel values that can be translated reliably into luminance and chromaticity coordinates (or alternatively into relative activation levels for the three color sensitive human photoreceptors, L, M and S cones). The custom built rail mounts the camera at a normal eye height of about 172 cm. The rail moves the camera in 30 precise increments of 3.3 cm along a meter long path. Additionally, the rail allows for the collection of images with horizontal disparity information, since the rail allows for precise lateral movements over the distance of half a meter (Figure 3.1). Although the disparity and color information is not used in this dissertation, the collection of image sequences is intended to be a contribution to the field, and thus an effort was made to gather information that might be useful for future research.

Since this work intends to link measured statistics to human perception, it is important that the sequences of images correspond to a real perceptual phenomenon experienced by humans. Normal walking speed is a little more than one meter per second. At this speed, if one takes a picture at three centimeter increments over the course of a full meter, that results in one full second of video at 30 frames/second. Given the speed of human judgments of heading, one second is a sufficiently large time interval over which to sample.

During shooting, the time interval between capturing images was on the order of a minute. There were two reasons for this long interval. First, the calibrated camera is a still-image camera and not a movie camera and therefore individual

Figure 3.1: **Photograph of the custom built camera rail.** This photograph shows the telescoping rail with the camera mounted and the rail extended a full meter. The rail allows the camera to be moved forward and back in precise increments over the distance of a meter. The rail telescopes so that no part of the rail is visible in the images. The rail also allows precise lateral movements over a distance of about half a meter so that images with disparity information can be collected.

frames had to be shot in sequence. Second, because of the need for precise camera motion, the telescoping rail was used to move the camera to the next location between each frame. Because of the long delay between frames, shooting locations and times were chosen carefully to insure that there was no motion in the scene due to wind and no sharp shadows in the scene, both of which can create motion artifacts in a stop-action movie. Visual inspection of the movies made it clear that motion captured in the image sequences was indeed due to the motion of the camera through the scene, and not artifacts.

For this analysis, in order to remove focus artifacts and other potential artifacts, the captured images are down-sampled to a third of their native resolution, resulting in image sequences (movies) that are 30 frames long and 750 pixels wide by

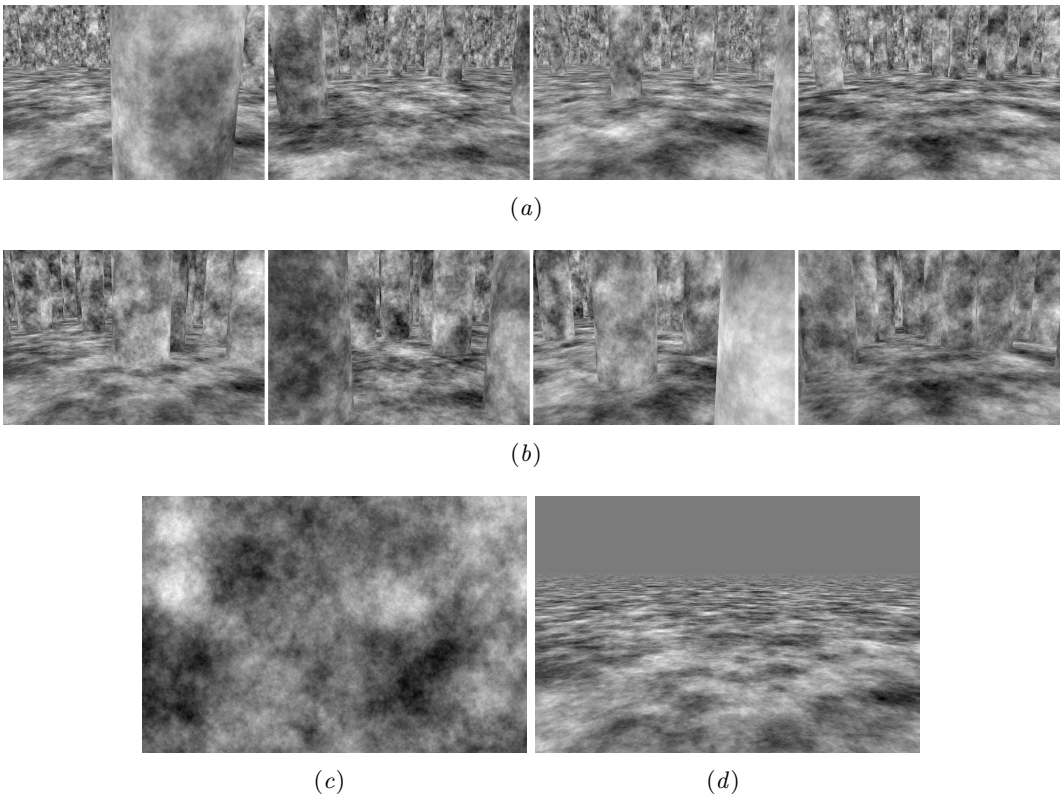


Figure 3.2: **Examples of the four different types of simulated scenes.** Shown above are single frames from movies generated using each of the four different environmental models. (*a*) Four frames from different movies using the forest model with a low density of cylinders. (*b*) Four frames from different movies using the forest model with a high density of cylinders. (*c*) Flat vertical plane. (*d*) Flat ground plane.

500 pixels high. The camera images corresponded to visual angles of approximately 21 by 14 degrees.

3.1.2 Ray Tracer

In order to build an accurate simulation of the world, the environmental model used in this dissertation is based on the range measurements of Huang et al. (2000) and the static image measurements of Field (1987) (Chapter 2). In order to replicate as much as possible the circumstances of the natural movies, a ray tracer (MegaPOV-Team, 2005; Persistence of Vision Pty. Ltd., 2004) was used to generate movies of

motion through the simulated environment. The visual angle of the camera and the motion of the camera through the scene was replicated so that each movie frame was 750 by 500 pixels corresponding to a visual angle of 21 by 14 degrees.

Four different kinds of scenes were generated at two different human walking speeds (0.9 and 1.5 meters/second) making eight different scene conditions. The four different kinds of scenes are shown in Figure 3.2: two "forest" scenes (*a,b*), one flat-wall scene (*c*) and one ground-plane only scene (*d*). The two forest scenes were based on the range statistics of Huang et al. (2000). The scenes consisted of a flat ground plane with a Poisson distribution of cylinders. Different forest scenes were generated by drawing random samples of cylinder locations. The cylinders had a radius of 0.6 meters and were distributed with a density of either 0.0538 cylinders/meter^2 (low-density) or 0.215 cylinders/meter^2 (high-density). This range of cylinder densities was chosen to span the reasonable range of forest densities.

To match measured image statistics, the reflectance of all the surfaces in all the scenes is a $\frac{1}{f}$ procedural texture Field (1987). This texture is the weighted sum of Perlin noise (Perlin, 2002) of different frequencies. The advantage of using a procedural texture is that it gives a continuous and smooth $\frac{1}{f}$ reflectance function at any viewing distance. Thus, using a procedural texture instead of texture mapping allows the texture to be rendered and anti-aliased accurately at any distance. In each movie, the model observer moved forward and gazed in the direction of translation, 10 degrees down from the horizon. The model observer's eyes were located 1.5 meters above the ground plane.

3.1.3 Fourier Analysis

To compare statistics from the natural and simulated movies, Fourier power spectra were computed. The spectra were computed by averaging the power of randomly selected samples of size 64 pixels by 64 pixels by 30 frames. Each movie was sampled 1000 times and each sample was filtered using a Welch window in space and time. This computation was carried out for each individual movie, for the ensemble of simulated movies and for the ensemble of natural movies. Power as a function of spatial frequency was binned radially. All spatial frequencies referred to in this chapter are radial spatial frequencies. These sampling and averaging techniques were chosen to mimic the methods used by Dong and Atick (1995a), making it

possible to fit the data to their model. One advantage of sampling from the movies is that it allows for comparison across movies of different dimensions. It is highly unlikely that a different sampling scheme or using the whole movie would change any of the qualitative results. In order to see how these measurements compared to those made by Dong and Atick, the data from the ensembles of movies were fit using Dong and Atick's model (see Appendix A for details).

Binning spatial frequencies radially removes one dimension, but three-dimensional data still remains: power as a function of radial spatial frequency and temporal frequency. In order to visualize this complicated space, four different plots will be used throughout this chapter: a color temperature plot of the full three-dimensional power (Figure 3.3a), a plot of power as a function of *temporal* frequency at different spatial frequency cross sections (Figure 3.3c), a plot of power as a function of *spatial* frequency at different temporal frequency cross sections (Figure 3.3d) and a plot of the 'frequency-adjusted' power as a function of velocity (Figure 3.3b). The dashed lines in plot (a) indicate the cross sections plotted in (c) and (d). To give some intuition to this complicated space, note that in Fourier space, an object that moves with a particular velocity will contribute to the power observed at multiple points in the space. These points will lie along a diagonal line specified by the relation $v = \frac{\omega}{f}$, where v is the velocity of the object in the image plane, and f and ω are the spatial and temporal frequency components. Thus, iso-velocity lines in the space in Figure 3.3a are simply straight lines running through the origin where the slope of the line corresponds to the velocity. One might expect that summing and plotting the power at each $\frac{\omega}{f}$ (velocity) might give a sense of how much of a particular velocity is present in a scene. This is not the case, however, because the static texture of every moving part of an image contains power in many different spatial frequencies. For this reason, Figure 3.3b is intended to correct for the average power at each spatial frequency and gives a sense of the distribution of speeds in the image sequences (for a more detailed explanation and the math behind this figure see Appendix A). For all of these plots, measured data are shown as points and fits from Dong and Atick's model are plotted as curves.

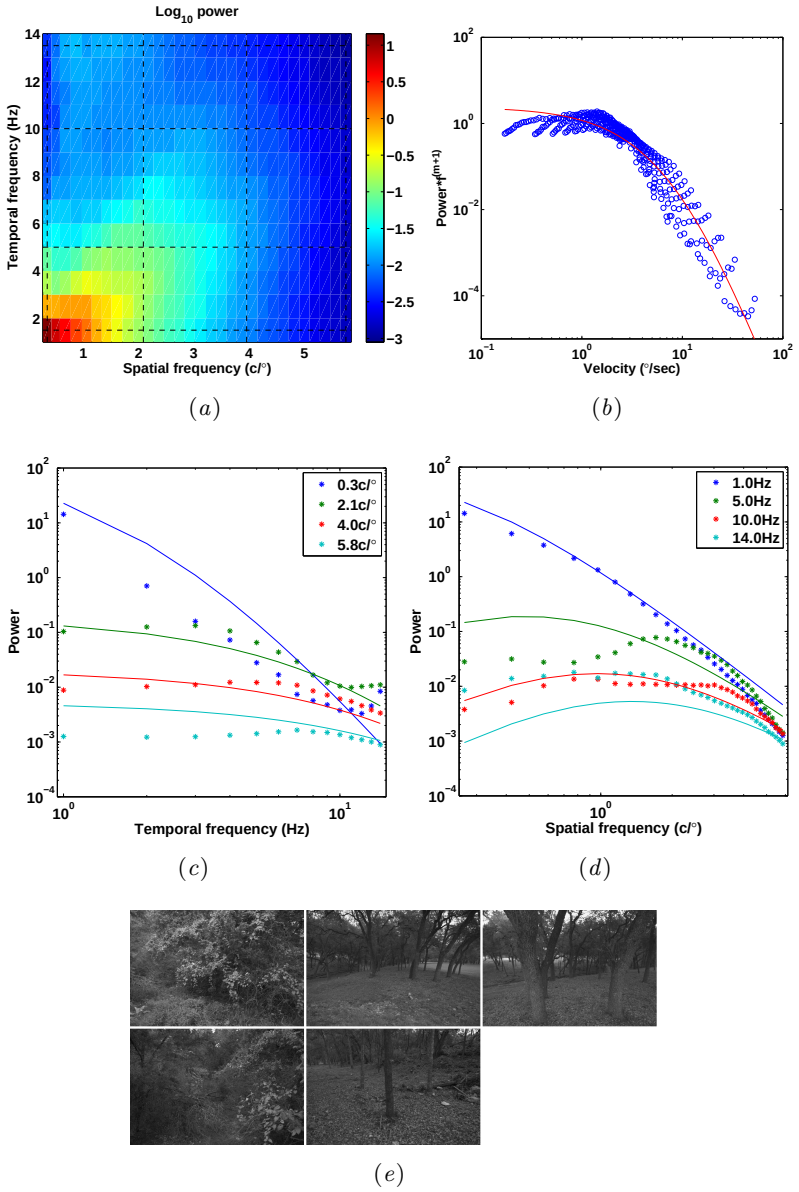


Figure 3.3: **Power spectrum of natural scenes.** Average Fourier power spectrum of all five natural image sequences. (*a*) The log power as a function of spatial and temporal frequency is plotted using color temperature, with a scale at the right. The dashed lines indicate the locations of the temporal and spatial cross sections shown in (*c, d*). (*b*) Power that has been adjusted by the average spatial frequency of the scene is plotted as a function of velocity. The solid lines in (*b,c & d*) indicate the fits for Dong and Atick's model. (*e*) Example frames from each of the image sequences are shown. Dong and Atick's model does not offer a satisfactory explanation for the observed spectrum (see Figure 3.4 for comparison). Additionally, this power spectrum differs from the spectrum for the ensemble of all simulated scenes (Figure 3.5)

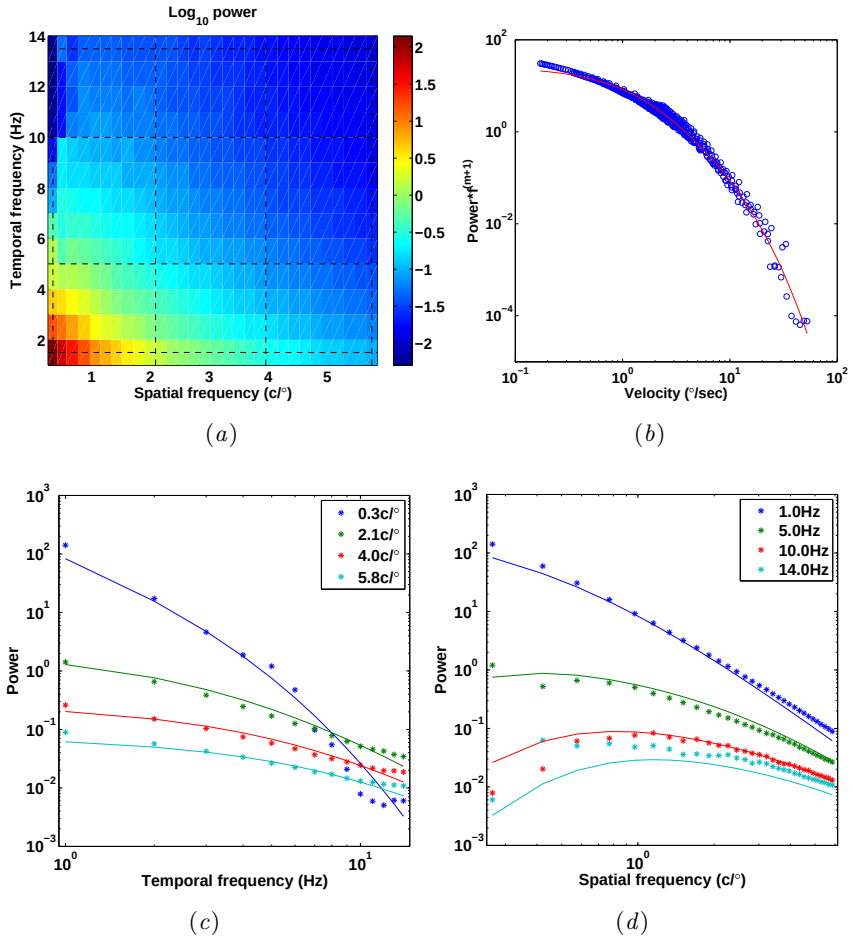


Figure 3.4: **Power spectrum of a replication of Dong and Atick's model.** Dong and Atick's model does not offer a satisfactory explanation for either the simulated scenes (Figure 3.5) or for the natural scenes (Figure 3.3).

3.2 Results

To compare statistics across simulated and natural scenes, a power spectrum was computed for the ensemble of all natural movies and for the ensemble of all simulated movies. Subsequently, to explore the effects of different environmental aspects, Fourier spectra were computed and compared for individual scenes.

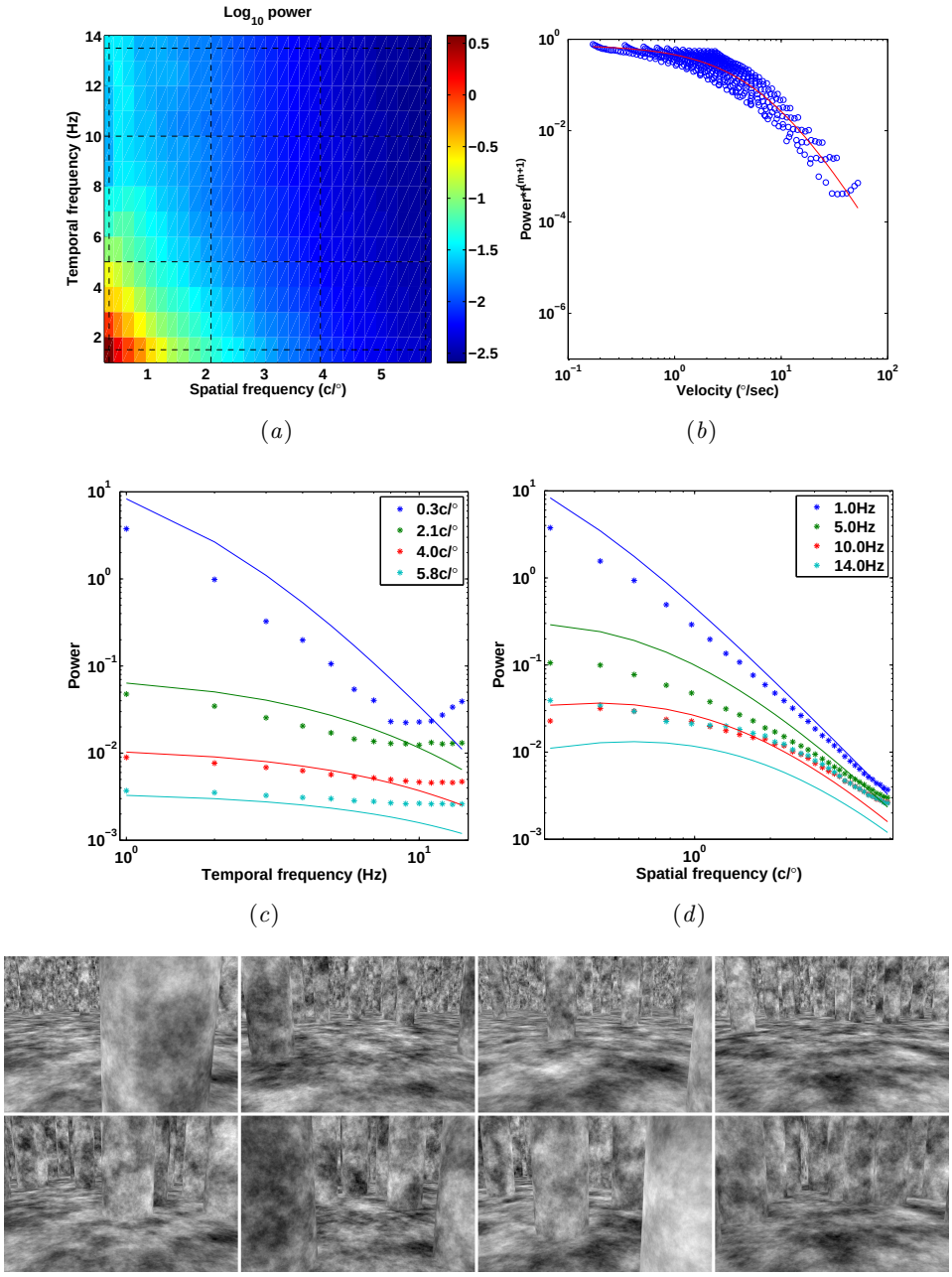


Figure 3.5: **Power spectrum of all simulated forest scenes.** This spectrum differs from both simulation of Dong and Atick's model (Figure 3.4) and the natural scenes (Figure 3.3).

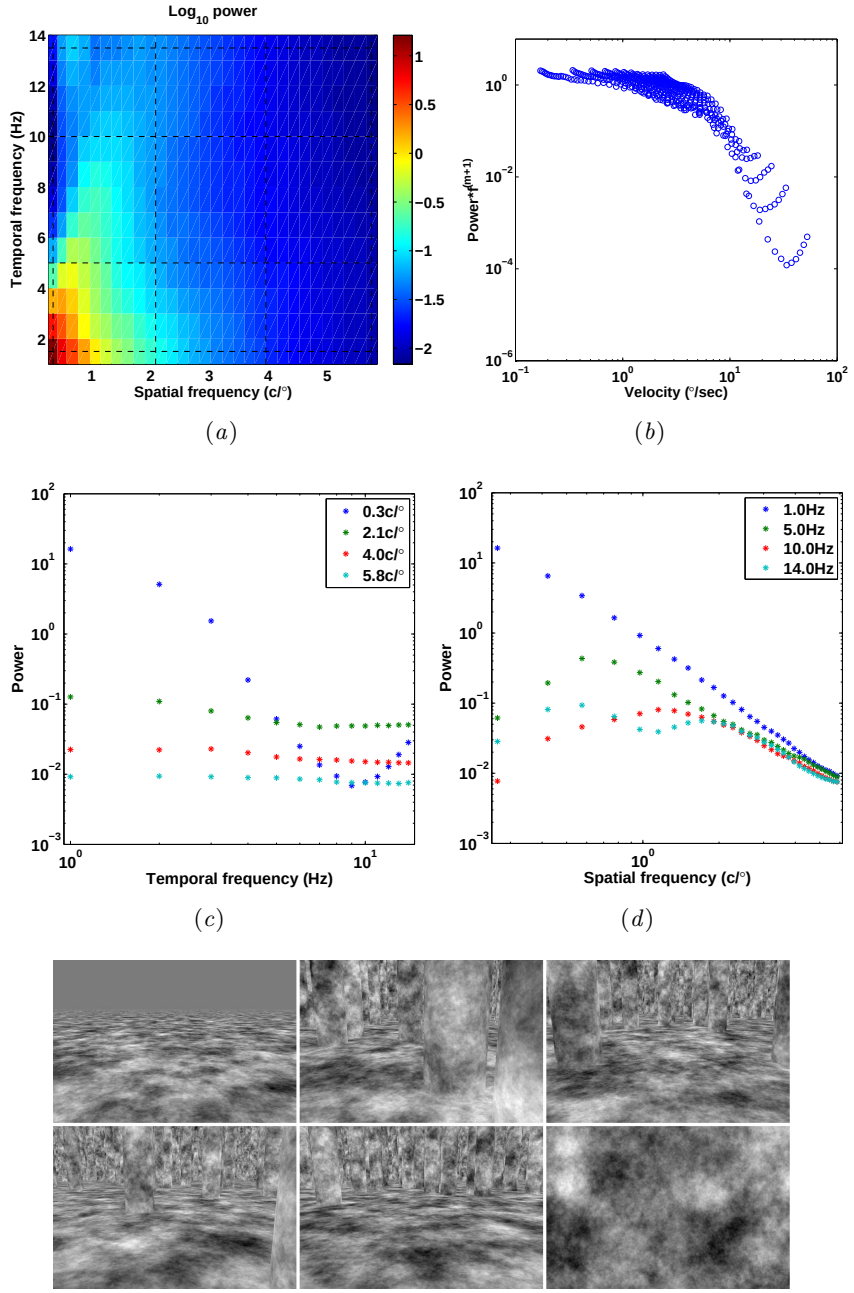


Figure 3.6: **Power spectrum of a custom ensemble of simulated scenes.** Scenes were picked by hand in an attempt to match the measured natural power spectrum show in Figure 3.3.

3.2.1 Ensembles of Scenes

Figure 3.3 shows the Fourier power spectrum for the complete collection of five natural image sequences. The solid lines in subplots (b), (c) and (d) show the fits of Dong and Atick's model to the measured data. This model does not explain the observed data well. In comparison, Figure 3.4 shows the power spectrum for a replication of Dong and Atick's model. For this figure, image sequences of translational motion were generated with velocities drawn from the probability distribution described by Dong and Atick. These image sequences were then binned and fit using the same procedures. These data serve as a self-consistency check, verifying that the procedures are capable of fitting the data. In addition, they serve as a baseline for comparison. If the assumptions of Dong and Atick's model hold true, then the data points in Figure 3.4b would fall on a single line and would not exhibit the scatter apparent in Figure 3.3b. Additionally, the model distribution of velocities (shown in Figure 3.4b) is quite different from the observed distribution in Figure 3.3b.

Figure 3.5 shows the Fourier power spectrum for the complete collection of eight simulated forest scenes (four low-density sequences and four high-density image sequences). Again, the Dong and Atick model does not do a good job of fitting the simulated data. It is clear from the scatter in Figure 3.5b that there is a significant amount of non-translational motion. Comparing Figure 3.5 with Figure 3.3 we can see that these simulated scenes also do not compare favorably to the ensemble statistics from natural scenes. Again, the distribution of velocities measured in Figure 3.5b is quite different from that observed by Dong and Atick.

It might be possible that the characteristics of the individual scenes in the ensemble of image sequences might have a strong effect on the nature of the statistics. In order to explore this possibility, a custom ensemble of artificial image sequences was hand picked in an ad hoc attempt to match the measured natural image sequences. Figure 3.6 resembles the measured natural statistics much more than either the complete ensemble of forest scenes (Figure 3.5) or the replication of Dong and Atick's model (Figure 3.4). Thus, it is not just the task that is important when measuring natural scene statistics; the specific properties of the measured scene also strongly influence the Fourier power spectra. This may seem trivial, but is contrary to the results of Dong and Atick. Additionally, the static power spectrum of natural images is so consistent across different images, it is somewhat surprising to find that

environmental differences affect the spatiotemporal power spectrum.

3.2.2 Individual Scenes

The fact that different ensembles of scenes have different statistics suggests that the Fourier power spectra of individual scenes should be examined as well. Figures 3.7–3.9 show the spectra of three individual natural scenes, while Figure 3.10 shows the spectrum of an individual simulated low-density forest scene, and Figure 3.11 a high-density one. Also, Figure 3.12 shows the power spectrum of an observer approaching a textured wall, and Figure 3.13 shows the power spectrum of an observer walking through an environment with only a textured ground plane (a forest scene with no trees).

The content of the natural scenes has a strong effect on the power spectrum. When comparing natural and simulated movies, the movies whose environmental make-up most resemble each other also have power spectra that resemble each other. For example, the power spectrum of Natural Movie 1 (Figure 3.7), in which the observer approaches a bush at a fairly uniform distance, is most similar to the power spectrum of the flat wall (Figure 3.12). Scenes like these, where objects are at a narrow range of distances, are going to exhibit narrow ranges of velocities. Both these spectra are dominated by a strong diagonal ridge in their power 3-D spectrum (red/yellow shape in both of the (a) subfigures). A diagonal ridge corresponds to power concentrated around a specific velocity, indicating that most of the power in these scenes is in a narrow range of velocities. Notice that the velocities for the simulated wall scene are higher (i.e. have a larger slope) than that for the natural movie. The particular velocities in each scene are a function of the specific relationship between the observer's walking speed and the distances of the objects in the scene. Comparing the (b) subfigures shows that the distribution of velocities in these two scenes is flat and then drops suddenly. Notice also that there is a small degree of scatter in both the (b) subfigures indicating that the motion in these scenes is fairly translational.

Looking at the power spectrum of Natural Movie 5 (Figure 3.9), in which the scene is dominated by ground plane and has only a few thin trees, we can see that it is similar to that of the simulated ground plane (Figure 3.13) and also to a movie from the low-density forest scenes with very few trees (Figure 3.10).

Finally, the power spectrum from Natural Movie 4 (Figure 3.8), which contains dense foliage at a range of distances, most resembles the power spectrum from a simulated movie which contains a dense distribution of trees (Figure 3.11). For the complete set of power spectra from all five natural movies and all 8 simulated movies, see Appendix C. These types of corroborations between the measured power spectra and the environmental structure, lend validity to the environmental model.

Under a particular set of assumptions (detailed in Appendix A), subfigure (b) shows the unnormalized distribution of image-plane velocities in the scene. These assumptions are the same as in Dong and Atick's model, with the exception of the power-law distribution of velocities. When these assumptions are true, all the data points in these plots will fall on a single line. Thus the more scatter in the points, the more the assumptions are not met. In comparing across scenes, the only assumption that changes significantly is the translational nature of the motion. Comparing subfigure (b) across all the individual scenes, both simulated and natural (Figure 3.7–Figure 3.11), it is possible to understand which types of scenes have more or less translational motion. As one would expect, the flat wall scene with no ground plane has the least amount of scatter (Figure 3.12b). The natural scene that most resembles the wall, also has very little scatter showing that the small amount of occlusions and sheering due to the leaves and branches in this scene does not create a great deal of non-translational motion (Figure 3.7b). In the other scenes, the more ground plane that is visible in the scene, the more scatter is apparent in subfigure (b). It is interesting to note that the presence of more trees (compare Figure 3.11b and Figure 3.10b) seems to lessen the scatter. This difference again indicates that the sheering and occlusion due to the trees does not add a significant amount of non-translation motion in comparison to the non-translation motion created by the ground plane. Again, the similarities between the power spectra and environmental make up across simulated and natural scenes help validate the environmental model.

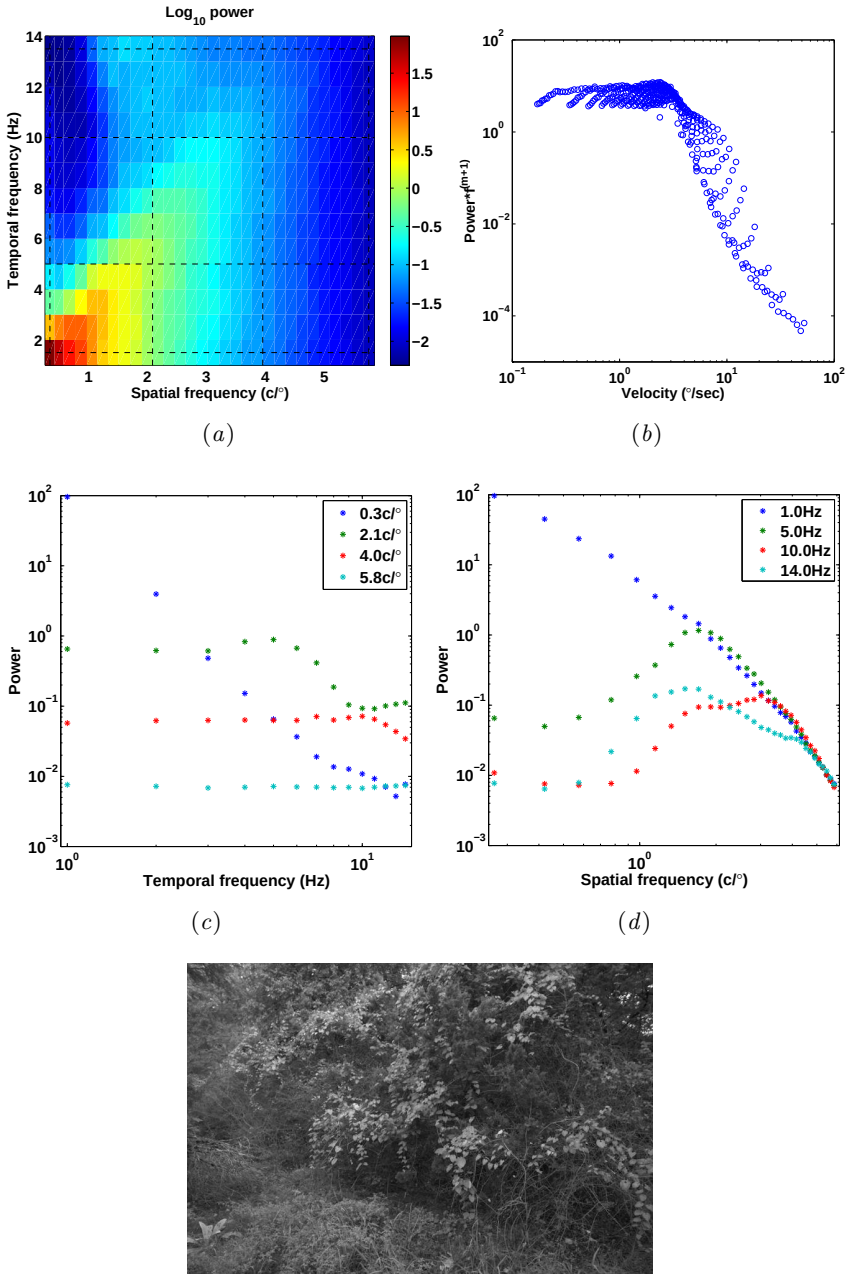


Figure 3.7: **Power spectrum of Natural Movie 1.** The power spectrum of this movie, in which an observer approaches a bush at a fairly uniform distance, is most similar to the power spectrum of the flat wall (Figure 3.12). This is an example of how simulated and natural scenes with similar properties have similar power spectra.

