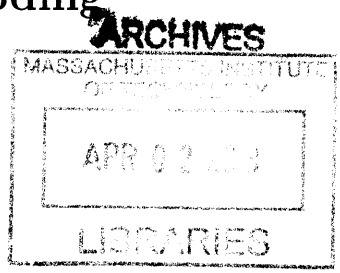
Ambiguous Statistics – How a Statistical Encoding in the Periphery Affects Perception



by

Alvin Andrew Raj

B.S., University of Washington (2006)
S.M., Massachusetts Institute of Technology (2008)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2013

Author .....
Department of Electrical Engineering and Computer Science
January 15, 2013

Certified by .....
Ruth Rosenholtz
Principal Research Scientist
Thesis Supervisor

Accepted by .....
Leslie A. Kolodziejski
Chair, Department Committee on Graduate Students

Ambiguous Statistics – How a Statistical Encoding in the Periphery Affects Perception

by

Alvin Andrew Raj

Submitted to the Department of Electrical Engineering and Computer Science
on January 15, 2013, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Recent understanding in human vision suggests that the periphery compresses visual information to a set of summary statistics. Some visual information is robust to this lossy compression, but others, like spatial location and phase are not perfectly represented, leading to ambiguous interpretations. Using the statistical encoding, we can visualize the information available in the periphery to gain intuitions about human performance in visual tasks, which have implications for user interface design, or more generally, whether the periphery encodes sufficient information to perform a task without additional eye movements.

The periphery is most of the visual field. If it undergoes these losses of information, then our perception and ability to perform tasks efficiently are affected. We show that the statistical encoding explains human performance in classic visual search experiments. Based on the statistical understanding, we also propose a quantitative model that can estimate the average number of fixations humans would need to find a target in a search display.

Further, we show that the ambiguities in the peripheral representation predict many aspects of some illusions. In particular, the model correctly predicts how polarity and width affects the Pinna-Gregory illusion. Visualizing the statistical representation of the illusion shows that many qualitative aspects of the illusion are captured by the statistical ambiguities.

We also investigate a phenomena known as Object Substitution Masking (OSM), where the identity of an object is impaired when a sparse, non-overlapping, and temporally trailing mask surrounds that object. We find that different types of grouping of object and mask produce different levels of impairment. This contradicts a theory about OSM which predicts that grouping should always increase masking strength. We speculate some reasons for why the statistical model of the periphery may explain OSM.

Thesis Supervisor: Ruth Rosenholtz
Title: Principal Research Scientist

To my late father, Richard.

Acknowledgments

I am immensely grateful to my doctoral thesis advisor, Ruth Rosenholtz, for being a terrific advisor – kind, patient, understanding, insightful, and posessor of a sense of humour that found my oddness amusing. She has taught me so much about science and life.

To Ted Adelson, my master's thesis co-advisor, I am also very thankful. His insight on vision, science, and how to think visually has been hugely influential, and his mentorship has been invaluable.

Many thanks to the rest of my thesis committee: Antonio Torralba, and Bill Freeman, for great insight into how to make this thesis at least an order of magnitude better with insightful comments, suggestions, and critiques.

In writing this thesis, my wife, Patrice Metcalf-Putnam, has been patiently supportive, and has motivated me with daily rewards of a puppy picture for each completed page. Thanks also to my mother, Julie, for being supportive and understanding. And to my brother, Allan, for teaching me about many odd and interesting things.

To Dieter Fox, Jeff Bilmes, Marina Meila and Amar Subramanya, thank you for getting me interested in research and for guiding me in my undergraduate studies. To Gregory Ferencko, thank you for helping me get acclimated to academics in the United States.

During my academic journey at MIT, I am privileged to have met and worked with brilliant collaborators: Kimo Johnson, Benjamin Balas, Jie Huang, Forrester Cole, and Livia Ilie. I also thank all my wonderful labmates for great discussions, lunch meetings, and the obligatory geeky escalation of ideas: Phillip Isola, Ce Liu, Yuanzhen Li, Lavanya Sharan, Nat Twarog, Roland Van den Berg, Krista Ehinger, Derya Akkaynak Yellin, Bei Xiao, Rui Li, Xiaodan Jia, Xuetao Zhang, Shuang Song. Many thanks to John Canfield for being a terrific lab administrator.

To my Harding roommates: Michael Bernstein, Oshani Seneviratni, Grace Woo, and Jenny Yuen, thanks for an unforgettable time, for making our cheap rental truly

amazing, and of course, bura.

My deepest gratitude to all.

Contents

List of Figures

Chapter 1

Introduction

Seeing is typically an active process. When watching a movie, reading, walking, or searching for keys, people move their eyes around to perform tasks. But sometimes, it isn't necessary to make many or any eye movements at all. Why do people need to move their eyes for some tasks, but not others? This thesis suggests the answer is that the periphery encodes enough summary information about the visual input so one does not need to make eye movements for some tasks, but the same information is insufficient for other tasks. Further, in this thesis, the type of summary information is hypothesized to be that of statistics computed on the visual field in overlapping regions that increase in size as they get further from the center of fixation. We show evidence supporting this hypothesis in visual search and in visual illusions.

1.1 Outline

In Chapter 2, the role of peripheral vision is revisited with recent understanding from human vision. This includes psychophysical studies on visual crowding, where researchers find that humans do poorly in identifying an object in the periphery when there are other objects flanking it. We examine a statistical model of peripheral vision that accounts for those results in crowding. Chapter 3 examines the statistical model discussed in Chapter 2. The representational capability of the model is investigated, and in addition, we propose an algorithm to visualize the information contained in

that statistical model.

In Chapter 4, we discuss visual search, a task where subjects are asked to find a target in a search display (i.e., find a tilted line among vertical lines). The key insight contributed in this thesis is that most of the search display is peripheral. Thus, peripheral vision is required to simply perform the task in pop-out search conditions (detecting a target at a glance), or is needed to guide eye movements to the target location. Given the prominent role peripheral vision plays in visual search, we examine whether the loss of information in the peripheral representation can explain why certain types of searches are easy or difficult. When people are asked to search for a target among some distractors, their reaction time depends on the type of search task involved. For example, the search for a tilted line among vertical is fast, but the search for a T among L is not. Puzzlingly, there are search asymmetries: Q among O is fast, but O among Q is slow. The discriminability of the target item to a distractor item has not sufficed as an explanation, because each target item is easily discriminable from any given distractor item. We show that this is the case, and that a measure of statistical discriminability of patches from search displays correlates well with search performance.

Beyond correlations, we may make quantifiable predictions of search performance. In Chapter 5, a quantitative model based on the experimental results in Chapter 4 is proposed, and its performance evaluated. The model is a variant of an ideal saccadic targeter that saccades to the most likely target location. Various considerations common to modeling visual search and human vision are incorporated into the model: memory, saccade length preferences, pooling region density, and whether a heuristic instead of an ideal model is used to infer the most likely target location. The modeling results correspond well to human performance.

Visual illusions are often studied because we can gain insight into the visual system, by investigating instances where it seems to be broken. If the peripheral visual system loses a large amount of visual information, as is suggested in this thesis, there are likely many instances of visual stimuli where the summary information provided by the periphery is misinterpreted. In Chapter 6, we investigate one particular class

of illusory stimuli – the Pinna-Gregory illusions. We show that a statistical view of the periphery predicts how various modifications of the illusion will affect the illusory strength, and that the visualization of the statistical information from the images exhibits many aspects of the illusions.

Thus far, we have discussed the periphery in the context of static stimuli and illusions. Another strange phenomena where performance in object identification is more severely affected in the periphery is that of object substitution masking (OSM). OSM describes a form of masking where the presence of a sparse, non-overlapping, and temporally trailing mask impairs the perception of an object when attention is distributed over a large region. The masking strength appears strongest in the periphery, though it is possible to elicit masking in the fovea as well. In Chapter 7, we show that different types of groupings affect masking strength differently, contrary to the prediction by Moore and Lleras [32] that stronger grouping should lead to stronger masking. This suggests that a lower-level explanation other than "object files" may underlie some of the results in OSM. We also suggest a line of future work, expanding the static model to a spatio-temporal model of peripheral vision as a potential explanation for this phenomena. Further modeling work and psychophysical experiments are required to test whether such a model could explain OSM.

In Chapter 8, we discuss some applications of visualizing the information available to peripheral vision to help design better user interfaces, and to understand why mazes are difficult or easy to solve. Some additional areas of future work are discussed, and we conclude with a summary of the contributions made in this thesis.

Chapter 2

Statistical Encoding in the Periphery

2.1 Outline

In this chapter, we discuss why it is important to understand the peripheral visual system, discuss what is known about the periphery, then consider a model of peripheral vision that explains the puzzling data about visual crowding in the periphery. The work presented in this chapter introduces research done in the Perceptual Science Group at MIT by Benjamin Balas, Ruth Rosenholtz, and Lisa Nakano [3], and is the background needed to understand the extension of that work that is presented in this thesis.

2.2 Why care about peripheral vision?

Why should anyone care about peripheral vision? In Figure 2-1 a circle in the center of the image is blacked out, roughly occupying the area that the fovea would occupy (about 2^o visual angle, assuming you were $25in$ from the image). The area occupied by the fovea is tiny compared to the periphery. The importance of this observation can be easy to overlook. Almost all of one's visual field is peripheral, and so understanding how visual information is encoded in the periphery is necessarily important

Figure 2-1: The fovea (blackened) occupies a very small area compared to the periphery, which constitutes everything else in the scene.

for understanding the human visual system.

There are many tasks which humans only need a single glance to perform, such as pop-out search [53], material perception [47], scene recognition [42], and animal vs no animal categorization [49]. In a single fixation, the visual area processed by the fovea is dwarfed by the periphery. Unless one is lucky enough to have fixated on a distinctive image feature that happened to be sufficiently informative for the task, the fovea would not contribute much useful information.

When looking at any particular scene, humans typically scan the visual environment by fixating on one location, then saccading to the next, and so on. Information is gathered and processed during each fixation, but not during saccades [30]. This points to a fairly discrete algorithm that the visual system uses to make sense of the visual environment.

The need to make saccades even in reading this manuscript implies that the underlying encoding of the visual field is not uniform. If it were uniformly encoded, saccades would not serve a purpose because no new information would be gained by

fixating at a new location.

Not only does the periphery occupy most of the visual field, but as a consequence, the periphery may additionally encode global structures that span a large area. These global features have proven to be important for scene gist recognition [34].

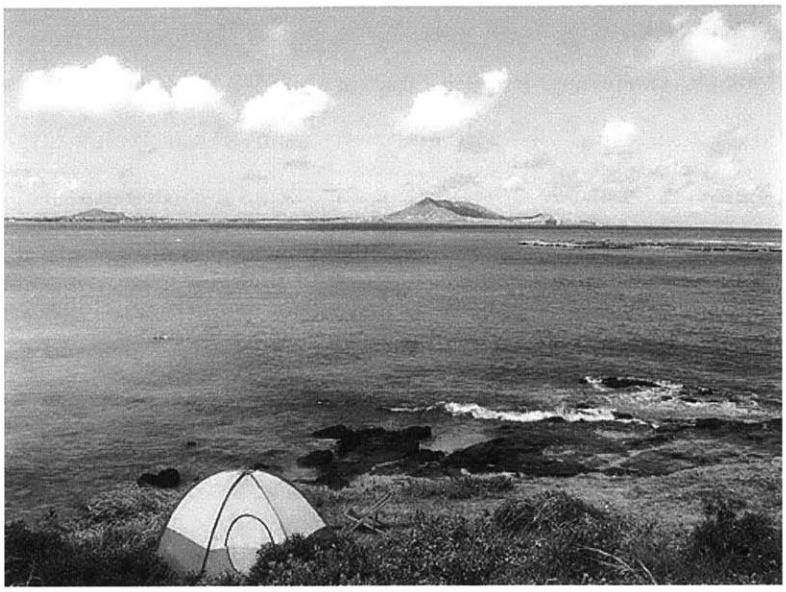
Larson and Loschsky [25] find that the peripheral vision system is more important for scene recognition than the fovea. In an gaze-contingent display, they either imposed an artificial scotoma to their subjects to simulate the loss of the fovea, or blocked off peripheral vision. Under one of those two conditions, subjects were asked to categorize a number of scenes. Scene recognition performance of subjects with a 10^o diameter scotoma in a 27^oX27^o display were only slightly worse than a control condition where the entire scene was visible. In contrast, subjects with the periphery blocked off required a 20^o diameter circle "fovea" visible to achieve similar performance.

2.3 Isn't it just about acuity?

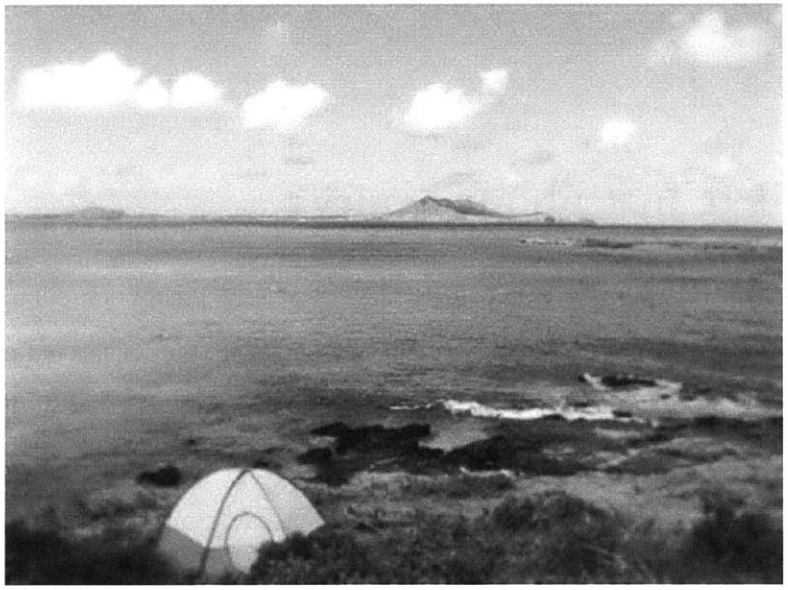
Even after accounting for the visual area the fovea and periphery occupies, Larson and Loschsky find that performance was uneven for foveal vs peripheral processing of information. Clearly, the encoding of visual information in the fovea is different than in the periphery.

This is unsurprising given our everyday experience of simply not being able to tell exactly what's "out there" in our peripheral vision. The encoding of visual information in the periphery has been studied from a large number of perspectives and has shaped much of the classic understanding of peripheral vision – that acuity and color discriminability is worse as distance from the fovea increases.

Many studies on the retina have show that the density of photoreceptors decrease as eccentricity increases [43]. Having this understanding of peripheral vision, one could then simulate the expected information loss from peripheral vision by blurring an image with an appropriately sized filter at each location, depending on its distance to the simulated fovea [1].

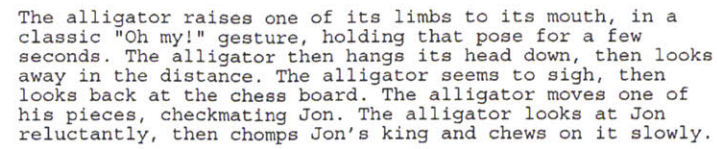


(a) Original Image

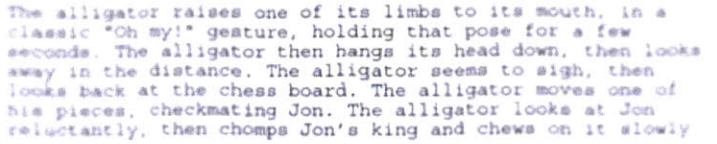


(b) Simulated Acuity Loss

Figure 2-2: In the simulated acuity loss (b) of the original image in (a), the fixation was placed in the middle of the image. The simulation assumes loss of acuity as measured in [56] and that the image's width occupies approximately 12^o v.a. horizontally (i.e., if you hold this image about 16 to 20 inches from your eyes). Notice that the simulation exhibits a lot of details in the scene, yet it is difficult to have been able to introspect all these details while fixating in the middle of the scene.



(a) Original Image



(b) Simulated Acuity Loss

Figure 2-3: In the simulated acuity loss (b) of the original image in (a), the fixation was placed in the middle of the image. The simulation assumes loss of acuity as measued in [56] and that the image's width occupies approximately 12^o v.a. horizontally. The text in the simulated acuity loss is still easily readable yet when trying to read the original text, one must make multiple fixations. This indicates that the loss of acuity is not sufficient to account for why one needs to make those fixations.

Figure 2-2 shows an example of simulated peripheral acuity loss of a beach scene. What one might notice is that not much information is lost at all. Looking at one's visual environment shows crisp details that are not blurry in the periphery. To make this point clearer, notice that in Figure 2-3, where peripheral acuity loss is applied to an image of some text, that essentially all the words are still legible, yet we still need to move our eyes to read it. This indicates that perhaps the encoding in the periphery loses more than just acuity alone.

2.4 Visual Crowding

If not acuity, what might underlie this lack of ability to identify things or read words in the periphery? There have been a large number of studies on a phenomenon known as visual crowding, where the ability to identify an object in the periphery is more difficult when it is flanked by distracting objects. See [27] for a literature review.

(a) Single Letter

(b) Three Letters

(c) Three Letters Spaced Apart

Figure 2-4: Crowding Demonstration

When fixating on the plus in Figure 2-4 a, the isolated letter is easy to identify, but in (b), when the same letter is flanked by two distracting letters , it is more difficult to determine its identity. Bouma found that when the letters are spaced beyond approximately 1/2 the central target's distance from fixation (eccentricity), the performance is improved once again, as demonstrated in Figure 2-4c [4]. Further, when subjects are asked about what they saw in these crowded letter displays, they reported not seeing the central letter at all, or that they saw letter-like shapes made up of mixtures of parts from several letters [26, 29].

This is puzzling behavior, in which the visual system seems to retain much of the details necessary to perceive letters or letter parts, but not encode the information necessary to keep track of the locations of those details. The ease of recognizing an isolated target indicates that crowding is not simply due to reduced visual acuity. Instead, the visual system seems to lose additional information about the stimulus. Some researchers attribute this effect to excessive feature integration [37], and propose

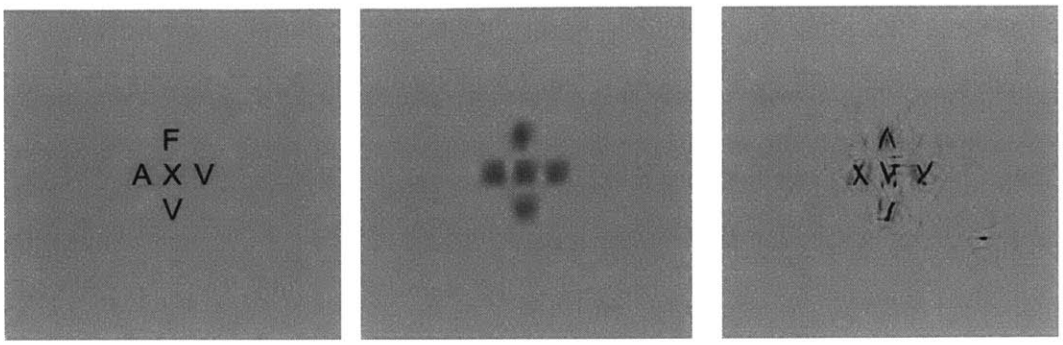
that the visual information is jumbled within pooling regions that grow linearly with eccentricity (radius .5 * eccentricity), and are elongated radially from the fovea [50].

Letters are not the only type of object that is subject to lower identification performance when surrounded by flankers in the periphery. There is evidence that color, orientation, hue, and size are all subject to crowding [55]. Sufficiently complicated objects may even crowd themselves [29]. Parsimony would prefer a simple explanation for this diverse set of stimuli that suffer crowding effects. It suggests a general mechanism that the peripheral visual system employs to process information, as opposed to a special mechanism that activates whenever the visual system detects more than a single object present and messes up the visual information.

2.5 Coarse Representation of Visual Information

Why might the visual system represent information in such a manner that allows these types of phenomena to occur? It is useful to consider how a vision system might be built with a constraint that there is a bandwidth limit for how much information may be processed through the pipeline at any given time. The human visual system seems to adopt an active vision approach to manage the information bottleneck.

To view a scene, humans actively fixate on areas that they want more detail on, while obtaining coarser information in the periphery, perhaps to guide the visual system in deciding where next to make an eye movement, and to give context for the details in the fovea. But what type of representation should the coarse encoding in the periphery use?



(a) Letters (b) Low Resolution (c) Statistical Summary

Figure 2-5: Comparison of a low resolution representation vs a feature statistics summary when limited to 1000 numbers.

Figure 2-5 shows what type of information would be contained in several types of compressed encodings limited to roughly 1000 numbers. In (b) the low-resolution encoding is able to preserve information about where the letters are in the display. In (c), a statistical representation based on a texture model proposed by Portilla and Simoncelli [41] makes location mistakes, but preserves information about letter fragments. In (c), it is difficult to know any particular letter's identity, but one can tell that the shapes in the image are letter-like stuff in the figure.

Balas, Nakano, and Rosenholtz [3] proposed that the visual system computes statistical summaries of visual information in local, overlapping pooling regions placed throughout the visual field. The model proposes that the visual system computes information equivalent to marginal and joint statistics of responses of V1-like oriented feature detectors across position, scale, and orientation based on the texture model proposed by Portilla and Simoncelli [41]. The details of the statistical encoding are discussed in detail in Chapter 3.

Balas et al measured subjects' ability to discriminate between the statistical summaries of stimuli in a 4-Alternative-Forced-Choice task. Please see the methodology in [3] for details. The performance in that task was compared against how well humans could discriminate between those types of stimuli in a crowding task. The types of stimuli tested included various letters in varying fonts, different types of objects, and some symbols. They found that the performance in the statistical discrimination

task predicted performance in the crowding task of those classes of stimuli. This suggests that the information loss due to this compression predicts what types of stimuli are subject to reduced identification performance in the periphery, and further, how much that performance will be affected.

2.6 Conclusion

It is important to study and understand peripheral vision because the periphery comprises most of the visual field, and behaves in a seemingly odd manner, as shown by experiments in visual crowding.

If we take the bottleneck in information processing into account, this odd behavior could be understood as a side effect of an information compression effort by the peripheral visual system. Balas et al [3] propose and show that the texture model in Portilla-Simoncelli [41] captures much of the nature of compression that visual information undergoes for a single pooling region in the periphery for a number of objects and letters.

This work in this thesis extends the work by Balas et al, applying the peripheral model they propose to explain a number of phenomena, and proposing an algorithm to visualize the information in all the pooling regions so the information contained in the periphery as a whole may be examined. In the next chapter, we investigate the statistical encoding used in the model.

Chapter 3

Visualizing Statistics

3.1 Outline

In this chapter, the details of the statistical model discussed in Chapter 2 are explored. Our choice of statistics is a working hypothesis of the information that the visual system extracts in the periphery. We examine some methods of modeling the visual appearance of textures by characterizing a texture as: a set of marginal statistics from a pyramid decomposition (Heeger-Bergen [16]), or joint statistics from a steerable complex pyramid decomposition (Portilla-Simoncelli [41]).

The Portilla-Simoncelli texture model represents many types of real and artificial textures well, but does poorly at representing spatial location, phase, and makes some mistakes about complex shapes and images in general. We argue that these areas of poor representation are shared by the peripheral visual system, and so the statistics used in the Portilla-Simoncelli representation are a good candidate for a working hypothesis of the information that the visual system extracts in the periphery. In addition, we argue that it would be difficult to adapt non-parametric models of visual appearance to model how the periphery represents information, because it would be difficult for those models to make these kinds of mistakes. In particular, the non-parametric approaches have difficulty in hallucinating visual "stuff" not present in the original image, for example, a white vertical line, when the original image contained only black verticals and white horizontals.

Balas et al [3] suggest that the Portilla-Simoncelli texture model captures the information contained in a single pooling region in the periphery, and so synthesizing images that shared the same texture parameters effectively visualizes the information that the pooling region contained. Their work showed that the model can predict the information contained in a number of types of letter array stimuli and some object arrays.

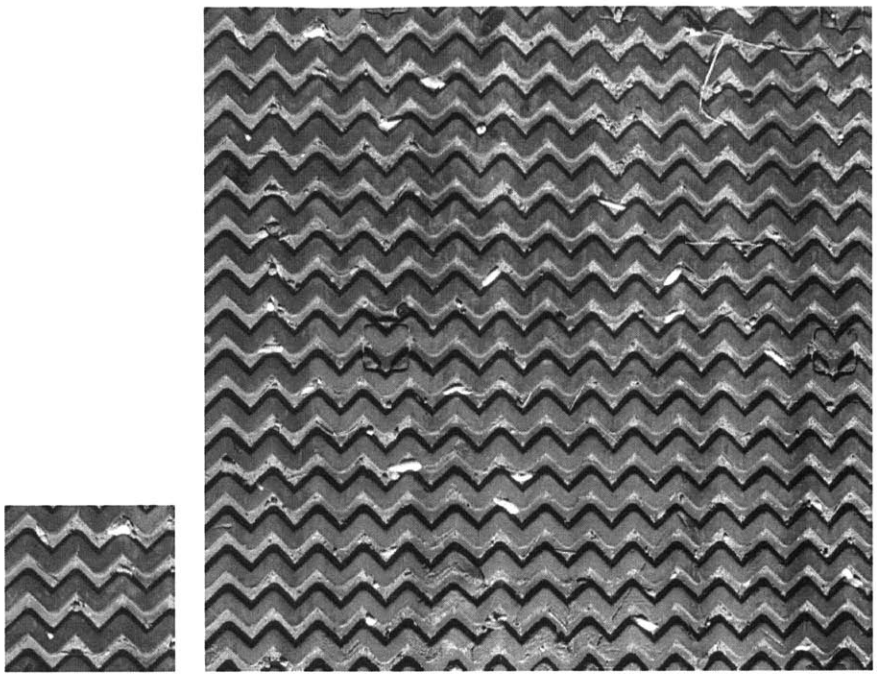
Extending that initial work, in this thesis, we provide additional support for the conjecture that the model captures the same visual information as the peripheral visual system does. In particular, we find that the model makes similar errors that humans make in regards to phase and contour perception. Further, we propose a method to visualize the information in all the pooling regions in the visual field simultaneously.

The work in this chapter presents work that I conducted under the supervision of Ruth Rosenholtz. My contribution was in developing the algorithm to visualize the statistics in the entire visual field.

3.2 Texture Models as Compressed Representations of Visual Appearance

Texture synthesis is a technique that aims to produce a new, arbitrarily sized image that looks like a sample texture image. For example, given the small patch in Figure 3-1(a), the goal is to produce more of the same visual appearance, like in (b). In order to produce larger images that share the visual appearance, the texture synthesis algorithm either explicitly or implicitly defines a model of texture that it uses to synthesize a new image to match the underlying model. We examine two parametric texture models that encode the marginal statistics (Heeger-Bergen [16]) and joint statistics (Portilla-Simoncelli [41]) of oriented subbands from a steerable pyramid decomposition of the sample image.

To succeed in producing images that have the appearance of the original sample



(a) Sample (b) Desired Output

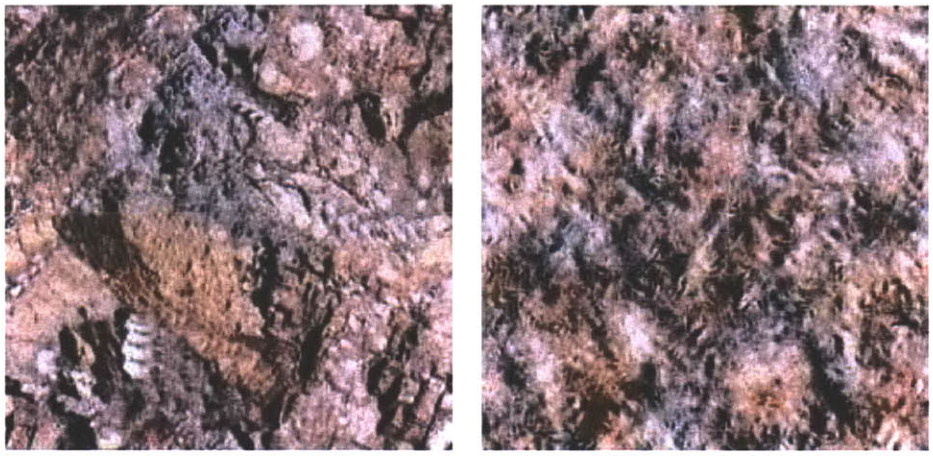
Figure 3-1: Texture synthesis aims to create arbitrarily sized images that share the visual appearance of a sample texture.

texture, the synthesis algorithm needs to produce the new image as though the same process that generated the sample texture also generated the new image. If the texture model manages to synthesize images that appear like the original sample, it has in a sense captured the nature of the stochastic process that generated the original sample. In the case of explicit, parametric texture models, the number of parameters in the model are typically much smaller than the number of pixels in the input image.

What visual information is lost if we only retain the parameters of a texture model? Because these are texture synthesis models, they provide methods for visualizing the information contained in the encoding. Consider the Heeger-Bergen texture model. It computes histograms of subbands in a steerable pyramid and synthesizes new textures iteratively matching the corresponding histograms from a random noise image. In addition to the subbands of the pyramid, the pixel histogram is also matched.

Figure 3-2 shows an example of a texture whose synthesis looks like the original,

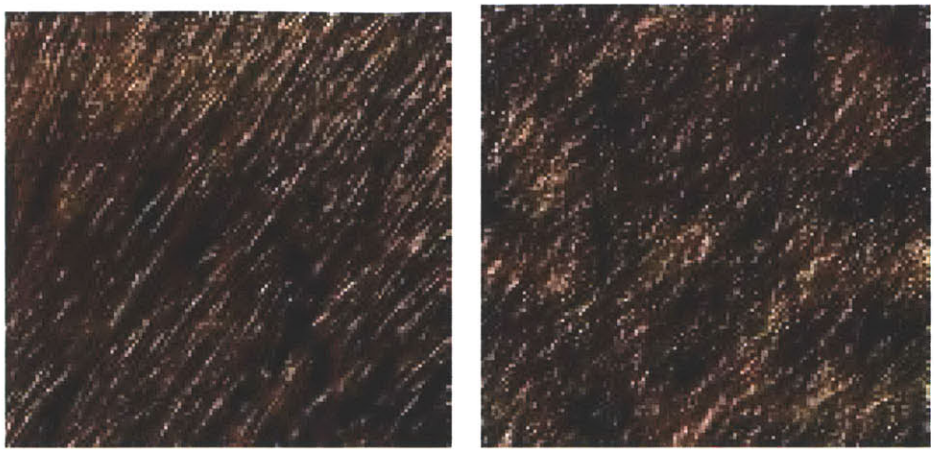
and Figure 3-3 is an example where it is very easy to distinguish the original from the synthesis. If images are represented as a set of histograms, then all images which share the same set of histograms form an equivalence class of images under that representation. Variations within that class correspond to the ambiguities that could arise from compressing the visual information to just those histograms. So, by looking at these syntheses, we can gain intuitions about the ambiguities inherent in this representation.



(a) Sample (b) Synthesized

Figure 3-2: Synthesis using the Heeger-Bergen method, by matching subband histograms in a steerable pyramid decomposition [16]

The Portilla-Simoncelli texture model on the other hand, computes joint statistics of subbands of a complex steerable pyramid. In particular, the shapes of the distributions of subband responses are measured in addition to joint statistics of the steerable complex pyramid. The joint statistics include correlations between orientations at any scale, correlations between neighboring scales, autocorrelations within subbands, and some phase statistics. Figure 3-4 show examples of texture synthesis with the Portilla-Simoncelli model. The syntheses seem to preserve extended structures better, as compared against Heeger-Bergen. Overall it seems to be able to synthesize images that are plausible extensions of the samples. It does not, however, fully preserve all the visual details of the original image, and it introduces irregularities that one wouldn't expect to observe in processes that generate those types of

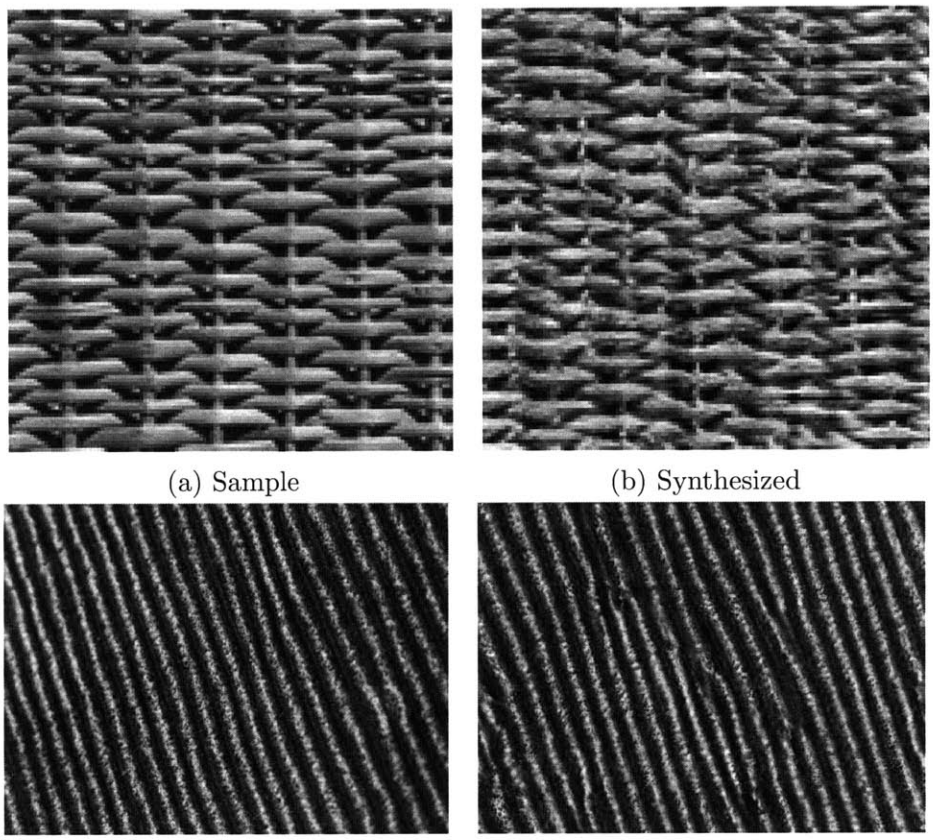


(a) Sample (b) Synthesized

Figure 3-3: Synthesis by matching subband histograms in a steerable pyramid decomposition [16]

images.

In the following, we use a slight variation of the texture model in [41]. For robustness, as modifying skewness and kurtosis often results in numerical instabilities on artificial images, we use histograms instead to represent the shapes of the marginal distributions of the subbands. We also allow the model to compute statistics only over a specified region in the sample texture, as well as enforce the statistics to match only in a specified area in the image being synthesized. To compute the statistics over a given area, we simply compute the various statistics, applying a weight to each location based on the mask. Synthesis is performed similarly, with normalization weights appropriately computed for correlation and autocorrelation modifications. To produce color syntheses, we use Independent Components Analysis [20] to obtain a decorrelated space in which to run three separate syntheses, then recombine the outputs after the synthesis step is complete.



(a) Sample (b) Synthesized

(c) Sample (d) Synthesized

Figure 3-4: Example Portilla-Simoncelli syntheses

3.3 Stimuli Not Accurately Represented by the Portilla-Simoncelli Texture Model

The human perception system makes errors. If the representation underlying the perceptual system itself is ambiguous on the same types of errors, then the representation is a likely explanation. Does the representation we propose exhibit ambiguities where human perception makes errors?

In chapter 2, the Portilla-Simoncelli statistics were presented as a feature-space where difficulty in categorizing the statistical representations of arrays with multiple objects predicted how difficult the crowding task using those stimuli would be. Figure 3-5 shows that the model does not have difficulty in representing a single letter, but when the stimulus is complex, ambiguities arise to the point where letter identities are

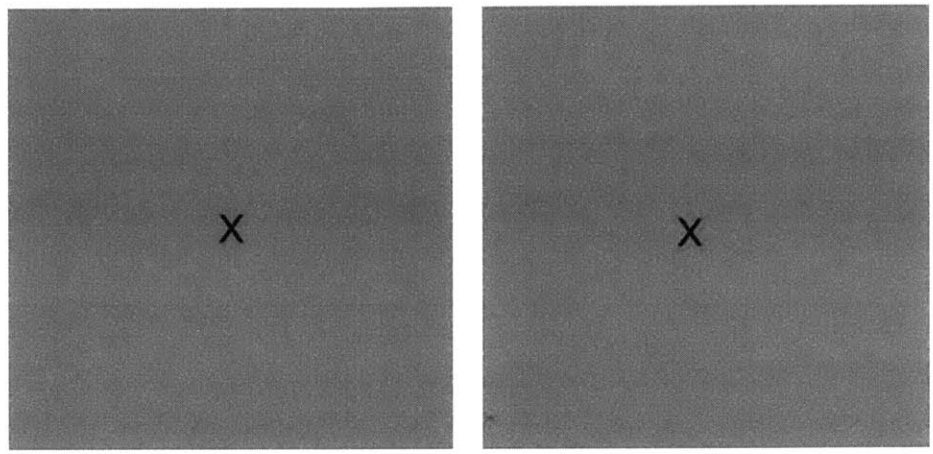
no longer easy to establish. It should be noted that these syntheses exhibit artifacts from the implementation of the synthesis procedure (the synthesis assumes image wrap around – i.e., top is connected to bottom, and left is connected to right).

In Figure 3-6 the syntheses do not exhibit many ambiguities about the contour, but in Figure 3-7 the model is not able to unambiguously represent the more complicated contour. In particular, we notice that it hallucinated a T junction when there were no such junctions in the original image.

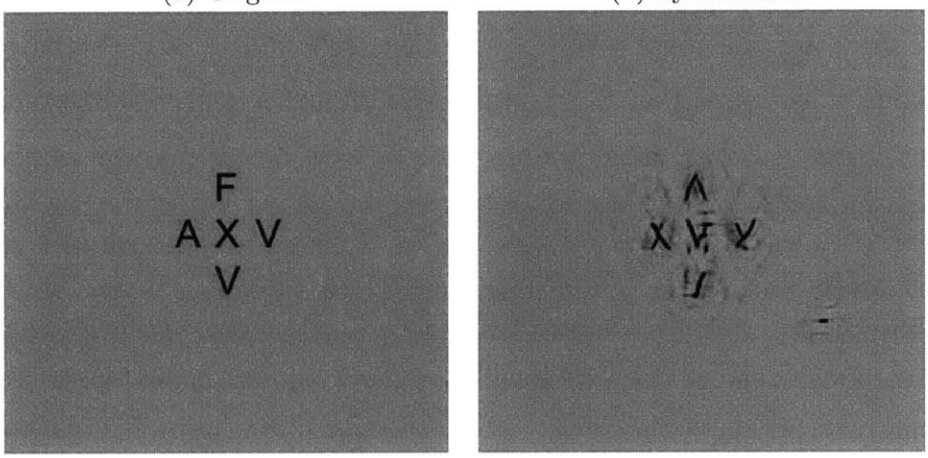
We consider another synthesis in Figure 3-8. When all elements are black, the statistics are unambiguous about the color of the various oriented line segments. (c-d) When there are black and white line segments, the model hallucinated a white vertical line segment even though the original image had no such combination. This may be an indication that ambiguities in the statistical description can explain why illusory conjunctions like those often reported in visual search occur. Note also that these illusory combinations of color and orientation are difficult to reproduce in non-parametric synthesis models, for example in [10, 9, 24]. This is because these non-parametric models tend to only use small (perhaps irregular shaped) parts from the original image in the synthesis algorithm, which makes it difficult to hallucinate parts that were never present in the original.

Clearly, the Portilla-Simoncelli model allows many ambiguities about the visual information it tries to represent in the examples presented in this section. However, by optimizing the model to capture texture appearance, the model also seems to have selected model parameters that allow ambiguities that fool the peripheral visual system. The errors the model makes are the types of errors that humans also make in the periphery. In visual crowding, people do poorly in identifying letter identities when the letter is flanked by surrounding letters. In addition, phase discrimination in the periphery is also known to be poor in humans [50, 33].

Additionally, Balas [2] shows that humans were not very good at parafoveal discrimination of real vs synthesized Portilla-Simoncelli textures. These lines of evidence suggest at least that the choice of the Portilla-Simoncelli texture model will suffice for a working hypothesis for the set of visual information that the peripheral visual



(a) Original (b) Synthesized



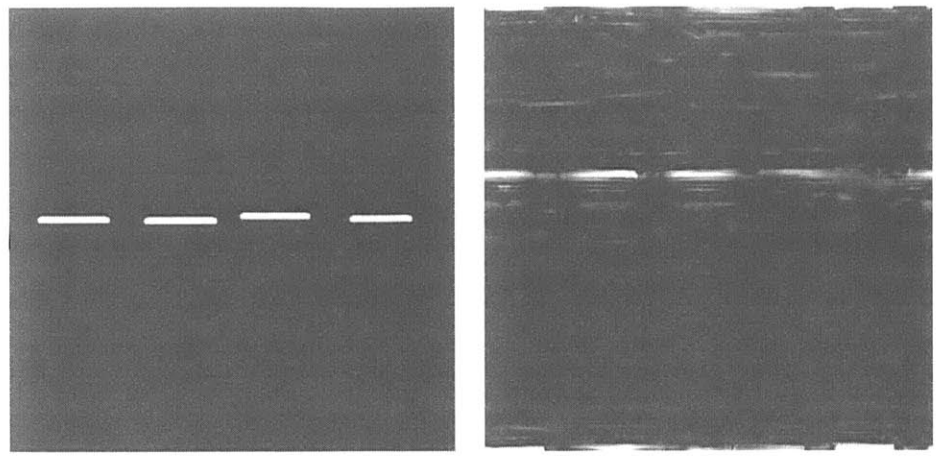
(c) Original (d) Synthesized

Figure 3-5: The ambiguities in representing a single letter are few, but in complex stimuli with multiple letters, the statistics do not sufficiently constrain the synthesis so that the letter identities are preserved in the representation.