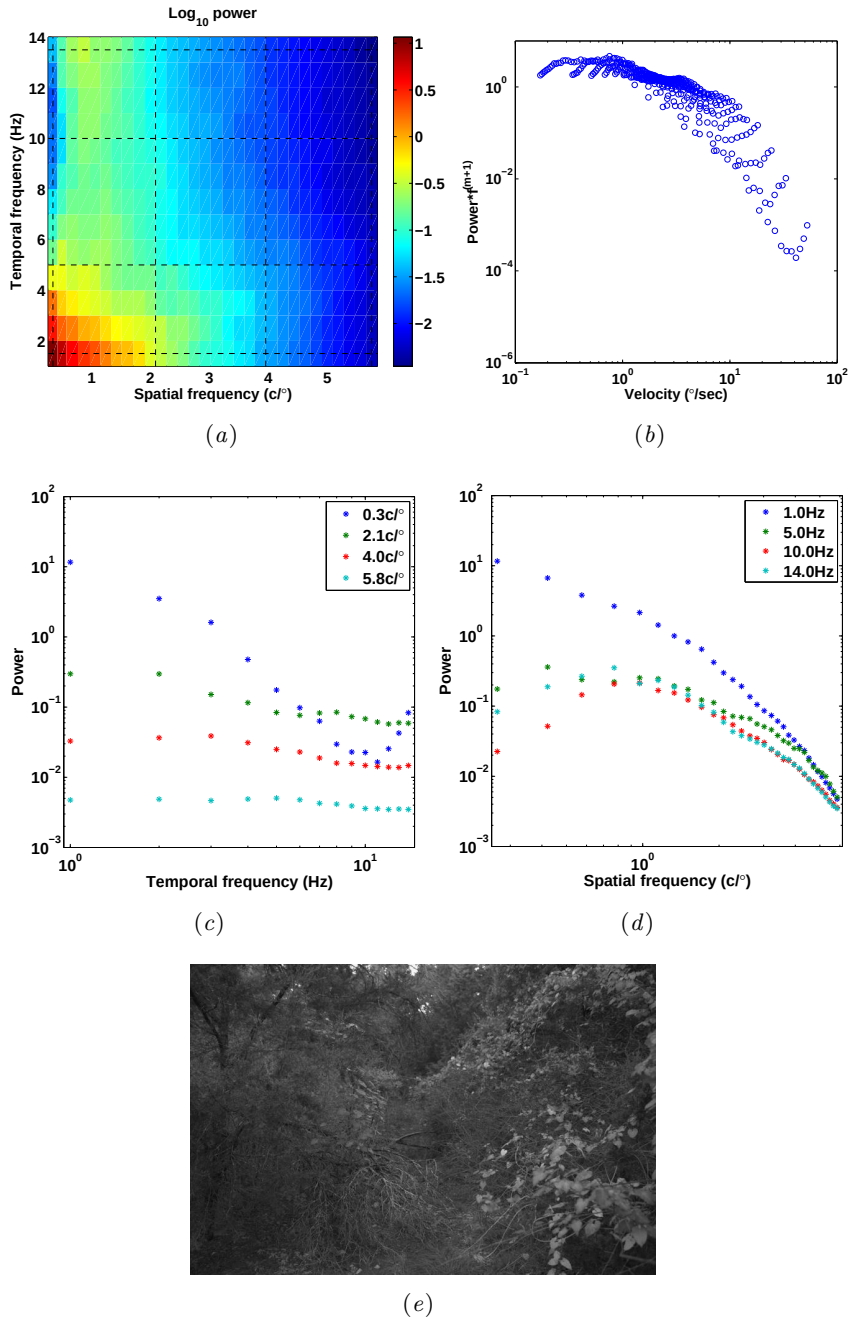


Figure 3.8: **Power spectrum of Natural Movie 4.** The power spectrum from this movie, which contains dense foliage at a range of distances, most resembles the power spectrum from a simulated movie which contains a dense distribution of trees Figure 3.11.

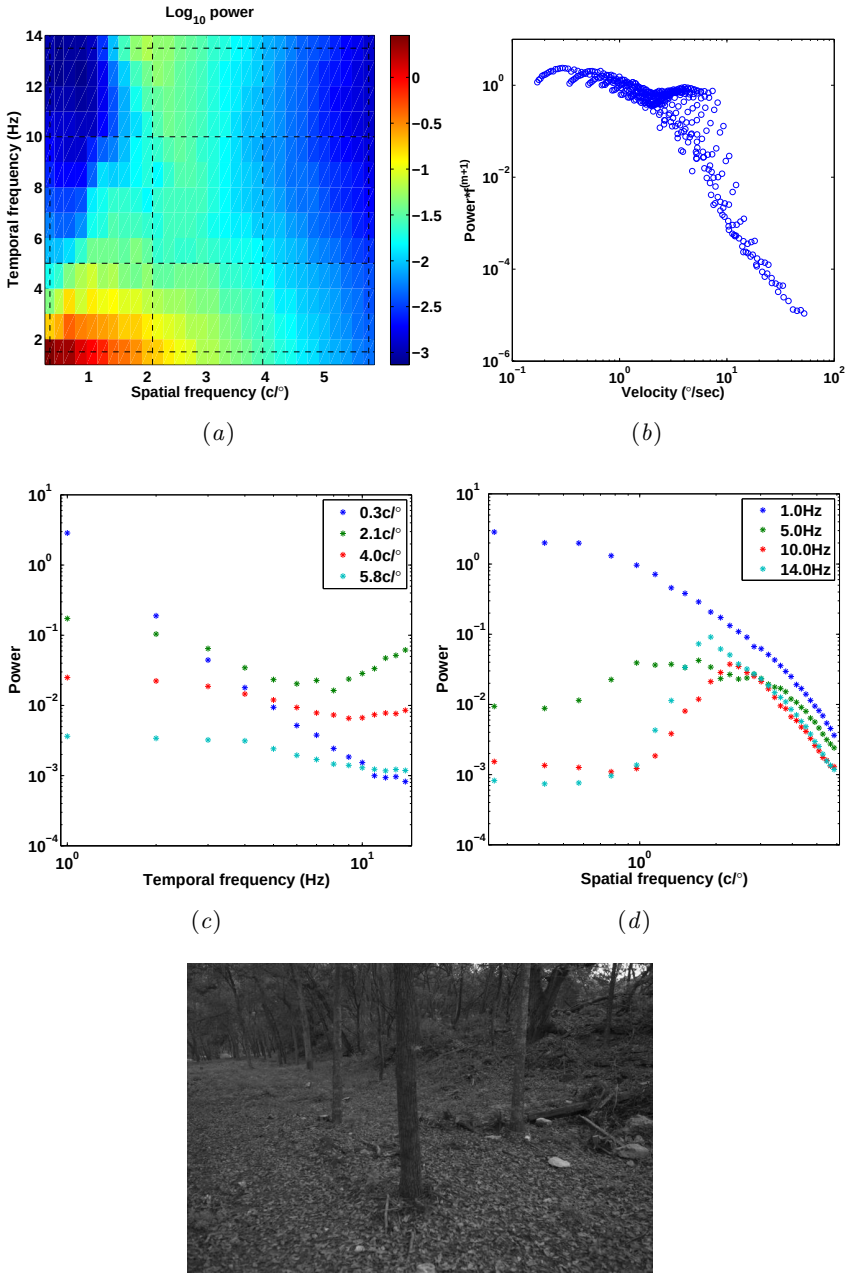


Figure 3.9: **Power spectrum of Natural Movie 5.** The scene in this movie is dominated by ground plane and has only a few thin trees. It is similar in its power spectrum and its environmental makeup to both the simulated ground plane (Figure 3.13) and to a low-density forest scene (Figure 3.10).

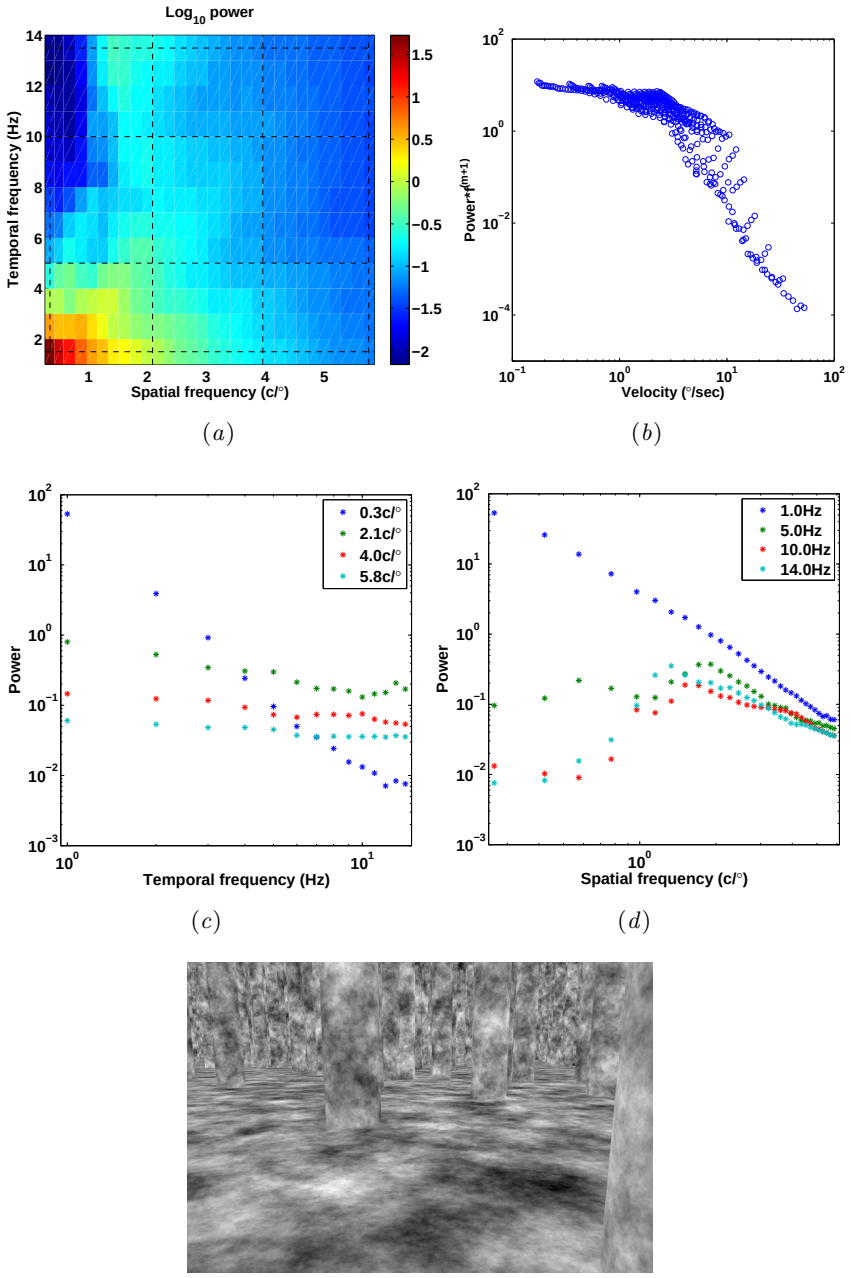


Figure 3.10: **Power spectrum of a low-density simulated forest scene.** Compare the power spectrum of this scene with the similarly configured Natural Movie 5 (Figure 3.9).

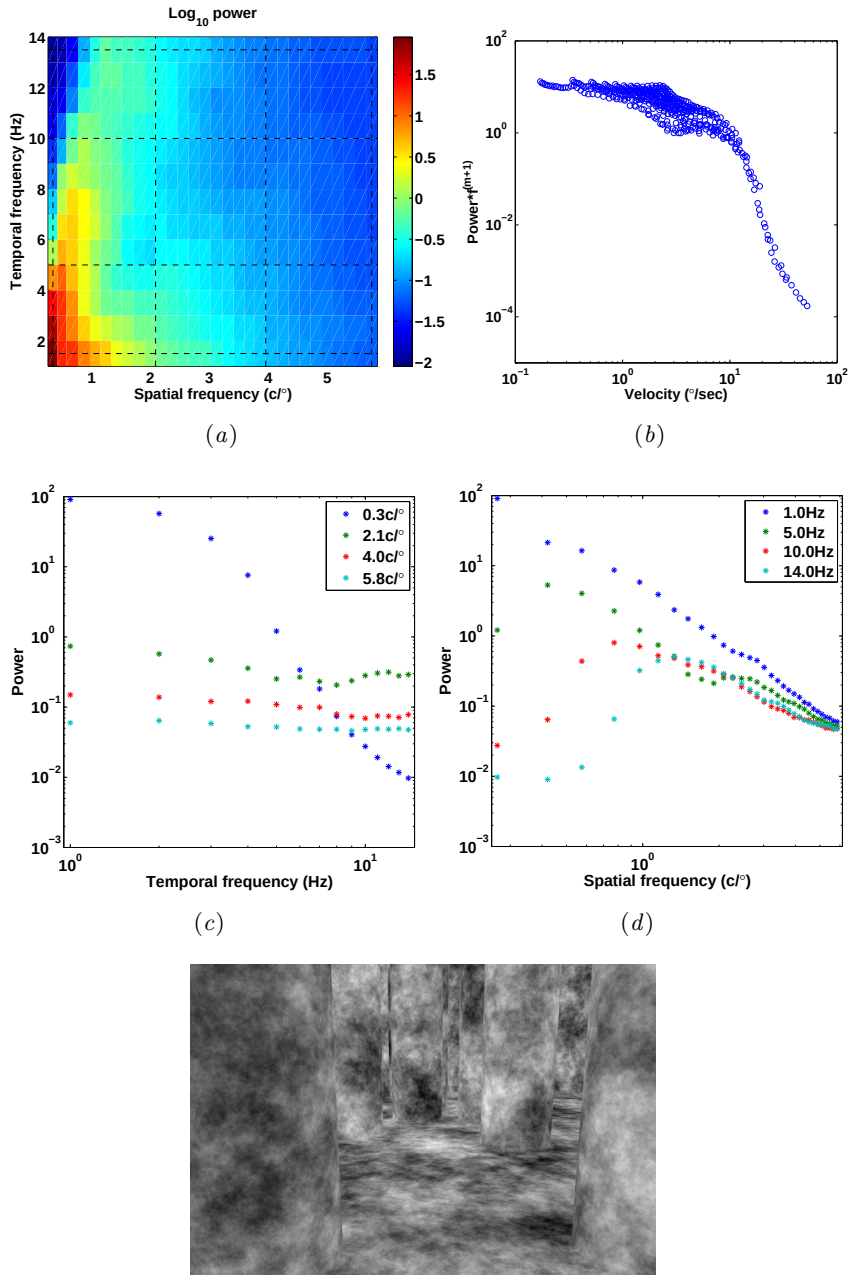


Figure 3.11: **Power spectrum of a high-density simulated forest scene.** Compare the power spectrum from this movie with that of Figure 3.8. Both have a high density of objects at a range of distances.

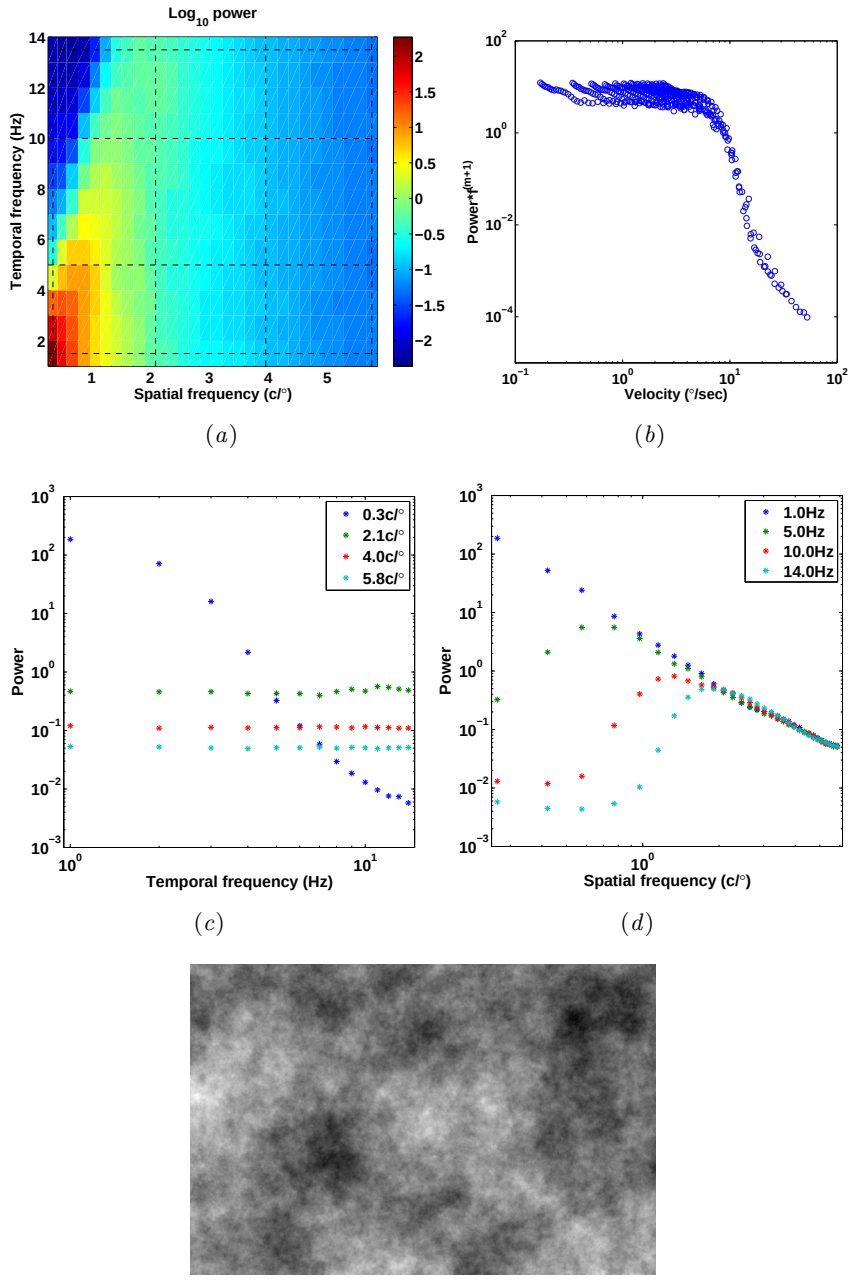


Figure 3.12: **Power spectrum of a simulated flat wall.** The power spectrum of this movie resembles the power spectrum from Natural Movie 1 (Figure 3.7), which has an environmental composition that is similar to a flat wall.

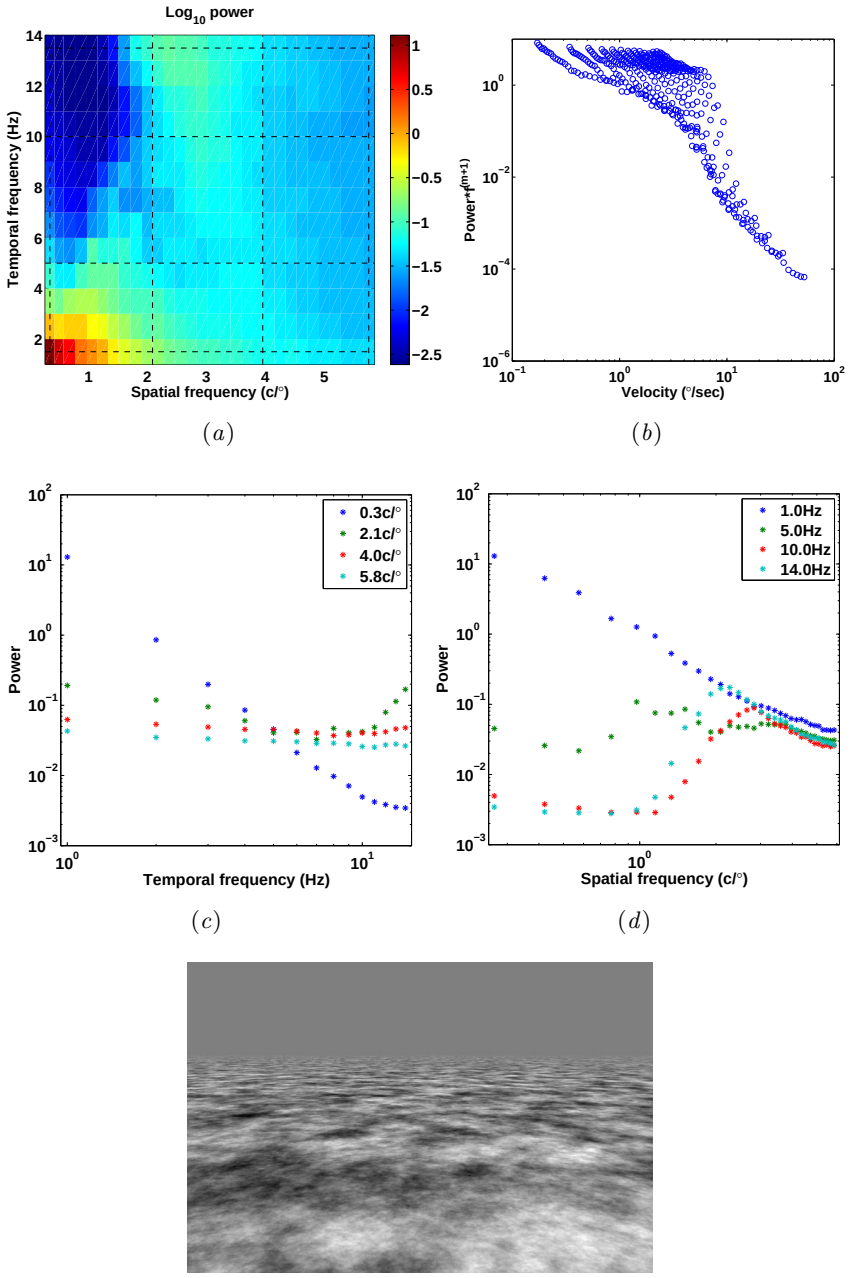


Figure 3.13: **Power spectrum of a simulated flat ground plane.** This scene's environmental makeup and power spectrum both resemble that from Natural Movie 5 (Figure 3.9).

3.3 Discussion

Motion image sequences were captured for real and simulated forest environments. The goal was to replicate the spatiotemporal retinal image experienced during navigation. Fourier power spectra were computed for individual motion sequences as well as for ensembles of sequences. They were then compared with the dominant current model of motion statistics, that of Dong and Atick (1995a). This model was found to be unsatisfactory in its ability to account for the measured power spectra.

There are two major assumptions made by Dong and Atick's model that explain this result. First, the amount of non-translation motion in our scenes was greater than that in Dong and Atick's measurements. Dong and Atick's movie sequences were a combination of custom video shot by the authors and clips from commercial movies. In retrospect, it seems likely that the motion in their movies was dominated by panning and other translational motions. The origin of all the motion in simulated and natural image sequences in this chapter was due to the first-person motion of the observer. The motion observed when moving across a ground plane is especially non-translational, but the occlusions and expansion due to objects in the scene also contribute to the non-translational nature of the image motion.

Second, the power-law distribution of velocities assumed by Dong and Atick does not seem to fit image motion created by first-person motion through forest environments. Not surprisingly, the measured distribution of velocities depends strongly on the environmental structure of the scene or ensemble of scenes being measured. Dense forest scenes will have a much narrower range of velocities and also a more uniform distribution of velocities. Power-law distributions are desirable computationally because they are invariant to changes in scale. The static power spectrum of natural scenes has previously been shown to be relatively scale invariant (Field, 1987). Now it is clear that the dynamic power spectrum of first-person motion is not scale invariant.

One might argue that using a larger sample of scenes (either natural or simulated) might cause Dong and Atick's model to fit better. This seems unlikely. Any set of scenes of first person translation are likely to have a large amount of visible ground, which will break the translational motion assumption. Additionally, the model's practicality is limited if it cannot account for the variability between

different types of environments.

All of the comparisons of power spectra presented here are qualitative in their nature. It would be desirable to make a more quantitative comparison using a model of the power spectra of natural scenes that alleviates the shortcomings of Dong and Atick's model. In particular, one would need a model that takes into account different distributions of velocities as well as the non-translational motions encountered during first-person motion. However, the assumption of translational motion comes not from the nature of the world, but rather from mathematical convenience. The translational motion assumption makes it possible to compute the necessary integrals for the model (Appendix A). How to contend with this mathematical difficulty is an avenue for future research.

3.4 Conclusion

In summary, image sequences were gathered in natural forest scenes and simulated in artificial ones. The Fourier power spectra of these scenes were compared with each other and with the current dominant model. The results show that it is important to gather image motion sequences in a task-dependent manner. Additionally, it is important to pay close attention to the environment in which image sequences were collected. Finally, the Fourier power spectra from individual simulated scenes, across a wide variety of types of scenes, showed marked similarity to the Fourier power spectra of individual natural scenes. Thus, the environmental model presented here seems capable of representing much of the variability and complexity of natural forest scenes. This forms a foundation for the subsequent work, since the validated environmental model can now be used to measure across-domain statistics that are too difficult to measure in natural environments.

Chapter 4

Designing Ideal Motion Sensors

In the previous chapter, the environmental model was validated in comparison to real-world statistics. Using this model, across-domain statistics can now be measured to explore how an ideal visual system ought to behave. Specifically, the following question is addressed: How should local motion detectors be designed to make accurate estimates when locomoting through the world?

4.1 Motivation

Motion detectors in computer vision and biological models of motion selective neurons commonly assume that local motion is well approximated by translational motion in the image plane. For such detectors, the accuracy of local velocity estimates depends on several factors including sensor noise, the extent to which the motion breaks the translational assumption, stimulus ambiguity (e.g. the aperture problem), and the spatial and temporal area over which information is pooled. Larger areas of integration typically contain more complex, non-translational motions. On the other hand, smaller areas of integration are more susceptible to the aperture problem and to sensor noise. For locomotion through an environment, there may be an ideal integration area for local motion detectors that will balance these constraints in order to maximize average accuracy. Furthermore, it is plausible that this optimal area depends upon the local image speed (magnitude of velocity). In this chapter, artificial scene statistical measurements are made to determine the optimal integration area for local image velocity detectors and how the optimal integration

area varies with local image speed. These results are relevant to biological vision research as well as to computer vision and robotics.

4.2 Methods

Image sequences were simulated for a variety of scene conditions using a ray tracer and local motion estimates were generated for a range of areas of integration (aperture sizes). Errors in the local motion estimates were computed as the difference between the ground-truth local motion and the estimated motion. The best aperture size for each speed for each scene condition was selected by averaging errors and picking the aperture size that minimizes average error.

4.2.1 Ray Tracer

Just as in the previous chapter, a ray tracer (MegaPOV-Team, 2005; Persistence of Vision Pty. Ltd., 2004) was used to generate movies for an observer translating through a model environment. Each movie consisted of six frames. At a 30 Hertz sampling rate, six frames corresponds to a movie duration of 200 ms, which is approximately the minimum time for a human fixation. Each movie frame was 316 by 252 pixels corresponding to a visual angle of 39 by 33 degrees. Samples from the movies were taken at different spatial aperture sizes. A valid location in the scene is defined as one which, for the largest aperture size, there is no clipping. The valid locations spanned 28 by 20 degrees.

This study employed the four types of scenes described in the previous chapter: high-density forest, low-density forest, wall and ground plane (Figure 3.2) as well as one additional forest scene type with thin trees. The motivation behind the additional thin-tree scene type was to increase the number of object edges in the high-density forest scene, keeping all other variables constant. To do this, the radius of the cylinders (trees) in the scene was halved to be 0.3 meters. In order to maintain an equal visual density of trees in the environmental model, the density was doubled to 0.431 cylinders/meter^2 (Huang et al., 2000). Doubling the density of cylinders and halving the width of the cylinders keeps the probability of there being a tree at a particular distance at a particular location constant. Each set of image sequences was generated at two different human walking speeds (0.9 and 1.5 meters/second).

The reflectance of all the surfaces in all the scenes was generated both with the typical $\frac{1}{f}$ texture and with a $\frac{1}{f^{1.5}}$ texture, containing less energy in the high spatial frequencies. As described in the previous chapter, the procedural textures are created as the weighted sum of Perlin noise (Perlin, 2002) of different frequencies. The advantage of using a procedural texture is that it gives a smooth reflectance function at any viewing distance. Procedural textures also avoid pixelation artifacts and texture boundaries that would be unavoidable if texture mapping were used.

In each movie, the model observer moved forward and gazed in the direction of translation, 10 degrees down from the horizon. The observer's eyes were located 1.5 meters above the ground plane. In the forest movies, cylinders were never placed closer than 0.3 meters to the observer. Image movies and range movies were rendered simultaneously using a feature of the ray tracer that allows 16 bit accuracy for range values spanning 0.3 meters to 300 meters. The range images were used in combination with the known observer motion to compute the ground-truth local projected motion at each image location.

4.2.2 Sampling

A *scene condition* is a scene type (flat wall, ground plane or one of three different forest types) combined with a surface texture ($\frac{1}{f}$ or $\frac{1}{f^{1.5}}$) and a walking speed (0.9 or 1.5 meters/second). For each scene condition, 100 movies were generated, each with a new random seed. The random seed determines both the random texture on all the objects and the random placement of the cylinders within the scene. For each of the scene conditions, 20,000 image locations were sampled. Each sample was obtained by randomly picking a movie and a valid location within that movie. At each sampled location, motion estimates were computed for a range of different aperture sizes (Figure 4.1). The estimated motion vector was then subtracted from the ground-truth motion vector at the center of the aperture and the magnitude of this difference vector was used as the error. Thus, errors are expressed in the same units as velocity, degrees/second. Errors were binned by the magnitude of the ground-truth retinal speed using equal quantile binning with 12 bins. In all conditions, the distribution of velocity amplitudes was found to be approximately log normal, implying that the quantile binning resulted in an approximately logarithmic speed binning.

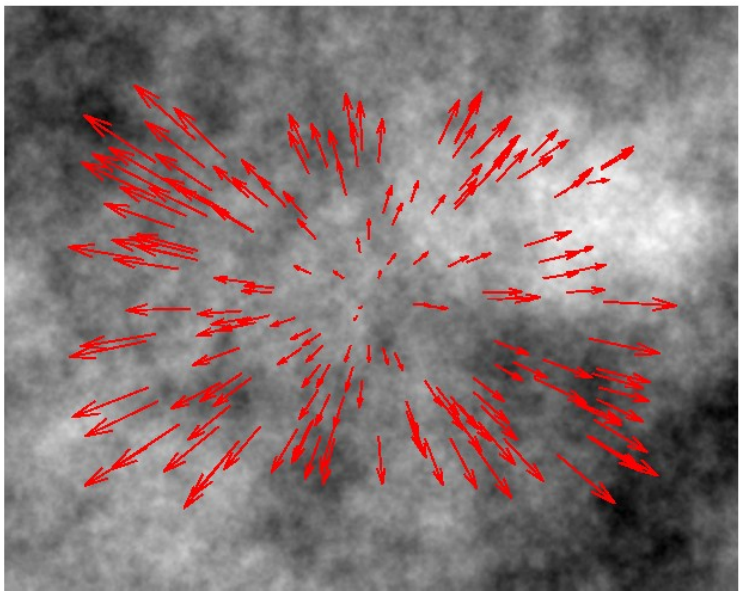


Figure 4.1: **Example frame with sampled estimated motions.** A frame from a flat wall movie is shown with a superimposed sample of estimated motions. Each of the shown estimated motions was computed for a single aperture size. At each sampled location, motions estimates (and errors) were generated for all aperture sizes. This figure demonstrates how errors were sampled to compute error as a function of aperture size.

4.2.3 Motion Estimation Algorithm

The aim of this study was to understand the distributions of velocities in natural environments and their effects on local motion processing, as opposed to designing an efficient or accurate computer motion detection algorithm. Therefore, the desired motion estimation algorithm is a simple one whose errors reflect the environmental and task constraints. The estimation algorithm used is based on the classic "optic flow constraint equation" (Ballard and Brown, 1982, Section 3.6.1, pages 102–103) which simply assumes that all changes in luminance across frames are due to motion. A single velocity vector for the whole space-time sample is computed using a gradient search on the squared error from the optic flow constraint equation. At each sampled location the images were windowed with a two-dimensional Gaussian function before they were given to the motion estimation algorithm. For the remainder of this chapter, "aperture size" is defined to be twice the standard deviation of the Gaussian window function. In separate tests, the algorithm was found to be

an accurate, unbiased estimate of ground-truth motion for motion where the image plane undergoes pure translation, as in rotation of the eye.