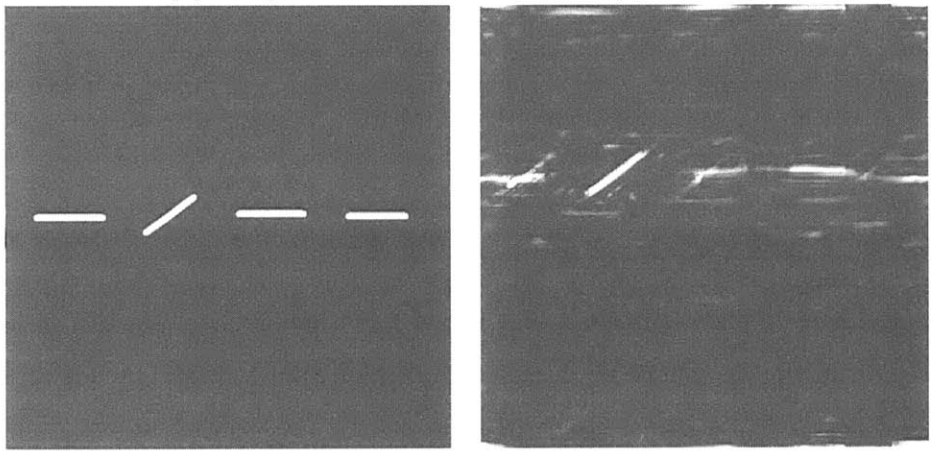
system computes.

3.4 Visualizing Statistics in the Entire Visual Field

We have noted that texture models can be viewed as compression algorithms for visual appearance and have identified the Portilla-Simoncelli model as one that seems to capture the same types of information that the human visual system does. However, the Portilla-Simoncelli model makes the assumption that the sample texture is generated from a stationary process (i.e., does not change depending on position), and



(a) Sample (b) Synthesized

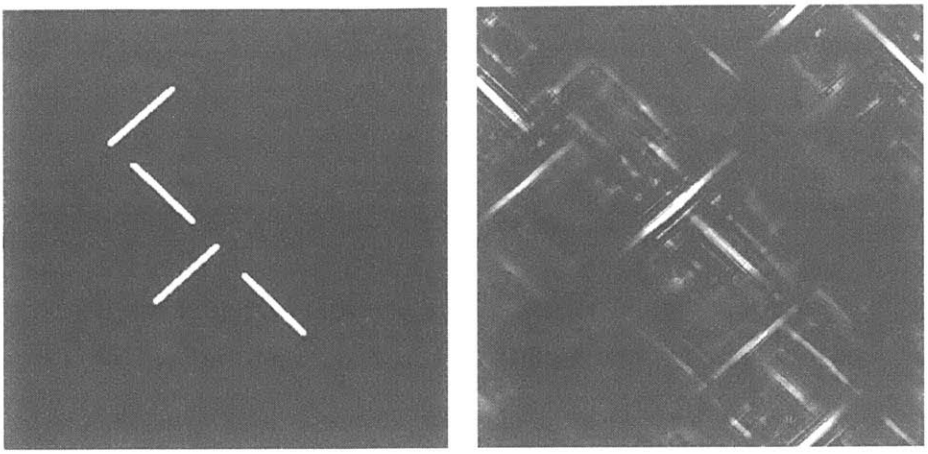


(c) Sample (d) Synthesized

Figure 3-6: Simple contours do not have representations that allow much ambiguity in the model

so compression is difficult to achieve in most natural images where this assumption is violated.

Balas et al suggested that these statistics are computed across the visual field, in local, overlapping pooling regions whose size grow with eccentricity [3]. The local pooling regions where these statistics are computed help relax the assumption of a global stationary process that generated the visual field. Without considering overlaps, the global stationary process assumption is reduced to a locally-stationary process, where locality is defined by the pooling regions. Allowing pooling regions to overlap, however, makes the locally-stationary assumption even weaker, as pooling



(a) Sample (b) Synthesized

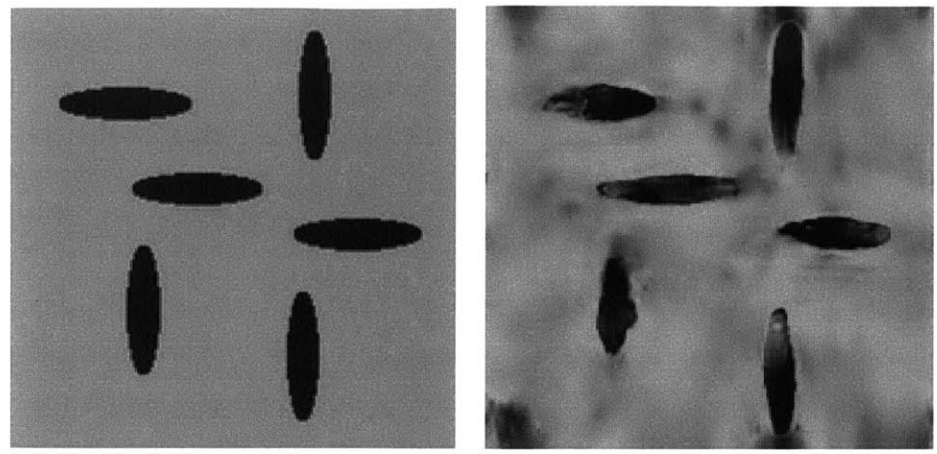
Figure 3-7: More complicated contours are difficult for the statistics to represent unambiguously. Note that the synthesis produced an illusory T junction in the middle, indicating that the model would have difficulty discriminating these rotated L-like structures from T junctions.

regions are allowed to influence each other.

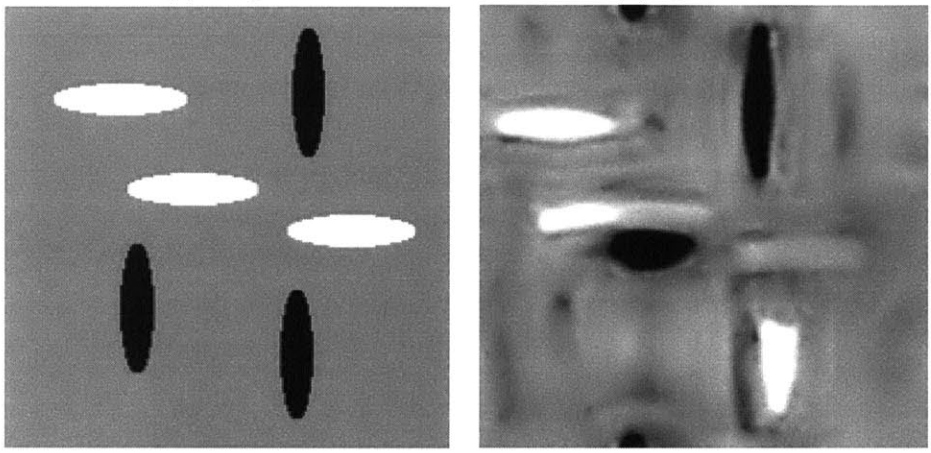
While Balas et al showed that it was useful to visualize the statistics of a single pooling region in the periphery [3], a natural question that follows is what about the entire visual field? The work on single pooling regions suggest that we might be able to gain additional intuitions about the global structures in stimuli that span a larger visual area, by extending that idea to visualize the local statistics in pooling regions across the entire visual field. How might we achieve this?

To visualize the peripheral representation of an image given the fixation point, we first lay out a number of pooling regions over the image. The pooling regions are placed in a log-polar grid, as depicted in Figure 3-9. The pooling regions are oval because it is the shape suggested by the research on visual crowding [50]. We place sufficient pooling regions to cover the entire image, plus some additional rings beyond, to allow the model to deal with the edges of the image as well. Having pooling regions with blank images as input,constrains those regions to be blank (blank regions are perfectly represented by the statistics). This constrains where where the algorithm will synthesize the non-blank visual "stuff".

The algorithm starts by seeding the synthesis with random noise. Then it copies



(a) Original (b) Synthesized



(c) Original (d) Synthesized

Figure 3-8: (a-b) When all elements are black, the statistics are unambiguous about the color of the various oriented line segments. (c-d) When there are black and white line segments, the model hallucinated a white vertical line segment even though the original image had no such combination

over a central region where the fovea is because there is little loss of information close to the center of fixation. Then, it iterates over all the pooling regions, constraining the local region to have the same statistics as the corresponding local region in the original image. We cannot offer any convergence guarantees at this point, but the model seems to converge after 50 to 100 iterations on the images we have tested it on.

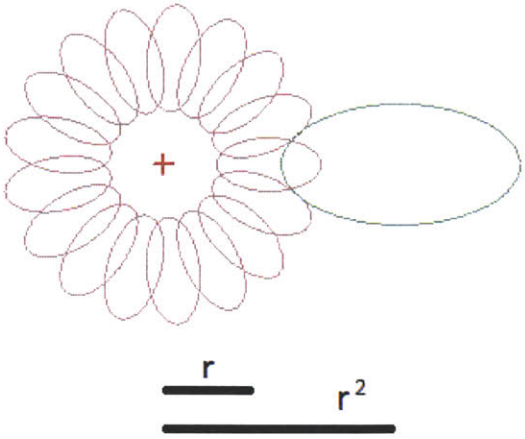


Figure 3-9: Pooling regions are placed in a log-polar grid.

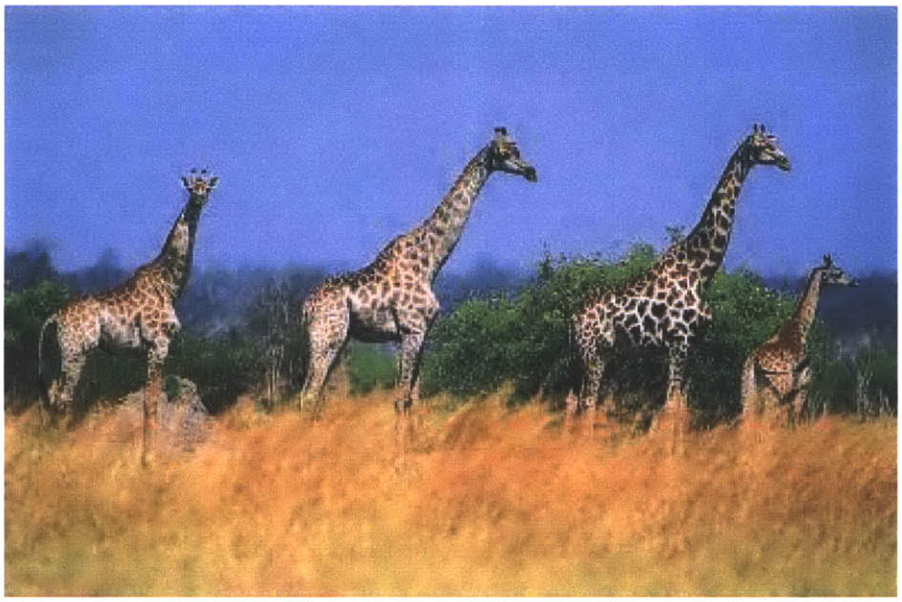
(a) Sample

(b) Synthesized

Figure 3-10: The texture tiling algorithm in progress

3.4.1 Example Syntheses

In the following examples, the fixation is in the middle of the image.



(a) Original

(b) Synthesized

Figure 3-11: Giraffe

(a) Original (b) Synthesized

Figure 3-12: Parachute

(a) Original

(b) Synthesized

Figure 3-13: Street scene

(a) Original

(b) Synthesized

Figure 3-14: Ducks

3.5 Related Work

Freeman and Simoncelli [14] independently published a similar model to the work presented in this thesis. Both models are inspired by the suggestion by Balas et al in

[3], and are approximately computing the same statistics. Some differences between their model and the model presented here are that they use different pooling region shapes and constrain statistics of overlapping pooling regions jointly.

3.6 Future Work

3.6.1 Convergence and Statistics

The synthesis procedure does not perfectly constrain random noise images to match the desired statistics. The distance between the synthesized image's statistics and the desired statistics vary greatly depending on the types of images. This makes it difficult to define a stopping rule. One possibility to address this problem is to better characterize what types of statistics are difficult for the synthesis algorithm to match by comparing the difference in the space of statistics for the image and the synthesized image, in a large dataset.

3.6.2 Space of statistics

Because we are able to compute the statistics for an arbitrary image patch, and these statistics seem to correlate well with peripheral vision perception, what can we say about natural image statistics, and what the space of statistics are on a natural image dataset? Using images taken from LabelMe [45], we computed Portilla-Simoncelli statistics from a million natural image patches at various scales from 600 images.

Figure 3-15 shows patches from those images. We place an image patch where its statistics' projection onto the two-dimensional subspace that optimally describes the highest variance of the dataset (using PCA [38]). To a first approximation, the first two principal components seems to describe orientation – horizontal patches appear near the bottom right and vertical patches in the top left. Further research is needed to better represent this space of statistics so that we may eventually predict the perceptual difference of two patches in the periphery given the statistics of those patches.

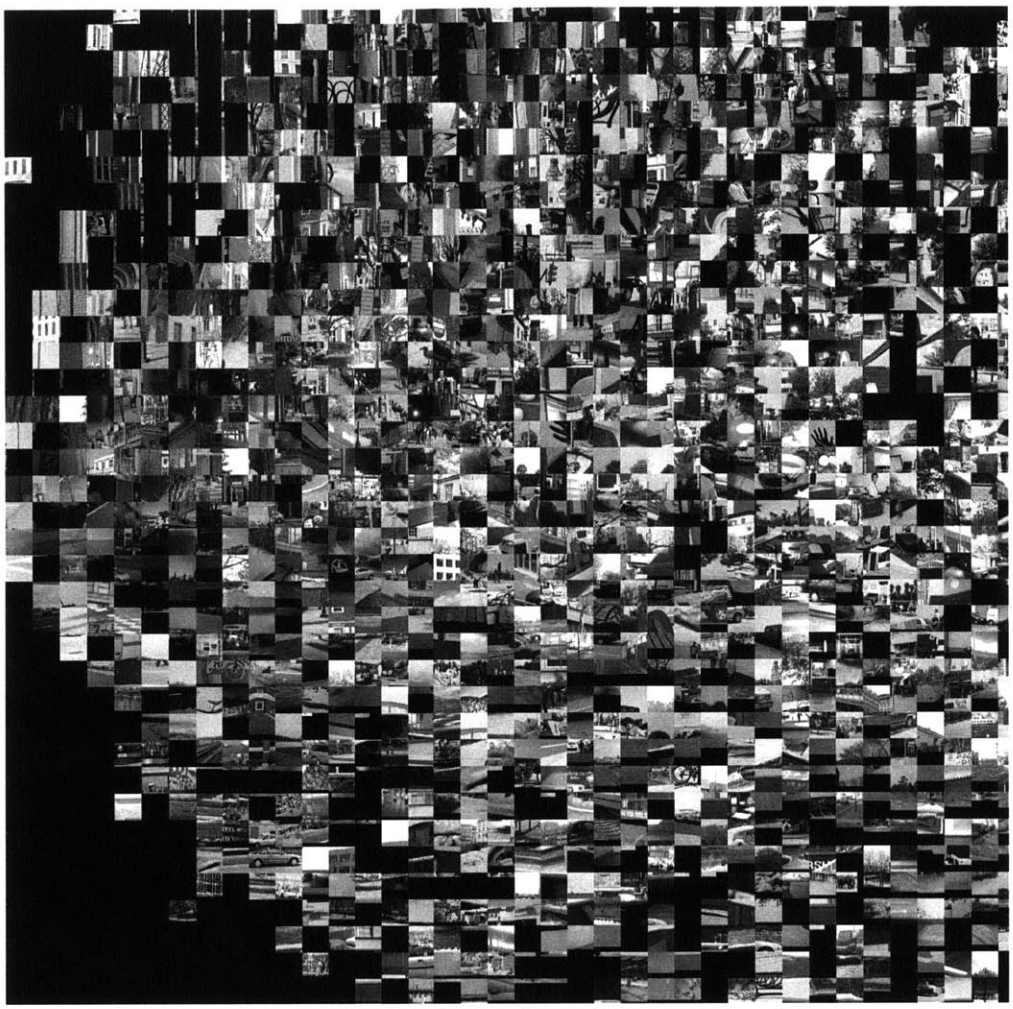


Figure 3-15: Space of statistics in natural images

Additionally, better understanding the space of statistics may help with problems where the human visual system seems to prefer certain solutions over others, when the statistics allow many interpretations. For example, the "Healing Grid" illusion [22] in Figure 3-16, if one stares at the center of the image for about 30 seconds, one tends to perceive a regular grid in the periphery. Our model currently cannot account for why this interpretation of the statistics is preferred over others. Perhaps a regularity prior on the possible interpretations (perhaps from natural image statistics) can account for these types of effects.

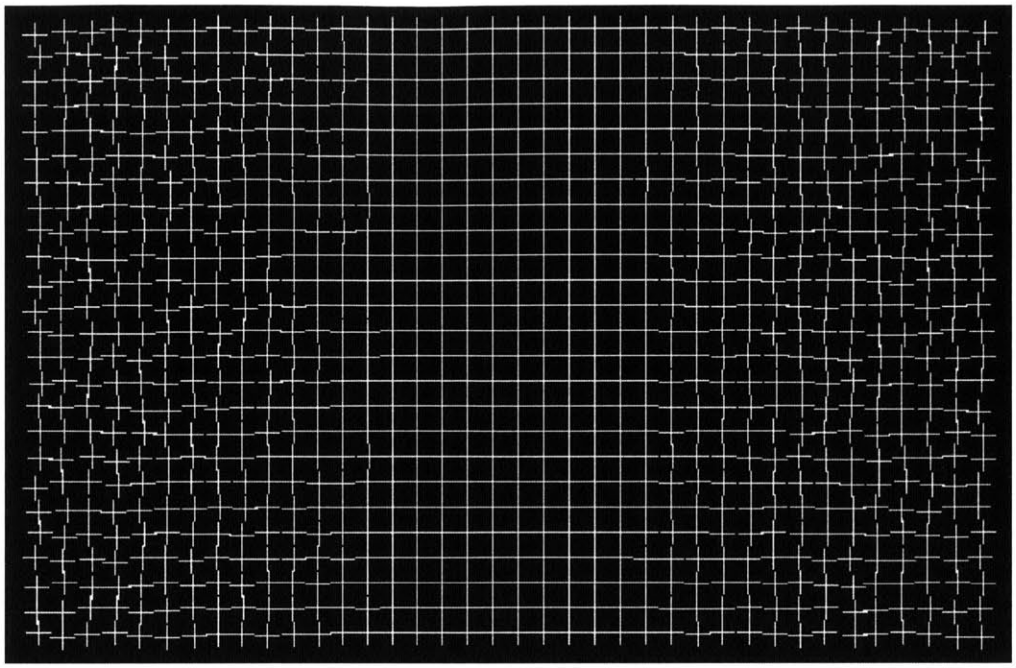


Figure 3-16: The "Healing Grid" illusion [22]. After staring at the center of the image for 30 seconds, the regularity appears to spread from the center of the grid. Perhaps priors in interpretations of ambiguous statistics drive this illusion.