Although the motion algorithm is unbiased for translational motion, errors and systematic biases occur when estimating the two-dimensional projection of three-dimensional velocities. These biases are important to understand when one is actually implementing a motion estimator. However, the focus of this research is to understand how properties of the environment influence motion estimation accuracy. Because any systematic bias in an estimator can be learned and removed it does not ultimately limit accuracy. For this reason, the biases were estimated at each speed bin for each class of scenes (forest, flat wall or ground plane) and all motion estimates were adjusted to remove these biases. Given the accuracy of estimations for translational motion and the correction for bias, it is likely that the measured errors in the motion estimates are overwhelmingly due to the complexities of the motion in the scene and/or to the aperture problem.

4.2.4 Best Aperture Size

For each speed bin, at each aperture size, the average error in the estimated velocity was computed over all the samples. Figure 4.2 shows an example plot of the average error as a function of aperture size for a particular speed bin. Error bars show the standard error across samples. The continuous line shown in these plots is an interpolation of the measured error points. Notice that the function is U-shaped, showing that for this condition and speed bin the environmental constraints balance against each other to give an ideal aperture size. Too small an aperture leads to larger errors because of a lack of information, what is known as the aperture problem. Too large an aperture leads to larger errors because there are too many different motions in the aperture. The best aperture size was calculated from the interpolated function. As an example, Figure 4.3 shows the same plot (error as a function of aperture size) for all 12 speed bins. Notice again that there is a well-defined optimal aperture size at each of the speed bins, indicated by a vertical red line. In subsequent figures in this chapter the best aperture size is plotted as a function of speed (Figure 4.4, for example). Error bars for the best aperture size were computed using a jackknife procedure (Efron and Tibshirani, 1993).

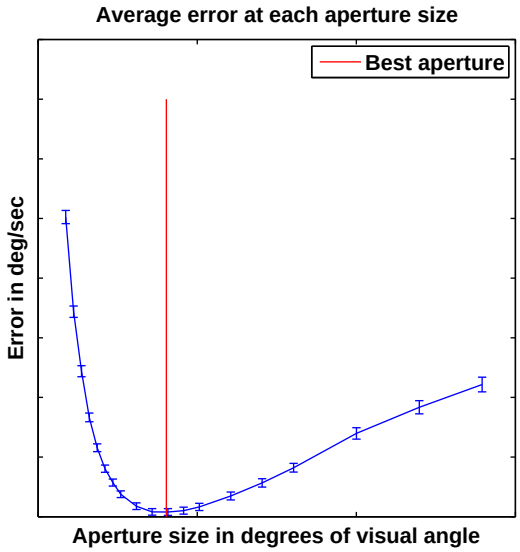


Figure 4.2: **Example plot of average error at each aperture size.** The average error for aperture size is U-shaped and has a clear minimum. The best aperture is marked with a vertical red line. Error bars show the standard error across samples. The plot is for a single velocity bin for the high-density cylinder scenes with a walking speed of 0.9 meters/secondand a surface texture whose amplitude falls with $\frac{1}{f^{1.5}}$. The error is computed as the length of the difference vector obtained by subtracting the estimated velocity vector from the ground-truth velocity vector. This figure demonstrates how error as a function of aperture size was used to estimate the best aperture size.

4.3 Results

Results for the optimal aperture size are discussed first for the forest scenes, then separately for the flat wall and flat ground plane scenes.

4.3.1 Forest Scenes

As can be seen in Figure 4.4 and Figure 4.5, image velocities range from 0 to 70 degrees/second. Ideal aperture sizes range from about 2 degrees at the slowest speeds to more than 12 degrees for the fastest. Also, ideal aperture size increases monotonically with local image speed. This increase in apertures size is apparent for all six forest conditions: both walking speeds, all cylinder densities and widths, and for both surface textures. When ideal aperture size is plotted as a function

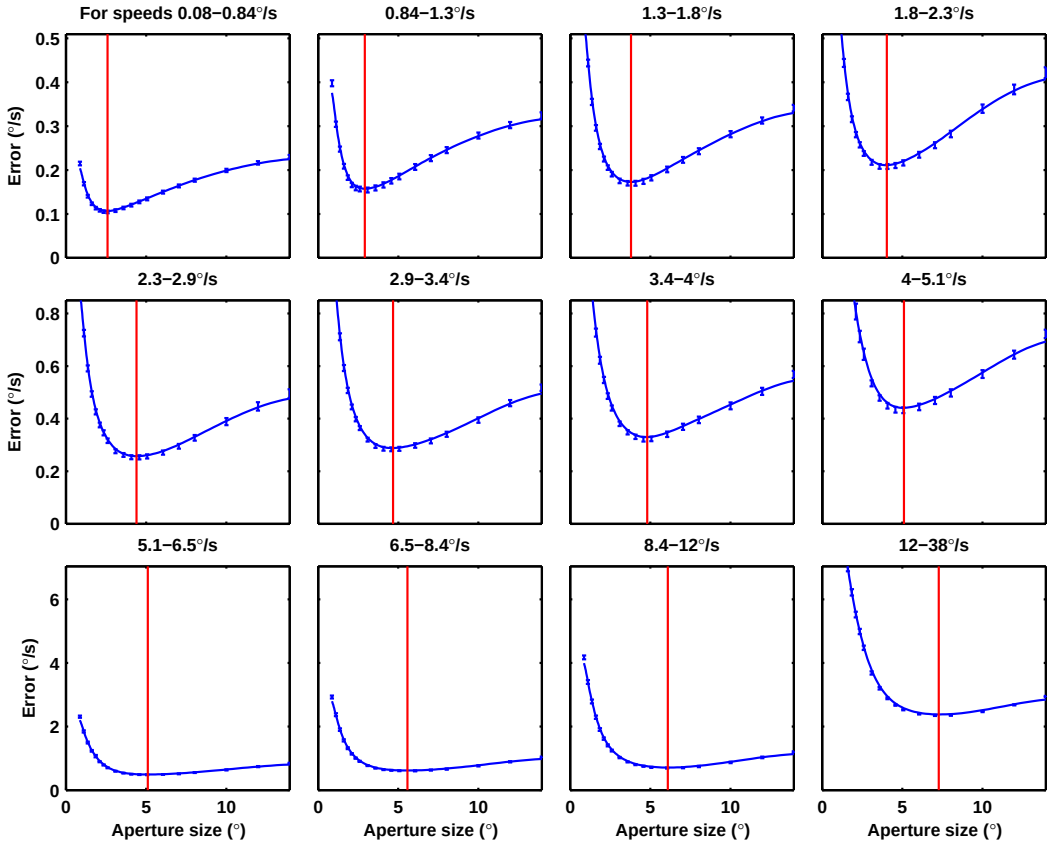
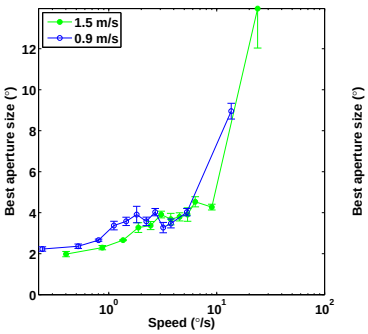
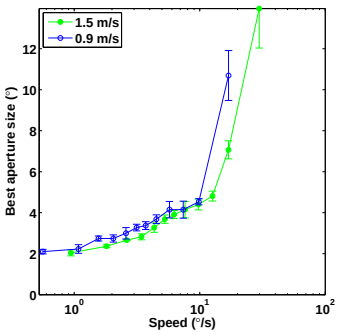


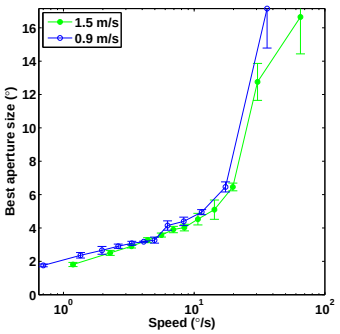
Figure 4.3: **Example plot of average error as a function of aperture size for different speed bins.** The speed range for each bin in degrees/second is indicated at the top of each plot. These plots are from the high-density cylinder scenes with a walking speed of 0.9 meters/second and a $\frac{1}{f^{1.5}}$ surface texture (the plot in Figure 4.2 is the last plot in the first row above). The minimum of each curve, shown with a red vertical line, is the optimal aperture size for that speed bin. Notice that the scale of the vertical axes (error) is different for each row. In order to have the U-shaped curve visible in each plot, it is necessary to change the scale because the magnitude of the errors increase as the speed increases. This figure demonstrates how the best aperture size as a function of speed was computed from the error functions.



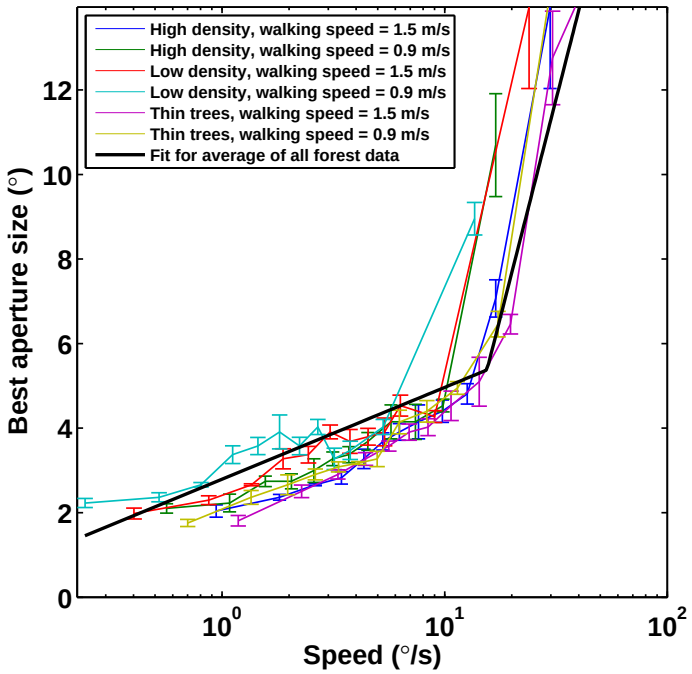
(*a*) Low density of trees



(*b*) High density of trees

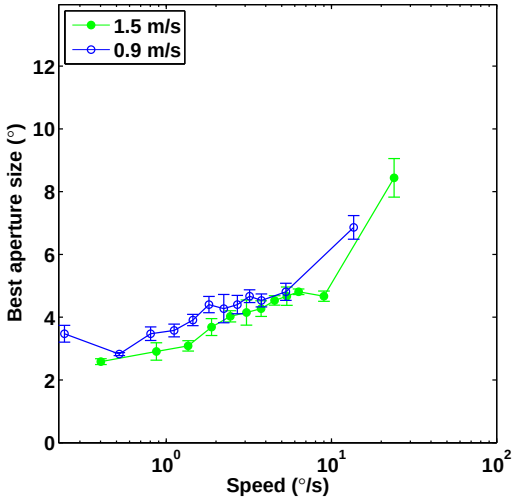


(*c*) High density of thin trees

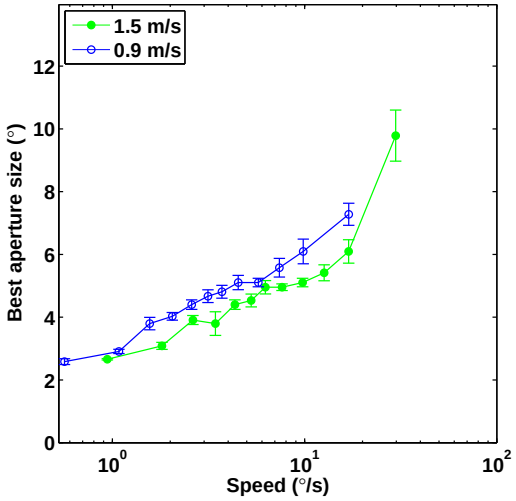


(*d*) Data from all forest conditions with $\frac{1}{f}$ texture.

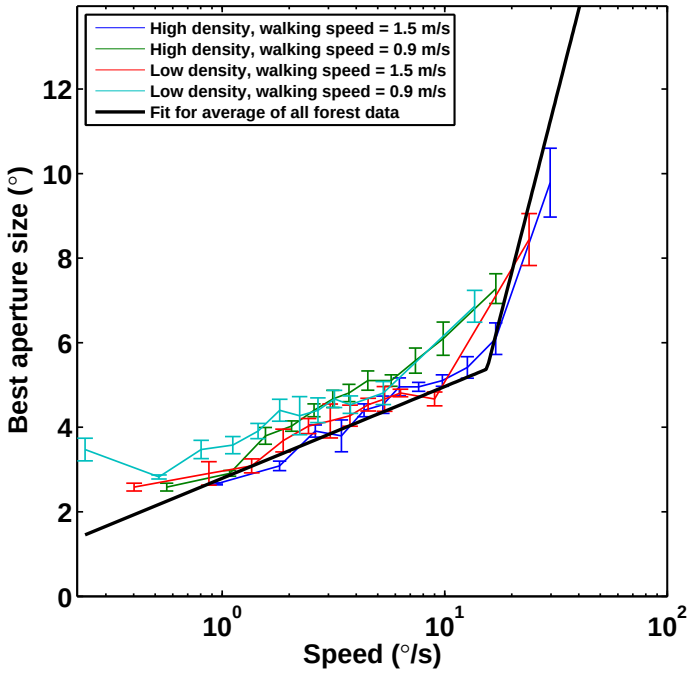
Figure 4.4: **Best width vs. speed for all forest scenes with $\frac{1}{f}$ texture.** (*a-c*) Plots are shown for both walking speeds for each type of forest scene with $\frac{1}{f}$ texture. (*c*) The thin trees condition has a tree density of twice that of the "high density" condition, but maintains a constant visual density of cylinders. (*d*) Combined data for all six forest conditions with $\frac{1}{f}$ texture. The solid line in black is a two-line fit of the best aperture size for the average of all forest data (includes both textures). Notice that the optimal aperture size monotonically increases with velocity and linearly increases with velocity for a significant range of speeds. This constraint is likely to be observed in the human visual system and is important to understand when designing machine vision systems.



(*a*) Low density of trees



(*b*) High density of trees



(*c*) Data from all forest conditions with $\frac{1}{f^{1.5}}$ texture.

Figure 4.5: **Best width vs. speed for all forest scenes with $\frac{1}{f^{1.5}}$ texture.** (*a,b*) Plots are shown for both walking speeds for each type of forest scene with $\frac{1}{f}$ texture. (*c*) Combined data for all four forest conditions with $\frac{1}{f^{1.5}}$ texture. The solid line in black is a two-line fit of the best aperture size for the average of all forest data (includes both textures). Notice that the optimal aperture size monotonically increases with velocity and linearly increases with velocity for a significant range of speeds.

of log speed, it becomes apparent that ideal aperture size increases approximately linearly with log speed for speeds less than about 20 degrees/second (Figure 4.4 and Figure 4.5). The combined data for all forest scenes was fit with two connected lines. Notice that this curve does not represent the average of the best aperture across different scenes. Rather, the best aperture is computed from the sum of the errors across all forest scene conditions. The formula for the fitted curve is:

$$y = \begin{cases} m_1 \cdot x + b_1 & \text{if } x < x_0 \\ m_2 \cdot (x - x_0) + b_2 & \text{if } x \geq x_0, \end{cases}$$

$$b_2 = m_1 \cdot x_0 + b_1,$$

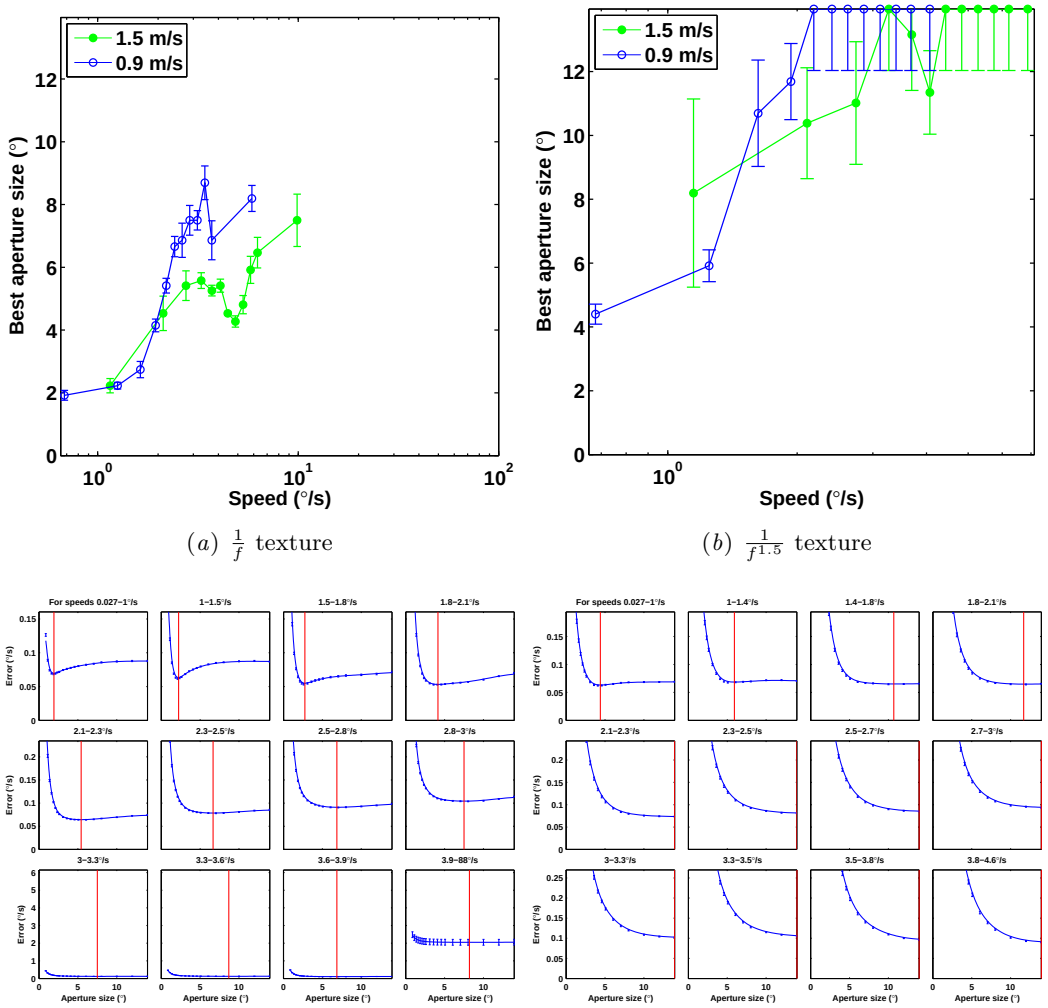
where $x = \log_{10}(\text{speed})$, y is the best width and there are four fitted parameters:

- The slope of the first line: $m_1 = 2.2$
- The slope of the second line: $m_2 = 20.6$
- The y-intercept of the first line (at speed = 1 so that $x = 0$): $b_1 = 2.8$ degrees.
- The speed cutoff between the two lines: $\text{speed}_0 = 15.4$ degrees/second, where $x_0 = \log_{10}(\text{speed}_0)$.

Qualitatively there are only small differences in the results across all the different forest conditions. A faster walking speed generally leads to slightly larger ideal aperture sizes for the same retinal speed. Similarly, a lower density of trees generally leads to slightly larger aperture sizes for the same retinal speeds. Surprisingly, the thin tree condition, which is equivalent to the high-density forest condition but adds more motion borders, does very little to change the ideal aperture size. It is not surprising that the $\frac{1}{f}$ texture, which contains more energy in the high spatial frequencies, has overall lower error curves and smaller aperture sizes at low speeds than the $\frac{1}{f^{1.5}}$ texture. At higher speeds, the $\frac{1}{f}$ texture has larger aperture sizes.

4.3.2 Flat Wall Scenes

For the flat wall scenes, the best aperture size is noisy and the optimal aperture sizes are large (Figure 4.6*a,b*). The reason for the noisy results becomes obvious when one looks at the average error as a function of aperture size (Figure 4.6*c,d*).



(*a*) $\frac{1}{f}$ texture (*b*) $\frac{1}{f^{1.5}}$ texture

(*c*) $\frac{1}{f}$ texture, 0.9 m/s walking speed (*d*) $\frac{1}{f^{1.5}}$ texture, 0.9 m/s walking speed

Figure 4.6: **Flat Wall.** (*a,b*) Best aperture size at each speed, for two different walking speeds for the flat wall scenes. (*c,d*) Average error as a function of aperture size for different speed bins for the 0.9 meters/second walking speed condition. For the flat wall condition the error function is flat above a certain aperture size. Plots in (*c,d*) were qualitatively similar for a walking speed of 1.5 meters/second. Notice that for these conditions for most of the speed bins, there is no pronounced minimum. This shows that the flat wall scenes do not significantly constrain the visual system's optimal aperture size.

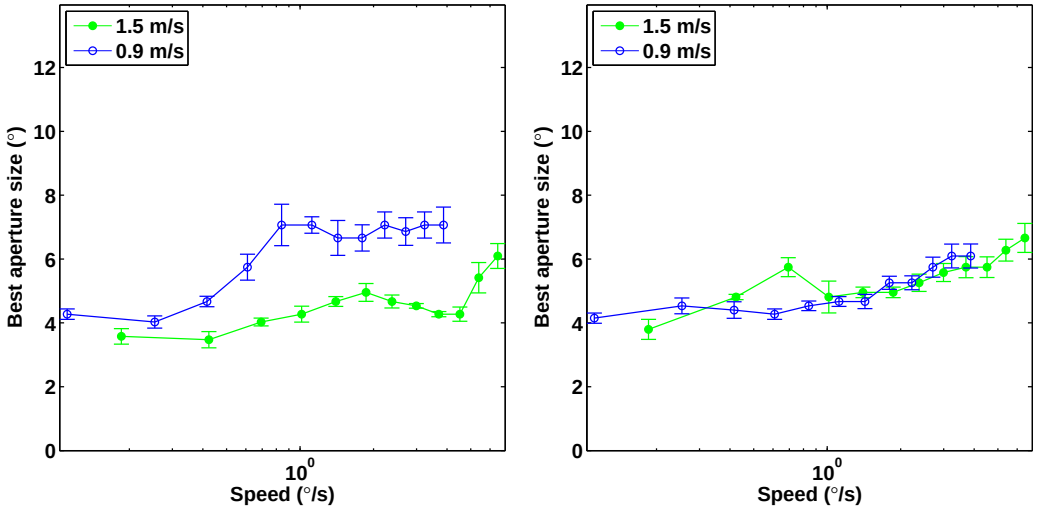
For the flat wall scenes, the error function is not U-shaped, except perhaps for the slowest speed bins. Above a certain minimum aperture size, there is no significant decrease in accuracy and therefore for this condition there is no single optimal aperture size. To understand this result, consider the nature of optic flow in this environment. There are no motion borders in the scene (objects) and the optic flow changes slowly across the plane. Thus, even for a large aperture, the average of the velocities will remain close to the velocity in the center of the aperture. The only exception is when the aperture overlaps the focus of expansion. Motion tends to be slow near the focus of expansion and hence at slow speeds somewhat more U-shaped error functions are observed, with minima at small aperture sizes. The implications of this result for a visual system are that flat wall scenes are likely have only a small effect on the optimal aperture size. If one were to design an optimal motion detector for a visual system that existed in a variety of environments, the constraints from the conditions where there is a significant U-shaped error function would quickly dominate over the flat wall scenes with the flat error functions.

4.3.3 Ground Plane Scenes

In the previous chapter, the ground plane scenes were found to have more non-translational motion than any other type of scene. It makes intuitive sense that more non-translational motion might give a translational motion estimator more problems. Consistent with this intuition, the ground plane scenes had a larger bias adjustment than the other types of scenes (this data is not shown). The ground plane scenes show a slight increase in aperture size as speed increases (Figure 4.7). In comparison with the forest scenes, the ground plane scenes have a larger aperture size at equivalent speeds. This is likely necessary to overcome the large amount of non-translational motion in the scene.

4.4 Discussion

The aim of this study was to determine how local motion detectors should be designed to optimally code the local image motion that occurs during self motion through the natural environment. More specifically, the following research question was posed: how should the integration area, or aperture size, vary as a function of



(a) $\frac{1}{f}$ noise (b) $\frac{1}{f^{1.5}}$ noise

Figure 4.7: **Ground plane.** Best aperture size at each speed, for two different walking speeds and for two different surface textures for the ground plane scenes. Like in the forest scenes, for these scene conditions, the ideal aperture size increases with increasing speed.

image speed. This question was answered by using a ray tracer to generate movies from the perspective of a model observer translating at two different walking speeds through five different classes of model environments that were textured with $\frac{1}{f}$ or $\frac{1}{f^{1.5}}$ noise. Three of the classes of model environments were based on the measured statistics of natural forest scenes. The other two classes of model world were simple textured planes lying either below the observer, as a ground surface, or frontal-parallel, as if the observer were walking towards a textured wall.

For most scenes and walking speeds, the ideal aperture size was found to increase monotonically with image plane speed. For the frontal-parallel plane scenes there was no ideal aperture size except at the slowest speeds; rather there was a minimum aperture size beyond which there were negligible changes in accuracy. Most importantly, for the forest scenes at a range of speeds the ideal aperture size increases approximately linearly with log speed for all five scene conditions.

How general are these results? In comparison to these artificial statistics, consider the best possible real-world measurements. Ideal measurements would include the positions and orientations of the eye of observers as they walk through

real forest and plains environments, as well as the calibrated image information and range data from the observers' point of view. Thus, there are two obvious inaccuracies in this model. First, it ignores the complex eye movements humans make when walking. Second, real forests and plains have leaves, branches, grass and other objects of varying size and complexity that create additional motion boundaries. However, given that these findings are robust to changes in walking speed, density of objects in the scene, and density of edges in the scene, it seems unlikely that more detailed scene data or tracking of eye movements would change the qualitative finding that optimal integration area increases monotonically with speed.

One might also question whether these results are dependent on the specific characteristics of the motion estimation algorithm. However, the algorithm is accurate and unbiased when the ground-truth motion is two-dimensional translation, and the algorithm corrects for the average bias due to more complex ground-truth motions. Thus, the algorithm's error variance is primarily the result of two factors: non-translational motion and stimulus ambiguity within the aperture. These two factors will affect any algorithm based on the assumption of two-dimensional translation (which includes most existing algorithms and models of local motion detection). Furthermore, qualitatively similar results (not shown here) were obtained without the average bias correction. Therefore, it seems likely that other local motion algorithms based on the assumption of two-dimensional translation would produce similar results.

One assumption made in this analysis is that the visual system being optimized must choose a single aperture size for a particular speed. It is certainly possible that a different constraint might lead to different results. It seems highly likely, in fact, that a visual system that had to choose two aperture sizes for a particular speed would have very different optimal aperture sizes. In fact, the most likely result would be that the two sizes are chosen on either side of the ideal single size for a given speed. Following this logic, it seems likely that a visual system that optimizes a distribution of aperture sizes at a particular speed, would have a distribution of aperture sizes around the optimal single size. Thus the results for optimizing a large population of aperture sizes for a particular speed is likely to be similar to the results for optimizing a single sensor.

These results have implications for computer and robotic vision. For the

forests scenes, if one used the optimal aperture size from the lowest speed bin, then the errors in the largest speed bin increase by over 350%. This is a worst case scenario. However, even if one chose a single reasonable aperture size from an intermediate speed bin, then the errors at the highest and lowest speed bins increase by 40%. Therefore, machine vision algorithms that assume translational motion might benefit from having motion detectors with a range of integration areas. VLSI vision systems in particular would benefit from an analysis like this, since they distribute sensors across the visual field and the integration area of those sensors is fixed. Calow et al. (2005) showed that using a local area of integration that increases with increasing eccentricity improves heading estimation performance. The results from this chapter give an ecological explanation to Calow et al.'s results. Since a moving visual system usually fixates towards it's direction of motion, the speed of motion is highly correlated with the eccentricity. Furthermore, it seems likely that tying area of integration directly to speed instead of eccentricity could lead to further improvements. Finally, robots that move through man-made flat environments may benefit from having their motion detectors designed in order to compensate for the large amount of bias created by the non-translational motion of the ground plane. One could imagine a computer-vision system that distinguishes ground plane from non-ground plane and corrects for the large bias due to the non-translational motion in the ground plane areas of the visual field.

Given these findings, one might expect that the receptive-field sizes of speed-tuned neurons in the cortex increase with the speed to which the neurons are tuned. There is evidence that the preferred speed of neurons in the middle temporal (MT) visual area of macaque monkeys increases with retinal eccentricity (Maunsell and Essen, 1983). Given that receptive-field size tends to increase with retinal eccentricity this result is consistent with the findings described in this chapter. However, a more definitive test would be to determine whether there is a positive correlation between the preferred speed and the receptive-field size at the same eccentricity. Unfortunately, no such study has been performed. Interestingly, several researchers did find a positive correlation between preferred speed and receptive-field size based on their psychophysical measurements of motion detection thresholds, as a function of stimulus size and eccentricity (Burr et al., 2006; van de Grind et al., 1986, 1983; van Doorn and Koenderink, 1984). However, the results are not clear cut and all these experiments used broadband stimuli. The question of whether receptive-field

sizes change with increasing speed for narrow-band stimuli is explored in detail in the next chapter.

4.5 Conclusion

Over a range of simulated environments, the ideal aperture size for local motion discrimination was found to increase monotonically with speed. For simulated forest environments over a range of speeds, the ideal aperture size increases linearly with log speed. This result makes predictions for cortical neurons involved in heading perception. Also, these results have implications for the design of computer and robotic motion detectors. This work demonstrates that the statistics of the environment have an influence on the optimal design of motion sensors.

Chapter 5

Psychophysical Measurements

The previous chapter described how receptive-field sizes for local motion discrimination ought to increase with the speed to which the detectors are tuned. The question remains, how are motion detectors in the human visual system designed? In particular, to what extent do the receptive-field sizes of motion sensitive neurons increase with increasing speed? To answer this question receptive-field sizes were estimated psychophysically as a function of speed. It was found that psychophysical receptive-field sizes increase with speed, but that this change in receptive-field size is better explained by the changes in spatial frequency than by the changes in speed. The hypothesis is therefore not correct for the specific conditions tested here, but it still bears looking at for future work.

5.1 Motivation

As described in Chapter 2, there is a large body of research attempting to estimate the receptive-field size of motion sensitive neurons using psychophysical methods (A selected reading list: Anderson and Burr, 1987, 1989, 1991; Anderson et al., 1991; Burr et al., 2006; Fredericksen et al., 1994, 1997; Georgeson and Scott-Samuel, 2000; Spillmann et al., 1987; Tadin and Lappin, 2005; Tadin et al., 2003; van de Grind et al., 1986, 1983; van Doorn and Koenderink, 1984; Watson et al., 1983; Watson and Turano, 1995). Although the methodologies of these studies differ, the basic theoretical concept behind the research remains the same. Neurophysiology and psychophysics provide evidence that the motion detection system of primates can be

modeled as a bank of motion detectors tuned to different spatiotemporal properties. In electrophysiological experiments an electrode is placed so that it measures the response of a motion sensitive neuron. Spatiotemporal stimuli are presented to the animal and the area of visual space that elicits responses from the neuron is known as that neuron's *receptive-field*. In psychophysical experiments it is impossible to measure the stimulation of a single neuron. Therefore, in order to use psychophysical methods to measure receptive-field size, researchers rely on one basic fact: that information is integrated more efficiently within a single neuron than it is across multiple neurons.

On an absolute scale, psychophysical performance will always improve as the area of the stimulus being presented increases. This improvement is simply because the amount of information has increased. Instead of just measuring absolute performance, researchers try to estimate the efficiency of human performance, that is, how well the subject is performing normalized by the amount of information presented. Consider the case where the stimulus is centered on a single neuron in an array of spatially distributed neurons. Assume, as well, that all the neurons in this array are tuned to the spatial and temporal frequency of the stimulus. Starting with a very small stimulus, as the stimulus size increases, as long as the stimulus size is smaller than the receptive-field size of the neuron being stimulated, the visual system's efficiency will stay constant. However, as the stimulus gets larger than the receptive-field size of a single neuron, more neurons are stimulated and the subject's efficiency declines. Thus researchers find the critical stimulus size at which efficiency starts to decline. This is the psychophysical receptive-field size.

Obviously, what happens in the visual system is much more complicated than the single neuron case described above. Neurons have overlapping receptive fields, there is jitter in the location and placement of the stimulus, and there are many different neurons that might be stimulated by a specific stimulus. Despite these complications and others, one would expect a similar phenomenon to take place. There are neurons in the visual system whose receptive-field properties best match the stimulus properties. As the size of the stimulus increases past the receptive-field size of these neurons, the efficiency of discrimination has to drop. It would be difficult to argue that the absolute psychophysical receptive-field size being measured is exactly the receptive-field size of a particular neuron in the visual system. However, the sizes must be strongly correlated. Therefore, changes in the psychophysical

receptive-field size that occur in response to changes to stimulus properties must reflect changes in receptive-field sizes of visual neurons. It is these qualitative changes that are measured in this chapter.