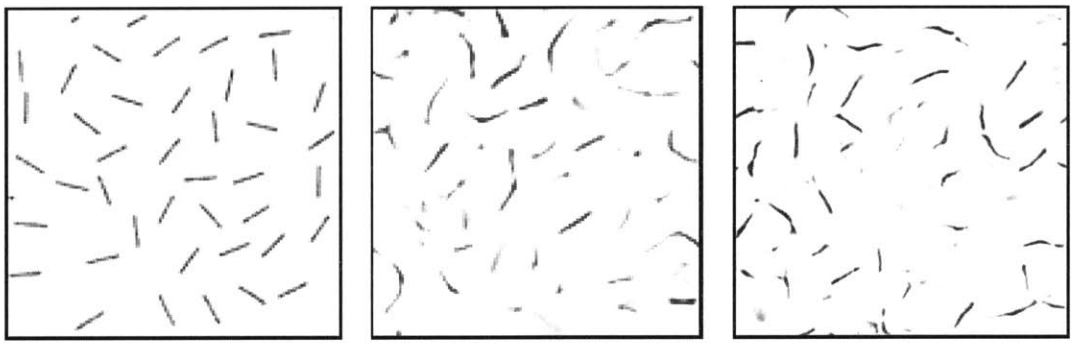
3.6.3 Machine Learning

Ideally, we should be able to calibrate and test the model without having to insert a human in the loop. Having a computer be able to simply take as input the statistical texture parameters and output the expected performance on a task is highly desirable. To that end, a machine learning approach seems very attractive.

We could theoretically use machine learning techniques to disciminate different types of textures using only the statistical representation. However, the space of the summary statistics has over a thousand dimensions, and so it would be trivial to separate minute differences in simple classes of images when the dimensionality is that high. Noise must be added to mimic the types of sensory and neuronal noise that occurs in the visual system. But this noise model must be learned, and at present, we do not have sufficient data, or a large and complex enough dataset to be confident that the resulting model will not be overtrained.

3.6.4 End-Stops

One particular type of visual information which the Portilla-Simoncelli model fails to represent well are the types of information that "end-stop" cells tend to respond to – variations in line ends, corners, and line segments [17]. The Portilla-Simoncelli model is not able to discriminate between these types of stimuli well. For example, in Figure 3-17, multiple oriented line segments (a) are not distinguishable from curves (b) to the standard Portilla-Simoncelli texture model. Visually inspecting the images suggest that curves that the synthesis produces and line segments from the original might be easily discriminable in the periphery, but an experiment is needed to verify this hypothesis. If true, this marks a situation where the statistics are insufficiently representing the visual information.



(a) Original (b) Synthesized (c) Additional Stats

Figure 3-17: The problem in representing end stops.

In the case that these types of stimuli are easily discriminable in the periphery, what types of statistics might be required to represent them well? The Hessian is a matrix that has all the second order partial derivatives of a function as entries. Applied to an image, it describes the local curvature of the image. Because of this, line segment ends can be identified by the determinant of the Hessian. Perhaps additional statistics describing the second derivatives of the image can help augment the set of statistics we use to better represent peripherally presented stimuli. This direction is related to Heitger et al's suggestion that second derivatives in combination with local-maxima detection are required to represent these types of visual information [17]. Some preliminary work to include second derivatives in the model is seen in

Figure 3-17 (c). It seems to yield some very small improvement, but more work is needed to make enough advances to fully address the problem.

3.6.5 Speed improvements

The algorithm, as currently implemented in MATLAB, is slow. It can take between 6 hours to 3 days to produce a full field synthesis of a moderately sized image (512x512). There are various ways of improving the speed of the algorithm which fall into two classes: algorithm improvements, and better implementation. Some model efficiency improvements include simultaneously enforcing constraints of multiple pooling regions where they overlap, so that the number of iterations needed to converge is reduced, and a better coarse-to-fine strategy can aid in speeding up convergence. The implementation itself may be sped up by massive parallelization on a GPU, and using C++ instead of MATLAB. The time it takes to run these experiments hampers model exploration and testing.

3.6.6 Further testing of model parameters

Freeman and Simoncelli [14] have done some work on determining some of the parameters in their version of the model. It is not clear from the literature exactly what the spacing of the pooling regions in the visual field should be. To test this more directly for our model, we need to investigate the model's representational capacity as we vary the pooling region placement and density. This effort is significantly slowed when we are only able to synthesize very few images at a time. The other parameters in the model (such as the size of the autocorrelation window, phase statistics, other correlations, etc), similarly, need to be tested in more detail.

3.7 Conclusion

In this chapter, we discussed additional reasons for why the Portilla-Simoncelli texture model is an acceptable working hypothesis for the information that is retained by the

peripheral visual system. In particular, we identified ambiguities in representing complicated stimuli with many types of junctions, and increased ambiguities when stimuli have elements on both sides of contrast of the background. In later chapters, we show that these types of mistakes underlie why certain types of visual search tasks are slow, and why some illusions occur.

Additionally, we present an algorithm to visualize the statistical parameters of all the pooling regions in the simulated visual field, and present the results of the syntheses on some number of examples. Some future lines of research are identified based on the work presented here.

Chapter 4

Visual Search

4.1 Outline

In Chapter 2, we noted that there is evidence that the visual system compresses visual information, and so it gives rise to ambiguities about what stimuli produced it. Because the periphery comprises most of the visual field, many tasks should have performance affected by these peripheral limitations.

Visual search is a task where subjects are asked to find a target among a number of distractors in a search display. For example, in Figure 4-1, the task is to find the O among the distracting Qs. Very little of the display is foveal, so peripheral effects would matter in trying to find the target. In many search tasks, items may be clustered together. But results from visual crowding indicate that identification performance is reduced for a particular target whenever it is surrounded by distracting items, which is the case in visual search displays with items clustered close to each other. Perhaps the peripheral mechanisms that cause reduced performance in crowding displays similarly have an effect in visual search.

In this chapter, we discuss the puzzle of search. The amount of time it takes to find a target in a search display depends strongly on what the target and distractors look like. Feature search (tilted line among vertical lines) is easy, configural search is difficult (Ts among rotated Ls), and there are asymmetries (Q among Os is easy, but O among Qs is difficult). The discriminability of individual target and distractors is

generally not predictive of search performance. Prior research has tried to explain these confusing results, resulting in various well-known theories like Feature Integration Theory [54]. When peripheral considerations are taken into account, however, we show that instead of the discriminability of single items, one should consider patch discriminability.

We replicate classic visual search results (measuring how quickly subjects find the target in five search tasks) and run a separate experiment to estimate the discriminability of the statistical information contained in target present patches as compared to target absent patches. The results show a strong correlation between statistical discriminability and search efficiency, lending evidence for the claim that the information contained in the statistical representation of the patches is predictive of search performance.

The work in this chapter presents research that I conducted in collaboration with Ruth Rosenholtz, Benjamin Balas, and Jie Huang [44]. In this chapter, I present parts of that research in which I had a direct involvement in. My specific contributions are in helping design, run, and analyze the search experiment, as well as running and analyzing the statistical discriminability experiment.

4.2 Puzzles of Visual Search

Figure 4-1 shows a typical search display, where the task in this case is to locate the "O" among the "Q" distractors. As one might expect, the choice of target and distracting items affects how easily one can find the target. The time taken to find the target is typically linear in the number of items in the search display, but the slope of the line varies depending on the choice of target and distractor.

Perhaps a simple and intuitive idea applies. Does the discriminability of a target item from a distractor item predict search rates? It turns out that this hypothesis can only explain the data in some limited cases. For example, Palmer et al find that it is easier to find an O among Xs than to find an O among Qs [36]. However, there are many cases where the discriminability of a single target item to a single distractor

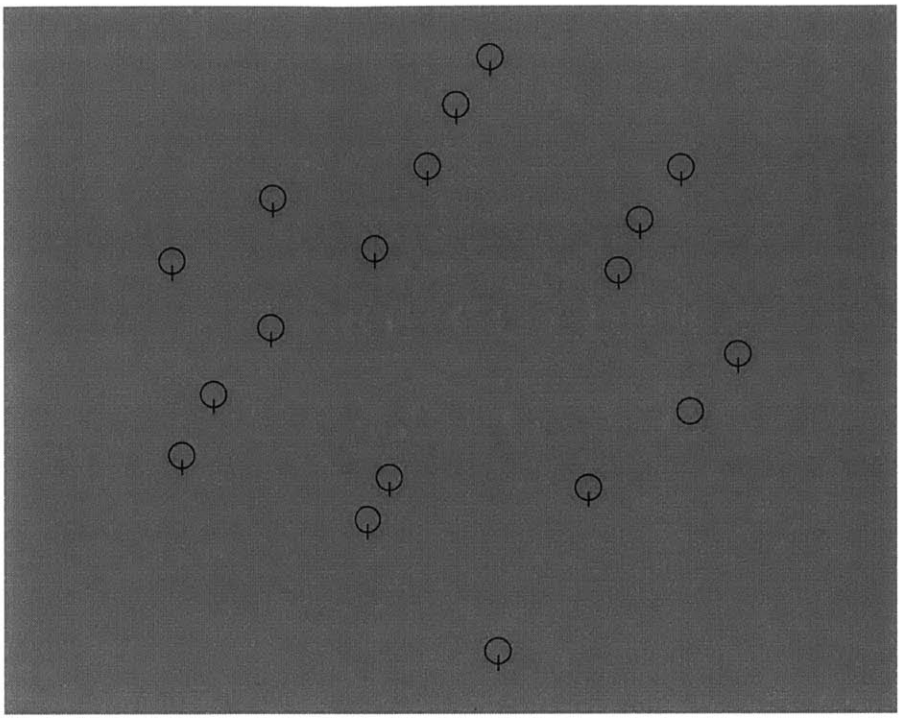


Figure 4-1: A typical visual search task: search for O

item is trivial, yet search performance ranges from easy to difficult. For instance, it is easy to discriminate a single white vertical bar from a vertical black bar or from a horizontal white bar, but the search for a white vertical bar among vertical black bars and horizontal white bars is difficult. The latter (more difficult) instance is a case of feature conjunction search [54]. Configural type searches, like for a T among Ls, is also difficult [60], even though it is trivial to discriminate a T from an L.

Problematic for the single item discriminability hypothesis is the existence of search asymmetries [53, 52, 59]. Clearly, the discriminability of a Q compared to an O is the same as that of an O to a Q, so one might predict that search times for those two situations would be similar, yet the search for Q among Os is much faster than the search for O among Qs [53].

Beyond the problems of single item discriminability not being able to predict search results, there are additional factors where search rates are unrelated to the choice of target or distractor. For example, how the items are placed in a search display, and whether/how items group together [57, 7].

Treisman is often credited, and deservedly so, for pioneering research on visual search. She suggested the field's first attempt at solving these puzzles: Feature Integration Theory (FIT) [54]. Later work by Wolfe built on those ideas to suggest a theory of Guided Search, which addressed behavior in target absent trials and how features may guide attention while searching for the target [58]. FIT proposes that the visual system first computes a number of features in the visual field in parallel. In the case of easy feature searches, finding the target is simply a matter of detecting the target feature in these parallel feature channels. But when multiple features are necessary to distinguish the target from the distractors, the visual field has to employ a slower serial attentional spotlight to bind features together at any particular location in order to test whether that location has the target properties.

FIT, therefore, can explain why feature search for a target like a tilted line among vertical lines is easy (simply detect the presence of an oriented line), but the conjunction search for a white vertical bar among black vertical bars and horizontal white bars is difficult (serial attention is needed to bind features at the various item locations). However, whether FIT correctly determines how easy a search condition is will depend on what features one incorporates into the model. Vision researchers have searched for the set of features that the visual system computes. However, many features had to be added in order to explain search results of puzzling experiments. Among the list of features identified include low-level computations such as color and orientation. But in addition, other features like 3D shape and reflectance were proposed as additional features that were computed in parallel when the search experiments suggested these conditions allowed fast visual search [11, 48].

In addition, FIT is better thought of as a theoretical framework of how the visual system might work in a visual search task as opposed to a model that can predict how quickly one can search for a target in an image. In general, the field lacks a model of visual search that can estimate the average search time given an arbitrary image and what the target looks like. Towards that goal, this chapter proposes a method to estimate the difficulty of search for an arbitrary search condition.

4.3 Relationship Between Peripheral Vision and Search

Most of the visual field is peripheral, and consequentially, so are most visual search displays. If the information in the periphery is ambiguous about where in the display the target is, it will be difficult to efficiently guide the fovea to find the target.

This insight is key in unraveling the puzzle of visual search. Our model of the peripheral visual system contains overlapping pooling regions throughout the visual field whose sizes grow linearly with its distance from the fixation (major axis radius $\approx$ .5 eccentricity). Because many pooling regions will be large, some will contain multiple items. Research on crowding indicates that when there are multiple items within a pooling region, discriminability is much lower than when there is only a single item. Perhaps the discriminability of patches with a target vs patches without underlie why some search conditions are easy or difficult. This discrimination task is essentially what the peripheral visual system is trying to solve when determining where to fixate next.

In Chapter 3, we noted that the representational ability of the statistical model we use has difficulty in representing complex stimuli with multiple rotated Ls and Ts. It also cannot accurately represent phase information, for example when there are bars with colors on opposite sides of contrast of the background placed in a fairly disorganized manner. The types of mistakes the model makes here seem indicative of search difficulty for T among L and for conjunction search.

To formalize the intuition, the model predicts that search will be easy if the summary statistics of a target-present patch are very different from the summary statistics of a target-absent patch. This requires us to estimate the separability or discriminability of those types of patches.

To achieve this, we could theoretically use machine learning techniques to estimate how easy it is to tell apart target-present and target-absent summary statistics. However, as discussed in Chapter 3, when the set of stimuli are this simple (artificial Ls and Ts on a gray background, as opposed to rich, naturalistic stimuli), it would

be trivial for any classifier to discriminate. An internal noise model of the statistics must be learned on an independent, rich, and large dataset in order to use machine learning on these types of stimuli.

Balas et al [3] propose a different method to obtain the separability of the summary statistics. They note that there are presently no reliable algorithms for mimicking human pattern recognition, however they can simulate human pattern recognition by using actual human observers. For each condition of interest, the information present in the statistics of some sample stimuli are visualized using a texture synthesis algorithm [41]. This is repeated a number of times to generate a set of images which share the same summary statistics. Humans are then asked to discriminate between synthesized images that share the same statistics as patches from one category, compared to patches from a different category. This discriminability provides a measure of the statistical discriminability of these categories of images. This methodology also accounts for human ability to use higher-level knowledge in the discrimination task. We apply the same methodology to estimate how discriminable the target-present patches are from target-absent patches using only the information from the statistical summary.

In order to judge whether the model is accurately predicting search results, we need to estimate how difficult visual search is on a number of different search conditions. Experiment 1 measures search efficiency on a number of classic visual search conditions. In Experiment 2, we estimate statistical discriminability of target present and target absent patches as described above. We show that the statistical discriminability strongly depends on the condition, and further, that it predicts how difficult the visual search task is.

4.4 Experiment 1: Classic Visual Search

In Experiment 1, subjects participated in five classic search tasks. Results for these tasks already exist in the literature, but to accurately compare across conditions, it is important to standardize the search displays, and minimize subject variances by using the same subjects to perform all five tasks.

4.4.1 Method

Ten subjects (six male) participated in the search experiment after giving informed written consent. Ages ranged between 18 and 40. All subjects reported normal or corrected-to-normal vision, and received monetary compensation.

Procedure

We tested five classic search conditions: Conjunction search (targets defined by the conjunction of luminance contrast and orientation), search for T among Ls, search for O among Qs, search for Q among Os, and feature search for a tilted line among vertical lines. Target and distractor items are shown in the left two columns of Figure 4-3.

Stimuli were presented on a 40 cm x 28 cm monitor, with subjects seated 75 cm away in a dark room. We ran our experiments in MATLAB, using the Psychophysics Toolbox [5]. Eye movements were recorded at 240 Hz using an ISCAN RK-464 video-based eyetracker for the purposes of quantitatively modeling the number of fixations to find the target which we discuss in Chapter 5. The search displays consisted of a number of items (set size), consisting of either all distractors (target absent trial) or one target and the rest distractors (target present trial). Target present and target absent displays occurred with equal probability.

Each search task had four set size levels: 1, 6, 12, or 18 total items. Stimuli were randomly placed on 4 concentric circles, with added positional jitter (up to 1/8 deg). The radii of the circles were 4, 5.5, 7, and 8.5 degrees of visual angle (v.a.) at a viewing distance of 75 cm. Example target-present stimuli for O among Qs is shown in Figure 4-1 for set size=18.

On each trial, the search display was presented on the computer screen until subjects responded. Subjects indicated with a key press whether each stimulus contained or did not contain a target, and were given auditory feedback. Each subject finished 144 trials for each search condition (72 target-present and 72 target-absent), evenly distributed across four set sizes. The order of the search conditions was counterbal-

anced across subjects, and blocked by set size.

4.4.2 Results

Search difficulty is quantified as the slope of the best-fit line relating mean reaction time (RT) to the number of items in the display. Only target-present trials when subjects correctly detected a target were included in this analysis. Figure 4-2 plots the mean reaction time on correct target-present trials against set size of search display, along with the best linear fit. These results are consistent with previously reported search studies.

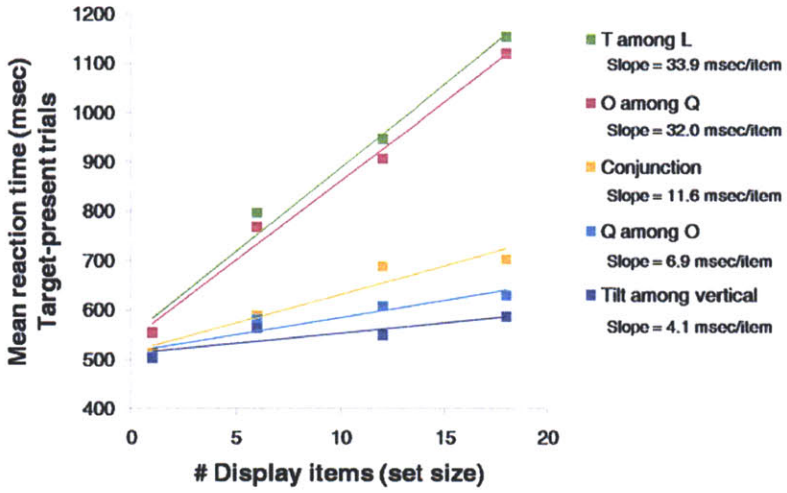


Figure 4-2: Mean reaction times (RTs) for correct target-present trials are shown, averaged across subjects, for each combination of condition and set size. The legend gives the slope of the RT vs. set size function for each condition, a typical measure of ease of search.

4.5 Experiment 2: Mongrel Discrimination

4.5.1 Subjects

Five subjects (four male) participated in this experiment. Their ages ranged from 18 to 45 years. Each reported normal or corrected-to-normal vision, and were compensated for participation.

4.5.2 Procedure

To measure the discriminability between target+distractor and distractor-only patches using only summary statistics, we used a similar methodology to Balas et al [3]. First, we generated 10 unique distractor-only and 10 unique target+distractor patches for the five visual search conditions described above (see Figure 4-3, columns 1 and 2). For each patch, we synthesized 10 new image patches that closely match the same summary statistics as the original patch (Figure 4-3, last 4 columns), using the Portilla-Simoncelli texture synthesis algorithm [41]. The resulting synthesized patches are nearly equivalent to the original input in terms of the summary-statistics measured by the model.

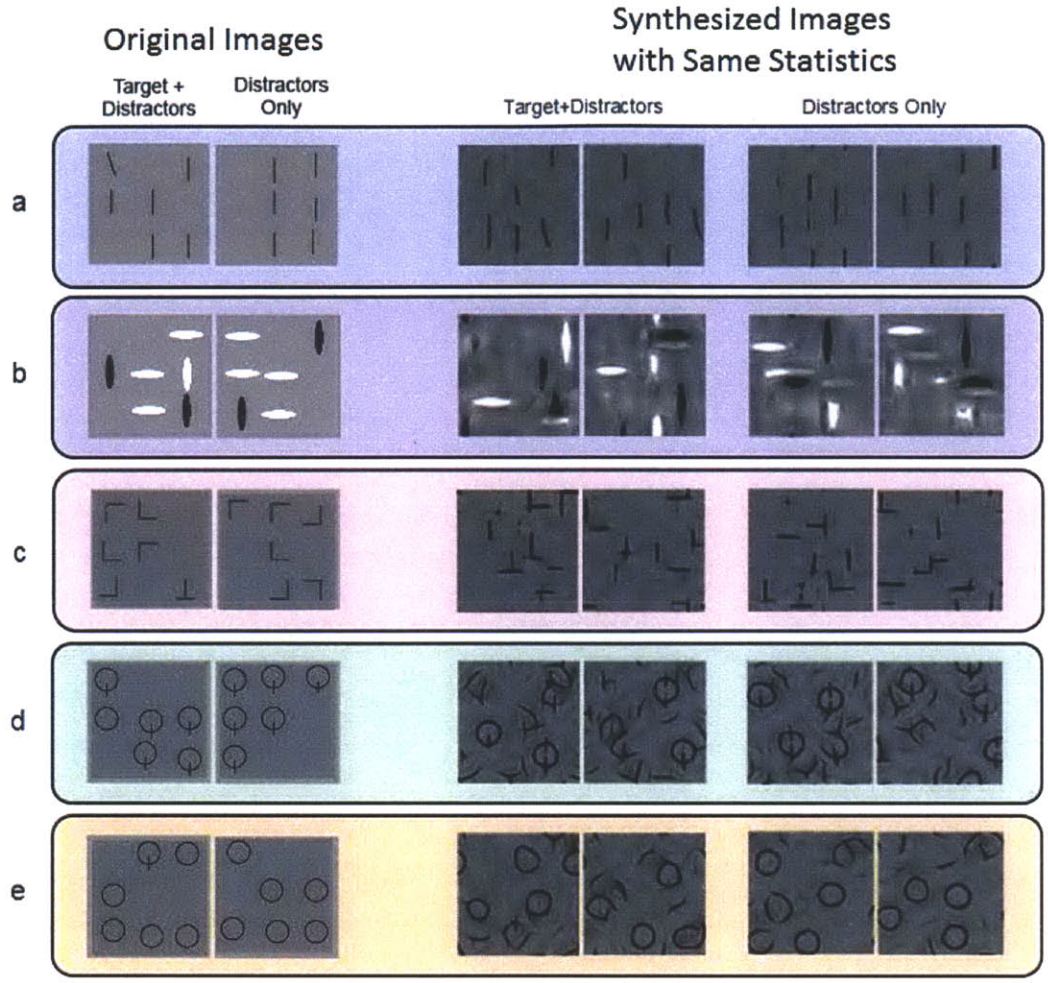


Figure 4-3: : Example target+distractor and distractor-only patches (columns 1 and 2) for five classic visual search conditions: (a) tilted among vertical; (b) orientation-contrast conjunction search; (c) T among L; (d) O among Q; and (e) Q among O. For each patch, we synthesized 10 images with approximately the same summary statistics as the original patch. Examples are shown in the rightmost 4 columns, at increased contrast, for visibility). In Experiment 2, observers viewed each synthesized image for unlimited time and were asked to categorize them according to whether they thought there was a target present in the original patch.

During each trial of the mongrel task, a mongrel was presented at the center of the computer screen until subjects made a response. Each mongrel subtended 3.8 x 3.8 degrees v.a. at a viewing distance of 75 cm. Subjects were asked to categorize each mongrel according to whether or not they believed a target was present in the original patch. We wanted to determine the inherent difficulty in discriminating tar-

get+distractor from distractor-only patches using summary statistics, and therefore chose to optimize observer performance at this task so subjects had unlimited time to freely view the syntheses. Observers viewed the mongrels at increased contrast, as shown in Figure 4-3.

Each of the five conditions (corresponding to one of our search tasks) had a total of 100 target+distractor and 100 distractor-only patches to be discriminated in this mongrel task, with the first 30 trials (15 target+distractor and 15 distractor-only) serving as training, to familiarize observers with the nature of the stimuli. Observers received auditory feedback about the correctness of their responses throughout the experiment.

4.5.3 Results

Performance of the mongrel task in each condition was described by discriminability, d', computed in the standard way, using the correct identification of a target+distractors mongrel as a Hit and the incorrect labeling of a distractor-only mongrel as a False Alarm,

$$d' = z(\text{Hit rate}) - z(\text{False Alarm rate})$$

where $z(p)$ indicates the z-score corresponding to proportion p. This measure of the discriminability of the mongrel images gives us an estimate of the discriminability of target+distractor from distractor-only patches based on their summary statistics, and from here on out we will refer to this d' as the statistical discriminability.

4.6 Discussion

Our model proposes that to a first approximation, discriminability based on summary statistics should predict whether a given search task is difficult or not. Specifically, the model predicts that when distractor-only patches have summary statistics similar to target+distractor ones, the corresponding search task should be difficult. To examine

our model's prediction, we carried out correlation analysis for each task's search reaction time slope and corresponding statistical discriminability.

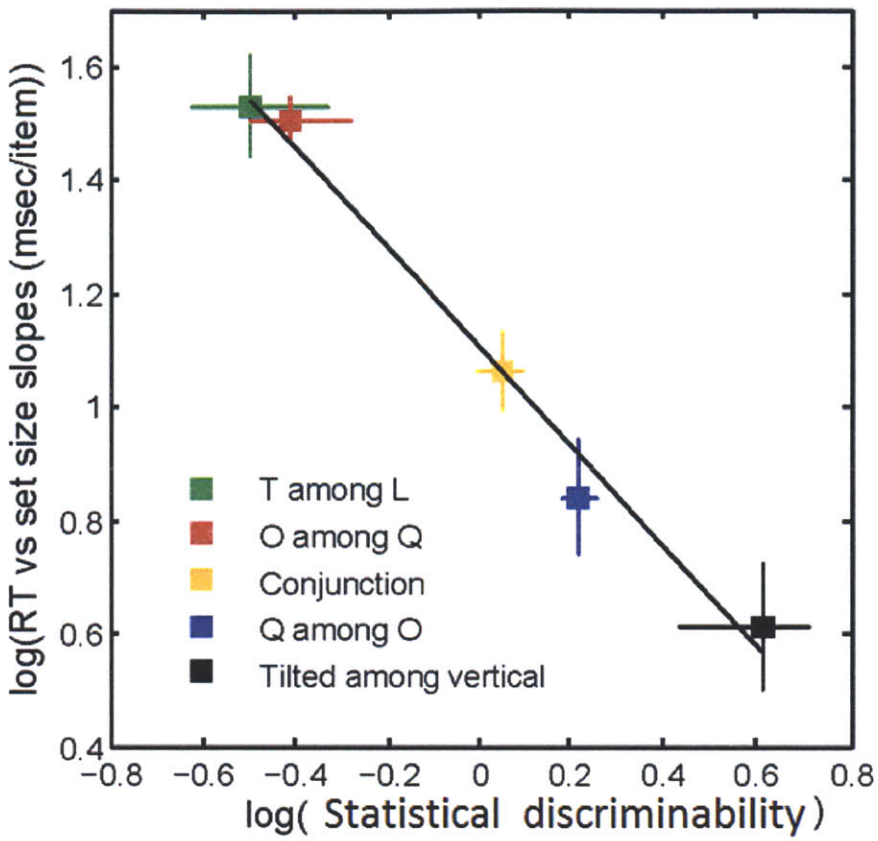


Figure 4-4: The correlation of the log of the search slopes to the log of the statistical discriminability ($R^2 = .99$)

Figure 4-4 plots log(search slope) on these five tasks versus $log(d)$ from our mongrel experiment. The data shows a clear relationship between search performance and our measure of the statistical discriminability of target+distractor from distractor-only patches ($R^2 = .99$, $p < 0.001$). The significant relationship echoes the insights gleaned from viewing the synthesized images and agrees with our predictions. When it is difficult to discriminate between target+distractor patch statistics and distractor-only statistics, search is slow; when the statistics are easy to discriminate, search is fast. The results of these experiments demonstrate the feasibility of thinking of visual search in terms of a summary statistic representation in peripheral vision.

4.6.1 Varying the number of items in each patch

It should be noted that the patches used in the statistical discrimination experiment are not necessarily the types of patches observed in pooling regions during a visual search trial. In particular, the number of items are not always six, but may vary from one to many. This variable was investigated by a collaborator, Jie Huang, and those results (along with the results presented here) are published in [44]. The findings are that in general, as the number of items increase, the statistical discriminability decreased for the various conditions. These discriminabilities based on the number of items are used by the quantitative model described in Chapter 5.

4.7 Future Work

While the statistical discriminability correlates well with human performance in search tasks, additional conditions should be tested to verify these initial findings. When sufficiently many experiments are conducted, we will eventually be able to better characterize the space of summary statistics and its relationship to perception of stimuli in the periphery, and soon bypass the need to run subjects on a synthesis discrimination task like in Experiment 2.

4.8 Conclusion

In this chapter, we observe that when we take peripheral limitations into account, the puzzles of visual search seems more straight-forward. Single item discriminability does not predict visual search performance because of the same reason that crowding effects occur. It makes sense that search asymmetries occur, because we are comparing target-present patches (Q+OOO) to target-absent patches (OOOO), and so asymmetries are to be expected. A simple model considering the limited discriminability in peripheral vision predicts how difficult five classic visual search conditions are.

Chapter 5

Modeling Visual Search

5.1 Outline

In Chapter 4, we showed that the statistical discriminability correlates well with the mean reaction time for humans to find the target in visual search tasks. If the difficulty of a search task is constrained by how discriminable stimuli are in the periphery, then ordinary free-viewing search involves gathering information in each fixation and saccading to a new location roughly every 200 ms until the target has been found. So, one can alternatively measure the difficulty of a free-viewing visual search task by counting the number of fixations a subject makes to find the target. In fact, mean reaction time is strongly correlated with the number of fixations subjects make in finding a target in visual search trials [61].

We would like to eventually be able to estimate how many fixations it would take on average to find the target in an arbitrary display. Towards that goal, we describe a model of visual search that can estimate the number of fixations needed to find a target, given the discriminabilities for the search condition being tested, the number of items in the display, and the experimental parameters for how items are placed. The purpose of this chapter is to evaluate whether a plausible model of visual search based on the peripheral limitations described in the previous chapter can be constructed.

There are a number of possible choices and parameters in such a model. Some

considerations about how the model should be constructed are discussed and tested. We show that it indeed is possible to create a model that can replicate human performance in the visual search experiment, but caution that more research and data is necessary to draw deeper conclusions.

The work in this chapter presents research I conducted in collaboration with Ruth Rosenholtz, Livia Ilie, and Jie Huang. Rosenholtz and Ilie started some initial work on this model, and Huang added considerations of number of items in a pooling region, as well as a preliminary (pit-stop) model of saccade limitations. All other considerations, analysis, designs, and model derivations presented in this chapter are my contributions.

5.2 Introduction

In Chapter 4, we have argued that the visual system's task in visual search is not to distinguish between individual targets and distractors, but rather to discriminate between sizable, crowded target+distractor patches and distractor-only patches. As has been repeated throughout this thesis, we argue that those patches are represented by a rich set of summary statistics. The experiments in Chapter 4 lent credence to this view of search, by demonstrating that statistical discriminability of crowded target+distractor and distractor-only patches can predict the qualitative difficulty of a set of classic search tasks.

Can a peripheral vision plus eye movements story account for search? Zelinsky has previously provided evidence that eye movements, rather than attention, may underlie natural search tasks [61]. In this chapter, we further test this hypothesis by developing a model that makes quantitative predictions of eye movements during visual search.

5.3 Modeling Visual Search

On each fixation, the model collects local summary statistics throughout the visual field and chooses the next fixation based on the available information. This process continues until the observer finds the target. In this chapter, we test the feasibility of a model of free-viewing search in which information is limited by a summary statistic representation in peripheral vision, and the primary method of gaining new information is to move ones eyes to a new fixation location.

We identified four questions about how such a model should work. (1) How does the model perform if it simply uses a heuristic to identify where next to saccade, as opposed to computing the maximum a posteriori solution of where next to saccade given the evidence? (2) Humans are known to make multiple saccades to move their eyes to a far away location [61]. How should the model implement this constraint, and does it affect the model results? (3) There is debate on whether or not memory is used in visual search. How does the model perform with and without memory? (4) How does changing the density of pooling regions affect the model?

5.3.1 Ideal vs Heuristic

Previous ideal observer models of search [35, 51, 21, 8] have assumed that the main limit on peripheral information, if any, was due to changing contrast sensitivity function with eccentricity (e.g. [21]). The main limiting factor in our search displays is visual crowding. The jumbling of features between neighboring items in crowding implies that in the presence of a target, multiple pooling regions may see that target. This situation is incompatible with previous ideal observers, and so we derive a new ideal observer model below.

Ideal model

We hypothesize that at a given instant, the visual system measures, in parallel, noisy estimates of the targetness from a number of overlapping pooling regions across the visual field, as shown in Figure 5-1. Targetness here is an abstraction of how much

the collected statistics in a pooling region resemble a target+distractor patch as compared to a distractor-only patch. Some pooling regions will be more discriminative than others, depending upon the degree of crowding – a function of both the search condition (i.e., which target and which distractors, as seen in Chapter 4) and the number of items in a pooling region.

For uncrowded pooling regions (numerosity = 1 or 0), we use a single $d\prime$ for all conditions, reasoning that for the five search conditions tested in the previous chapter, acuity was a minimal issue, and thus all uncrowded items were equally discriminable with a $d\prime$ of 5. We use this same large $d\prime$ to generate observations for empty pooling regions, reasoning that the observer should easily be able to tell that a pooling region contains no search items, and thus no target. Note that this assumption has more recently been called into question and is the subject of some ongoing research. This is discussed in the section on Future Work for this chapter 5.5.

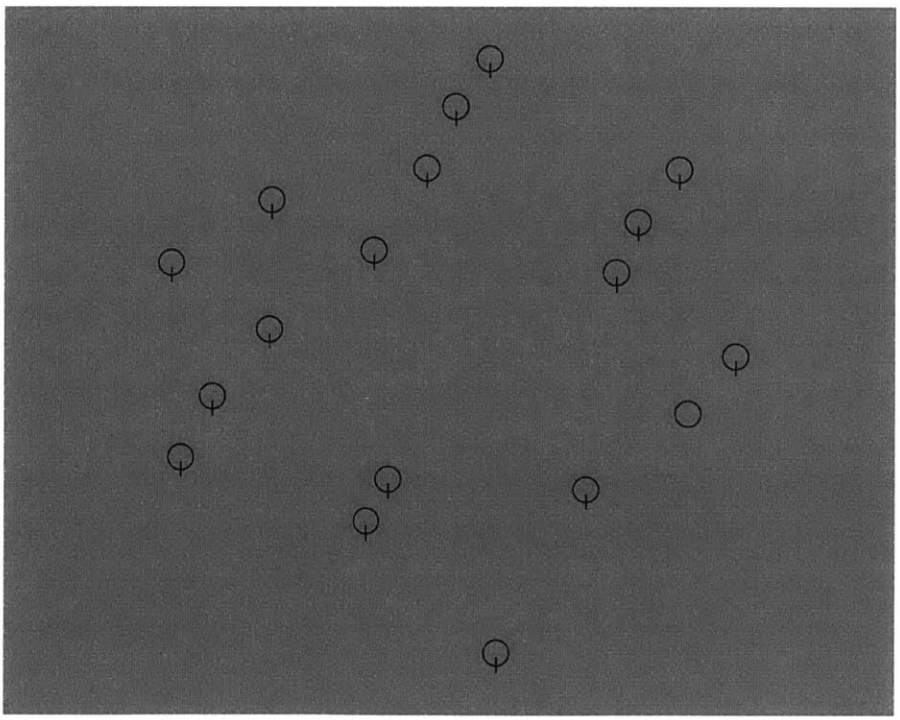


Figure 5-1: The model measures noisy estimates of "targetness" from overlapping pooling regions across the visual field.

Given measures of targetness, where should the observer fixate next? Visual search

researchers have developed two main kinds of ideal observers for predicting fixations. An ideal saccadic targeter, as in [51], always makes fixations to the location with the maximum a posteriori (MAP) probability of containing a target. An ideal searcher, as in [21], maximizes the probability of identifying the target on the next fixation. As a result, an ideal searcher sometimes fixates at a location between neighboring candidate targets so as to simultaneously resolve them both. In this thesis, consider two variations of the saccadic targeter: simply moving to the location with the highest average targetness of overlapping pooling regions (heuristic), and the ideal saccadic targeter that moves to the location with the highest posterior probability of being the target location given the evidence.

After choosing the location to saccade to next, this process repeats until the model fixates sufficiently close to the target to unambiguously recognize it; say, within 1 degree of the target. Note that our model does not know the locations of the items in the display, but instead has to infer the target location only from the observations in the periphery.

Derivation of Ideal Model

The model takes as input, a vector of noisy observations, one for each of the N pooling regions: $O = \{o_1, o_2, ..., o_N\}$. As per the ideal saccadic targeter model [51], we select the maximum a posteriori location of the target given the observations.

$$argmax_{(x,y)} P(\text{Target location:} T = (x,y) \mid \text{Observations:} O = \{o_1, o_2, ..., o_N\})$$

Applying Bayes' Rule, and canceling terms which are constant with respect to (x, y), we get:

$$= argmax_{(x,y)} \frac{P(O = \{o_1, o_2, ..., o_N\} \mid T = (x,y)) P(T = (x,y))}{P(O = \{o_1, o_2, ..., o_N\})}$$

We assume that the observations of the different pooling regions are independent, conditioned on whether the pooling region contains a target, and on the numerosity

of the pooling region. This means we can simplify the above equation to:

$$= argmax_{(x,y)} \prod_{j=1}^{N} P(o_j \mid T = (x, y))$$

We can further simplify this equation by dividing the set of pooling regions into those that do ($C_{(x,y)}$) and do not ($\bar{C}_{(x,y)}$) contain the location (x,y).

$$= argmax_{(x,y)} \prod_{j \in C_{(x,y)}} P(o_j \mid T = (x, y)) \prod_{j \notin C_{(x,y)}} P(o_j \mid T = (x, y))$$

The observations from pooling region, j, not containing the target are normally distributed with a unit variance, and mean of $\mu_{\text{absent},j} = a_j$. Observations from any pooling region, j, that contains the target is distributed with a mean of $\mu_{\text{present},j} = a_j + d_j\prime$. Here, $d_j\prime$ is the discriminability between target-present and target-absent patches for pooling region, j, which is a function of the numerosity of the pooling regions as well as the search condition. The offset, a_j, is unknown, but as we will later show, the model predictions are independent of the choice of a_j. By plugging in these conditional probabilities and taking the log, we obtain:

$$= argmin_{(x,y)} \sum_{j \in C_{(x,y)}} (o_j - a_j - d_j\prime)^2 + \sum_{j \notin C_{(x,y)}} (o_j - a_j)^2$$

We can implement this ideal observer model by evaluating this equation for every hypothesized target location (x, y) and selecting the hypothesis that yields the minimum sum. However, all locations that share the same pooling region membership also share the same value in the sum in that equation. For example, if locations (x_1, y_1) and (x_2, y_2) are both within pooling regions m and n, and in no other pooling regions, then $P(T = (x_1, y_1) \mid O) = P(T = (x_2, y_2) \mid O)$. To the model, these two points are equally good (or bad) choices for the next fixation. We address this in our implementation by directing the model to fixate at the center of mass of the set of points yielding the MAP solution.

Further Intuitions about the Ideal Saccadic Targeter

The MAP decision rule is described in the previous segment. In running the model, one generates observations for each pooling region. For a pooling region, j, that contains the target, we can write the observation $z_j + a_j + d_j\prime$, where z_j is a standard normal random variable. The observation for a pooling region, k, that does not contain the target can be written $z_k + a_k$. A real observer will not know a priori which pooling regions contain a target, but by plugging in these observations, we can gain additional insight into the implications of the MAP decision rule. Pooling regions fall into four categories, according to whether the presence or absence of a target in that region is consistent (=) or inconsistent () with a given candidate target location, (x, y): (1) (T, =) Pooling regions containing the target, and also the candidate location, (x, y). (2) (no T, =) Pooling regions containing neither the target nor location (x,y). (3) (T, $\neq$) Pooling regions containing the target, but not (x, y), and (4) (no T, $\neq$) Pooling regions containing (x,y) but not the target. Splitting the MAP decision rule into these four categories yields the following:

$$= argmin_{(x,y)} \sum_{j \in (T,=)} (z_j + a_j + d_j\prime - a_j - d_j\prime)^2 + \sum_{j \in (noT,=)} (z_j + a_j - a_j)^2 + \sum_{j \in (T,\neq)} (z_j + a_j + d_j\prime - a_j)^2 + \sum_{j \in (noT,\neq)} (z_j - d_j\prime)^2$$

The terms corresponding to consistent pooling regions (containing either both the target and (x, y) or neither) are identical. Since z_j is a standard normal variable, and thus symmetric about 0, the same is true for inconsistent pooling regions. Therefore, collapsing across target presence, we obtain:

$$= argmin_{(x,y)} \sum_{j \in Consistent} z_j^2 + \sum_{j \in Inconsistent} (z_j - d_j\prime)^2$$

All the offsets, a_j, cancel out in the computation. Thus, the unknown a_j terms do not contribute to the model. When determining the (x, y) that minimizes the sum in this equation, each pooling region consistent with a hypothesized target location (x, y) incurs a small noise penalty, in that any value sampled from the standard

normal random variable, z_j, is squared, and counts as "evidence" against that (x, y) position being the target location. An inconsistent region, on the other hand, on average incurs a greater penalty due to that pooling region containing point (x, y) but not the target, or vice versa because the value $d_j\prime$ is added to z_j and is squared. For small $d\prime$, the contribution of the two types of terms (consistent vs inconsistent) is similar, which makes it difficult to discriminate between regions containing the target and those which do not. As $d\prime$ increases, inconsistent pooling regions are more heavily penalized, making it easier to distinguish candidate locations likely to contain the target from those that are not.