As described above, there are clearly difficulties in claiming that a set of psychophysical measurements correspond to a physiological receptive field. In support of this line of study, Spillmann et al. (1987) attempted to corroborate psychophysical measurements with electrophysiological. Spillmann et al. estimated receptive-field sizes psychophysically in both humans and monkeys and then measured receptive-field sizes using electrophysiological techniques in monkeys. They found a good corroboration between the different measurements. Also, the psychophysical receptive field is a consistently measurable phenomenon and is therefore scientifically interesting independent of whether it has a direct physiological correlate. While acknowledging the inherent difficulties in tying a psychophysical phenomenon with a physiological one, to make this dissertation more readable, the term "receptive-field sizes" will be used with the understanding that it refers to *psychophysically estimated* receptive-field sizes.

Despite the large body of experimental work described in Chapter 2, it is still not completely clear how receptive-field sizes vary with speed. There are studies that use narrow-band stimuli and vary the spatial frequency of the stimulus but fix the temporal frequency (Anderson and Burr, 1987, 1989; Anderson et al., 1991, for example). These studies consistently show that the receptive-field size decreases as the spatial frequency of the stimulus increases. As we pointed out in Chapter 3, the velocity of a sine wave or a Gabor patch is inversely proportional to the spatial frequency. More specifically, $v = \frac{\omega}{f}$, where v is velocity, ω is the temporal frequency and f the spatial frequency of the wave form. Thus for a fixed temporal frequency the receptive-field size does increase with increasing velocity, simply because the spatial frequency is decreasing.

The question remains, however, whether the increasing receptive-field size is just a dependency on spatial frequency or whether the receptive-field size will change with velocity when the temporal frequency is manipulated as well. Several researchers have specifically examined how receptive-field sizes vary as a function of velocity (Burr et al., 2006; van de Grind et al., 1986, 1983; van Doorn and Koenderink, 1984). They all find that to some extent receptive-field sizes seem to grow with increasing velocity. However, the results are not clear cut. Over some ranges

of velocities and for some stimuli the receptive-field size is unchanging. More importantly, all these studies use broadband stimuli, stimuli that contain many different spatial frequencies. It is not obvious how receptive-field sizes will change with velocity when using narrow-band stimuli. An alternative explanation is that when the velocity of a broadband stimulus increases, high spatial frequency information is lost, leading to a change in the effective spatial frequency that is perceived by the neurons involved in motion discrimination. receptive-field sizes of neurons are known to be tightly integrated with spatial frequency, and thus the changing spatial frequency content of the stimulus will induce a change in the receptive-field size. To truly understand how receptive-field sizes change with velocity, it is necessary to measure how receptive-field size changes with changing spatial and temporal frequency for narrow-band stimuli. This chapter describes just such an experiment.

There are several methodologies that have been used for estimating psychophysical receptive-field sizes. Many studies fit two lines with different slopes to the measured contrast threshold (or sensitivity) curves and find the critical point at which there is a change in slope (Anderson and Burr, 1987, 1991, for example). This methodology was avoided for several reasons. First, finding the location of a transition between slopes is a difficult parameter to estimate accurately. Second, the methodology requires some additional assumptions about how information is integrated in the visual system. It assumes that information is integrated equally well at all sizes below the critical size. It is possible that efficiency declines below the critical size. The results in this chapter ultimately show that this does happen.

The methodology used in this dissertation is one that measures efficiency more directly, by selecting a task for which there is a known ideal observer. An ideal observer is a mathematical model that computes the best possible performance for a task with a noisy stimulus (see Appendix B for a derivation of the ideal observer used in this experiment). Previous ideal observer analyses of motion receptive-field size have simply attempted to find the most efficient stimulus across a variety of dimensions like stimulus size, spatial frequency, temporal frequency, duration and carrier speed (Watson et al., 1983; Watson and Turano, 1995). In contrast, this experiment parameterized the study to understand how the maximum efficiency in one dimension changes as a function of two others, spatial and temporal frequency. The methodology gives high quality data, but requires many measurements and makes for a large parametric study.

In summary, for any specific spatial and temporal frequency, the experiment measures which size stimulus is the visual system most efficient at detecting. That stimulus size should be strongly correlated with the size of the neural receptive fields primarily responsible for coding that spatial and temporal frequency.

5.2 Methods

The goal of the experiment was to estimate the stimulus width at which the human visual system can most efficiently detect motion. In order to disambiguate the influence of speed and temporal frequency on receptive-field size, measurements of receptive-field size were made at two different spatial frequencies, $f = 0.5$ or 1.0 cycles/degree, and three different temporal frequencies, $\omega = 4, 8,$ or 16 Hertz. Since $v = \frac{\omega}{f}$, that corresponds to four distinct velocities, $v = 4, 8, 16$ or 32 degrees/second. The visual stimulus used in the experiment was a Gabor patch, a drifting sine wave multiplied by a stationary Gaussian envelope of variable width. Using a Gabor patch makes it possible to independently manipulate the size, spatial frequency and temporal frequency of the visual stimulus. The space-time Gabor patch is described by

$$I(x, y, t) = A \cos(2\pi f x + 2\pi\omega t)\, e^{-\left(\frac{x^2+y^2}{2\sigma^2}\right)},$$

where $I(x, y, t)$ is the signal intensity at a particular space-time location, A is the signal amplitude, f is the spatial frequency, ω is the temporal frequency, and σ determines the standard deviation of the size of the stimulus envelope. The stimulus width is defined as 2σ, which is the width at which the Gaussian envelope is at half height (also known as the half-width). The amplitude of the signal was adjusted to create different contrast conditions so that a psychophysical function (percent correct at different contrasts) could be measured. Gaussian random noise was added to the signal. This noise was clipped at plus or minus two standard deviations and was fixed at an RMS contrast of 10%.

Observers sat in a darkened room and viewed the monitor using a chin rest at a distance of 296 cm. The monitor resolution was 640 by 480 pixels at 120 Hz, and the width of the monitor was 40 cm yielding square pixels that spanned approximately 43 seconds of arc per side. The stimuli were presented for 258 ms (31 frames at 120 Hz). Because of the large viewing distance, each noise square was two by two

screen-pixels wide. Stimuli were generated using Matlab and displayed on a calibrated monitor. The three video card outputs (red, green and blue) were combined using an attenuator circuit and attached to the green gun of the monitor allowing for 12-bit linear luminance steps. The mean luminance of the monitor was $19.2\,\mathrm{cd/m^2}$ and the luminance range spanned from $0.4\,\mathrm{cd/m^2}$ to $38.0\,\mathrm{cd/m^2}$. The chromaticity coordinates of the monitor were $x = 0.2814$, $y = 0.6036$. The temporal frequencies of the Gabor patches were chosen carefully so that the first and last frames of the stimuli were identical and the frames were in sine phase to eliminate any possible spatial cues as to the direction of motion. The maximum signal contrast was constrained to 59% in order to avoid any clipping due to the limited luminance range of the monitor.

A time-line for a single trial of the experiment is shown in Figure 5.1. A fixation cross was presented until the observer indicated readiness by pushing a button. Mean luminance was presented for 258 ms followed by the stimulus, signal and noise, for 258 ms. The stimulus signal was randomly chosen to be a rightward or leftward moving Gabor. After the stimulus, a fixation cross was presented until the observer indicated with a button press which direction of motion was observed, left or right. Feedback on whether the response was correct was given in the form of a tone and the process was repeated until the end of the block.

For each experimental condition (spatial frequency, temporal frequency and width) subjects were run in a pilot staircase procedure intended both to familiarize subjects with the task and to gather estimates of the contrast thresholds for each observer in each condition (data not shown). Subsequently, subjects were run in blocked experiments with five different contrast levels at each of 12-15 widths for a given spatial and temporal frequency. The five contrasts were chosen so that the expected performance of the subjects evenly span 60% correct to 90% correct. The expected performance calculation was based on Weibull fits of the subjects' performance in the pilot experiment. In a given experimental session, subjects were run in a single spatial frequency, temporal frequency and width condition. Two 32-trial blocks were run at each contrast and all contrast blocks were counterbalanced to control for fatigue and practice effects. Across sessions, width conditions were also counterbalanced to avoid practice effects. For the smallest width conditions, because of the limited signal contrast, it was sometimes impossible to get a reasonable range

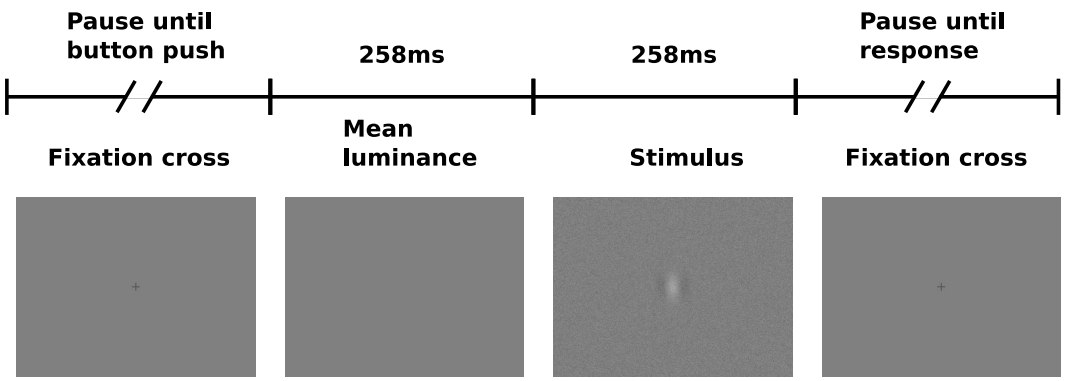


Figure 5.1: **Time-line of a single experimental trial.** A schematic diagram is shown above representing the time course of a single experimental trial. Reading left to right, a fixation cross was presented until the subject pressed a button. Then, mean luminance was displayed for 258 ms, followed by the stimulus moving either to the right or the left. Finally, a fixation cross was presented and the user indicates by a button press which direction of motion was perceived. A feedback tone indicating whether the subject's response was correct was given on each trial. Blocks of trials are run at different signal contrasts so that the contrast threshold can be estimated at each stimulus condition.

of performance data. Width conditions were ignored when the total performance of the observer was below 60% across all contrast levels or if the observer performance never exceeded 70% in any single contrast level.

Psychometric functions were fit for each condition using a Weibull function. For a particular spatial frequency and temporal frequency, the fit was performed across all width conditions at the same time giving a single slope, β, and a different threshold, α, for each width (Figure 5.2). Parametric bootstrap estimates of the thresholds were used to calculate expected standard deviations on the thresholds.

The Weibull fits gave an estimated contrast threshold at each stimulus width. (Figure 5.3*a*). This contrast-width curve was fit with

$$\log(\text{contrast}) = \frac{\alpha}{\beta + \log(\text{width})} + \gamma. \tag{5.1}$$

This function is intended to be strictly descriptive. It is simply a mathematical form that is flexible enough to give a good interpolation of the contrast-width function. From visual inspection (Figure 5.3*a*, green line) it was clear that the chosen function accomplished its purpose adequately. From the contrast threshold data,

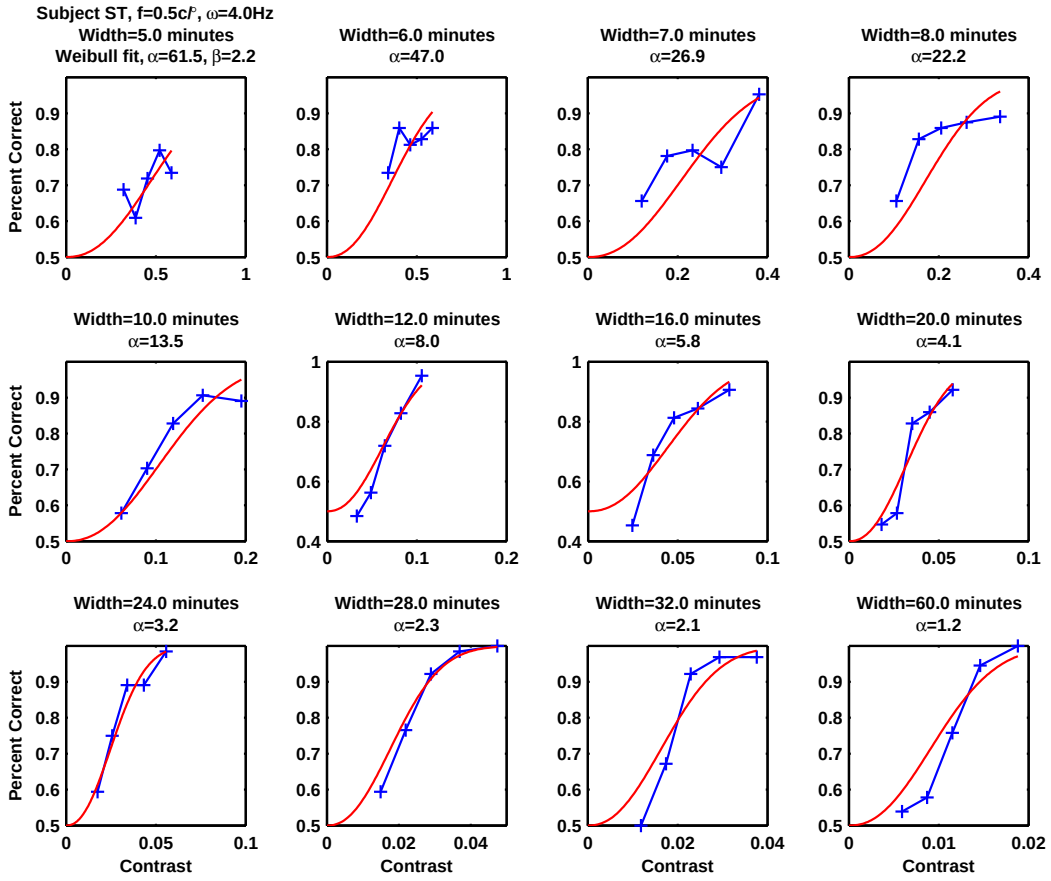


Figure 5.2: **Weibull fits of threshold contrast.** As an example, raw performance data plots are shown for subject ST at spatial frequency 0.5 cycles/degree and temporal frequency 4 Hz at a range of different widths. Each blue cross indicates the average performance for two blocks of 32 trials. The solid red curves indicate the maximum likelihood Weibull fit to the data. The slope parameter, β, is fit across all width conditions while the threshold parameter, α, is fit for each condition individually. This plot shows how the raw performance data are converted into contrast thresholds at each width condition. This contrast threshold can then be converted into an efficiency using the ideal observer (Figure 5.3)

an efficiency can be computed based on the expected performance of the ideal observer. The complete equations and a derivation of the ideal observer are presented in Appendix B. The efficiencies for the measured contrasts and for the interpolated contrast curve (Figure 5.3*b*) were computed for each condition. Subsequently, the optimal stimulus width was estimated by picking the width on the fitted contrast curve that gives the highest efficiency. A jackknife procedure was used to give an

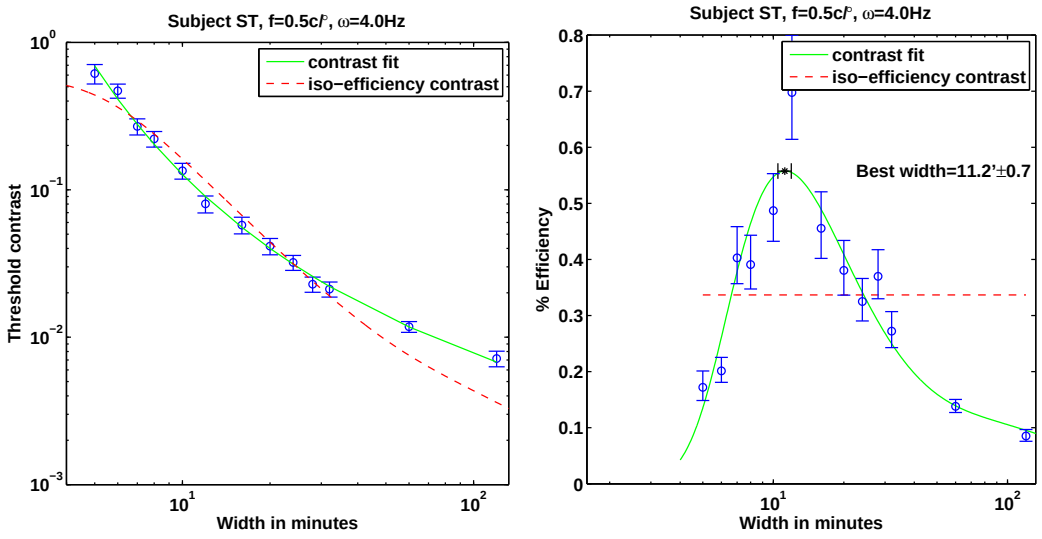
estimate of the error of the best width (Efron and Tibshirani, 1993). For most conditions, these estimated errors are quite small, providing further evidence that the contrast function used provides a good interpolation of the data.

5.3 Results

Efficiency data for each of the three subjects is shown in Figures 5.4–5.6. Notice that in most of the conditions, the efficiency functions have an upside-down U-shape implying that there is a width at which subjects are performing optimally. This study is the first to show reliably that efficiencies decline as stimulus width decreases past the optimum. With a few exceptions, most of the previous research in this area assumes that integration is equally efficient for all widths below a critical cutoff. This false assumption can lead to less reliable estimates of receptive-field size. However, measuring psychophysical thresholds with stimuli that are smaller than 10 minutes of arc is a difficult task. The large number of samples, the small pixel size and the high frame rate used in this experiment allowed for accurate measurements of performance at several widths below the optimal width giving more reliable measurements of the true optimum.

It is clear from Figure 5.6 that subject JW had much more difficulty performing the task than either of the other subjects and had much more variability in her performance. JW's absolute performance level was much lower than the other subjects and her efficiency curves show much less pronounced peaks. For the $f = 0.5\,\text{cycles/degree}$ and $\omega = 8\,\text{Hertz}$ condition no stable best width was found. Comparing the efficiency at the optimum width with the efficiency at the largest width, one can see that the other two subjects had 4-5 fold increases in efficiency for all the $f = 0.5\,\text{cycles/degree}$ conditions. In comparison, JW had approximately a two-fold increase in efficiency. For this reason, subject JW was not run in the second spatial frequency condition.

The measured psychophysical receptive-field sizes (best widths) are summarized in Figure 5.7. When the optimal stimulus widths are plotted as a function of speed, it appears as if there is a correlation between receptive-field size and speed.



(a) (b)

Figure 5.3: **Example plots of contrast threshold and efficiency as a function of stimulus width.** (a) Threshold contrast (α from Figure 5.2) is plotted in blue as a function of width for subject ST at a spatial frequency of 0.5 cycles/degree and a temporal frequency of 4 Hz. Blue error bars indicate bootstrap standard deviations for the Weibull fits. The line in green indicates the fit of the contrast data using Equation 5.1. For comparison, the dashed red line shows an iso-efficiency line. As one would expect the subject's absolute performance improves rapidly as width increases. However, in comparison to the iso-efficiency line, one can see that for very large and very small widths, the subject is performing less efficiently. For some width close to 10 minutes of arc, the subject is most efficient. (b) The performance data from (a) is shown, converted into efficiencies based on the ideal observer. As in (a), the blue data points indicate the subject's contrast thresholds converted into efficiencies. The error bars indicate bootstrap standard deviations from the thresholds fits, converted into units of efficiencies. The green line shows the fit from Equation 5.1 converted into units of efficiency. The dashed red line shows the same iso-efficiency curve from (a) which is now, naturally, a straight line at a single efficiency level. The black star indicates the stimulus width where the subject's performance is most efficient, as predicted by the contrast fit. The horizontal black error bar indicates a jackknife error measure of the optimal width. These plots are included to help illustrate how the ideal observer is used to convert threshold contrast into efficiency. The efficiency data make it possible to find the optimal aperture size for each condition. Subsequent results presented in this chapter are all of the form of subfigure (b).

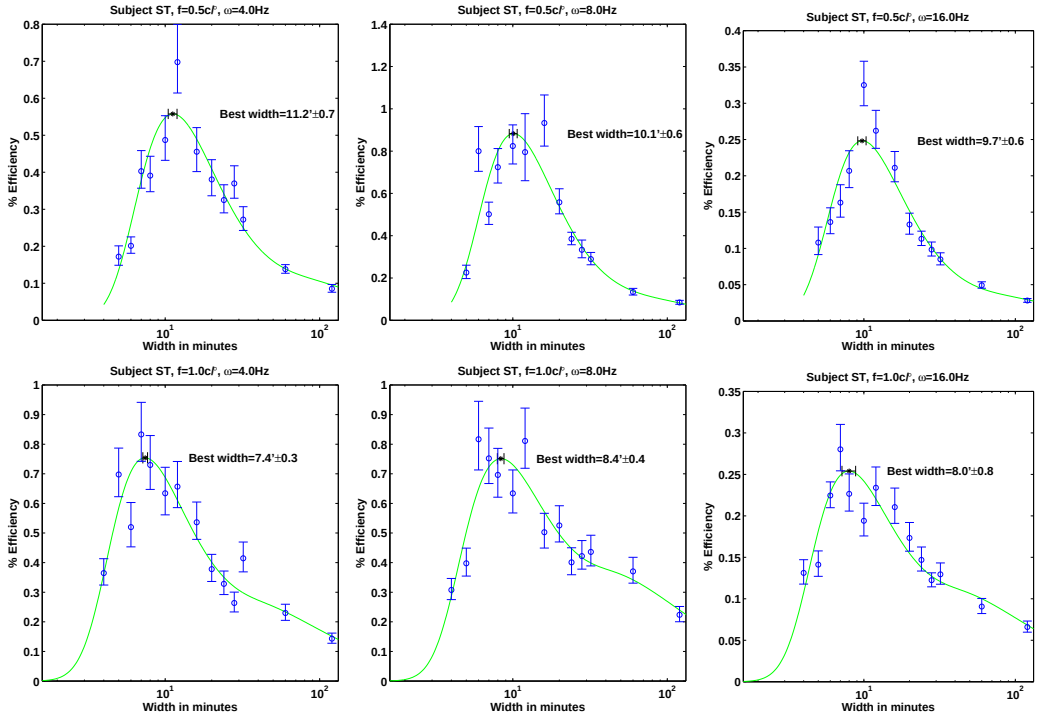


Figure 5.4: **Efficiency data from subject ST.** The top row shows results for a spatial frequency of 0.5 cycles/degree, the bottom for 1 cycles/degree. The columns show results for temporal frequencies of 4, 8 and 16 Hertz. These plots are included to show the raw efficiency data along with the interpolation and the estimated optimal width. The optimal widths are plotted in Figure 5.7.

However, when the width data are plotted as a function of temporal frequency it becomes clear that all the observed changes in widths simply result from changes in different spatial frequency. For a fixed spatial frequency, changes in the temporal frequency of the stimulus (and therefore in the speed of the motion) cause no significant changes in the best widths. The single exception to this is for the data from subject JW which are unreliable. The changes in speed are large and the results are consistent for both the other subject in both conditions.

5.4 Discussion

This chapter showed that the receptive-field size of neurons involved in motion discrimination does not change with speed. In contrast the simulations in Chapter 4

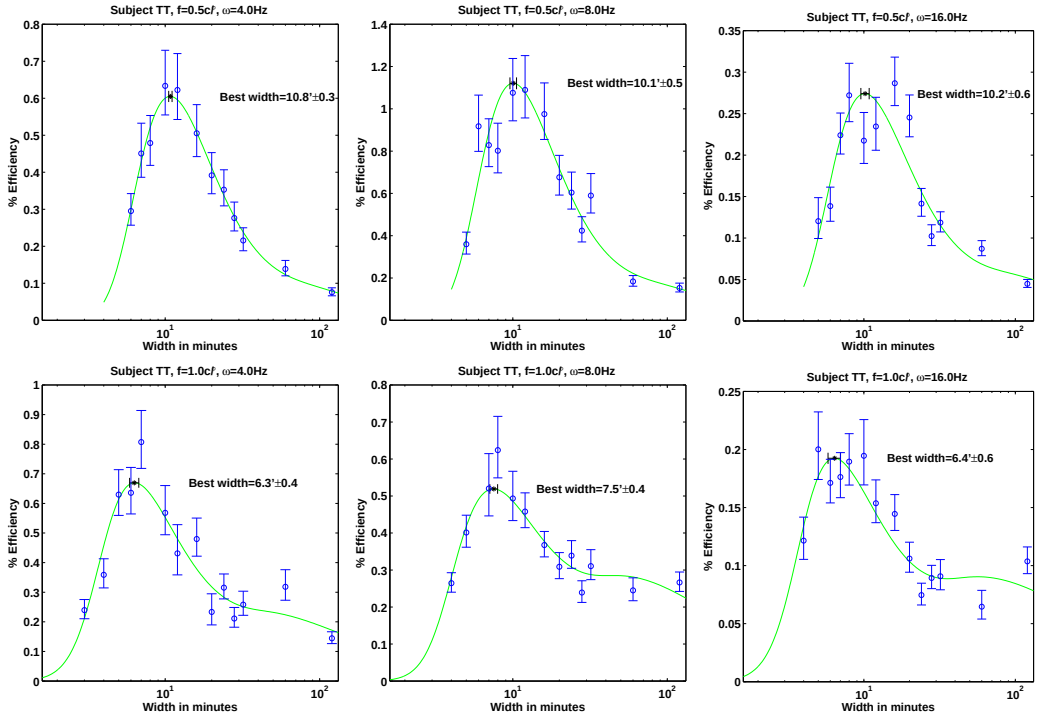


Figure 5.5: **Efficiency data from subject TT.** The top row shows results for a spatial frequency of 0.5 cycles/degree, the bottom for 1 cycles/degree. The columns show results for temporal frequencies of 4, 8 and 16 Hertz. These plots are included to show the raw efficiency data along with the interpolation and the estimated optimal width. The optimal widths are plotted in Figure 5.7.

found that the optimal aperture size of motion discrimination sensors changes with speed. There is a discrepancy between the theoretical and experimental results, but there are a number of reasons why this might occur.

All of the previous psychophysical experiments that show a connection between speed and receptive-field size used broadband stimuli (Burr et al., 2006; van de Grind et al., 1986, 1983; van Doorn and Koenderink, 1984). Similarly, the theoretical analysis from Chapter 4 also used broadband stimuli. In contrast, the experiment in Chapter 5 used narrow-band stimuli. The experimental results described in Chapter 5 support the intriguing possibility that cells involved in motion discrimination that are tuned to specific spatial frequencies, like V1 cells, do not alter their receptive-field size with speed, whereas higher level cells involved in broadband motion discrimination, like MT cells, do alter their receptive-field size with speed.

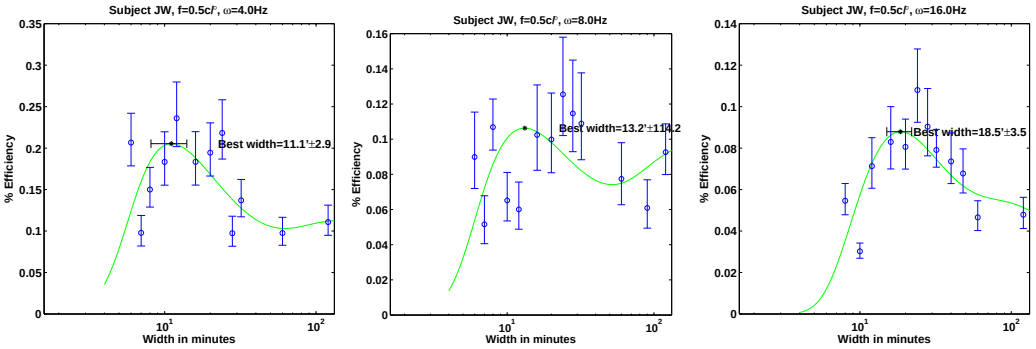


Figure 5.6: **Efficiency data from subject JW.** The left, middle and right plots show results for subject JW at a temporal frequency of 4, 8 and 16 Hertz, respectively. Subject JW had a difficult time performing the task so the subject was only run in a single spatial frequency condition, 0.5 cycles/degree. Notice that for the ω=8 Hz condition, there is no stable best width. Also, compare JW's absolute efficiencies with the performance of subjects TT and ST in Figure 5.4 and Figure 5.5. For these reasons, JW's data is deemed unreliable. These plots are included to show the raw efficiency data along with the interpolation and the estimated optimal width. The optimal widths are plotted in Figure 5.7.

In other words, lower-level frequency-specific cells have receptive-field sizes that are dependent on their spatial frequency tuning, not their speed tuning, and higher-level frequency-independent cells have receptive-field sizes that are directly dependent on their speed tuning.

One might ask, why not perform the psychophysical experiment using broadband stimuli? One reason is mentioned above; there are several previous studies that have already touched on this question. However, it would still be worthwhile to perform an experiment focused specifically on the relationship of receptive-field size to speed. However, designing an ideal observer to measure efficiencies for a broadband motion discrimination task is extremely difficult. Other, less rigorous, methodologies might have to be used.

Alternatively, one might ask why not repeat the theoretical analysis with narrow-band stimuli? There are obvious problems with building an environmental model that only contains narrow-band stimuli, foremost is that natural environments always contain broad distributions of spatial frequencies. A better question to ask is: What is the optimal aperture size of narrow-band motion sensors in the human visual system when exposed to natural broadband stimuli. In order to study this question, one would have to model the behavior of them human visual system in

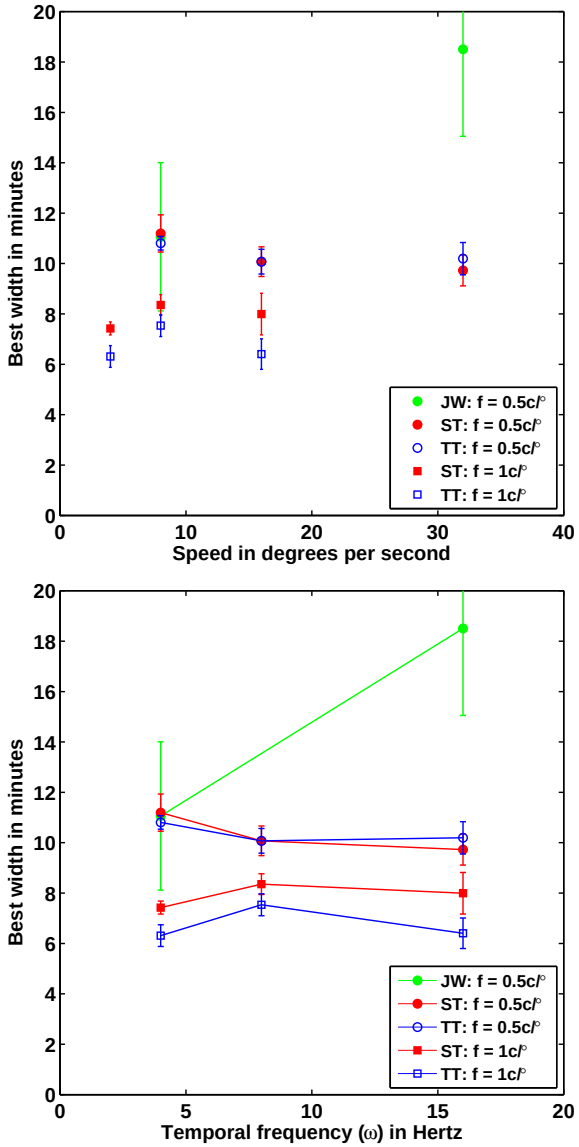


Figure 5.7: **Psychophysical receptive-field size.** The top plot shows the best width (receptive-field size) as a function of speed. The bottom plot shows the same data as a function of temporal frequency. The data in green, from subject JW, is deemed unreliable (Figure 5.6). From the top plot, it appears as if there is a correlation between width and speed. On closer inspection of the bottom plot, it is clear that width changes as a function of spatial frequency, but remains fixed as a function of temporal frequency implying that there is no direct connection between receptive-field size and speed under these conditions. Thus the hypothesis from Chapter 4 is not confirmed.

some detail. One would have to include the behavior and noise of the retinal inputs all the way through to the behavior and noise of the narrow-band motion sensor. This is a useful research path, but should be taken with care, since many theoretical assumptions would have to be made about the behavior of the visual system.

There is another possible explanations for the experimental results: retinal location. The experiment only examined human performance at the fovea. In contrast, the theoretical analysis described in Chapter 4 was agnostic in regards to retinal location. For the human visual system this analysis corresponds to integrating performance across the whole retina. Since humans often track objects during locomotion, the motion signal at the fovea can be quite different from the motion in other parts of the retina. If the distribution of motion signals at the fovea is different than for the rest of the visual field, is the optimal aperture sizes for motion discrimination at the fovea different? This is a question for future research, both theoretical and experimental.

On the theoretical side, in order to perform a comparable computational analysis of ideal aperture size that takes into account retinal location, one would need to incorporate a model of human eye motions during locomotion. Modeling eye motion is an interesting but difficult research vein. To build an accurate model of eye motion, one would want to base it in real-world statistics and validate it in comparison to real measurements. One simple possibility is to model human eye fixations during walking as a distribution of locations relative to the direction of motion. On the other hand, since we know that humans tend to fixate and track objects, it would be more desirable and likely to be more accurate to have a model of human fixations based on image or scene properties. Clearly measuring and modeling human eye movements during locomotion is a large research endeavor by itself.

On the experimental side, it is an open question whether the fovea's optimal aperture size for motion discrimination is different from the rest of the visual field. As was observed above, when taking eye movements into account, the motion signal falling on the fovea is likely to be different from the average motion independent of retinal location, but the motion in more eccentric locations might not change significantly. This fact makes it likely that the current theoretical analysis is accurate for extra-foveal locations, but might differ for the fovea. It would be interesting to repeat the psychophysical experiment at more eccentric locations in the retina.

When studying receptive-field size as a function of retinal location, there is an additional possible confound due to cortical magnification. Cortical magnification is a term used to describe the correlation between receptive-field size and retinal eccentricity. For many different types of visual neurons, receptive-field sizes at the fovea are small and the size of the neuron's receptive field increases as the location of the neuron's receptive field gets further from the fovea. Additionally, for V1 neurons at any retinal location, the spatial frequency tuning and receptive-field size are highly correlated and the temporal frequency tuning tends to be fixed. Recall that $v = \frac{\omega}{f}$, where v represents velocity and ω and f represent temporal and spatial frequency, respectively. Thus for motion sensitive neurons with a fixed temporal frequency tuning, and a spatial frequency tuning that changes with cortical magnification, there will be neurons tuned to small velocities with small receptive fields in the fovea and neurons with larger receptive fields tuned to larger speeds in the periphery. Looking at the whole visual field, on average, this distribution of cell sizes will create a situation where receptive-field sizes increase with the speed to which the neuron is tuned simply because of the properties of V1 neurons and cortical magnification. This observation leads to an interesting research question: Should receptive-field sizes change with speed at a fixed retinal location? This is an open question from a computational and an experimental point of view.

5.5 Conclusion

This chapter described an experiment designed to test the hypothesis developed in the previous chapter, that the receptive-field sizes of human motion discrimination units increase with speed. The experiment was the first to measured the optimal stimulus size for motion discrimination as a function of both temporal and spatial frequencies. In addition the experiment was the first to show a decline in efficiency as stimulus sizes decreased past the optimum. The hypothesis turned out to be false for the specific conditions in this experiment. It was found that for foveal narrow-band stimuli, the psychophysical receptive-field size does not increase directly with speed but with spatial frequency. Despite the negative result, the experiment furthers our understanding of the human visual system and leads to more experimental and theoretical work.

Chapter 6

Conclusion

This chapter starts by summarizing the contributions of this dissertation. Next, the implications of the work are discussed and future research projects are described. Finally, the dissertation concludes with a high level view of the simulated scene statistics approach.

6.1 Summary of Contributions

The major goal of this dissertation was to understand how the environmental properties of the world influence the design of visual systems. This goal was pursued in the context of a particular task, navigation during locomotion, and in a particular type of environment, forest scenes. To meet this goal, three connected research projects were undertaken. First, an environmental model was developed and image statistics were measured in natural scenes and compared to simulated statistics from the environmental model. Second, a computational analysis was performed that simulated across-domain statistics in an effort to understand how the chosen environment and task influence the integration area of visual motion sensors and how that area changes with the speed of motion. Third, a psychophysical experiment was conducted to understand how the receptive-field size of motion discrimination units in humans changes as a function of speed. The significant findings and contributions that arose out of these research projects are described below.

- A novel theoretical methodology, artificial scene statistics, was developed and implemented. The motivation behind this methodology is to use simulations

to overcome the difficulties of making accurate and detailed physical measurements. The central insight of this approach is the importance of developing and validating environmental models. This is a broad theoretical approach with applications to both computer vision and psychology.

- A repository of natural movies of an observer moving through a forest environment was created and will be made available online. There are other repositories of natural images that are valuable resources to the computational neuroscience community, most notably that of van Hateren (2005). However, the movies collected for this dissertation are the first such collection tailored to a particular task, i.e. navigation through a natural forest environment. In order to increase the value of the collection to researchers, as much information as possible was collected concurrently with the movies. The motion of the camera between frames is known precisely, the images were created with a calibrated camera so that accurate luminance and chromaticity values are known at each pixel location and disparity images were created along with the movie image sequence. These movies can serve as a resource to researchers studying the statistics of natural environments and tasks.

- Fourier power spectra of natural movies during first person locomotion were computed. These power spectra turned out to be very different from previously measured natural movie power spectra (Dong and Atick, 1995a). Unlike static scenes, the statistics of these scenes are highly dependent on the environmental structure of individual scenes. These measurements have implications for understanding the design of motion detection systems in human and computer visual systems.

- An environmental model for simulating locomotion through forest scenes was developed based on previously measured scene statistics and validated against new measurements. This model can be a useful tool for researchers studying the visual statistics of natural scenes in both human and computer vision. Code for automatically generating artificial movies of translation through the model forest environment is available upon request.

- A computational analysis was performed that showed that optimal aperture sizes for motion discrimination changes with speed. This result offers hy-

pothesis for neural coding in biological visual systems. Also, this result has applications for the design of motion systems in computer vision. Additionally, the result offers ecological explanations for psychophysical and physiological phenomena.

- An experimental methodology for psychophysically measuring ideal stimulus size was developed. The methodology is sensitive enough to measure efficiencies below the size for which humans are most efficient.

- Psychophysical measurements were made showing that efficiency declines as stimulus size is decreased past the optimal size for a foveal motion discrimination task. This is an important result that contributes to our understanding of the connection between electrophysiological receptive-field sizes and psychophysically estimated receptive-field sizes.

- Psychophysical measurements were made that indicate that receptive-field size does not change with speed for foveal motion discrimination using narrowband stimuli. This result contributes to our understanding of the structure and function of motion perception in the human visual system.

6.2 Discussion and Future Work

Because the reported work is interdisciplinary, it has implications for future work in a number of different areas. This section will be divided into future work in five different areas: image statistics, optimal aperture size estimation, psychophysics, physiology and computer vision.

6.2.1 Image Statistics

The image domain measurements described in Chapter 3 exclusively examined Fourier power spectra. Future analyses for measuring image motion statistics in real and simulated scenes could include statistics other than Fourier power spectra. An advantage of the Fourier power spectrum is that it captures an innate invariant across *static* scenes: the power spectrum falls as $\frac{1}{f^2}$ for most natural scene images. Results from this dissertation showed that this invariant property does not exist for spatiotemporal power spectra. There might, however, be a different statistic that

does capture some invariant property of natural movies. For example, there may be an alternative way to calculate Fourier power spectra that contains more information. The power spectra calculated in this dissertation were binned radially. There are clearly systematic differences in horizontal and vertical motion in the sampled movies. It is possible that a different method for calculating power spectra that included horizontal and vertical frequencies might actually simplify the data and find invariants across scenes, despite adding another dimension. For example, a wavelet analysis might be able to capture regularities in the expansion motion that are not apparent in radially averaged power spectra.

While all of the comparisons of power spectra presented in Chapter 3 were qualitative in their nature, one could create a quantitative comparison. Indeed, it would be desirable to have a model of the power spectra of natural scenes that alleviates the shortcomings of Dong and Atick's model. This would require a model that takes into account different distributions of velocities as well as the non-translational motions encountered during first-person motion. However, the assumption of translational motion comes not from the nature of the world, but rather from mathematical convenience. The translational motion assumption makes it possible to compute the necessary integrals for the model (Appendix A). How to contend with this mathematical difficulty is an avenue for future research.

6.2.2 Optimal Aperture Size

There are a variety of possible future research projects extending the theoretical analysis from Chapter 4, in which ideal aperture size was found to increase with increasing speed. One incremental research project would be to remove the assumption that the motion sensor has a circular receptive field. One could parametrize both the size and shape of the motion sensor aperture. Is the optimal motion receptive field elongated in the direction of motion? Does the elongation increase or decrease with speed? These are all interesting questions for future simulations.

A more ambitious extension of the optimal aperture size analysis would be to include human eye motions in the analysis. This would first require construction and validation of a model of human eye motion. If one had a model of human eye motion during locomotion, one could begin to answer the questions brought up in the discussion above. Due to fixation effects, is the ideal aperture size of motion sensors

at the fovea different from more eccentric sensors? How does the ideal aperture size change with speed at a non-foveal location?

In general, there are two basic routes for further research in simulated scene statistics: expanding and building on the environmental model, and exploring further questions about the optimal design of sensors. These two research paths are interdependent because studying additional aspects of optimal sensor design would likely necessitate adding detail to the environmental model. For example, instead of investigating the ideal area of spatial integration, one might investigate the ideal duration of temporal integration. Just as the ideal area depends largely on the statistics of the spatial derivatives of motion in the image sequence, one might expect that the ideal duration depends on the temporal derivatives of motion in the image sequence. To study this question one would, again, need detailed information about eye gaze durations and locations during locomotion.

6.2.3 Psychophysics

The experimental results from Chapter 5 were performed specifically in the fovea and for broadband stimuli. This still leaves open questions about the organization of motion receptive fields across the retina. It would be useful to repeat the psychophysical experiment at more eccentric locations in the retina. In addition the task performed in the experiment was a left-right motion discrimination task. Some might argue that this task is too different from the general motion vector error computed in the theoretical analysis. Another useful experiment would involve measuring optimal aperture size for an angular motion discrimination task, as opposed to a left-right discrimination task. The challenge in this instance is to design an experiment that controls against all possible confounds.

As described in Chapter 5, estimating receptive-field sizes of neurons psychophysically is an active area of research. Despite this activity, the theoretical foundations behind this research have never been fully fleshed out. Although it is clear that information is integrated more efficiently within a single neuron than across neurons, this does not necessarily lead to the conclusion that a whole visual system's efficiency will be dominated by the receptive field properties of a single receptor. It seems likely that the visual system's optimal stimulus size will be closely related to the receptive-field size of individual receptors, but this is a hypothesis that

ought to be explored theoretically. It may not be possible to develop this theory analytically, but a reasonable simulation would not be difficult. The question is: How would the efficiency of a model visual system change as a function of stimulus size in relation to the receptive-field size of the neurons? One could easily include in this simulation some more realistic details including overlapping receptive fields, a noisy stimulus location and sensor noise. It is likely that a simulation like this would predict a lower efficiency as the stimulus size decreased below the receptive-field size, as was observed in Chapter 5.

6.2.4 Physiology

There are several interesting physiological experiments suggested by the results in this dissertation. One could build on the research of Maunsell and Essen (1983), who measured single cell response properties of neurons in the middle temporal (MT) visual area in macaque monkeys. They found a positive correlation between speed tuning and eccentricity. Based on the theoretical analysis in Chapter 4, it seems possible that the correlation between speed tuning and eccentricity is due to a correlation between speed and receptive-field size. This raises three interesting physiological questions. First, at a given eccentricity do the receptive-field sizes of MT neurons correlate with speed? Next, how does the distribution of speed tuned neurons change at different eccentricities? Finally, is the distribution of motion receptive-field sizes different in the fovea than in other visual areas? Since the results presented in Chapter 5 suggest that the speed-size dependence exists for broadband cells, like in MT, but not for narrow-band cells, like in V1, all of the previous questions can be asked about V1 cells in addition.

6.2.5 Computer Vision

The emphasis of this dissertation has been on psychophysics and computational neuroscience, but the work has applications and extensions in computer vision as well.

The theoretical analysis in this dissertation showed that the optimal area of integration for motion discrimination increases with the speed of the motion being detected. This result has obvious applications to VLSI vision systems in particular, since they distribute sensors across the visual field and the integration area of those

sensors is fixed. Calow et al. (2005) showed that using a local area of integration that increases with increasing eccentricity improves heading estimation performance. Given the findings presented here, It seems likely that tying area of integration directly to speed instead of eccentricity could lead to further improvements.

Another finding from this dissertation applicable to computer vision involves ground plane scenes. This dissertation revealed that there is a much greater degree of non-translational motion and bias when estimating motion on the ground plane. This finding is particularly applicable to robots that move through human constructed flat environments. These robots may benefit from having their motion detectors designed in order to compensate for the large bias created by the non-translational motions of the ground plane. One could imagine a computer-vision system that distinguishes ground plane from non-ground plane and corrects for the large bias due to the non-translational motion in the ground plane areas of the visual field.

The simulated scene statistics methodology can be applied to many areas of computer vision research. For example, one could build accurate environmental models of robot environments and validate them against measured statistics. One could then use the model to tune and optimize computer vision algorithms. It is not unusual for computer vision researchers to use simulated environments to test the performance of their algorithms. However, the simulated scene statistical approach emphasizes three things that are not common in computer vision research: building environmental models based on measured statistics, validating environmental models based on real-world measurements, and using the simulated statistics to optimize the parameters of the computer vision system.

6.3 Conclusion

The natural systems approach can be summarized in the following set of steps:

1. Identify and characterize a natural task.
2. Measure and analyze the relevant environmental properties.
3. Build a computational model for the task and environment.

4. Perform experiments to test hypotheses generated from the computational model.

The artificial scene statistics approach fits into this framework by expanding on steps 2 and 3 above. To measure and analyze the environmental properties relevant to the task, one can build an environmental model based on measured scene properties and validate it against measured scene properties. With that environmental model, one can then use it to simulate across-domain statistical measurements or to help simulate a computational model of the task in the environment.

This dissertation describes a course of research that develops the simulated scene statistical approach. The motivation for this approach stems from the fact that measuring the statistics of natural visual environments during task performance can be a valuable tool to help understand visual computation. In particular, measuring across-domain statistics is especially powerful, but can be prohibitively difficult. Using artificial scene statistics helps overcome the difficulties of measuring complete across-domain statistics in natural environments. An important emphasis of this approach is to build an abstract environmental model that is based in previously measured real-world statistics and validated in comparison to additional real-world statistics.

In sum, this dissertation presents an approach to the study of vision that has applications to psychophysics, neuroscience and computer vision. The emphasis on accurate and validated environmental models for simulating scene statistics can help improve our understanding of the structure and function of the human visual system and also help us build more accurate and robust computer vision systems.

Appendix A

Modeling Fourier Power Spectra

This Appendix describes the math behind the modeling of the Fourier power spectra described in Chapter 3. The majority of this derivation is an adaptation of the Fourier model of Dong and Atick (1995a).

A.1 Fourier Transform of Translational Motion

Consider the case of strictly translational motion. Some notation:

- (x, y, t) – Space time coordinates in the image plane.
- (f_x, f_y, ω) – Frequency coordinates.
- $s_0(x, y)$ – The static spatial light intensity.
- $S_0(u, v)$ – The Fourier transform of s_0.
- $R_0(u, v)$ – Static power spectrum.
- $s(x, y, t)$ – Full image function.
- $S(f_x, f_y, \omega)$ – Fourier transform of s.
- $R(f_x, f_y, \omega)$ – Full power spectrum.

Assuming that the whole image is translating at velocity $\vec{v} = (v_x, v_y)$, then the moving light intensities are described by $s(x, y, t) = s_0(x - v_x t, y - v_y t)$. Taking the

Fourier transform of $s(x, y, t)$:

$$\begin{aligned} S(f_x, f_y, \omega) &= \iiint s(x, y, t) e^{-i2\pi(xf_x + yf_y + t\omega)} dx dy dt \\ &= \iiint s_0(x - v_x t, y - v_y t) e^{-i2\pi(xf_x + yf_y + t\omega)} dx dy dt \end{aligned}$$

Changing variables: $x' = x - v_x t$, $y' = y - v_y t$, $dx' = dx$, $dy' = dy$

$$\begin{aligned} S(f_x, f_y, \omega) &= \iiint s_0(x', y') e^{-i2\pi((x' + v_x t)f_x + (y' + v_y t)f_y + t\omega)} dx' dy' dt \\ &= \int e^{-i2\pi(v_x f_x + v_y f_y)} e^{-i2\pi\omega t} \left\{ \iint s_0(x', y') e^{-i2\pi(x' f_x + y' f_y)} dx' dy' \right\} dt \\ &= S_0(f_x, f_y) dt \int e^{-i2\pi(v_x f_x + v_y f_y)} e^{-i2\pi\omega t} \end{aligned}$$

Using the shift theorem:

$$S(f_x, f_y, \omega) = S_0(f_x, f_y) \delta(\omega - v_x f_x - v_y f_y)$$

This result holds as well for the power spectrum. Writing the same equation in vector notation for the power spectrum:

$$R(f_x, f_y, \omega) = R_0(f_x, f_y) \delta\left(\omega - \vec{v} \cdot \vec{f}\right)$$

To compute the average Fourier power of a series of translational motions draw from a distribution of velocities, $P(\vec{v})$, we integrate across velocities:

$$R(f_x, f_y, \omega) = \int_{\vec{v}} R_0(f_x, f_y) \delta\left(\omega - \vec{v} \cdot \vec{f}\right) P(\vec{v}) d\vec{v}$$

Notice that there is an implicit assumption here that there is an average static spectrum, R_0 that does not depend on the velocity. This is a reasonable assumption for natural scenes, since numerous studies have shown that the average static amplitude spectrum is proportional to $\frac{1}{f^n}$. Unless the surface reflectance of objects changes as one moves through the world, this assumption will hold true. If we also assume that $P(\vec{v})$ is rotationally invariant, then the Fourier spectrum must also be rotationally invariant. We can then simplify the equation and rewrite it in terms of

a singe frequency component, f_x, and velocity component $P_x(v_x) = \int P(V)dv_y$. For simplicity we will just write, f, v and $P(v)$ from now on.

$$R(f,\omega) = \int_v R_0(f)\delta(\omega - vf)\, P(v)dv$$

Changing variables: $v' = fv$, $dv' = fdv$

$$\begin{aligned} R(f,\omega) &= \int_{v'} R_0(f)\delta(\omega - v')P(\frac{v'}{f})\frac{1}{f}dv' \\ R(f,\omega) &= \frac{R_0(f)}{f}\int_{v'} \delta\big(\omega - v'\big)\, P(\frac{v'}{f})dv' \\ R(f,\omega) &= \frac{R_0(f)}{f}P(\frac{\omega}{f}) \end{aligned}$$

We can model the static spectrum as $R_0(f) = \frac{K}{f^m}$ and use $v = \frac{\omega}{f}$ to solve for $P(v)$:

$$P(v) = \frac{K}{f^{m+1}}R(f,\omega)$$

Thus, if the assumptions in this model hold true, we can plot the average probability distribution of velocity amplitudes in a set of scenes. One can think of $\frac{R(f,\omega)}{f^{m+1}}$ as being the power adjusted by the average static spatial frequency. In each of the power spectrum figures in Chapter 3, subplot (b) shows the spatial frequency adjusted power as a function of velocity. To generate a plot like this, the exponent m is fit from the average static power spectrum of the set of scenes. Plotting the spatial frequency adjusted power as a function of $\frac{w}{f}$ shows the approximate distribution of image-plane speeds, $P(v)$.

It is important to keep in mind all the assumptions that enter this model.

- All motion is translational relative to the image plane.
- The distribution of motion velocities in the image plane is rotationally invariant.
- The static power spectrum has a consistent average that is independent of the motion in the scene.

Each of the assumptions is likely to be broken to some extent in the real and in the simulated scenes. Despite this, the frequency-adjusted probability plots can serve

as approximations for the distribution of image plane velocities in the scene. The less these assumptions are broken, then the less scatter there is in these plots. For example, compare the scatter in Figure 3.4*b*, a simulation based on translational motion, with the scatter in Figure 3.13*b*, a flat ground plane which contains a large amount of non-translational motions.

A.2 Dong and Atick's Model

Dong and Atick (1995a) used the model described above, with the additional assumption that about the shape of the distribution of velocities in the world (rather than in the image plane). They assume that motions in the world are translational relative to the image plane, and are drawn from a power-law distribution of velocities:

$$P(u) \sim \frac{1}{(u+u_0)^n}$$

This distribution has a mean velocity, $\bar{u} = u_0/(n-2)$. The velocities occur at a random distance r uniformly drawn from the range r_1 to r_2. The model predicts the power spectrum will be:

$$R(f,\omega) = \frac{K\bar{u}}{2f^{m-1}\omega^2}\left[\frac{n-2}{(x+1)^{n-1}} - \frac{n-1}{(x+1)^{n-2}}\right]_{(\omega r_1)/(fu_0)}^{(\omega r_2)/(fu_0)}$$

(Dong and Atick, 1995a, p.354 eq.19)

The model has four free parameters:

- $\bar{u}$ – mean velocity
- n – exponent in power distribution
- r_1, r_2 – the range of distances over which motion occurs

Notice that K and m are fit from the static power spectrum and are therefore not considered to be 'free' parameters. To fit this model, the sum of the squared errors in the frequency-adjusted velocity space ($P(v)$ space) were minimized using the simplex algorithm in Matlab$^{\text{TM}}$. Function fits using this model are shown in Chapter 3 and in Appendix C.

Appendix B

Ideal Observer Model

This Appendix describes the ideal observer model used in Chapter 5. Most of this appendix is a discrete version of the derivations found in "Detection of Signals in Noise" (Whalen, 1971, p.153–161).

To start with, we define some notation:

- i – Pixel location. This includes the spatial and temporal location combined into a single parameter. (Often this is (x, y, t), but for the sake of simplicity, it is described here with just a single index).
- n_i – Noise at location i (which is assumed to be normally distributed, zero mean, with standard deviation σ_n).
- s_i^a, s_i^b – Signal a or b at location i.
- $r_i = n_i + s_i$ – Received stimulus (noise and signal) at location i (an a or b superscript may indicate which signal is actually present).
- $\Delta s_i = s_i^a - s_i^b$ – Ideal observer match template.
- $G^a = \sum_i r_i^a \Delta s_i + thresh$ – The ideal observer will compute its response variable, G, by multiplying the observed signal r_i with the match template Δs_i and summing over all locations. When the response variable exceeds the $thresh$, the ideal observer reports that signal a is detected otherwise signal b is detected.

B.1 Derivation of the Optimal Decision Rule

Our observed signal is

$$r_i = \left\{ \begin{array}{c} s_i^a \\ or \\ s_i^b \end{array} \right\} + n_i$$

We want to design a decision rule that chooses between two hypothesis (signal a is present or signal b is present). Lets say there are m possible pixel values ($1 <= i <= m$). The optimal decision rule is to use the likelihood ratio, $\lambda(\mathbf{r})$, and compare it to a threshold, λ_0.

$$\lambda(\mathbf{r}) = \frac{p^a(r_1, r_2, \cdots, r_m)}{p^b(r_1, r_2, \cdots, r_m)}$$

Let's characterize these probability distributions. Since n is a Gaussian random variable, and is independent at each i, the likelihoods are also Gaussian and can therefore be characterized by their mean and standard deviation.

$$\begin{aligned} \mu_\lambda &= E\{r_i\} = E\{s_i + n_i\} = s_i \\ \sigma_\lambda^2 &= E\left\{[r_i - E\{r_i\}]^2\right\} = E\{n_i^2\} = \sigma_n^2 \end{aligned}$$

where σ_n^2 is the variance of the noise n. Therefore the likelihood functions are:

$$\begin{aligned} p^a(r_1, \cdots, r_m) &= \left(\frac{1}{2\pi\sigma_n^2}\right)^{m/2} \exp\left[-\sum_{i=1}^{m} \frac{(r_i - s_i^a)^2}{2\sigma_n^2}\right] \\ p^b(r_1, \cdots, r_m) &= \left(\frac{1}{2\pi\sigma_n^2}\right)^{m/2} \exp\left[-\sum_{i=1}^{m} \frac{(r_i - s_i^b)^2}{2\sigma_n^2}\right] \end{aligned}$$

The likelihood ratio is then:

$$\begin{aligned} \lambda(\mathbf{r}) &= \frac{p^a(\mathbf{r})}{p^b(\mathbf{r})} = \frac{\exp\left[-\sum_{i=1}^{m} \frac{(r_i - s_i^a)^2}{2\sigma_n^2}\right]}{\exp\left[-\sum_{i=1}^{m} \frac{(r_i - s_i^b)^2}{2\sigma_n^2}\right]} \\ \lambda(\mathbf{r}) &= \exp\left[-\frac{1}{2}\sum_{i=1}^{m}\left(\frac{2r_i s_i^b}{\sigma_n^2} - \frac{2r_i s_i^a}{\sigma_n^2} - \frac{(s_i^b)^2 - (s_i^a)^2}{\sigma_n^2}\right)\right] \geq \lambda_0 \end{aligned}$$

Taking the log gives us the log-likelihood ratio. After rearranging terms, we get:

$$\sum_i r_i s_i^a - \sum_i r_i s_i^b + \frac{1}{2}\sum_i \left[\left(s_i^b\right)^2 - (s_i^a)^2\right] \geq \left(\sum_i \sigma_n^2\right) \ln(\lambda_0)$$

If the cost of both kinds of errors (false alarms and false positives) is equal and the priors on the signals are equal, then $\lambda_0 = 0$ and our decision rule becomes choose signal a if:

$$G = \sum_i r_i s_i^a - \sum_i r_i s_i^b + \frac{1}{2}\sum_i \left[\left(s_i^b\right)^2 - (s_i^a)^2\right] \geq 0$$

B.2 Performance of Ideal

The task in our psychophysical experiment was to discriminate between two Gabor signals in noise moving in opposite direction. We are interested in the performance of the ideal observer (matched filter) in the same task. The prior on the signals is equal and we want to maximize percent correct (equal cost for both kinds of errors). We will describe the calculations for an arbitrary signal, s, in white noise. Gabors are just a special case.

In order to know the average performance of this ideal observer, we need to compute the mean and variance of G under our noise conditions. From symmetry arguments we can see that the mean is the same but of opposite sign if signal a or b is present. The variance of G will be the same regardless of which signal is present.

From above, the optimal decision rule is, choose signal a if:

$$G = \sum_i r_i s_i^a - \sum_i r_i s_i^b + \frac{1}{2}\sum_i \left[\left(s_i^b\right)^2 - (s_i^a)^2\right] \geq 0$$

Let's look at this under the assumption that signal a is present:

$$\begin{aligned}
G^a &= \sum_i r_i^a s_i^a - \sum_i r_i^a s_i^b + \frac{1}{2}\sum_i \left[\left(s_i^b\right)^2 - (s_i^a)^2\right] \\
&= \sum_i \left(n_i + s_i^a\right) s_i^a - \sum_i \left(n_i + s_i^a\right) s_i^b + \frac{1}{2}\sum_i \left[(s_i^b)^2 - (s_i^a)^2\right] \\
&= \sum_i \left[n_i s_i^a + (s_i^a)^2 - n_i s_i^b - s_i^a s_i^b + \frac{(s_i^b)^2}{2} - \frac{(s_i^a)^2}{2}\right] \\
&= \sum_i \left[\frac{(s_i^a)^2}{2} - s_i^a s_i^b + \frac{(s_i^b)^2}{2} + n_i s_i^a - n_i s_i^b\right] \\
&= \sum_i \left[\frac{1}{2}\left(s_i^a - s_i^b\right)^2 + n_i s_i^a - n_i s_i^b\right]
\end{aligned}$$

Thus the expected value of G^a is:

$$\begin{aligned}
E\{G^a\} &= E\left\{\sum_i \left[\frac{1}{2}\left(s_i^a - s_i^b\right)^2 + n_i s_i^a - n_i s_i^b\right]\right\} \\
&= E\left\{\sum_i \frac{1}{2}\left(s_i^a - s_i^b\right)^2 + \sum_i n_i s_i^a - \sum_i n_i s_i^b\right\} \\
&= E\left\{\sum_i n_i s_i^a - \sum_i n_i s_i^b\right\} + \frac{1}{2}\sum_i \left(s_i^a - s_i^b\right)^2
\end{aligned}$$

Since n_i has zero mean, we can see that the terms containing n_i end up with an expected value of zero. Therefore:

$$E\{G^a\} = \frac{1}{2}\sum_i \left(s_i^a - s_i^b\right)^2 = \frac{1}{2}\sum_i \Delta s_i^2$$

And by symmetry arguments,

$$E\{G^a\} = -\frac{1}{2}\sum_i \Delta s_i^2$$

The performance of this detector depends on the variance of the response variable:

$$Var\{G^a\} = E\left\{[G^a - E\{G^a\}]^2\right\}$$

Let's start by just looking at the term inside the brackets (from the equations above):

$$\begin{aligned} G^a - E\{G^a\} &= \sum_i \left[\frac{1}{2}\left(s_i^a - s_i^b\right)^2 + n_i s_i^a - n_i s_i^b\right] - \frac{1}{2}\sum_i \left(s_i^a - s_i^b\right)^2 \\ &= \sum_i n_i \Delta s_i \end{aligned}$$

So the variance is then:

$$\begin{aligned} Var\{G^a\} &= E\left\{\left(\sum_i n_i \Delta s_i\right)^2\right\} \\ &= \sum_i \sum_j \Delta s_i \Delta s_j E\{n_i n_j\} \end{aligned}$$

Since the expected value of the product of two zero-mean random variables is zero, the only places where the expected value is non-zero is where $x = y$:

$$\begin{aligned} &= \sum_i \sum_j \Delta s_i \Delta s_j E\{n_i n_j\} \delta_{ij} \\ &= \sum_i \Delta s_i^2 E\{n_i^2\} \end{aligned}$$

Since n_i is a zero-mean distribution with standard deviation σ_n,

$$Var\{G^a\} = Var\left\{G^b\right\} = \sum_{i=1}^{m} \left(\Delta s_i^2 \sigma_n^2\right) = \sigma_n^2 \sum_{i=1}^{m} \Delta s_i^2 \tag{B.1}$$

Now, we can characterize the discriminability of the two signals by their d', the difference in the expected means of the decision variable divided by the standard deviation of the decision variable:

$$d' = \frac{\mu_a - \mu_b}{\sigma_G} = \frac{E\{G_a\} - E\{G_b\}}{\sqrt{Var\{G\}}} = \frac{\sum_i \Delta s_i^2}{\sqrt{\sigma_n^2 \sum_i \Delta s_i^2}}$$

$$d' = \frac{\sqrt{\sum_i \left(s_i^a - s_i^b\right)^2}}{\sigma_n} \tag{B.2}$$

Because our signals are rendered on a computer CRT with discrete pixels, our signal is inherently discretized. For each signal in our experiment, we compute the d' using the discrete sums in equation B.2.