Heuristic model

The heuristic model is much simpler than the ideal model. It simply decides to saccade to the location with the highest average targetness of all the pooling regions that contain that location. There may be other heuristics one could imagine, for example, selecting the center of the pooling region that produced the strongest response. However, this choice requires many more pooling regions in order to localize items well. It is also impractical to test all possible heuristic models, and so we test a choice that is intuitive and simple.

While the ideal model is independent of the offset of the targetness value, the heuristic model is not. Because we have several values for $d\prime$ depending on the number of items in a pooling region, we place the centerpoint between the target present and target absent distribution at 0. Then the heuristic model simply decides to saccade to:

$$argmin_{(x,y)} \frac{\sum_{j \in C_{(x,y)}} o_j}{\mid C_{(x,y)} \mid}$$

where $\mid C_{(x,y)} \mid$ corresponds to the number of pooling regions that contain (x, y).

5.3.2 Saccade Length

Beyond the limitations on search performance due to the available information in the periphery, the visual system may operate with additional constraints on saccades. Previous research has reported that one requires more eye movements to acquire a stimulus at larger eccentricities (8 degrees v.a.) than closer to fixation (4 degrees v.a.) [61]. This is also supported by the phenomenon of normal hypometria (not making a long enough saccade to reach a target location) often observed in saccades to distant targets [13].

Without any constraints on saccade length, the model would make a single saccade to the MAP location, regardless of the distance. To account for the observation that the visual system prefers shorter saccades, we considered several methods for imposing a saccade limitation: (1) no limitations; (2) a "pit stop" model that forces the saccader to only saccade within 3, 7, or 11 degrees; or (3) an exponential cost model which applies a penalty that is a function (exponential distribution density function) of the saccade length. Figure 5-2 illustrates the various saccade model choices.

In the "pit stop" model, whenever the model chooses a location that is too far from the current fixation, it fixates at a number of intermediate locations until it reaches the planned location, such that no eye movements exceed the limit on saccade length. For instance, if the desired saccade location is 10 degrees away, and the saccadic limit is 7 degrees, the model will make a saccade in the direction of the MAP location, but only travel 7 degrees in that direction. The model will then make its next saccade the rest of the way to the desired location.

The exponential cost model applies a multiplicative cost of saccading to any location before deciding on the location to saccade to next. The curves are given by the equation $\lambda e^{-\lambda x}$. The three curves indicate three different choices of parameter for the exponential cost function. The parameter controls the strength of inhibition for the various saccade lengths.

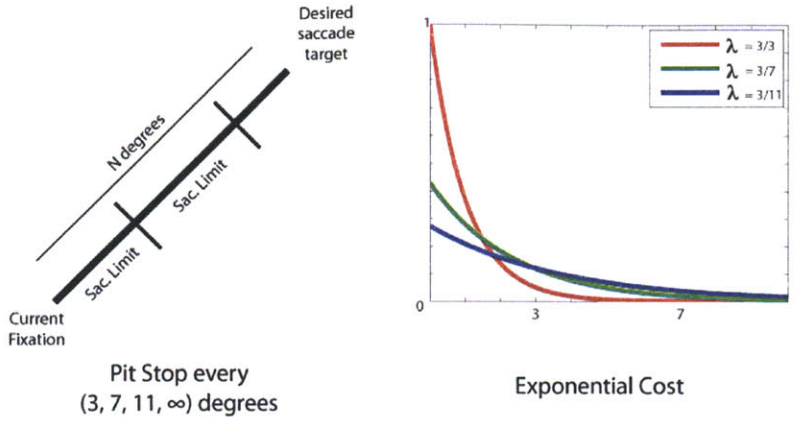


Figure 5-2: Various methods of imposing a saccade cost

5.3.3 Memory

Do humans use memory of previous fixations to guide search? An open issue is what, if any, memory is available to the visual system during search. This issue has been the source of some debate. For some serial search tasks, with a moderate number of items (≤ 16), search appears to be memory-less: Horowitz and Wolfe [19] showed no costs in search efficiency when search items are relocated continuously during search, suggesting no memory for search. However, with greater set sizes, there is evidence suggesting that memory for locations may play a role in search [23].

We look at the effect of allowing the model to recall observations from the last K fixations for $K = 0, 1, 4, \infty$. This was done by letting the model have access to the past K observations when deciding where next to saccade. In the case of $K = \infty$ memory, observations can be perfectly integrated over time by maintaining a map of the posterior probability (i.e., likelihood of each pixel (x, y) being the target, and not of every item), and updating it each time the model receives new observations.

5.3.4 Pooling Region Density

We know little about the number and overlap of the pooling regions that the visual system may process in parallel. Fewer pooling regions, which overlap less, would lead

to less information available to the model, and as a result lead to predictions of more fixations required to find the target.

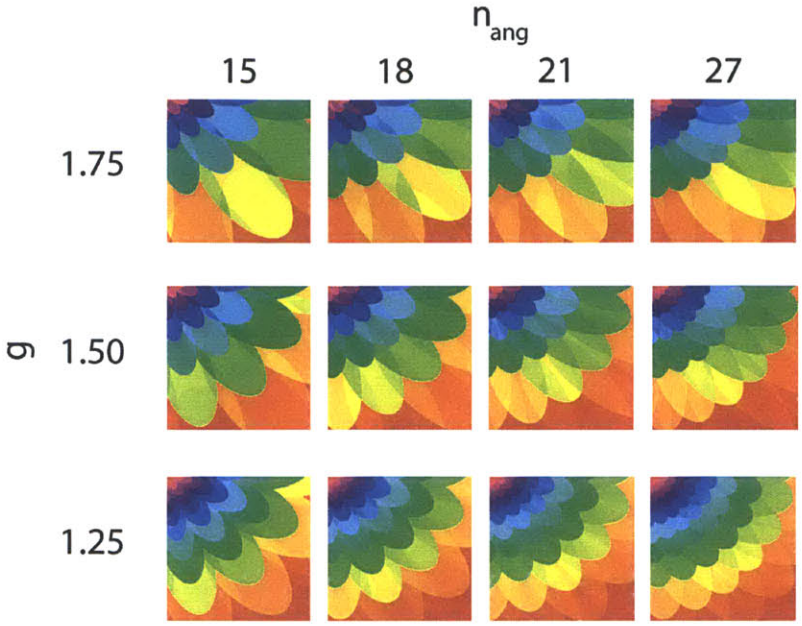


Figure 5-3: Visualization of pooling regions in a patch from the visual field, as the pooling region placement parameters are varied.

Pooling regions in this model are laid out in a log-polar grid. Two parameters control how densely the pooling regions span the visual field: (1) n_{ang} number of pooling regions in each ring, and (2) g the ratio between two successive rings' distance. Figure 5-3 depicts the density of the pooling regions of a patch as the two variables are varied. We observe how the model performs with that set of parameter choices.

5.4 Experiment

Using the discriminabilities of the five classic visual search tasks measured in the previous chapter, we evaluate how the model performs, as we vary the model choices enumerated above. For each set size, condition, and parameter setting, we run 1000 Monte Carlo simulations of the model to estimate the number of fixations needed to find the target in that situation. Model results are compared against subjects' eye

movement data from the previous chapter.

5.4.1 Results and Discussion

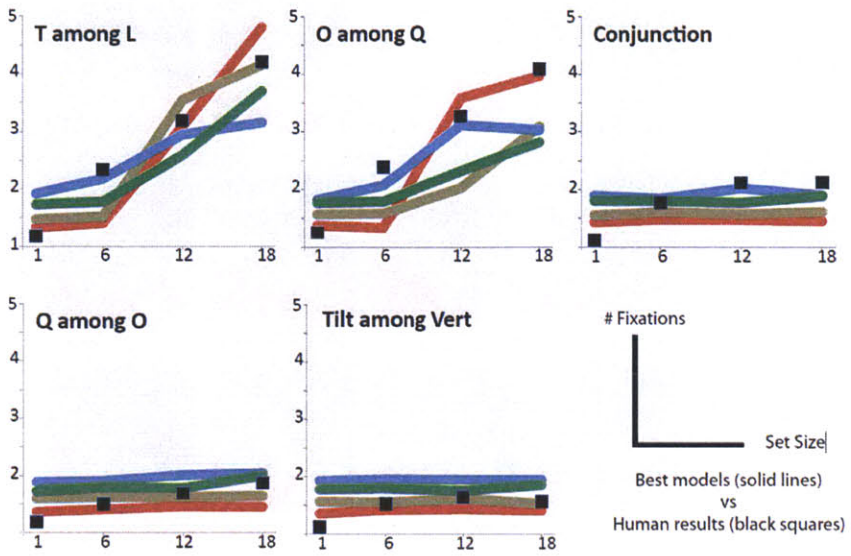


Figure 5-4: Results of some of the best fitting models

We can find some parameters in the model to match human data. Across all the possible models choices, the models that best mimicked human performance are shown in figure 5-4. The top models have in common some parameter choices: using the smooth exponential cost for saccades ($\lambda = 1$), and using the heuristic rather than the ideal integration rule. They varied most in whether the last observations were remembered or not, and in pooling region density.

There are only 20 data points, but many possibilities for a choice of model, in addition to all the implicit choices in the model that had already been made. The number of data points are simply too few to do any quantitative analysis about what the true parameters should be. However, we are still able to perform some qualitative analysis of how the parameters affect model performance. We can get a sense of how the model predictions vary with the various choices for the model. This is described below.

5.4.2 Ideal vs Heuristic

The mean squared error between the model data and the predictions of the best performing ideal model was 0.6, while the best performing heuristic model had error 0.2. The lower error for the heuristic model indicates that the heuristic based model performs more like humans do. When model performance was averaged across all other parameters, the ideal decision model found the target in 1.2 fixations (averaged over search conditions), while the heuristic found the target in 1.3 fixations. The ideal model is too efficient.

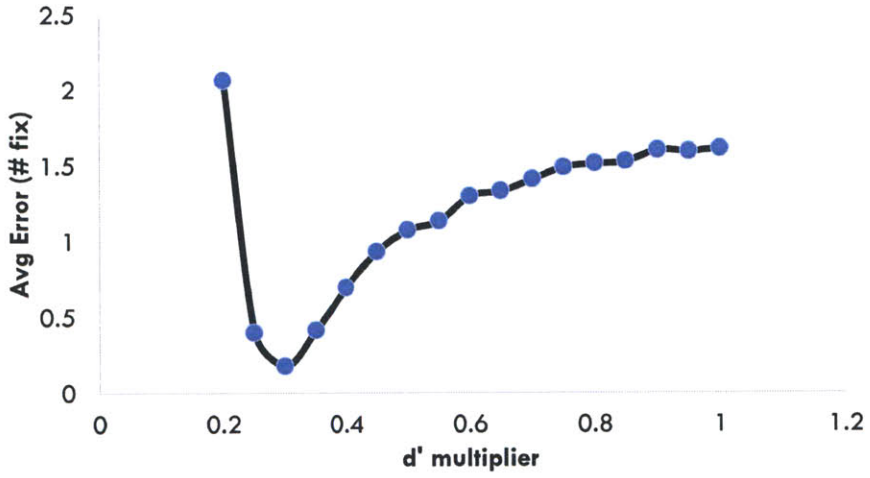


Figure 5-5: Error of the ideal decision model as experimental $d\prime$ is scaled

Perhaps the ideal model is too efficient because the $d\prime$ measurements were artificially high. Figure 5-5 shows the mean squared error as the experimental $d\prime$ values were multiplied by a scaling factor, so the ideal decision model is more viable if we are able to account for a .3 scaling factor for the experimentally derived $d\prime$ values. Uncrowded pooling regions may need to be estimated in a separate experiment. One can make mistakes in viewing a single item peripherally, especially for complicated stimuli. There is evidence that some objects are subject to self-crowding [29], and perhaps some items in our search displays are as well. To be certain, experiments are necessary to test this possibility.

5.4.3 Saccade Constraints

Figure 5-6: Normalized Search Time for the Various Saccade Rules

Figure 5-6 shows the normalized search time (average number of fixations needed to find the target, averaged across all other model parameters). The graph indicates that when the pit stop model can saccade to about 11 degrees or more (up to ∞) search is extremely efficient, but when the pit stop model can saccade to 3 degrees away, the model is much slower. Saccading at most 7 degrees makes the pit-stop model perform similarly to the exponential model. The choice of parameter in the exponential model has little effect on performance.

5.4.4 Memory

Figure 5-7 shows the normalized search time (average number of fixations needed to find the target, averaged across all other model parameters). The graph indicates that the mean effect of memory, averaged across all other model parameters, is not strong. This may be because subjects tended to find the target in about 4 fixations in the worst case. If the model finds the target quickly, then there will not be a large difference between remembering observations from a few previous fixations, as compared to an infinite number of fixations. We may find a more important role for

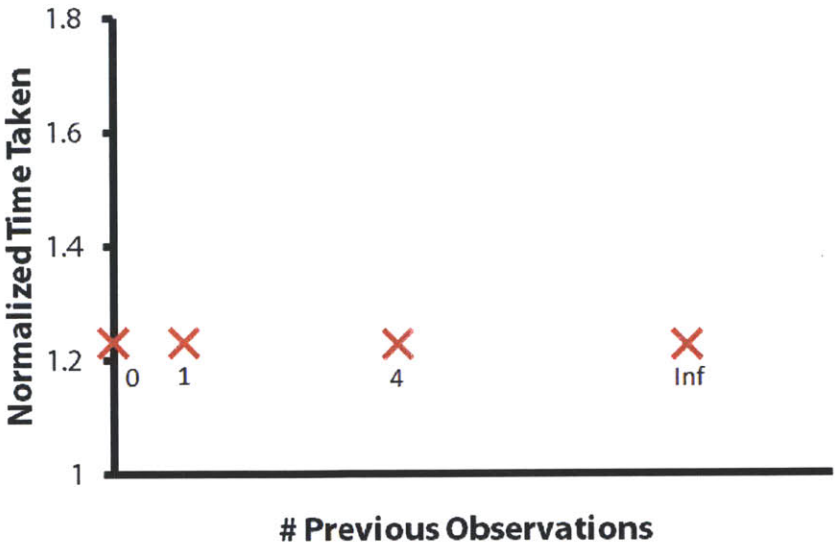


Figure 5-7: Normalized Search Time for the Various Amounts of Memory

memory in visual search tasks that are more difficult than the ones tested.

5.4.5 Pooling Region Density

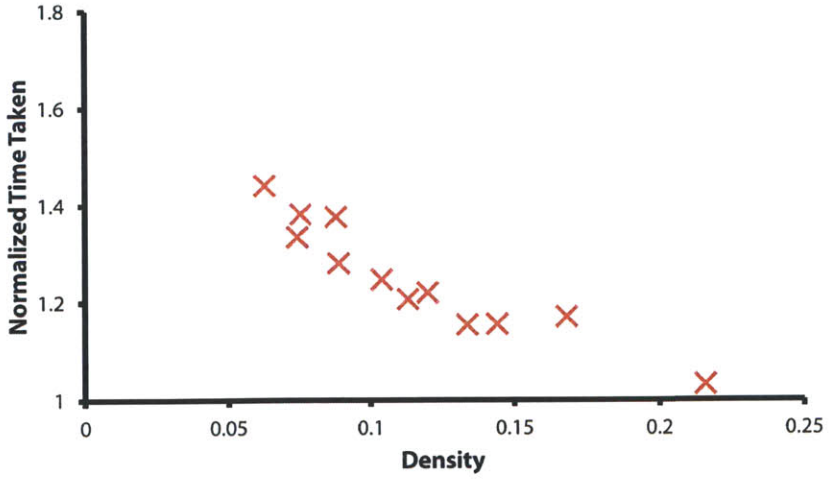


Figure 5-8: Normalized Search Time for the Various Amounts of Pooling Region Density

Figure 5-8 shows the normalized search time (average number of fixations needed to find the target, averaged across all other model parameters). The graph indicates

that as the density of pooling region increases, the efficiency in finding the target improves as well.

5.5 Future Work

As mentioned in the previous chapter, more visual search conditions will help verify the model. Testing different conditions can help provide a larger dataset against which to verify the model. Varying item placements can distinguish between the various models of pooling region placement and help narrow down choices for those model parameters. The role of memory can be better tested by running visual search experiments that are more difficult; the more fixations needed to find a target, the greater the divergence will be between the performance of models that do and do not use memory. Additionally, the d' for pooling regions with only one item need to be experimentally measured.

In the models presented, a location is either contained in a pooling region or not, regardless of whether the location is in the middle of the pooling region or on its edge. It seems unlikely that the visual system functions over pooling regions that have hard boundaries. A weighted ownership model might be a better description of how the visual system functions.

5.6 Conclusion

In this chapter, we introduced a search model, and take a step towards a general purpose model that can work on arbitrary images and targets. A derivation of an ideal saccadic targeter is shown, and various modeling considerations are evaluated on the data from the previous chapter. Our modeling results show that it is possible to construct a model in which fixations are guided by summary statistic information, gathered in parallel across the visual field from overlapping pooling regions. We additionally showed various effects on search efficiency when parameter choices are varied.

Chapter 6

Visual Illusions

6.1 Outline

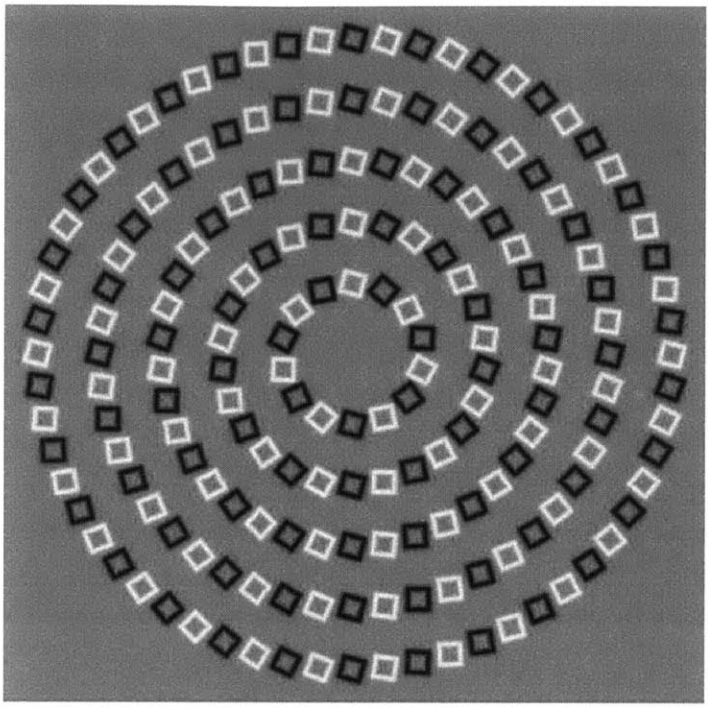
In this chapter, we explore another instance where statistical ambiguities in the periphery can affect perception. If the peripheral representation allows for ambiguities that make it difficult for the visual system to know what is out there, the visual system may make mistakes. We would expect to find illusions due, at least in part, to the limitations of peripheral vision.

Consider the Pinna-Gregory illusions [40] shown in Figure 6-1 (a) and (b). These images consist of black and white squares arranged in concentric circles. This is easily proved by tracing the circles with a pen. However, the percept is not of concentric circles. In Figure 6-1a, the squares seem to be arranged in a spiraling vortex. Rotating the squares in every other ring yields an illusion of intertwining curves seen in Figure 6-1b. We shall refer to Figures 6-1a and 6-1b as the spiral illusion and intertwining illusion, respectively.