Efficiency is defined as the square of the ratio of the observer's d' with that of the ideal:

$$Efficiency = \left(\frac{d'_{observer}}{d'_{ideal}}\right)^2$$

There are some additional notational conveniences we can add. Let's define the energy in the signals to be

$$\mathbf{E} = \frac{1}{2}\sum_i \left[\left(s_i^b\right)^2 - (s_i^a)^2\right]$$

And the cross-correlation is:

$$\rho = \frac{1}{\mathbf{E}}\sum_i s_i^b s_i^a$$

It can be shown that

$$\begin{aligned} E\{G^a\} &= \mathbf{E}(1-\rho) \\ E\left\{G^b\right\} &= -\mathbf{E}(1-\rho) \\ Var\{G\} &= 2\sigma_n^2\mathbf{E}(1-\rho) \end{aligned}$$

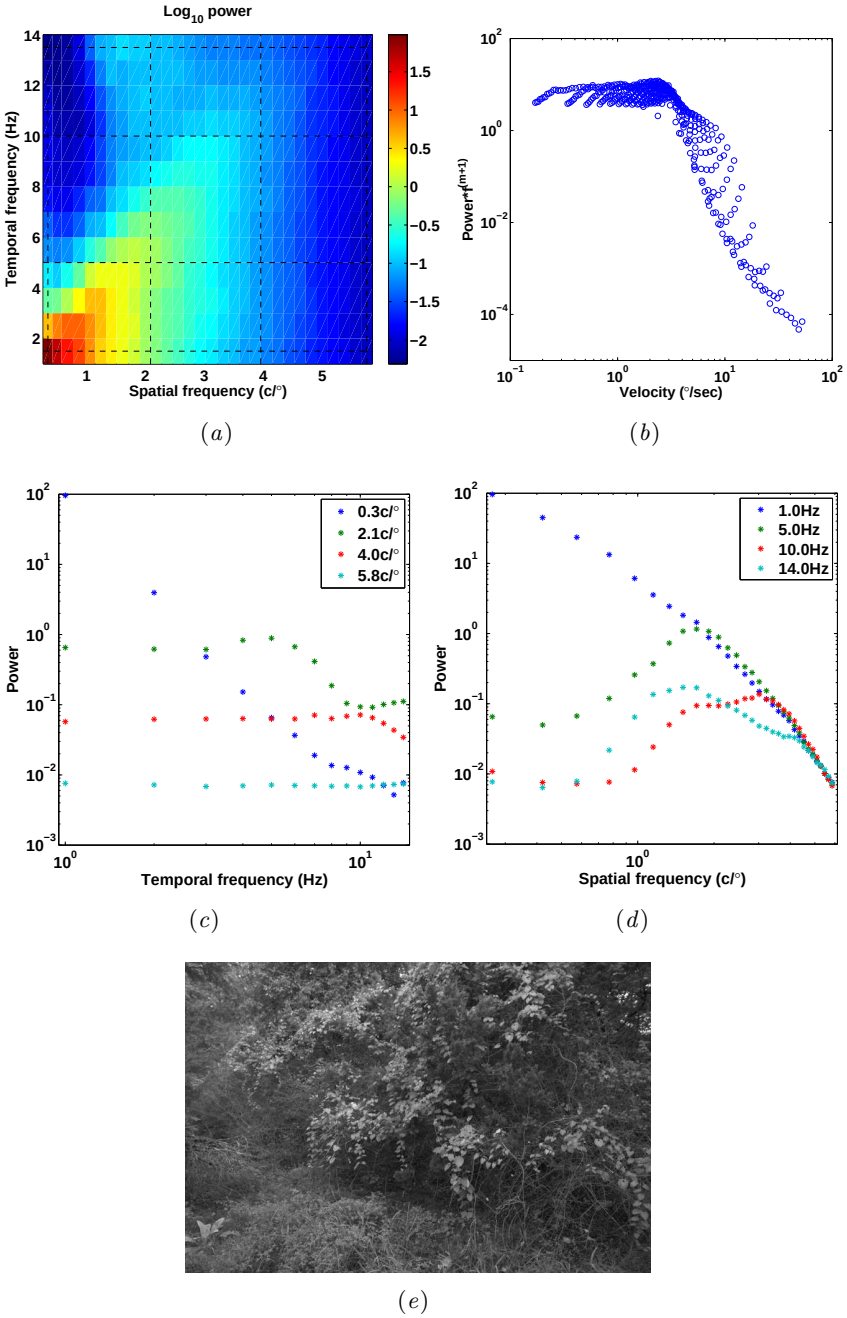
And therefore:

$$d' = \frac{2\mathbf{E}(1-\rho)}{\sqrt{\sigma_n^2\mathbf{E}(1-\rho)}} = \sqrt{\frac{2\mathbf{E}(1-\rho)}{\sigma_n^2}}$$

Appendix C

Data Plots

For reference, this Appendix contains the complete set of plots of the power spectra from the artificial and simulated scenes described in Chapter 3. This includes Fourier power spectra plots for all five natural scenes and all eight simulated scenes. Note that the data figures from Chapter 3 are replicated here.



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

Figure C.1: **Power spectrum of Natural Movie 1.**

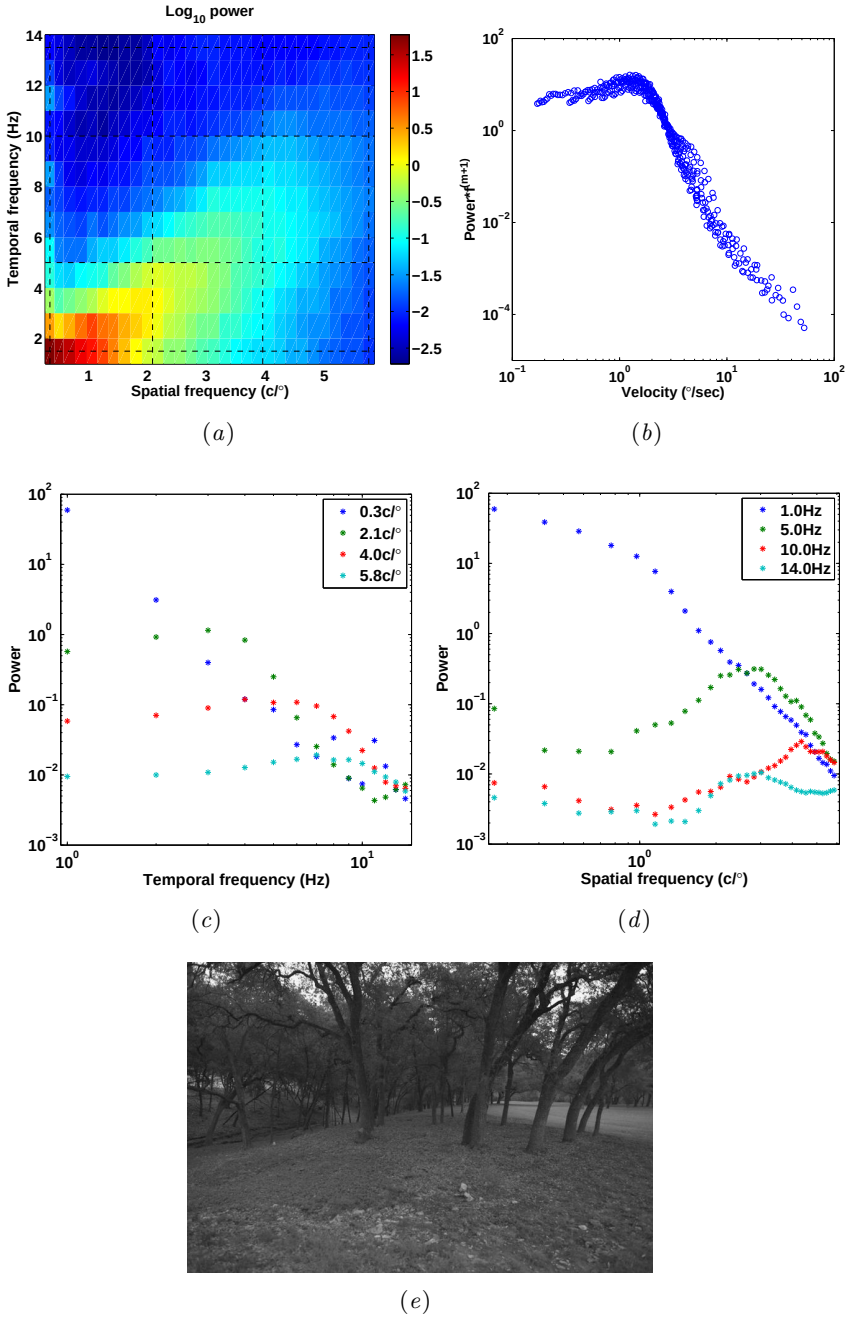
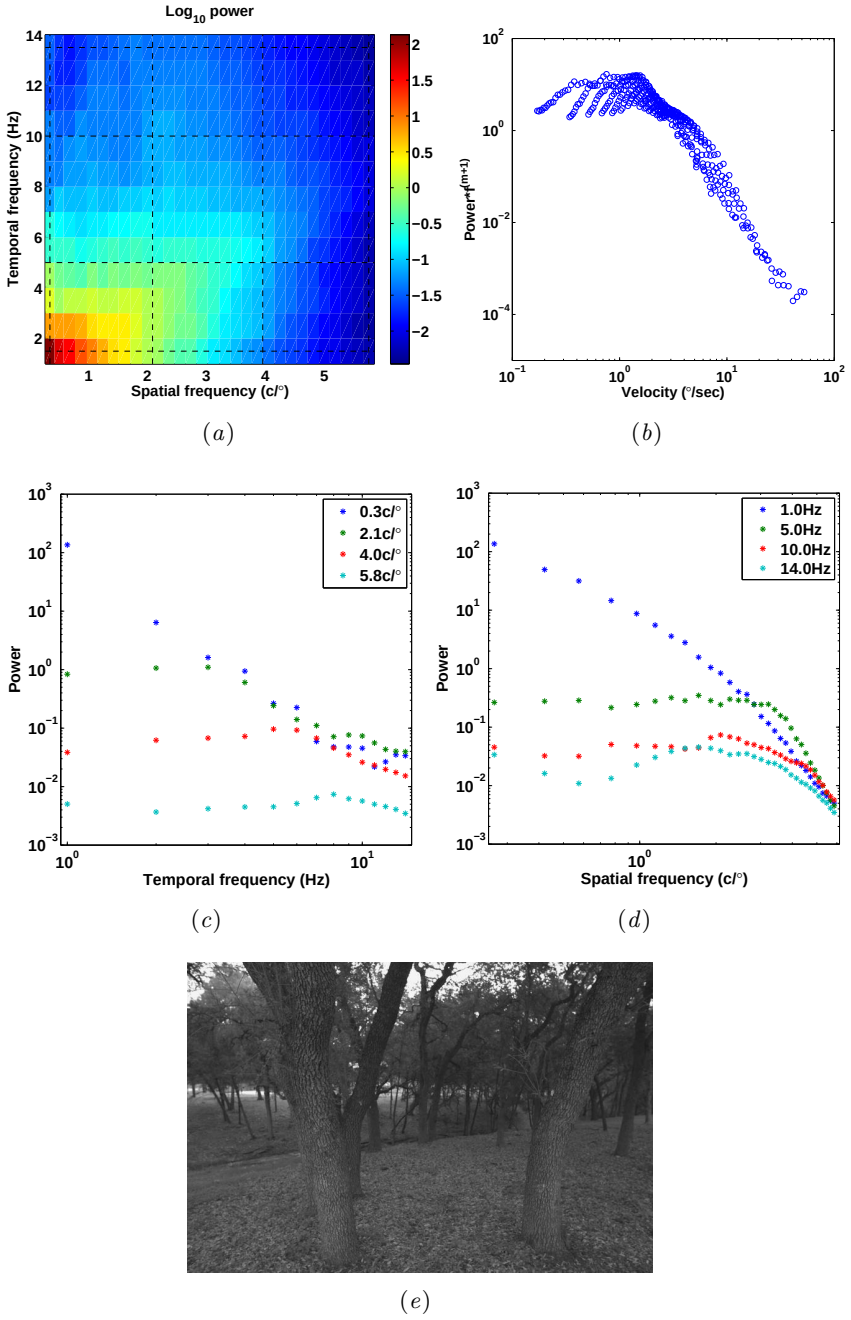


Figure C.2: **Power spectrum of Natural Movie 2.**



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

Figure C.3: **Power spectrum of Natural Movie 3.**

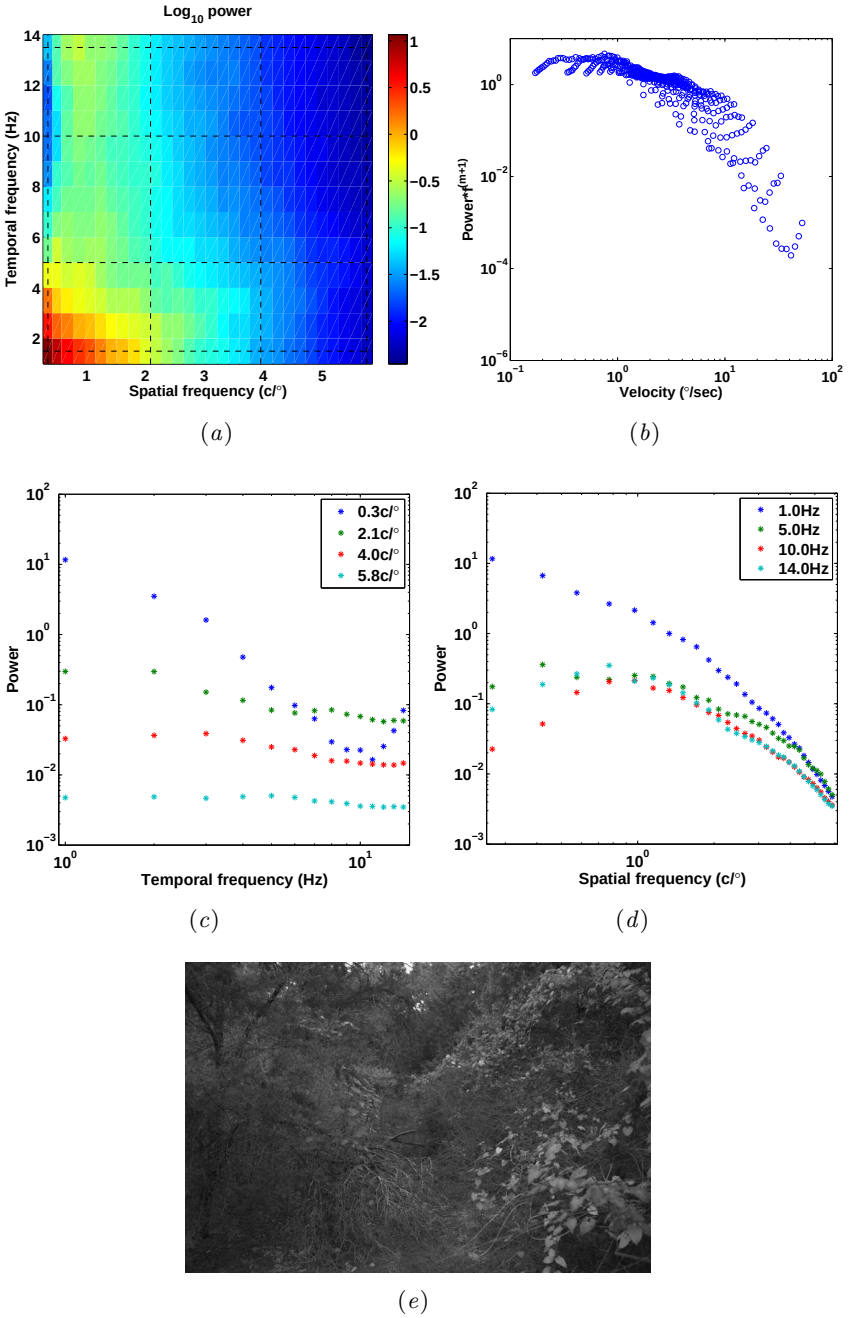


Figure C.4: **Power spectrum of Natural Movie 4.**

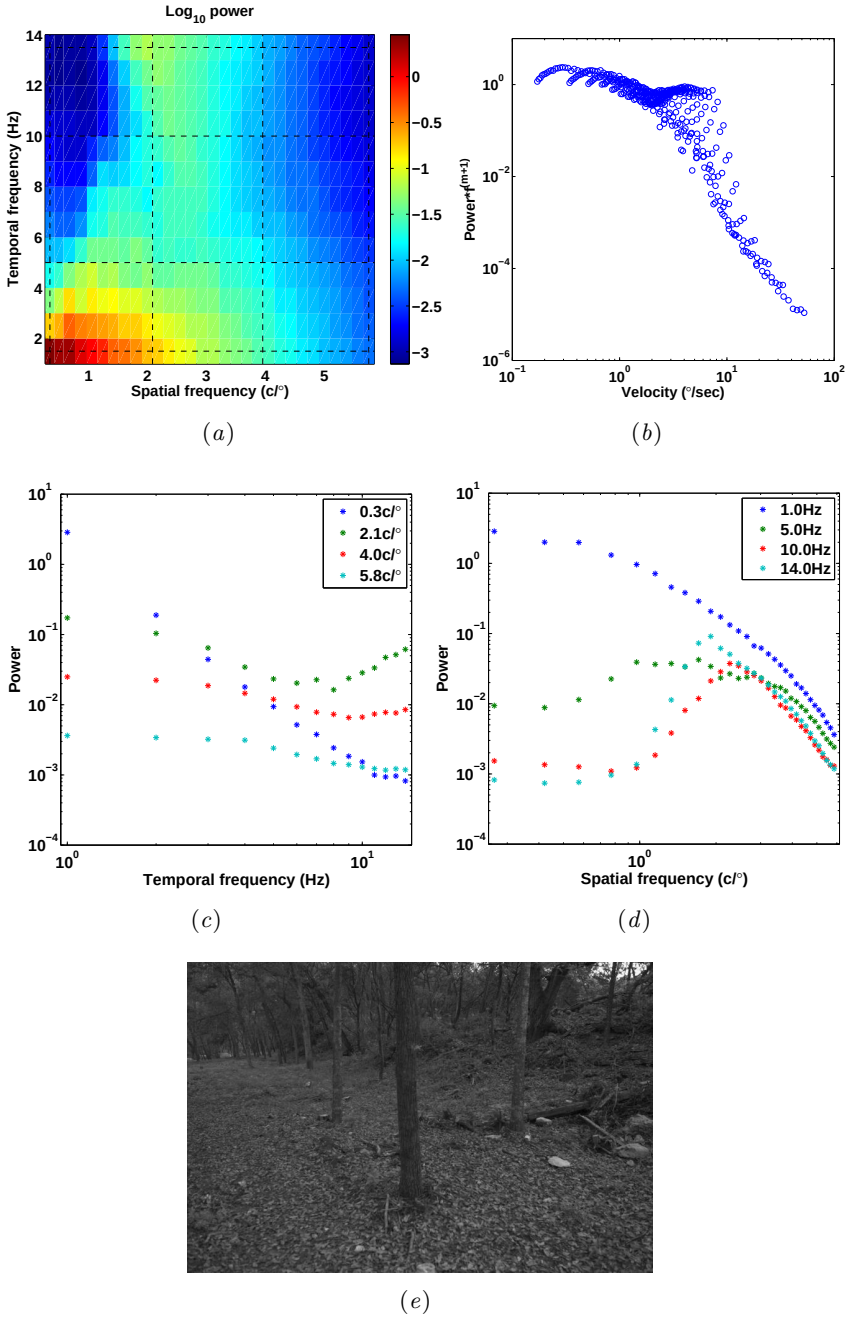


Figure C.5: **Power spectrum of Natural Movie 5.**

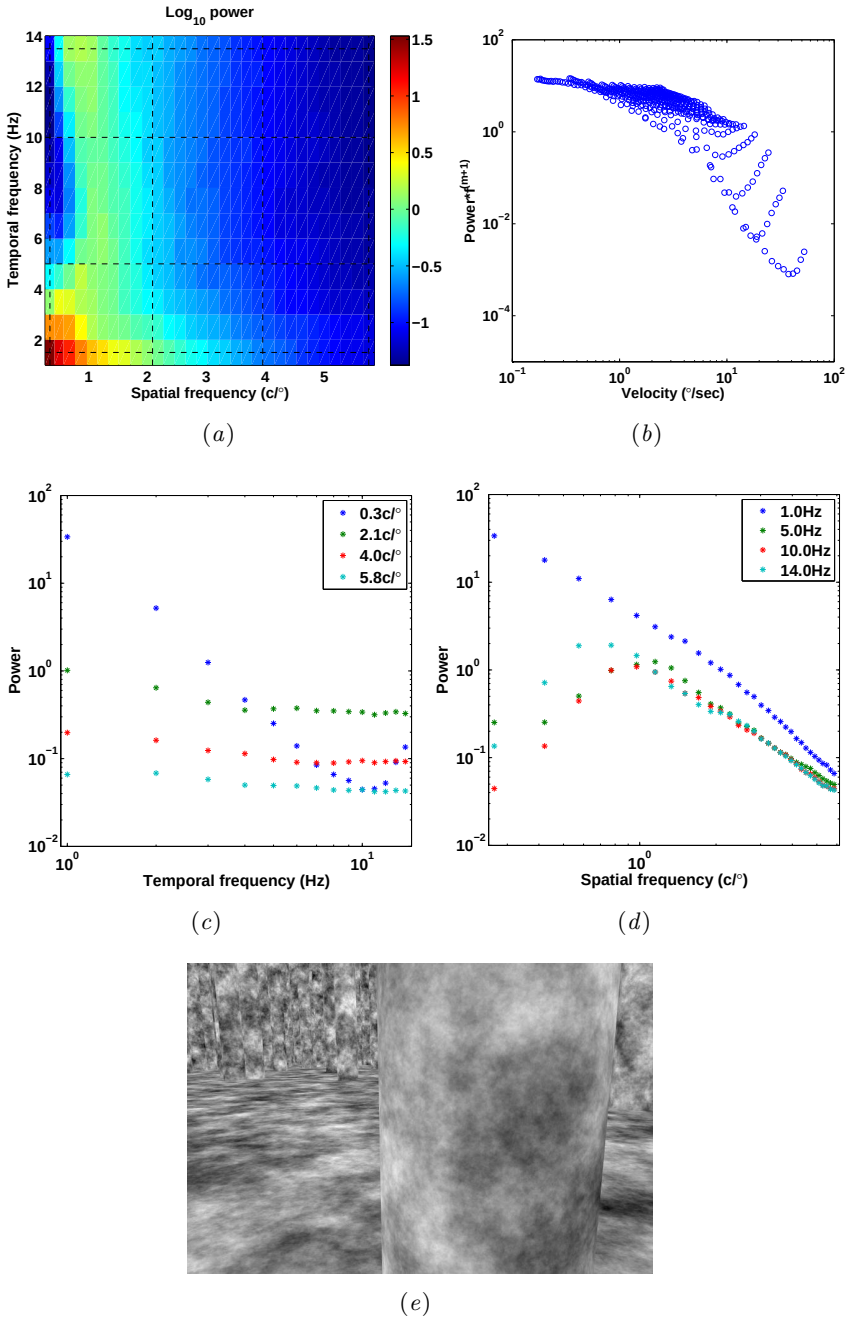


Figure C.6: **Power spectrum of low-density forest scene 1 at a 0.9 meters/second walking speed.**

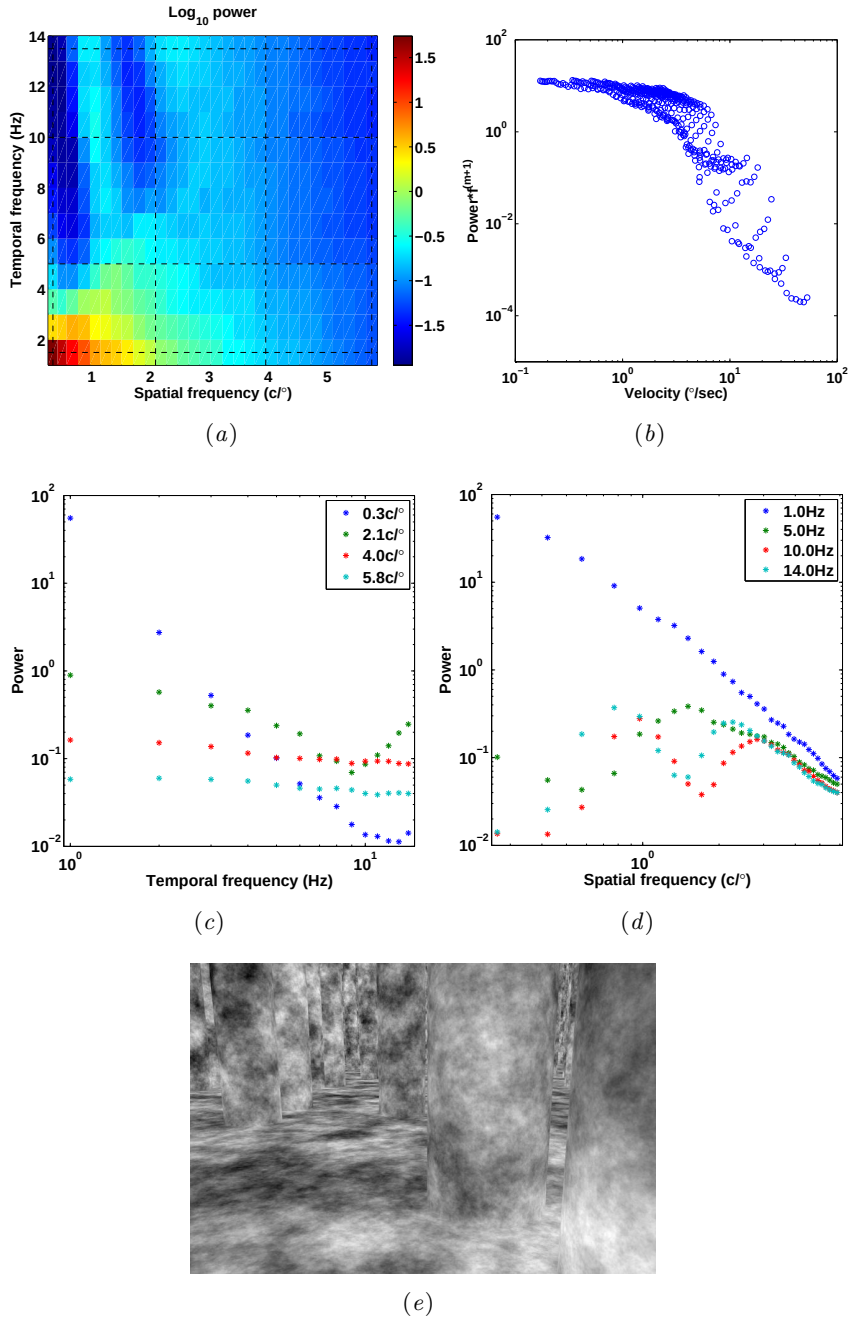


Figure C.7: **Power spectrum of low-density forest scene 2 at a 0.9 meters/second walking speed.**

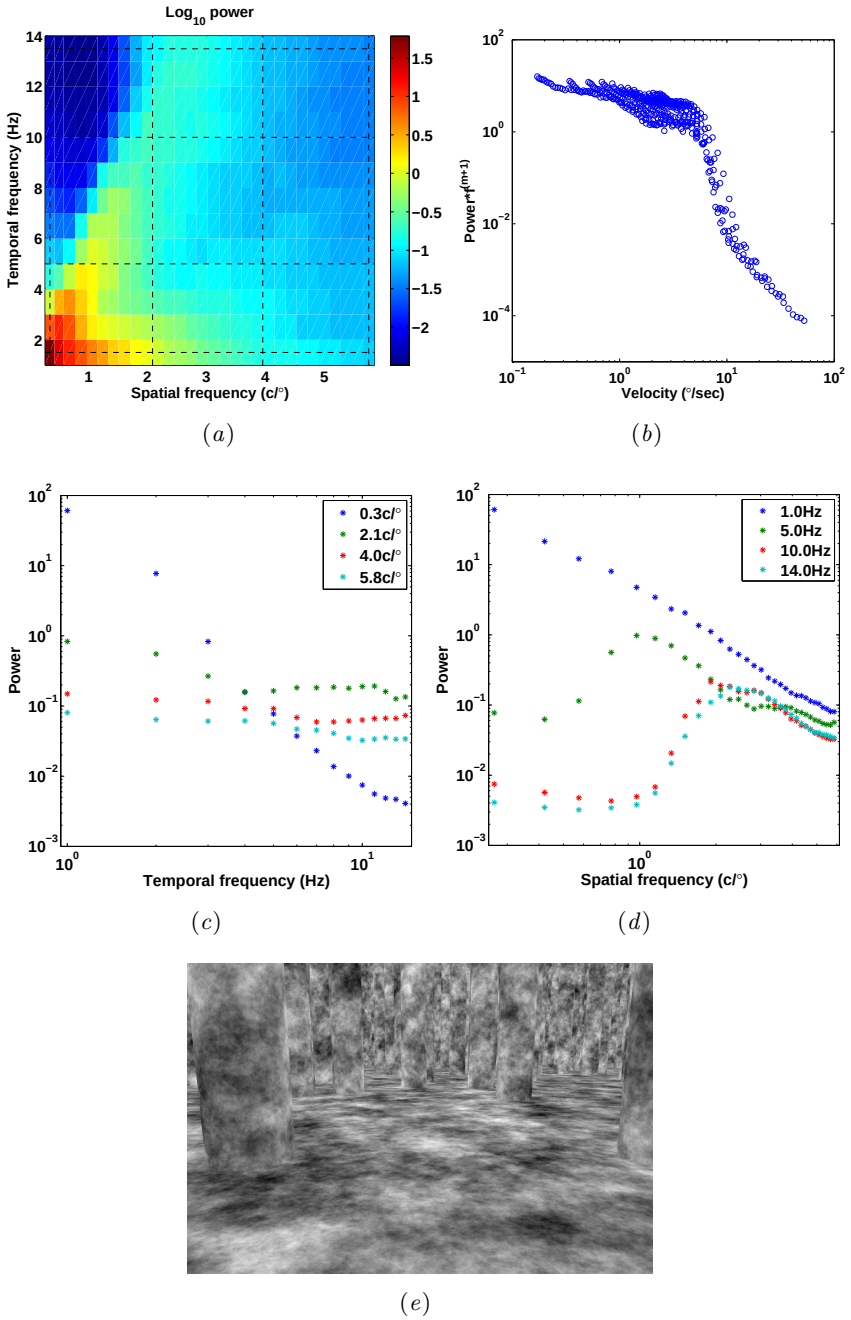


Figure C.8: **Power spectrum of low-density forest scene 3 at a 0.9 meters/second walking speed.**

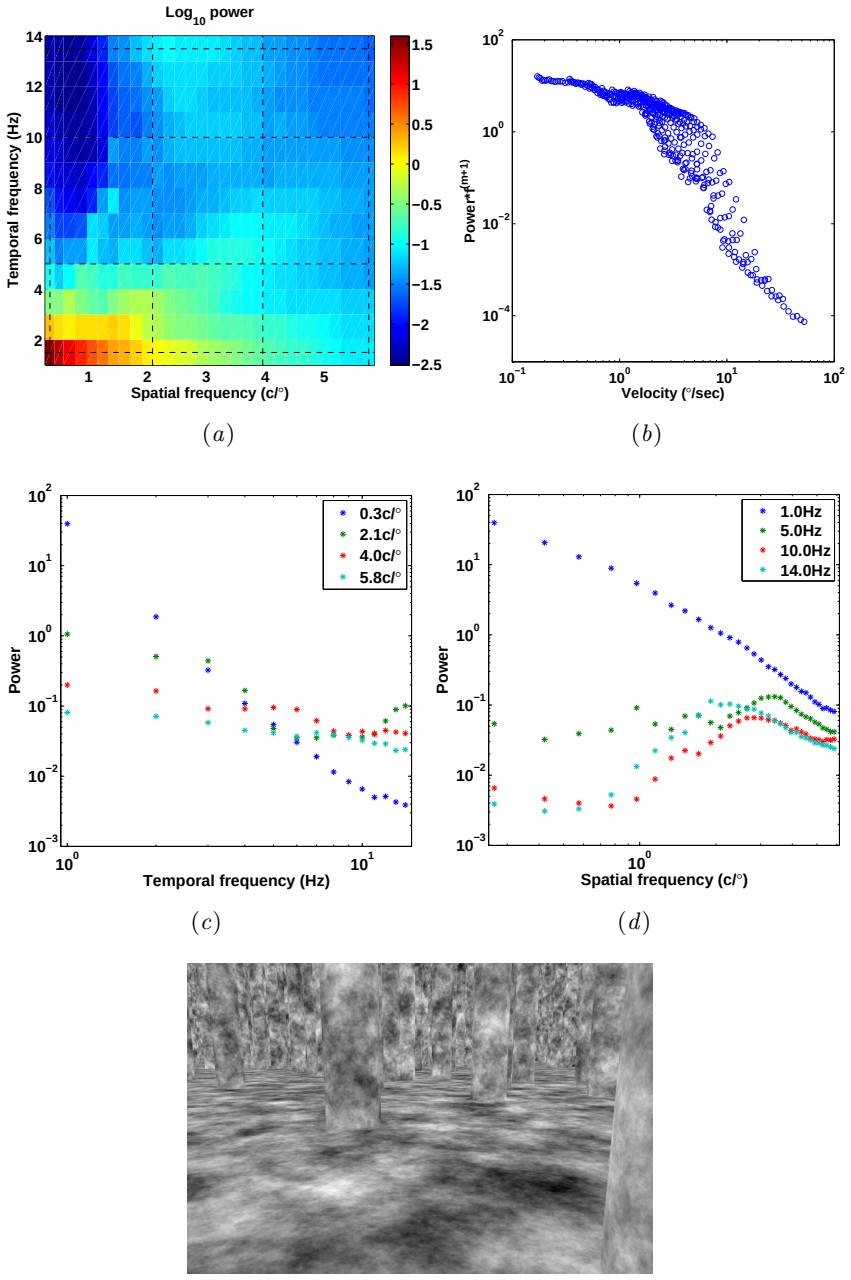
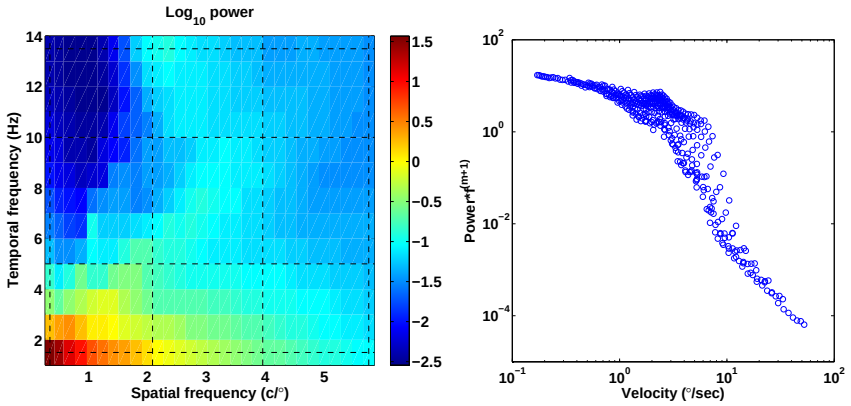
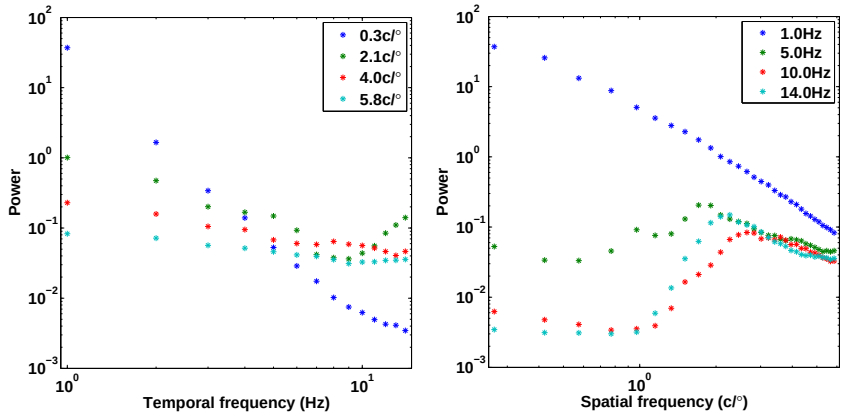