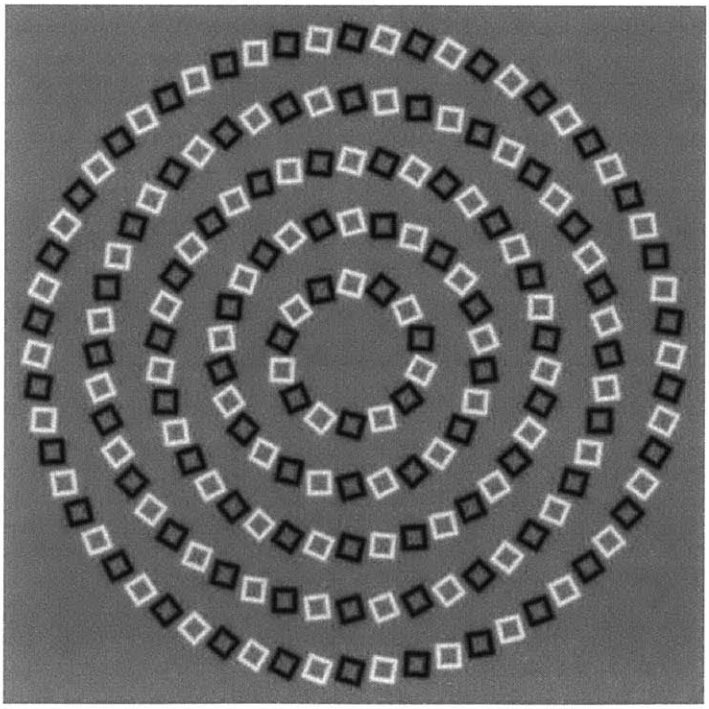
We show that our statistical model of peripheral information predicts how polarity, item width, and angle of tilt affect the perception of the illusion. These predictions are tested in an experiment that queries how illusory each image is within a dataset. We find that the results agree with the predictions. Furthermore, when we visualize the peripheral information as described in Chapter 3, we observe many qualities from our perception of the illusion. While we are not able to provide a full account of all aspects of this illusion, we are able to provide many intuitions that lead to predictions about these illusions.

All the work presented in this chapter is research I conducted myself, under the primary supervision of Ruth Rosenholtz and occasional discussions with Benjamin Balas.

6.2 Pinna-Gregory Illusions



(a) Spiraling Illusion



(b) Intertwining Illusion

Figure 6-1: Pinna-Gregory Illusions

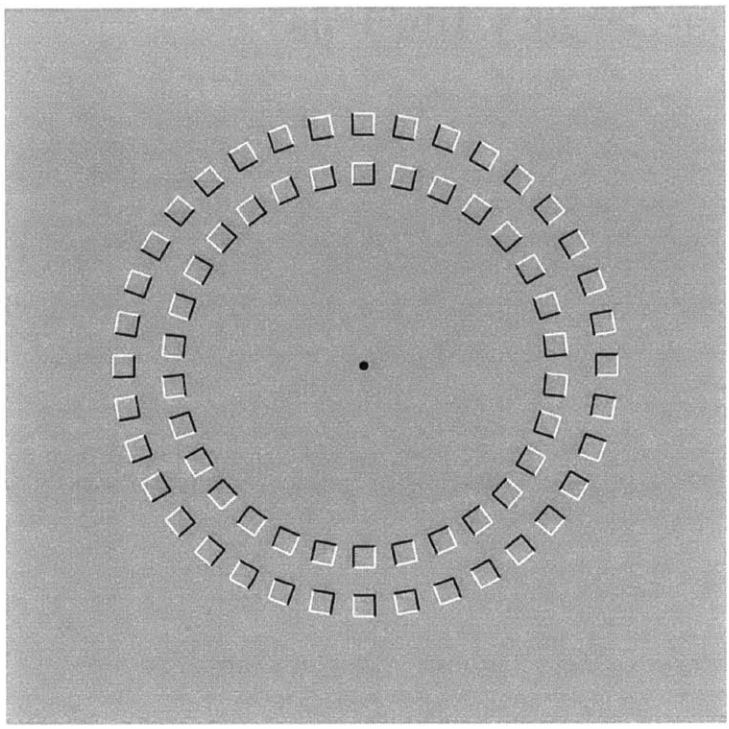


Figure 6-2: The Pinna Illusion.

One might notice that there are illusory motions as eye movements are made, or when moving towards or away from these figures. Pinna has also discussed a motion illusion [39] in which there are two rings of squares, and when an observer moves towards the figure, the inner and outer rings rotate in opposite directions (see Figure 6-2). He suggests that peripheral motion detection mechanisms in conjunction with grouping mechanisms underlie both the Pinna illusion as well as the illusory motion observed in the spiraling and intertwining illusions. While motion can play a role in the illusions in Figure 6-1a and 6-1b, there remains a strong illusory percept even with minimal eye or head movements. For readers interested in the aspects of the illusion that are motion-related, we refer them to [39] for further details. Here we focus on only static aspects of the illusion.

What might underlie these illusions? Pinna and Gregory [40] proposed that the illusory percepts were due to global integration elicited by the Gestalt factor of good continuation. We agree in principle that the illusions are probably related to good continuation. However, this explanation is somewhat unsatisfying without further

specifying a model of good continuation and demonstrating that it does in fact predict the percepts. A Gestalt-based understanding of these illusions should specify how grouping by similarity, proximity, and good continuation interact to produce these percepts. Beyond this suggestion of the role of good continuation, to our knowledge no other attempts have been made to explain these illusions.

For both the spiral and intertwining illusions, the illusory percept is reduced at fixation, which should be accounted for by any explanation of the phenomena. This suggests that these illusions may be a by-product of peripheral visual processing mechanisms. Can the model of peripheral vision we have been developing in this thesis account for the illusory percepts in the Pinna-Gregory illusions?

The general model of peripheral vision predicts that images which share highly similar summary statistics are difficult for the human visual system to discriminate in the periphery. This prediction hints at a possible explanation for the illusion. Perhaps the statistics in the peripheral pooling regions of the spiraling and intertwining illusion are easily confused with statistics consistent with actual spiral-like contours, or intertwining-curves-like contours. When you have seen these statistics before, they were much more likely to come from a spiraling-vortex-like pattern or from intersecting contours, both of which happen with relative frequency, and were less likely to have come from concentric circles with a bunch of alternately colored tilted squares, which almost never happens. This is, however, difficult to test directly because it is hard to be sure exactly what the percept is.

6.3 Prior work

Fermuller and Malm [6] propose that uncertainty in visual processes cause bias in the estimation of lines and their intersections. In most situations, they argue that these biases are not noticeable, but they are highly pronounced in some illusions like the Zollner illusion. This uncertainty leads to ambiguity in where image features are spatially located, resulting in an illusory percept. In particular, if the process of edge localization is thought of as finding zero crossings of the derivative of the image, then

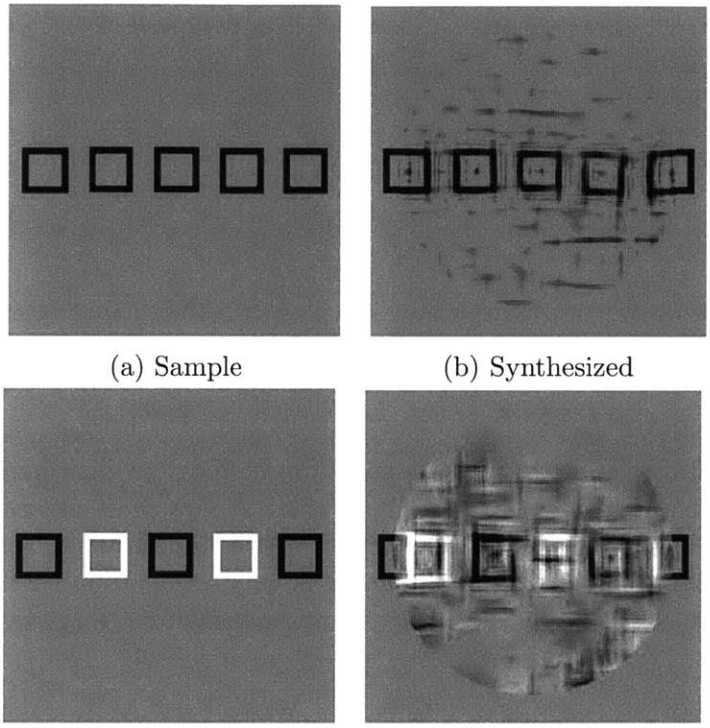
the edge location may vary depending on the scale of the derivative filters applied to the image. Noise and uncertainty in the neural processing were approximated by blurring the image with some Gaussian kernel whose width depended on the elements of the illusion they were analyzing. They showed that the uncertainty in estimating lines and junctions corresponded to the illusory percept. While their analysis was applied to the family of tilt illusions (Zollner, Fraser, etc), one can extrapolate and predict that the same uncertainty might underlie, or be at least related to the Pinna-Gregory illusion.

Fermuller and Malm argue that uncertainties are the underlying cause of the misperceived tilts, but in their account, the sources of the uncertainty are under-specified. They propose that the eventual perception of a line tilted in some direction was due to solving for the line in a least squares estimation problem, from the detected edge elements. In this manuscript, we supplement the Fermuller and Malm account by suggesting that the largest source of uncertainty is the periphery, and we additionally suggest that the statistical model we propose captures the nature of that uncertainty.

We are not the first to propose that the same peripheral mechanisms that underlie crowding may also underlie some illusions. Shapiro et al suggested that acuity loss cannot account for some peripheral motion illusions, and instead suggest that the perceived illusory motions are due to feature blur in the periphery [46]. Feature blur, as they have defined, refers to the combination of different features. In particular, they suggest that the motion energy of the original stimulus (first order motion energy) and the motion energy of the full-wave rectified contrast of the image (second order motion energy) are combined in the periphery so the motion signals are inseparable to their original feature sources. This blurring of features in the periphery, they note, is similar to the excessive feature integration account of crowding previously discussed, which we note, is related to the formulation of our statistical model.

6.4 Polarity

Recall from Chapter 3, that the periphery does not accurately represent phase in stimuli that have elements whose colors are on positive and negative sides of polarity of a background, and are relatively balanced. This suggests that perhaps the squares being black and white allow more ambiguous interpretations in the periphery.

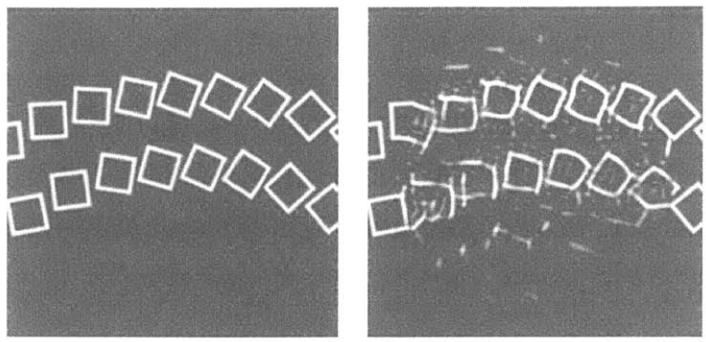


(a) Sample (b) Synthesized

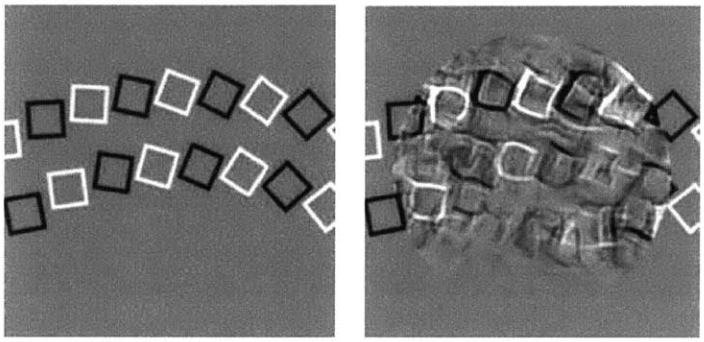
(c) Sample (d) Synthesized

Figure 6-3: The statistics in the black squares image in (a) are not ambiguous, which is why the synthesis in (b) reflects a good replication of (a). But when polarity variations are introduced in (c), the statistics of the black and white squares image show some ambiguity, as seen in the synthesis in (d). Some squares have both black and white edges, and there is more noise in the image. This suggests that the statistics allows some phase ambiguity and do not accurately represent the visual information in the original image.

Figure 6-3 shows that some ambiguities in the statistics may arise even in fairly simple contours, so long as polarity is varied. In more complicated contours with both a curved contour and tilted squares (Figure 6-4), variations of polarity in the image make it difficult for the set of statistics our model is based on to accurately represent



(a) Sample (b) Synthesized



(c) Sample (d) Synthesized

Figure 6-4: The statistics in the white squares image in (a) produce syntheses that are fairly unambiguous as seen in (b), but the statistics allow more errors when polarity is varied on the lines in (c), as can be seen from the visualization of those statistics in (d). Notice that the synthesis hallucinates a connection from the bottom line to the top line.

the image. The syntheses generated for Figures 6-3 and 6-4 were constrained to agree with the original image outside the central circular window, but were unconstrained in a central circular region. See Appendix B for more example syntheses of these patches.

When all the squares are the same color, there is less ambiguity. This leads to the model's first prediction: we should see less illusion when polarity is not alternated along contours. Indeed, turning all the squares white as in Figure 6-5 makes the illusory percept disappear. Instead, we perceive only concentric circles.

If polarity is important for this illusion, we can test it further by running an experiment to determine the relative illusory strength of variants of the Pinna-Gregory illusions that we modify to study the effect of polarity.

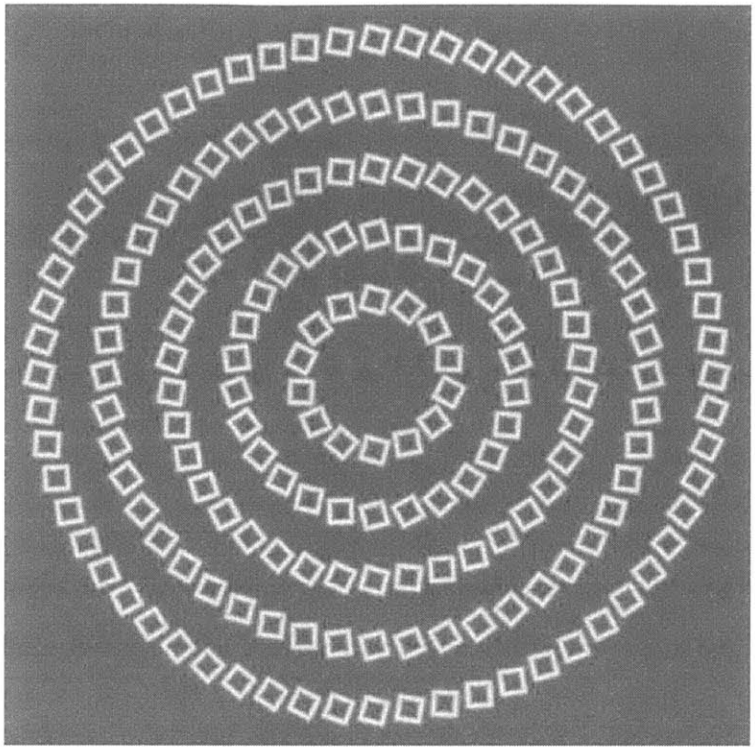


Figure 6-5: White Squares diminish the illusory effect of the intertwining stimulus

6.4.1 Experiment 1: Effects of Polarity

This experiment was conducted through the Mechanical Turk website.

Subjects

Subjects participated after indicating they consented to the task, and were compensated for their time. Subjects who reported vision that was not normal or corrected-to-normal were dropped from the study. Subjects on the Mechanical Turk website were allowed to participate in either one or both sets of images.

Intertwining Set: Thirty subjects participated, ages 20 to 65.

Spiraling Set: Thirty subjects, participated, ages 20 to 63.

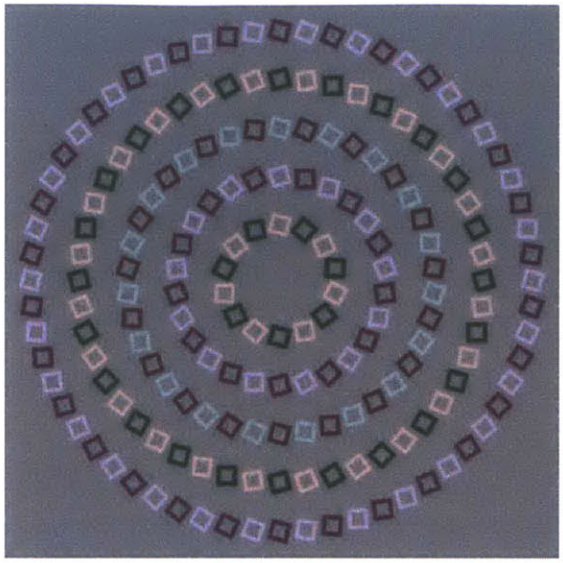
Method

The experiment was presented to subjects through their web browser. Subjects were asked to indicate which of two displayed images were more illusory. More specifically,

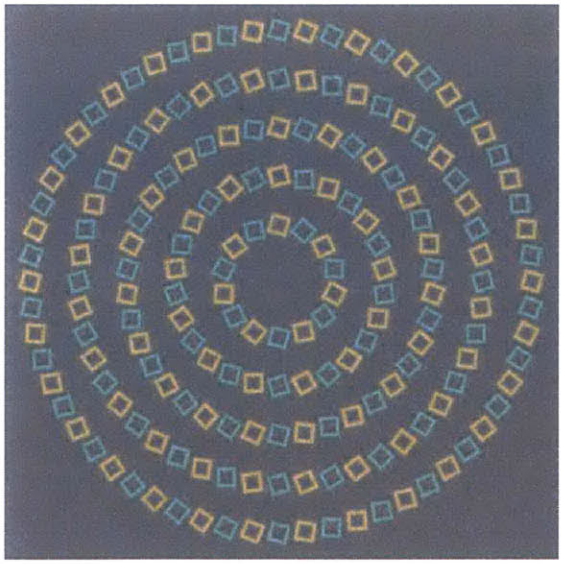
the subjects were asked, "The two images below are both made up of concentric circles. Please use your judgment to decide which looks more illusory (or which looks most like it's *not* just made up of concentric circles)." All possible image pairs within a given set were presented to the subject in a randomized order.

As with any experiment run over the internet, measures must be taken to ensure the quality of responses from participants. Subjects whose browsers were not able to display the stimulus without scrolling or were not able to display stimuli in color were not allowed to participate. Additionally, we inserted two types of quality measures: that of internal consistency, and ability to follow the instructions given. To measure internal consistency, each pair was presented twice, with the order reversed the second time that pair was shown. We also presented pairs from a standard set of images where there is no doubt which image should be selected as the more illusory of the pair in order to measure how well subjects were able to follow instructions. Subjects (a) whose scores for internal consistency fell below 75 percent in addition to responding too quickly, or (b) had reported the correct answer for less than 90 percent of the gold standard set were dropped from the analysis. Criteria (a) attempts to detect subjects who did not make a serious attempt to answer the question, and criteria (b) attempts to detect subjects who were randomly guessing.

In the intertwining set, 8 subjects were dropped for not meeting these standards, and in the spiraling set, 9 subjects were dropped. Figure 6-6 and Figure 6-7 show some of the images in the intertwining and spiraling set, respectively. Appendix A lists all the images used in these sets.

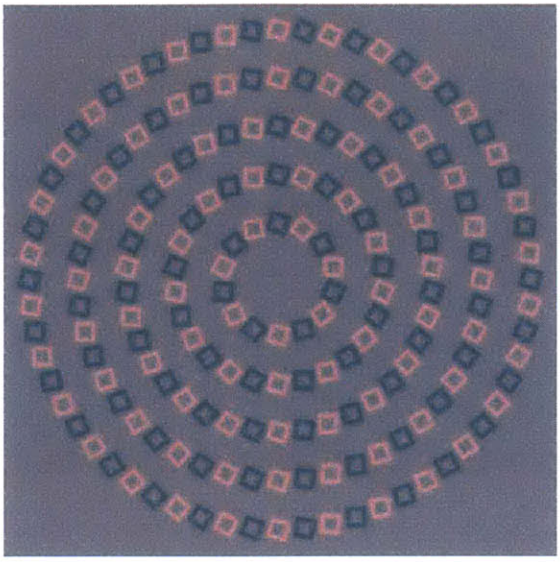


(a) A multi colored square image that alternates polarity along the rings

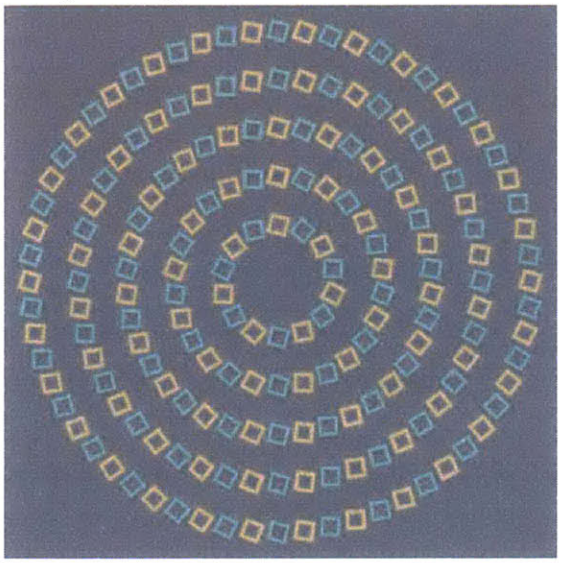


(b) A two color image whose colors are both on the positive side of polarity

Figure 6-6: One pair from the intertwining tilts polarity set



(a) A two color image that alternates polarity along the rings



(b) A two-tone image whose tones are both on the positive side of polarity

Figure 6-7: One pair from the spiraling tilts polarity set

Gold Standard Test

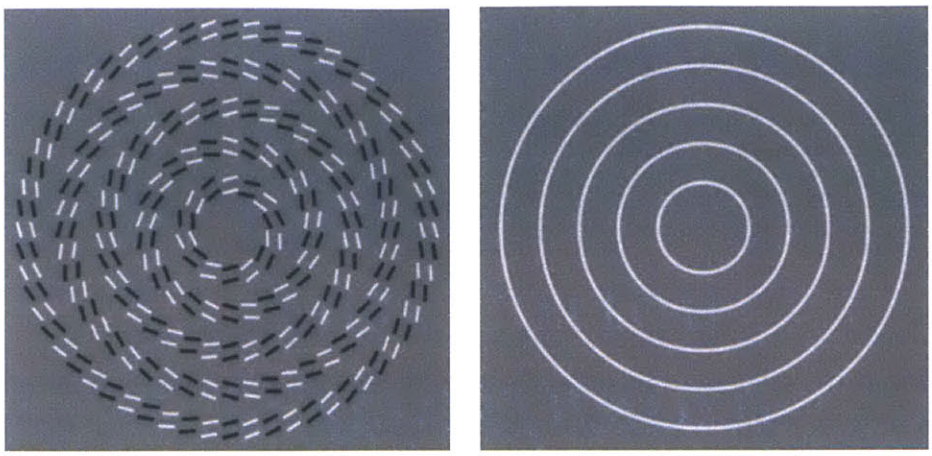


Figure 6-8: A typical pairing from the gold standard questions. It should be obvious which image looks more illusory.