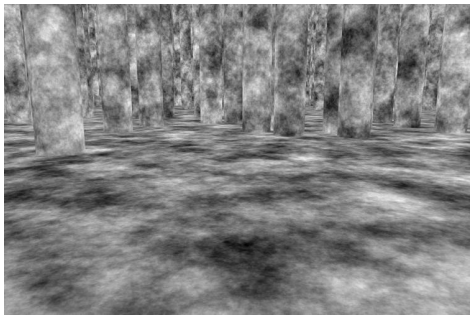
Figure C.9: **Power spectrum of low-density forest scene 4 at a 0.9 meters/second walking speed.**



(a) (b)



(c) (d)



(e)

Figure C.10: **Power spectrum of low-density forest scene 5 at a 0.9 meters/second walking speed.**

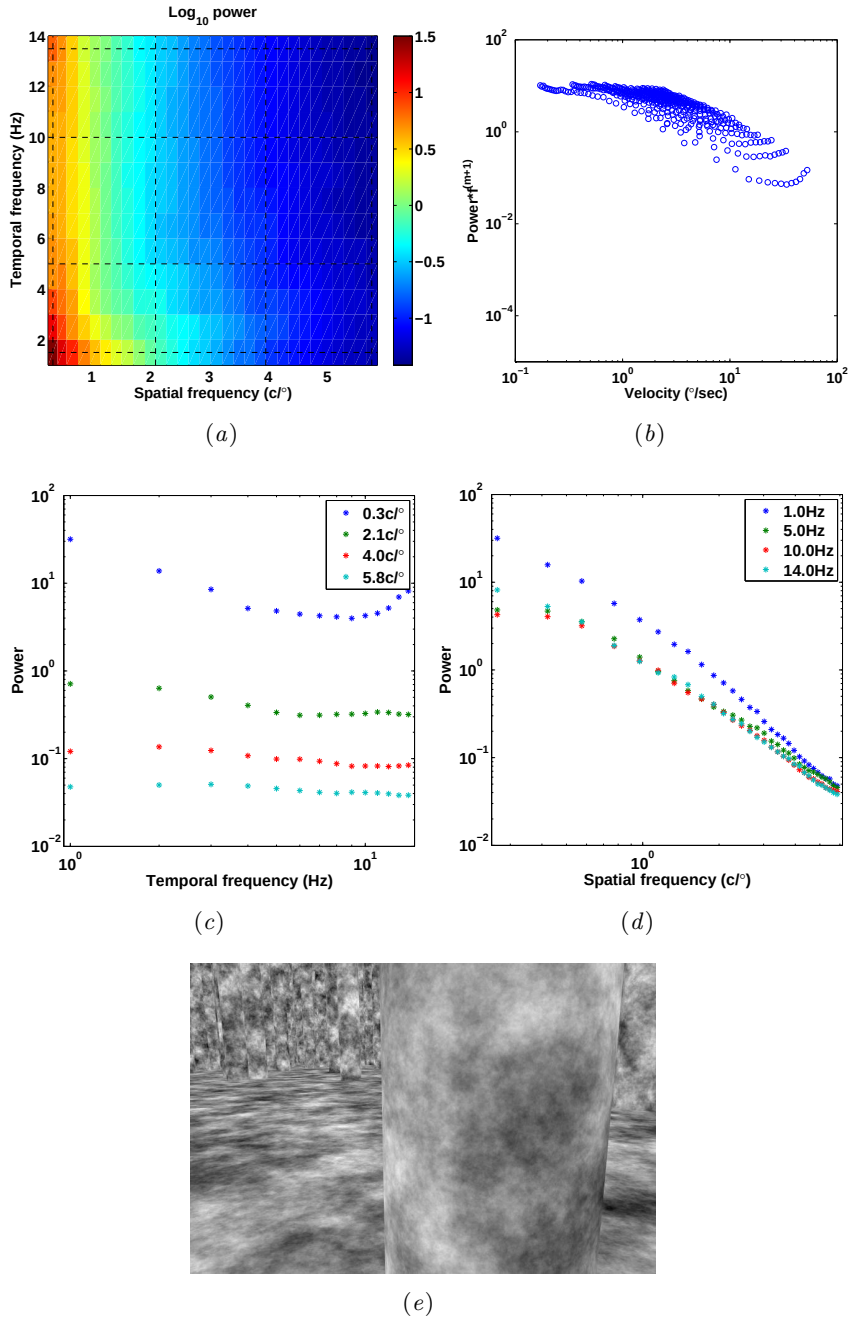


Figure C.11: **Power spectrum of low-density forest scene 1 at a 1.5 meters/second walking speed.**

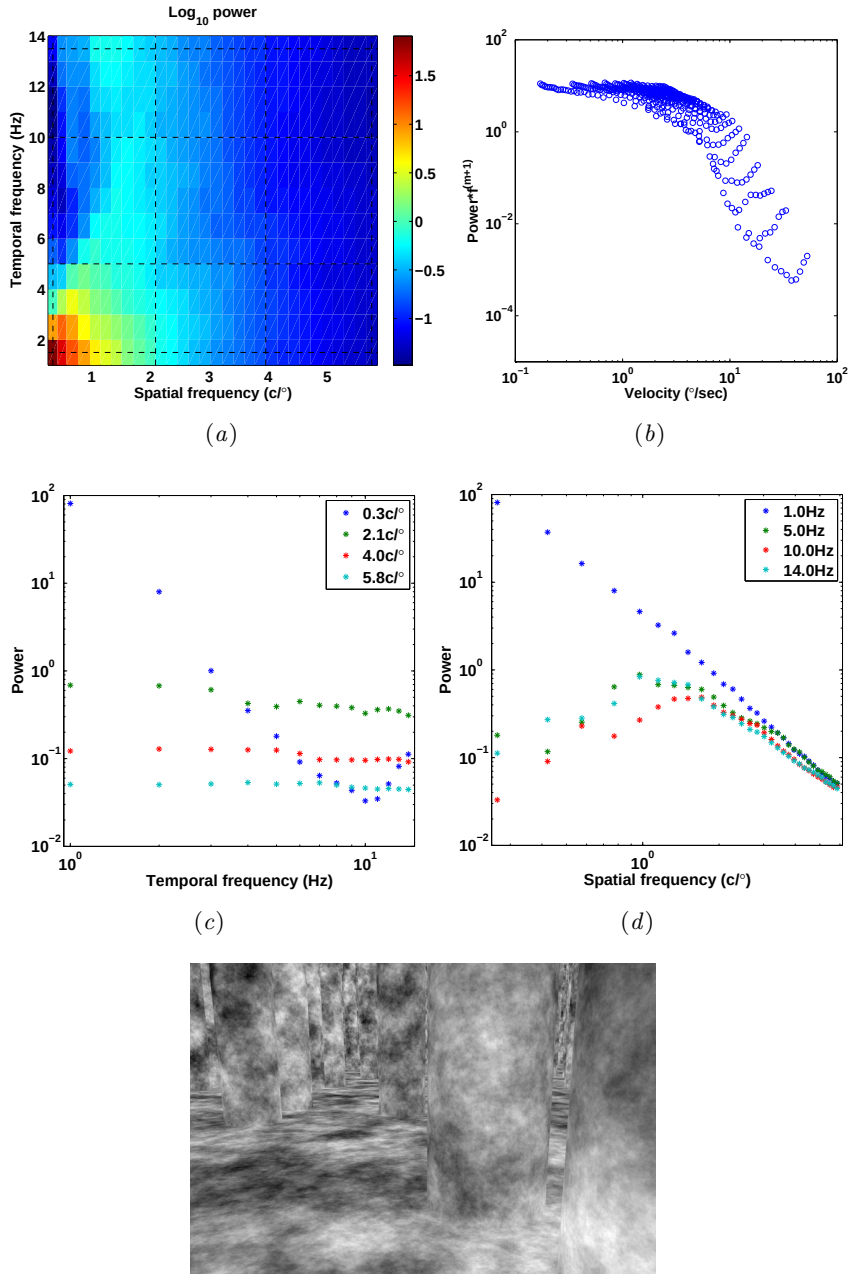


Figure C.12: **Power spectrum of low-density forest scene 2 at a 1.5 meters/second walking speed.**

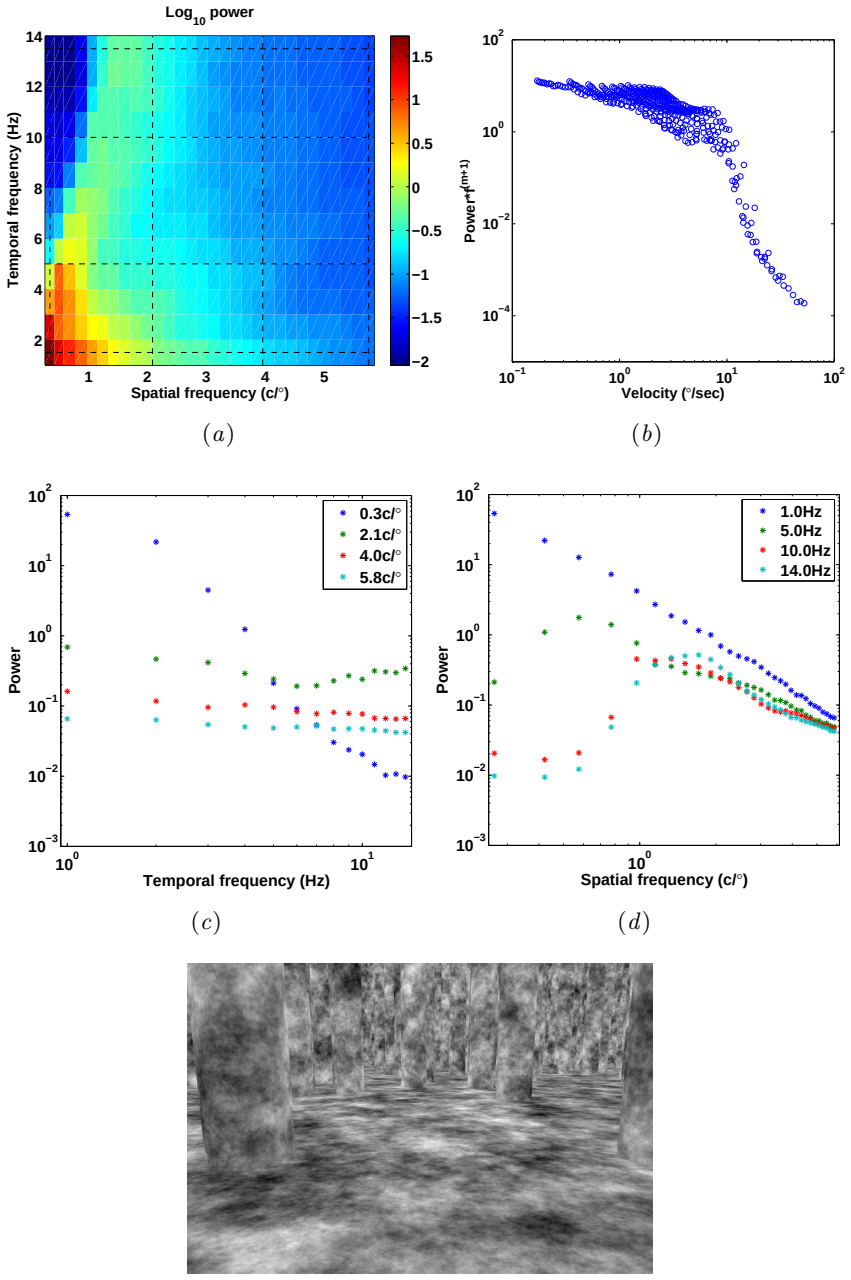


Figure C.13: **Power spectrum of low-density forest scene 3 at a 1.5 meters/second walking speed.**

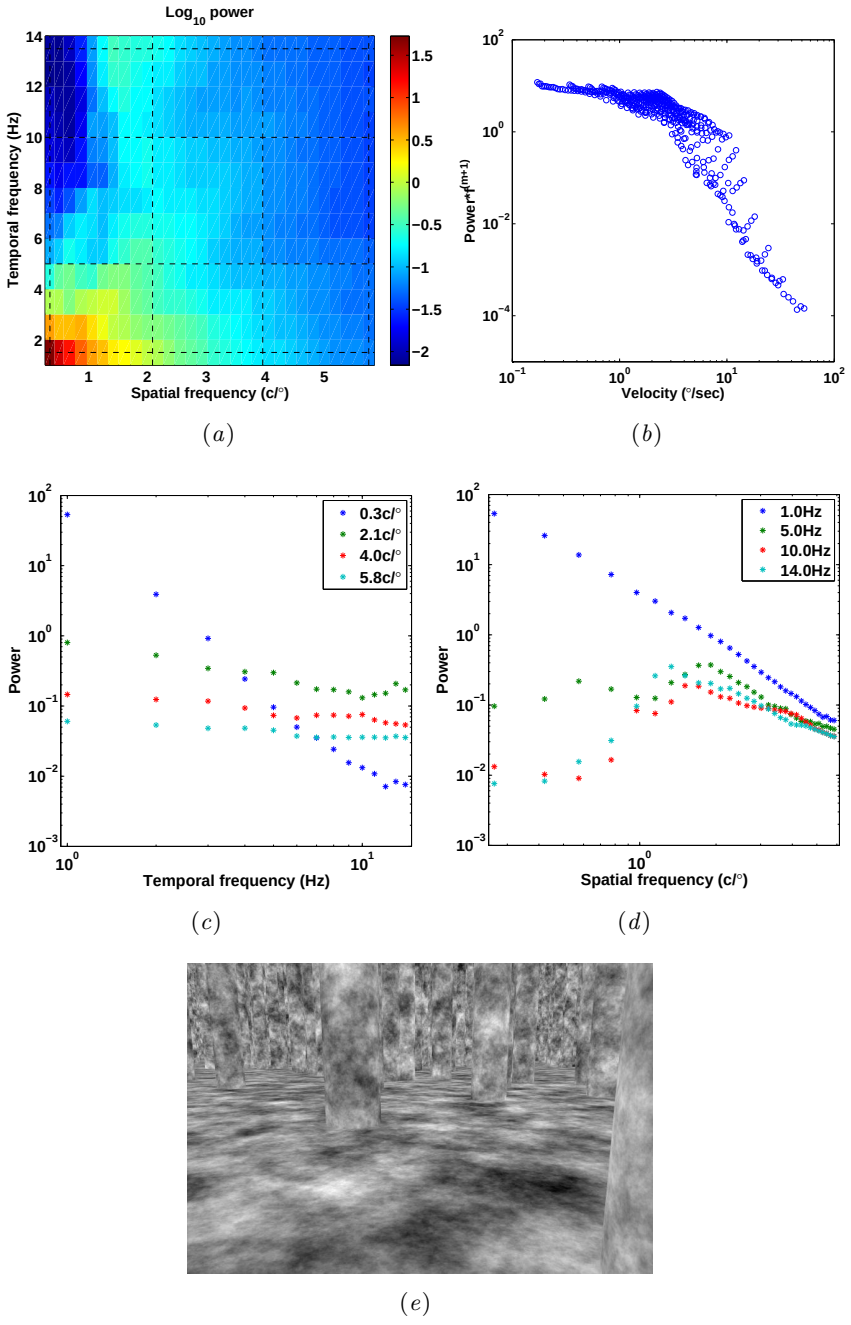


Figure C.14: **Power spectrum of low-density forest scene 4 at a 1.5 meters/second walking speed.**

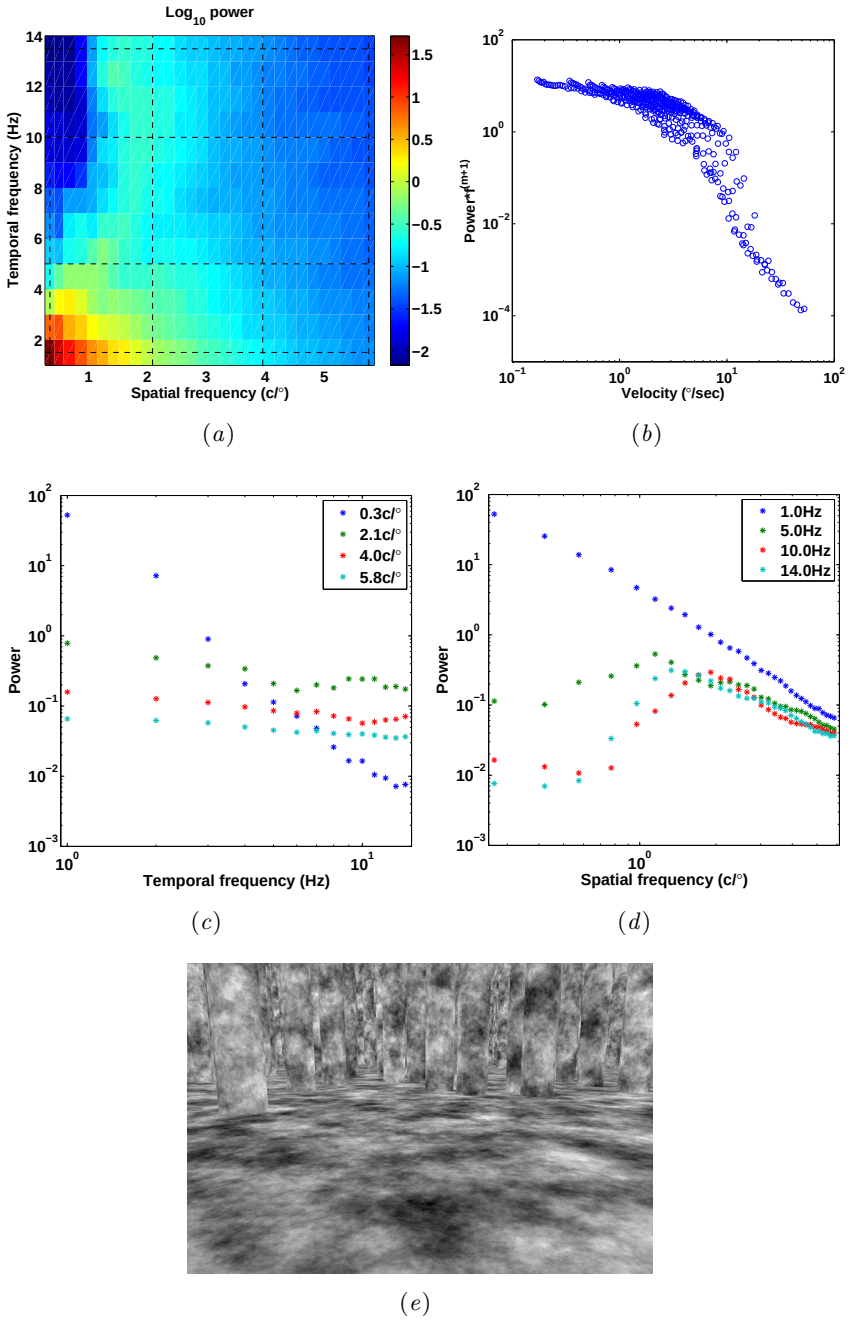


Figure C.15: **Power spectrum of low-density forest scene 5 at a 1.5 meters/second walking speed.**

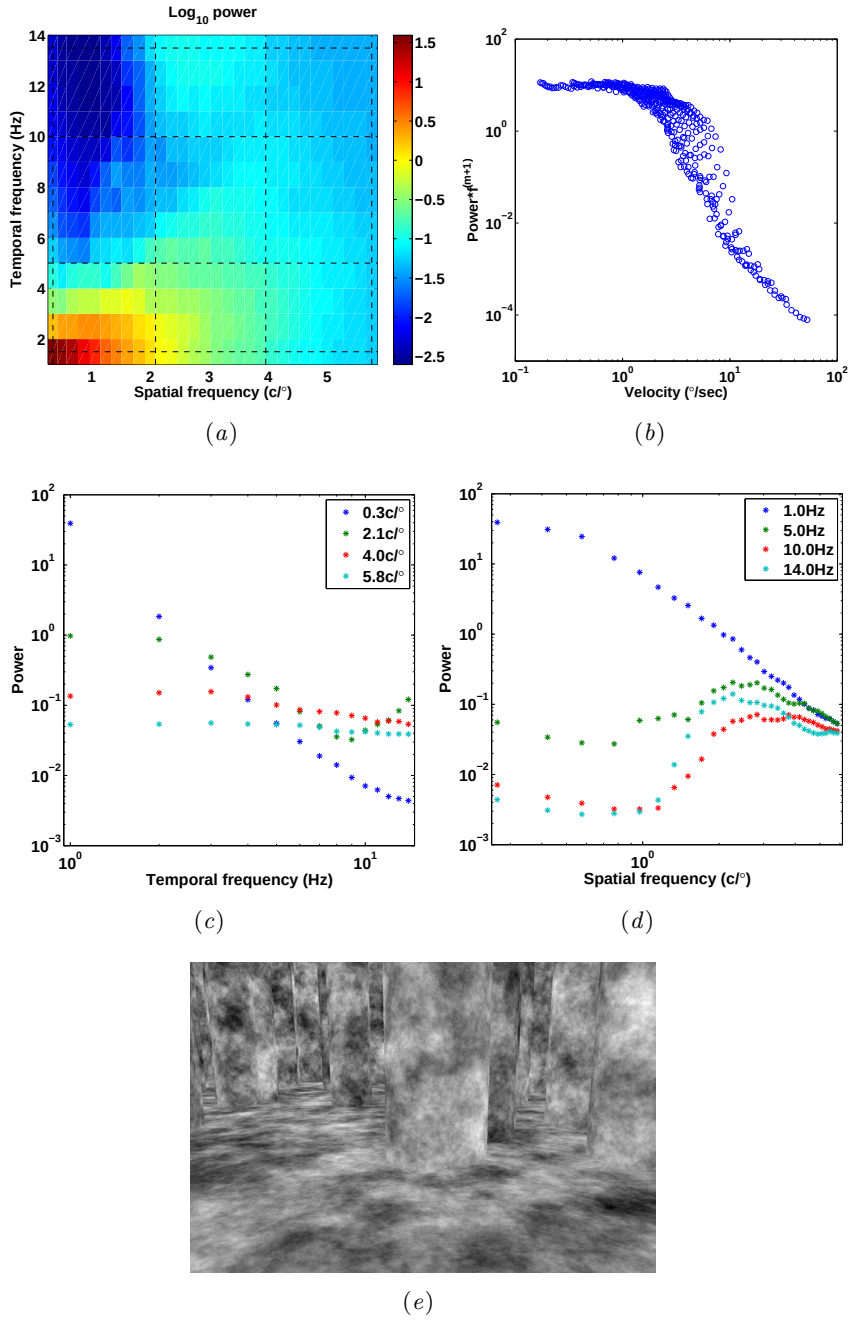


Figure C.16: **Power spectrum of high-density forest scene 1 at a 0.9 meters/second walking speed.**

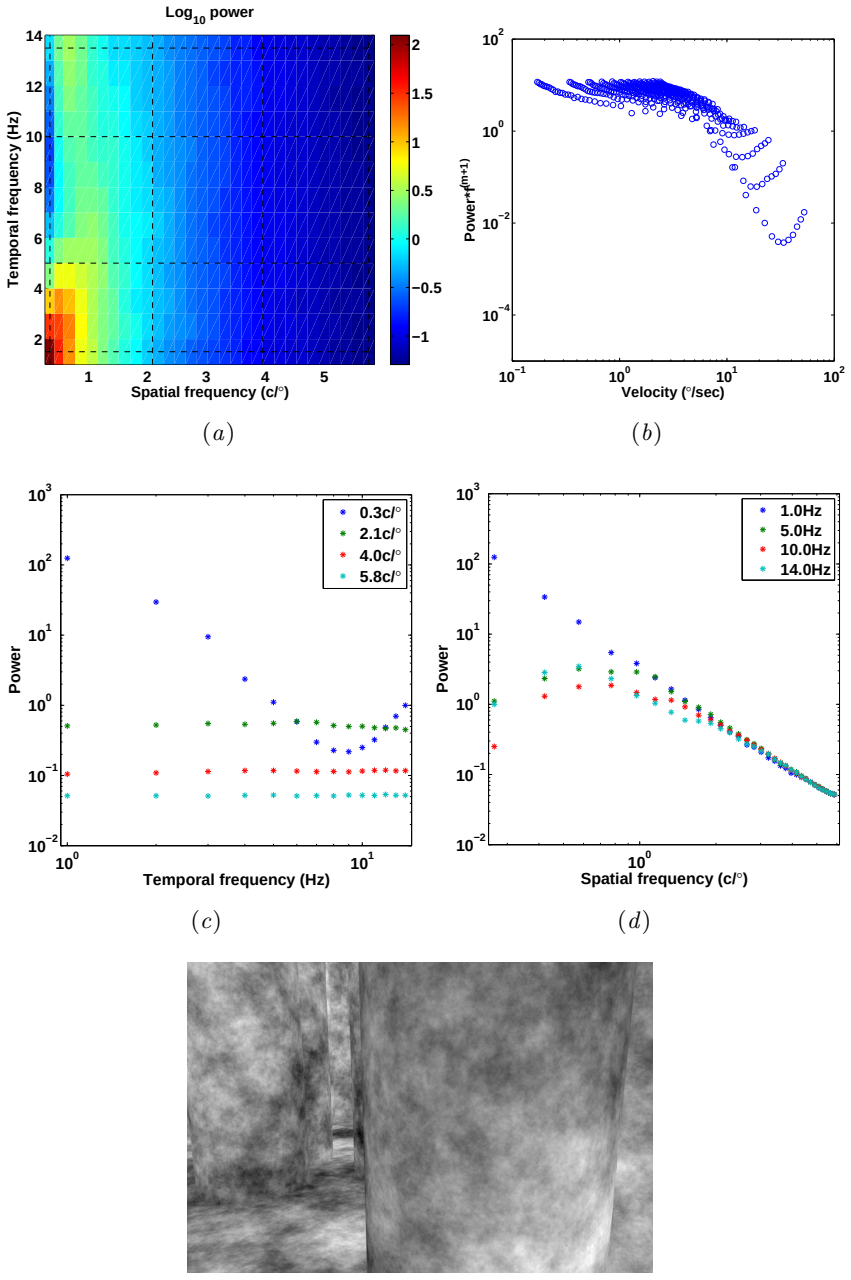


Figure C.17: **Power spectrum of high-density forest scene 2 at a 0.9 meters/second walking speed.**

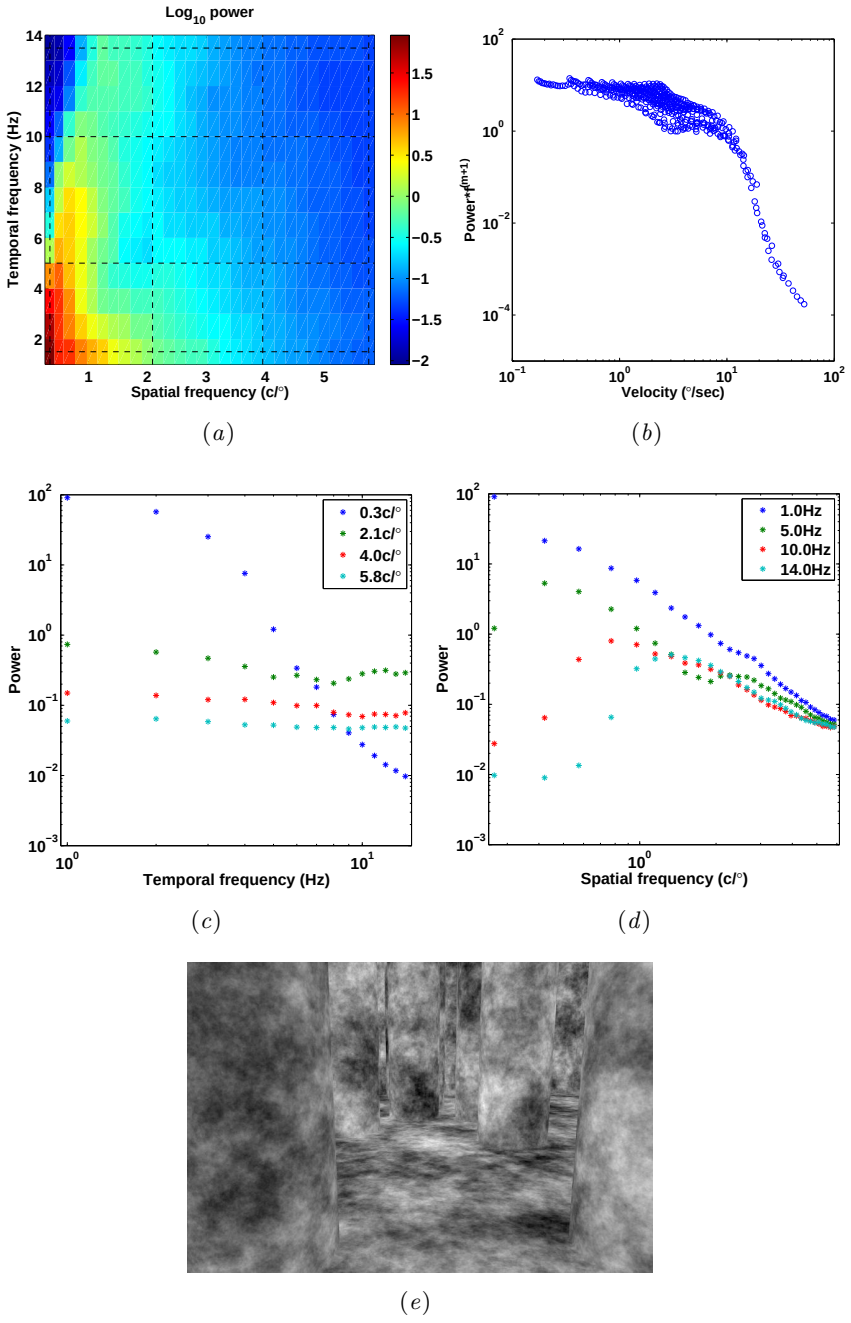


Figure C.18: **Power spectrum of high-density forest scene 3 at a 0.9 meters/second walking speed.**

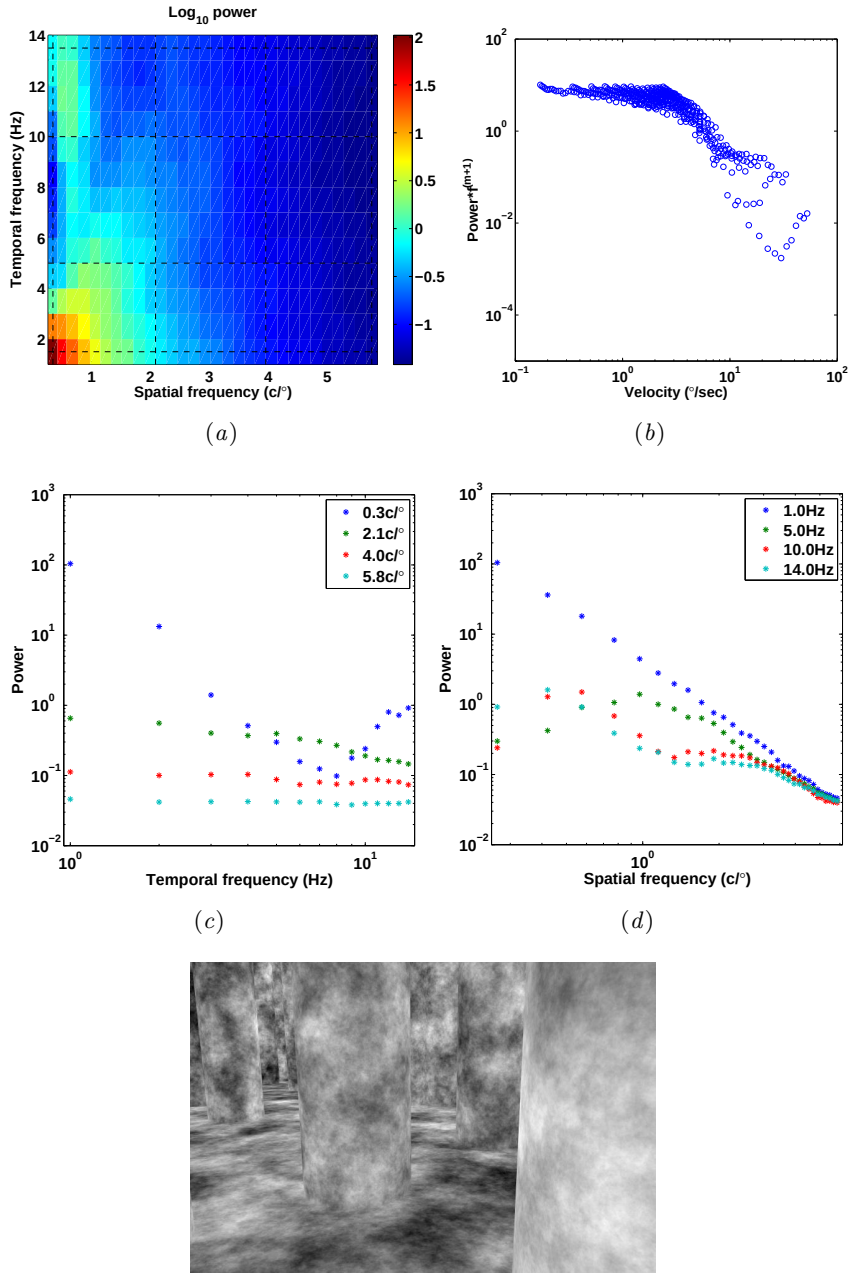


Figure C.19: **Power spectrum of high-density forest scene 4 at a 0.9 meters/second walking speed.**

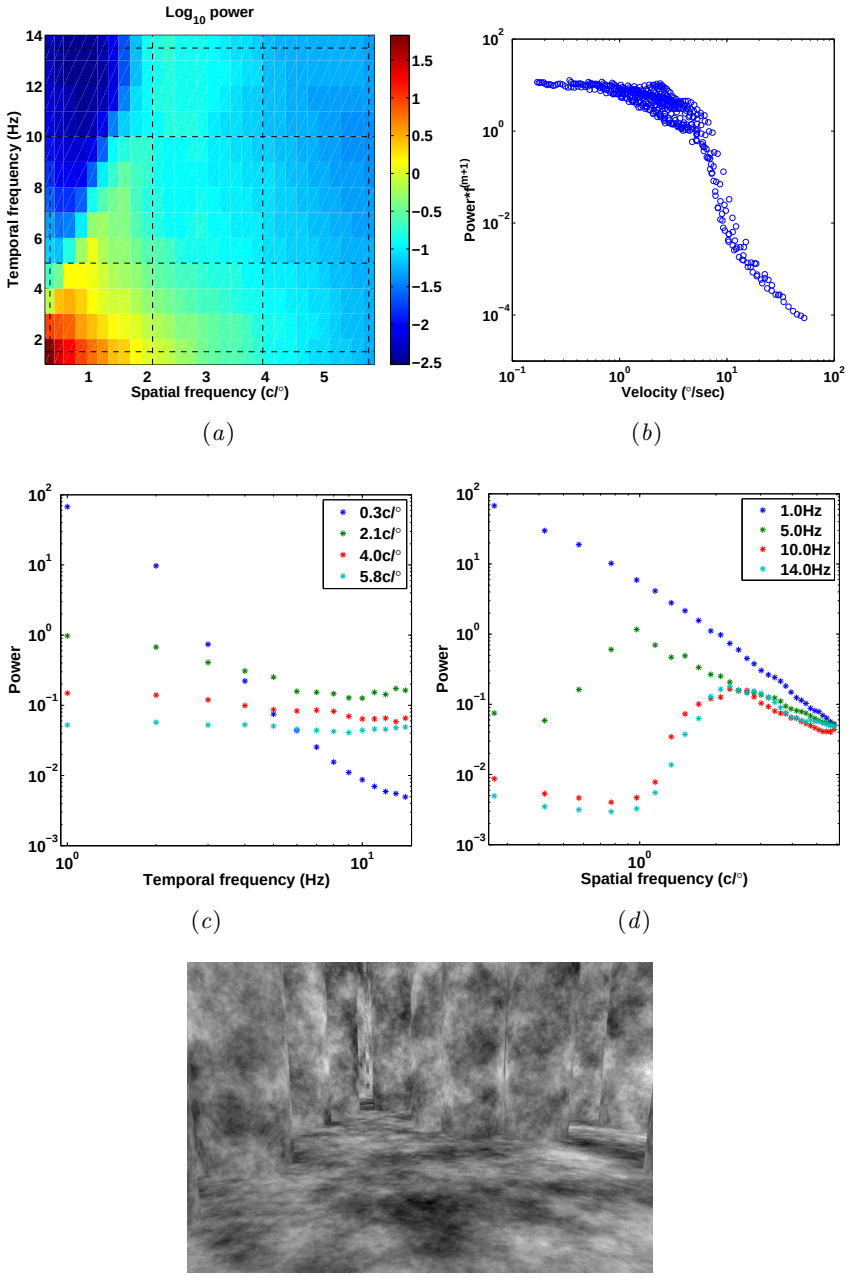


Figure C.20: **Power spectrum of high-density forest scene 5 at a 0.9 meters/second walking speed.**

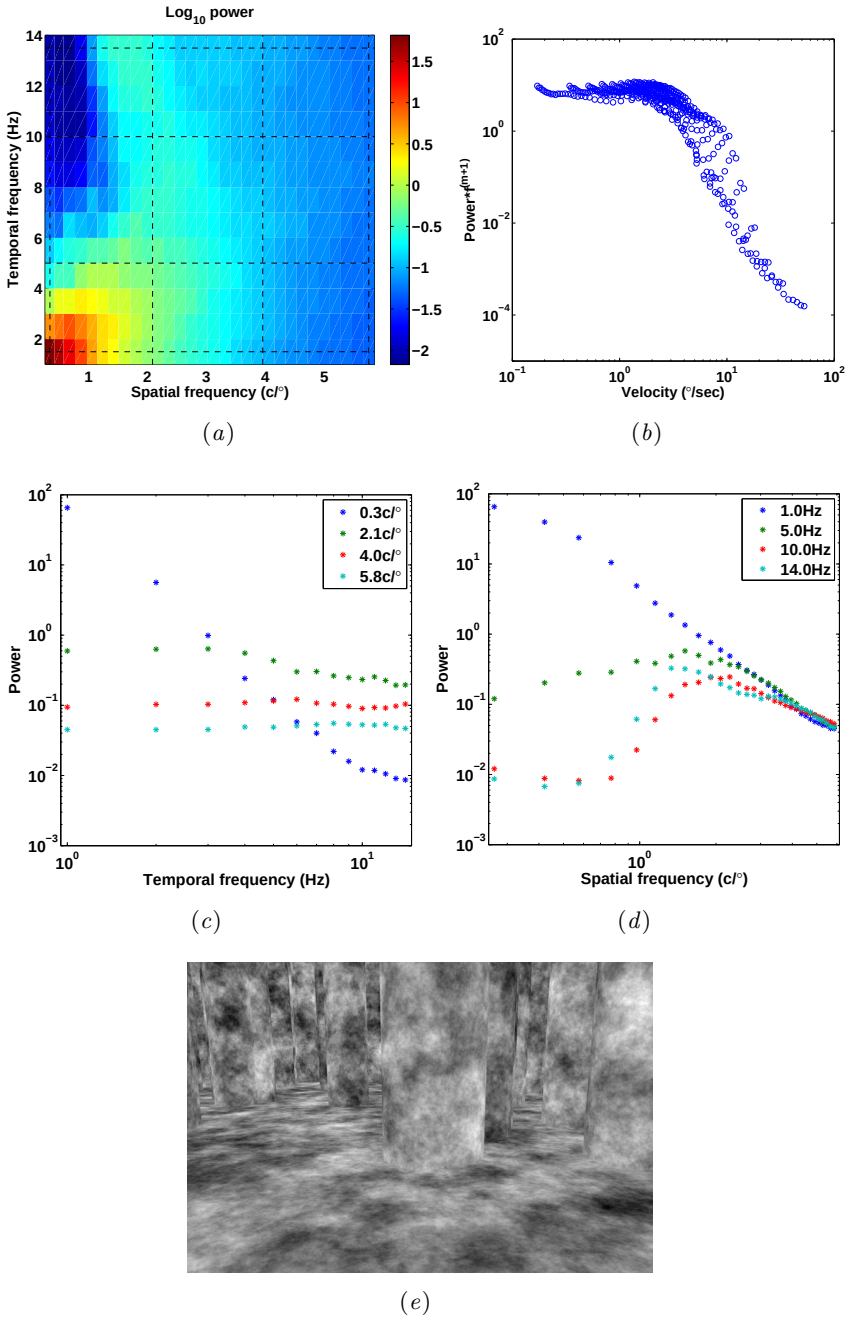


Figure C.21: **Power spectrum of high-density forest scene 1 at a 1.5 meters/second walking speed.**

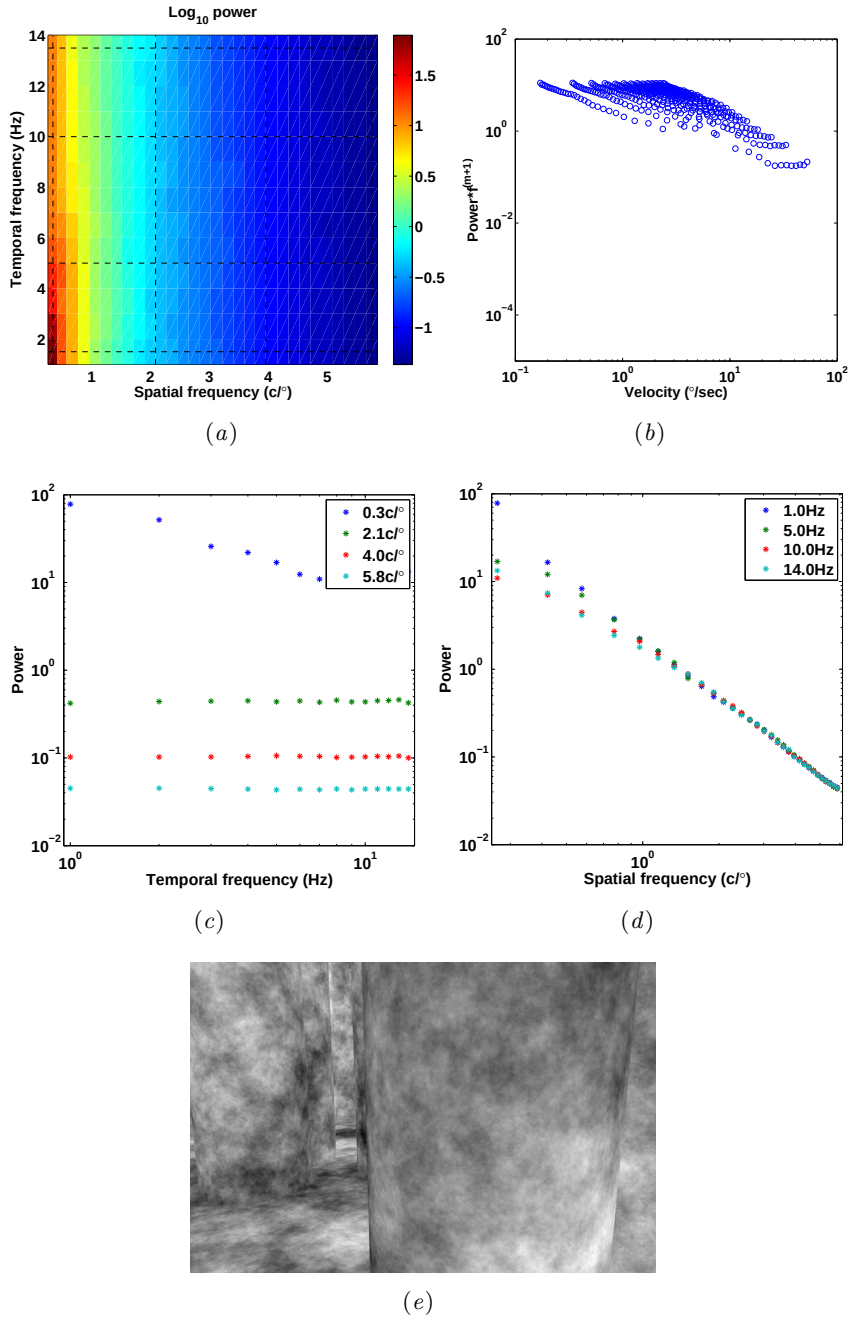


Figure C.22: **Power spectrum of high-density forest scene 2 at a 1.5 meters/second walking speed.**

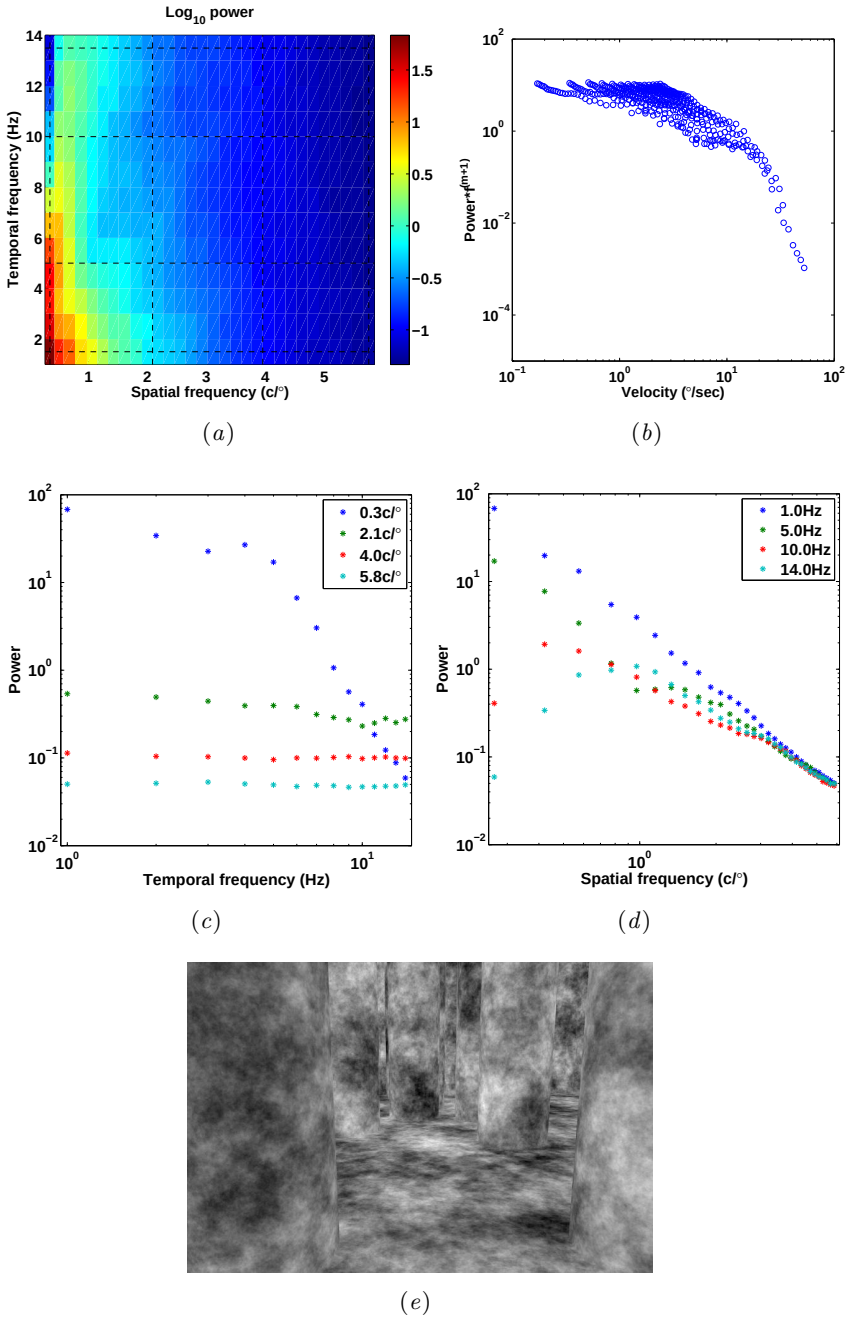


Figure C.23: **Power spectrum of high-density forest scene 3 at a 1.5 meters/second walking speed.**

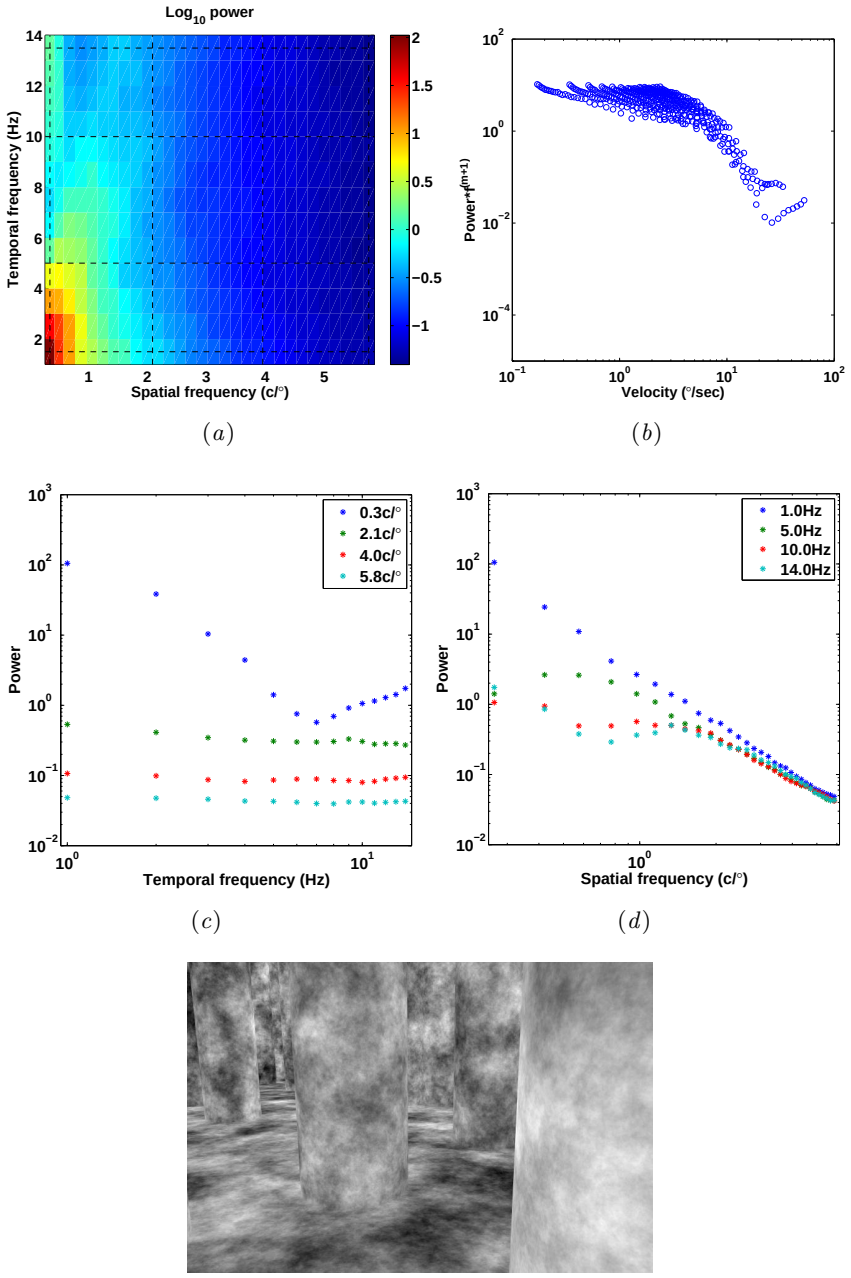


Figure C.24: **Power spectrum of high-density forest scene 4 at a 1.5 meters/second walking speed.**

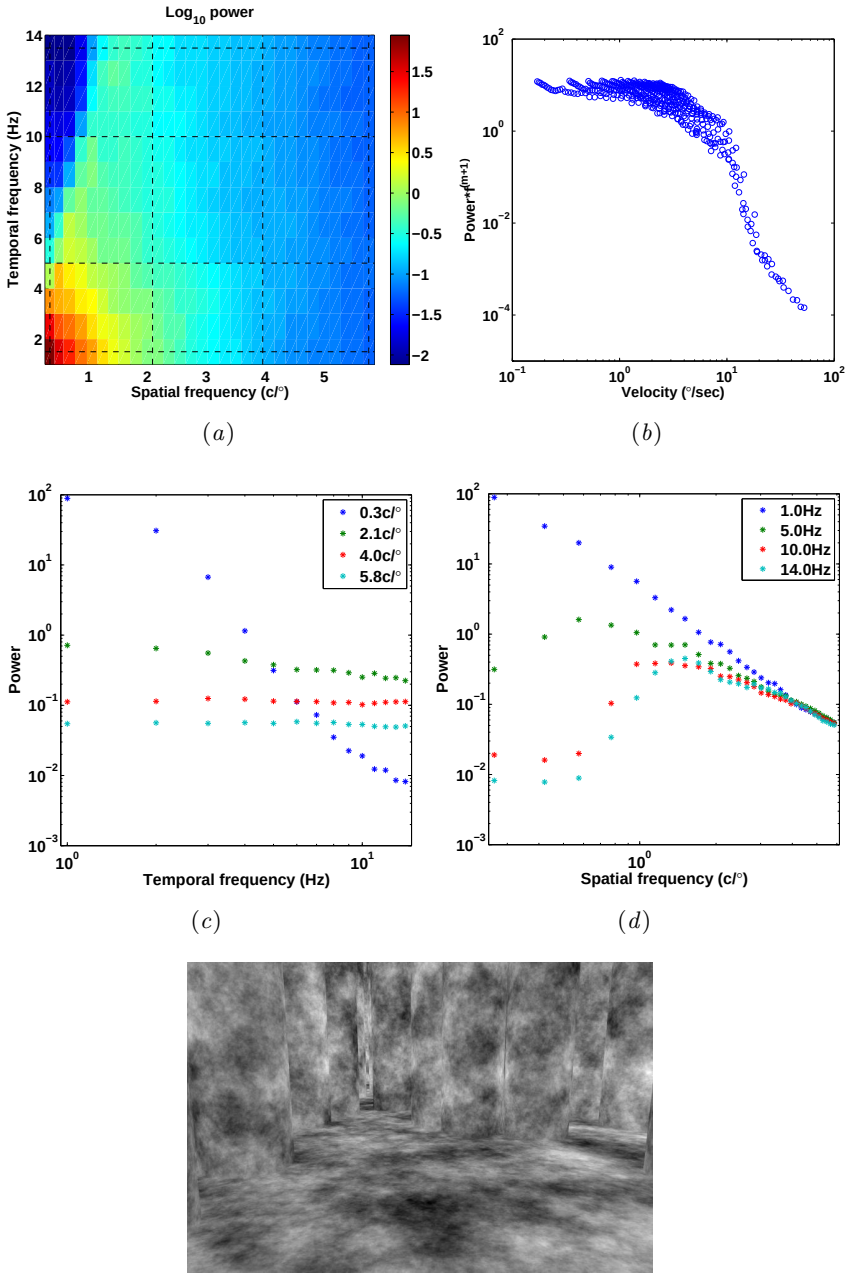


Figure C.25: **Power spectrum of high-density forest scene 5 at a 1.5 meters/second walking speed.**

Bibliography

Anderson, S. J., and Burr, D. (1987). Receptive field size of human motion detection units. *Vision Research*, 27(4):621–35. 19, 60, 62, 63

Anderson, S. J., and Burr, D. (1989). Receptive field properties of human motion detector units inferred from spatial frequency masking. *Vision Research*, 29(10):1343–58. 19, 60, 62

Anderson, S. J., and Burr, D. (1991). Spatial summation properties of directionally selective mechanisms in human vision. *Journal of the Optical Society of America*, 8(8):1330–9. 19, 60, 63

Anderson, S. J., Burr, D., and Morrone, C. M. (1991). Two-dimensional spatial and spatial-frequency selectivity of motion-sensitive mechanisms in human vision. *Journal of the Optical Society of America*, 8(8):1340–51. 19, 60, 62

Attneave, F. (1954). Some informational aspects of visual perception.. *Psychol Rev*. 4

Ballard, D. H., and Brown, C. M. (1982). *Computer Vision*. Prentice Hall Professional Technical Reference. 47

Barlow, H. B. (1961). Possible principles underlying the transformations of sensory messages. In Rosenblith, W. A., editor, *Sensory Communication*, 217–234. Cambridge, MA: MIT Press. 4

Boeddeker, N., Lindemann, J. P., Egelhaaf, M., and Zeil, J. (2005). Responses of blowfly motion-sensitive neurons to reconstructed optic flow along outdoor flight paths. *J Comp Physiol A Neuroethol Sens Neural Behav Physiol*, 191(12):1143–55. 4, 5, 12, 13

Burr, D., McKee, S., and Morrone, C. M. (2006). Resolution for spatial segregation and spatial localization by motion signals. *Vision Research*, 46(6-7):932–9. 19, 20, 58, 60, 62, 71

Calow, D., Kruger, N., Worgotter, F., and Lappe, M. (2005). Biologically motivated space-variant filtering for robust optic flow processing. *Network (Bristol, England)*, 16(4):323–40. 13, 58, 82

Calow, D., and Lappe, M. (2007). Local statistics of retinal optic flow for self-motion through natural sceneries. *Network (Bristol, England)*, 1–32. 13, 14, 15

Dong, D. W., and Atick, J. J. (1995a). Statistics of natural time-varying images. *Network: Computation in Neural Systems*, 6(3):345–358. 2, 4, 7, 10, 22, 26, 42, 77, 84, 87

Dong, D. W., and Atick, J. J. (1995b). Temporal decorrelation: a theory of lagged and nonlagged responses in the lateral geniculate nucleus. *Network: Computation in Neural Systems*, 6(2):159–178. 2, 4, 11

Dror, R. O., O'Carroll, D. C., and Laughlin, S. (2000). The role of natural image statistics in biological motion estimation. *Lect Notes Comput Sc*, 1811:492–501. 2

Efron, B., and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. New York: Chapman & Hall. 48, 68

Field, D. J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A, Optics and image science*, 4(12):2379–94. 17, 25, 26, 42

Fredericksen, R. E., Verstraten, F. A., and van de Grind, W. A. (1994). Spatial summation and its interaction with the temporal integration mechanism in human motion perception. *Vision Research*, 34(23):3171–88. 19, 60

Fredericksen, R. E., Verstraten, F. A., and van de Grind, W. A. (1997). Pitfalls in estimating motion detector receptive field geometry. *Vision Research*, 37(1):99–119. 19, 60

Geisler, W. S. (2008). Visual perception and the statistical properties of natural scenes. *Annual Review of Psychology*, 59(1). 4, 9

Georgeson, M. A., and Scott-Samuel, N. E. (2000). Spatial resolution and receptive field height of motion sensors in human vision. *Vision Research*, 40(7):745–58. 19, 60

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*. Boston: Mifflin. 9

Huang, J., Lee, A. B., and Mumford, D. (2000). Statistics of range images. In *IEEE Conference on Computer Vision and Pattern Recognition*, 324–331. 4, 17, 25, 26, 45

Langer, M., and Mann, R. (2003). Optical snow. *International Journal of Computer Vision*. 16

Maunsell, J. H., and Essen, D. C. V. (1983). Functional properties of neurons in middle temporal visual area of the macaque monkey. i. selectivity for stimulus direction, speed, and orientation. *J Neurophysiol*, 49(5):1127–47. 58, 81

McCane, B., Novins, K., Crannitch, D., and Galvin, B. (2001). On benchmarking optical flow. *Comput. Vis. Image Underst.*, 84(1):126–143. 16

MegaPOV-Team (2005). Megapov 1.2.1. `http://megapov.inetart.net/index.html`. 25, 45

Olshausen, B. (2001). Sparse codes and spikes. 12

Perlin, K. (2002). Improving noise. In *SIGGRAPH '02: Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, 681–682. New York, NY, USA: ACM. 26, 46

Persistence of Vision Pty. Ltd. (2004). Persistence of vision^{TM}raytracer (version 3.6). `http://www.povray.org/`. 25, 45

Potetz, B., and Lee, T. S. (2003). Statistical correlations between two-dimensional images and three-dimensional structures in natural scenes. *Journal of the Optical Society of America*, 20(7):1292–303. 4, 5

Roth, S., and Black, M. J. (2007). On the spatial statistics of optical flow. *Int J Comput Vision*, 74:33–50. 13, 15

Simoncelli, E. P., and Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annu Rev Neurosci*, 24:1193–216. 9

Spillmann, L., Ransom-Hogg, A., and Oehler, R. (1987). A comparison of perceptive and receptive fields in man and monkey. *Human neurobiology*, 6(1):51–62. 19, 60, 62

Tadin, D., and Lappin, J. S. (2005). Optimal size for perceiving motion decreases with contrast. *Vision Research*, 45(16):2059–64. 19, 60

Tadin, D., Lappin, J. S., Gilroy, L. A., and Blake, R. (2003). Perceptual consequences of centre-surround antagonism in visual motion processing. *Nature*, 424(6946):312–5. 19, 60

van de Grind, W. A., Koenderink, J. J., and van Doorn, A. J. (1986). The distribution of human motion detector properties in the monocular visual field. *Vision Research*, 26(5):797–810. 19, 20, 58, 60, 62, 71

van de Grind, W. A., van Doorn, A. J., and Koenderink, J. J. (1983). Detection of coherent movement in peripherally viewed random-dot patterns. *Journal of the Optical Society of America*, 73(12):1674–83. 19, 58, 60, 62, 71

van Doorn, A. J., and Koenderink, J. J. (1984). Spatiotemporal integration in the detection of coherent motion. *Vision Research*, 24(1):47–53. 19, 20, 58, 60, 62, 71

van Hateren, H. (2005). Natural stimuli collection. World Wide Web electronic publication. 77

van Hateren, J. H. (1993). Spatiotemporal contrast sensitivity of early vision.. *Vision Res.* 11

van Hateren, J. H., and Ruderman, D. L. (1998). Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex. *Proc Biol Sci*, 265(1412):2315–20. 2, 11

Watson, A. B., Barlow, H. B., and Robson, J. G. (1983). What does the eye see best? *Nature*, 302(5907):419–22. 19, 60, 63

Watson, A. B., and Turano, K. (1995). The optimal motion stimulus. *Vision Research*, 35(3):325–36. 19, 60, 63

Whalen, A. D. (1971). *Detection of Signals in Noise*. New York, NY, USA: Academic Press, Inc. 88

Yang, Z., and Purves, D. (2003). A statistical explanation of visual space. *Nature Neuroscience*. 4

Vita

Tal Tversky was born to Amos and Barbara Tversky in Palo Alto, California on August 11, 1971. He lived in Jerusalem, Israel from 1972 to 1976, after which he returned to Palo Alto. In 1989, he moved to New Haven, Connecticut where he studied physics at Yale University, eventually graduating with a B.S. in 1993. After graduating, he biked home from school; from New Haven, Connecticut to Palo Alto, California. There, he started working as a software engineer at a small start-up company called Molecular Applications Group. In 1997 he left Palo Alto for Austin, Texas to study Computer Science at the University of Texas. He is currently living in Austin with his wife, Jenna Martin, and two daughters, Adele and Flora, and working at Apple Inc. as a data mining scientist.

Permanent Address: Tal Tversky
1004B Charlotte St.
Austin, TX
78703

This dissertation was typeset with LaTeX 2_ε[1] by the author.

[1] LaTeX 2_ε is an extension of LaTeX. LaTeX is a collection of macros for TeX. TeX is a trademark of the American Mathematical Society. The macros used in formatting this dissertation were written by Dinesh Das, and extended by Bert Kay, James A. Bednar, and Ayman El-Khashab.