Pairs from the gold standard test were randomly inserted in trials for the experiment. The pairs in this test are selected so that there is a very obvious answer to which of the pair is more illusory. Figure 6-8 shows an example of such a pairing. Appendix A lists all the images used in the gold standard test.

Results and Discussion

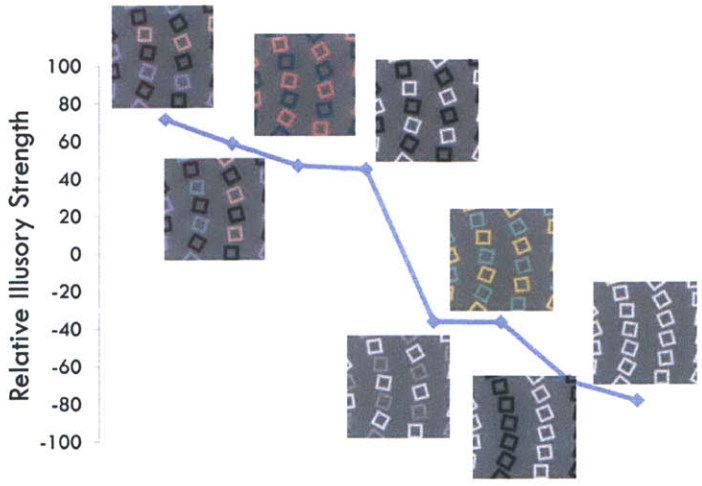


Figure 6-9: Results from the polarity experiment on the intertwining images

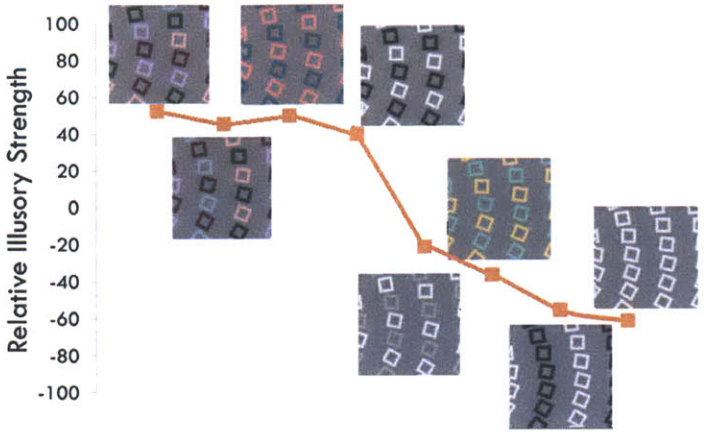


Figure 6-10: Results from the polarity experiment on the spiraling images

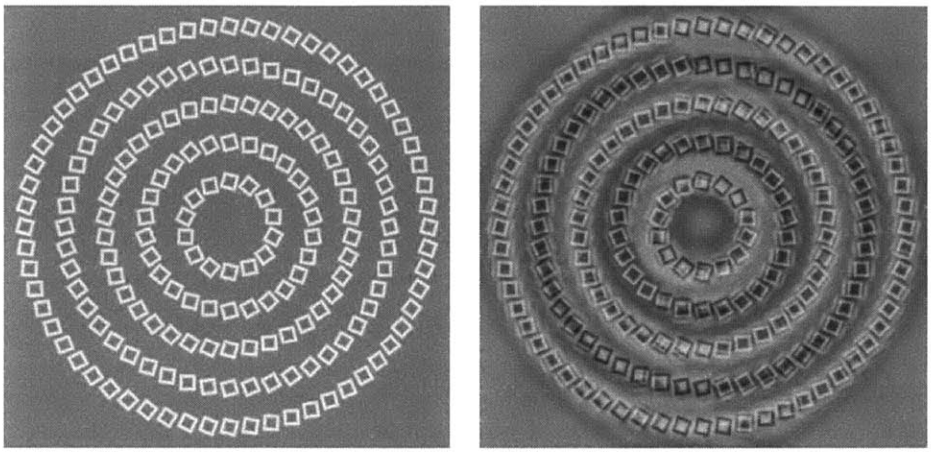
The results are shown in Figures 6-9 and 6-10. All possible pairs from the respective image set were presented to subjects. The relative illusory strength of an image is the percentage of times that image was judged to be more illusory than a competing image (across all subjects). By inspecting the two figures, one can see that there seems to be no difference in the ranking of images between the spiraling case and the intertwining tilts. Images which varied polarity along rings were rated more illusory than those that did not. This result was predicted from the statistical understanding of how peripheral vision works. So, we can conclude that polarity, as expected, helps induce the illusory perception.

6.5 Square Width

The visual system should be more likely to make mistakes when it is hard to distinguish one type of contour (squares along a circular path) from another (the area between circles). When the widths of the squares increase, the responses to orientations in the image occupy a larger area, and can lead to changes in how things are grouped in the image. In particular, it may lead to grouping between rings. This suggests we can point to a specific part of the statistical computation, orientation magnitudes, to see what types of ambiguity arise as the widths of the squares vary.

The statistics we use are computed by first decomposing the image using a steer-

able complex pyramid [41]. If there is ambiguity in the local magnitude of the oriented subbands, there will be ambiguity in the further compressed statistical representation. The local magnitude of the oriented subbands correspond to, roughly, orientation maps of the image that are agnostic about color.



(a) White squares with intertwining tilts

(b) Visualization of orientation information

Figure 6-11: (a) White squares with Intertwining Tilts. (b) shows a visualization of the magnitudes of the oriented subbands in the steerable complex pyramid of the white squares stimuli in (a).

To visualize the types of ambiguity present in the local magnitudes of the complex steerable pyramid, we create a new synthesis algorithm that imposes the oriented magnitudes onto a seed image, iteratively modifying them until the oriented magnitudes of the synthesized image matches the original image. The synthesis procedure is agnostic about color, and seeks only to preserve orientation structures. We use a blank gray image as the seed instead of random noise so as not to insert spurious orientation cues that may artificially inflate the nature of the ambiguity contained in the magnitudes of the oriented subbands.

For example, Figure 6-11 shows a synthesis of the white-squares image. Notice that the encoding allows black and white squares even though there were no black squares in the original image, but it synthesizes the colors so that in this case, polarity isn't alternated in the synthesized image. More important for the following analysis, observe that the contours between rings are blank, so it is unambiguous where the

rings and blank spaces are. The squares within each ring in the visualization do not alternate in polarity.

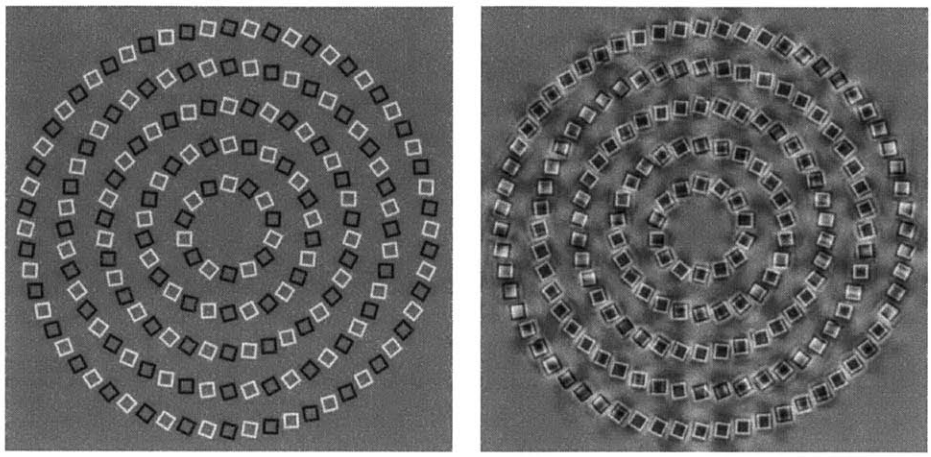


Figure 6-12: Width 1.0

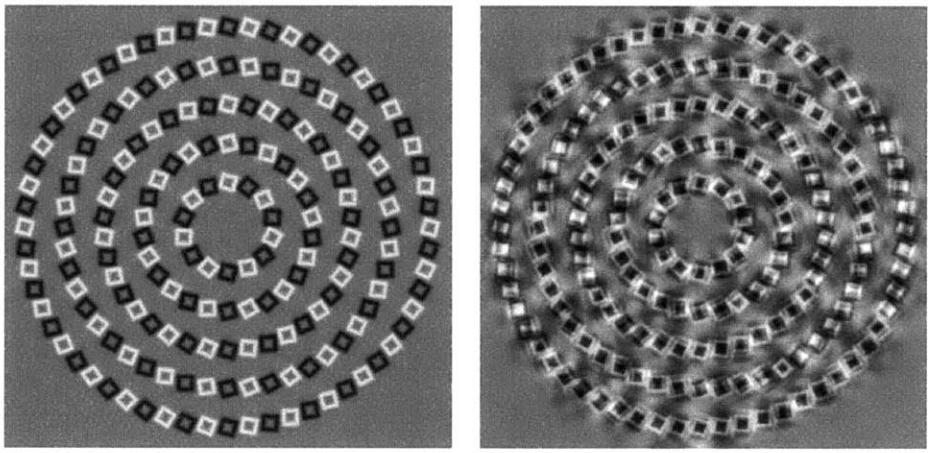


Figure 6-13: Width 2.0

Figures 6-12, 6-13, and 6-14 show that as the width of the squares increase, so does the confusability of the blank region between the rings, and the rings themselves. This demonstrates that the orientation computations themselves in the steerable complex pyramid give rise to ambiguities in where the contours in the image lie, and which parts of the images are actually rings or blank areas. Because the statistics are computed over the steerable pyramid, it can only encode less information than the steerable pyramid itself. Without encoding the additional phase information, this representation does not encode the original ring contours well. This analysis

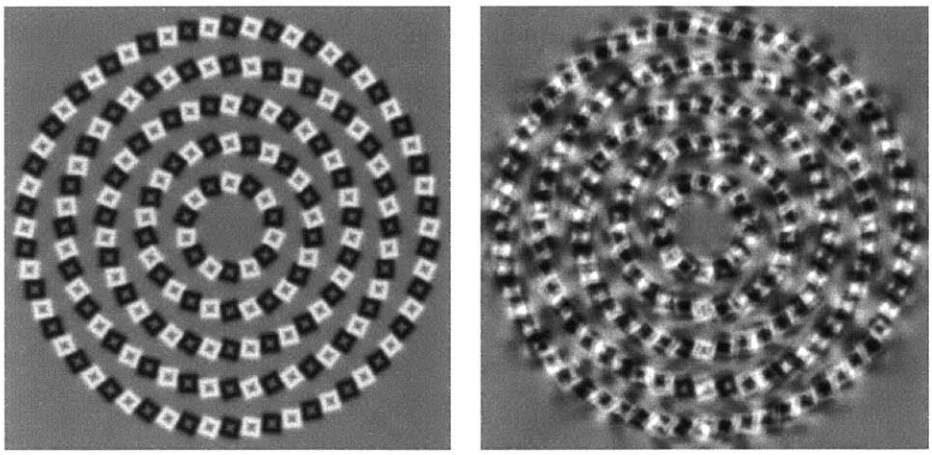


Figure 6-14: Width 3.0

indicates that as width increases, statistics are more ambiguous, and will lead to a stronger illusory percept. We test this hypothesis by running another Mechanical Turk experiment.

6.5.1 Experiment 2: Effects of Square Width

This experiment was also conducted through the Mechanical Turk website.

Subjects

Subjects participated after indicating they consented to the task, and were compensated for their time. Subjects who reported vision that was not normal or corrected-to-normal were dropped from the study. Subjects on the Mechanical Turk website were allowed to participate in either one or both sets of images.

Intertwining Set: Thirty subjects participated, ages 19 to 65.

Spiraling Set: Thirty subjects, participated, ages 20 to 65.

Method

The methodology is identical to that of Experiment 1, except different images are used. Please refer to the section on square width in Appendix A for all the images used in these sets. In the intertwining set, 9 subjects were dropped for not meeting

the standards as mentioned in Experiment 1, and in the spiraling set, 9 subjects were dropped.

Results and Discussion

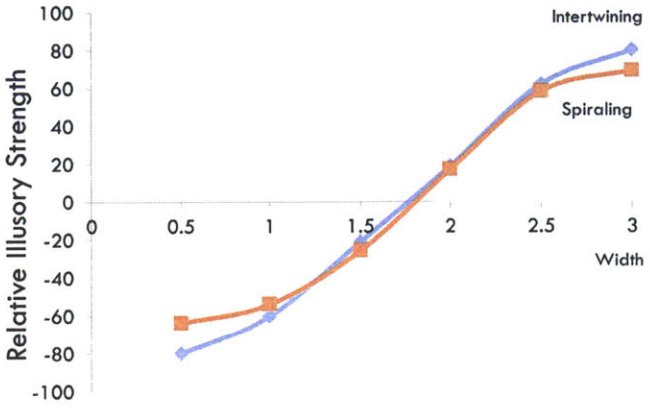


Figure 6-15: Results from the width experiment on the intertwining and spiraling images

The results are shown in Figure 6-15. As in Experiment 1, all possible pairs from the respective image set were presented to subjects and the relative illusory strength was computed similarly. Using spiraling or intertwining tilts did not alter the ranking of how illusory the various images were. As predicted, when the width increased, so did the perceived illusory strength. The width of the squares affects the illusory perception in a particular direction, and the representational ambiguities about where orientations are in the image explain why.

6.6 Square Tilts

By changing the tilt of the squares, one can alter the percept from a spiraling vortex to one of intertwining curves. When all the squares are aligned with the tangent of the rings, the illusory percept is reduced (Figure 6-16). The tilts, therefore, play an important role in this illusion.

To gain intuition, we inspect a non-oriented subband of the illusion in Figure 6-17. Notice that centers of black squares respond in the same way that edges of white

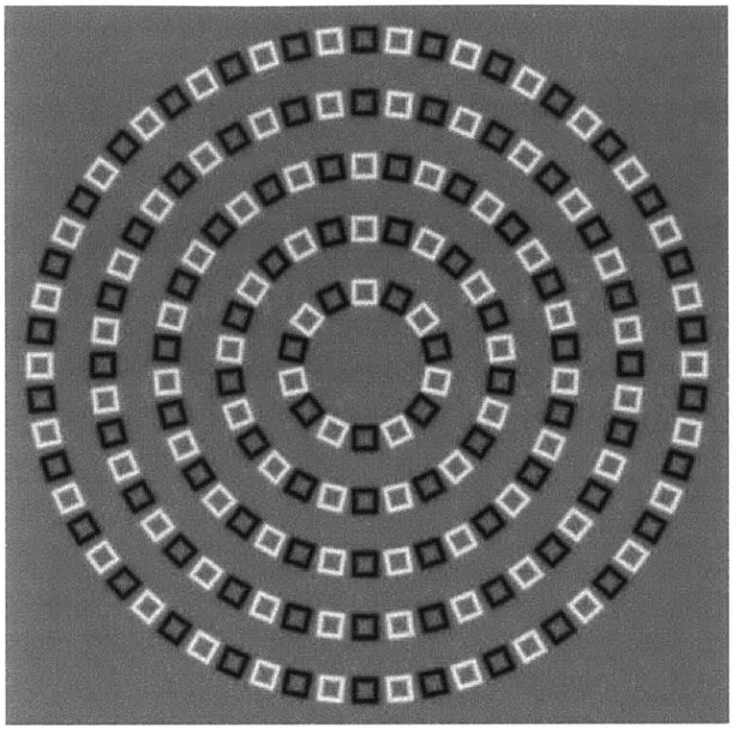


Figure 6-16: When squares are aligned to the tangent of the ring they lie on, there is reduced illusory percept.

squares do. Parallel to the edges of the squares are edges of opposite color. Speculatively, the oriented subband responses "bridge" the blank gap between two rings with the appropriate oriented responses in that region. Additionally, the appropriate angle of tilt can align the center of a square to the edge of the neighboring square, leading to an illusory percept of there being long connected tilted lines along each ring, as in Figure 6-18.

The visual system may count this orientation structure as evidence for grouping across the gap between the rings because those types of orientation measurements usually indicate a curve from one ring to the next. If true, then the "optimal" tilt angle, i.e., tilt angle that maximizes the illusory percept, occurs when the center of the square is aligned with the neighboring square's edges. We can investigate this by looking at the oriented filter responses of the image to identify whether there are smooth curves through the center of a square and the edges of its two neighborings squares.

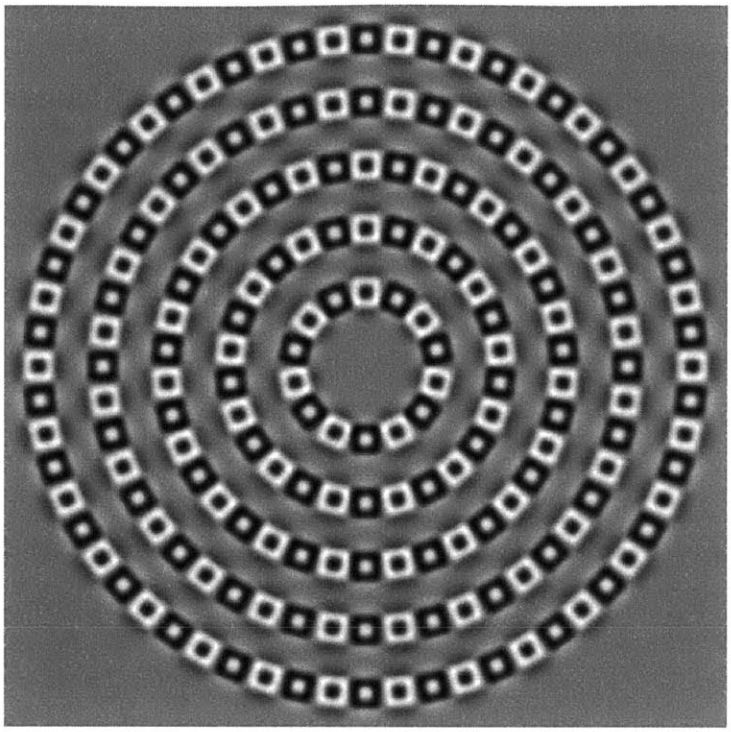


Figure 6-17: Non-oriented subband of Figure 6-16

Figure 6-19 shows patches taken from tilting squares by 5 to 45 degrees from the tangent. The columns are, respectively, the original patch, the oriented subband (or first derivative), and the detected lines (thresholded second derivative). Notice that the lines detected are longest at around 20 degrees. We can quantify this by plotting the mean line length against the tilt angles, as in Figure 6-20. In both, we generally see the length of the lines are highest between 15 and 25 degrees. Does this correspond to perceived illusory strength?

6.6.1 Experiment 3: Effects of Tilt Angle

This experiment was also conducted through the Mechanical Turk website.

Subjects

Subjects participated after indicating they consented to the task, and were compensated for their time. Subjects who reported vision that was not normal or corrected-

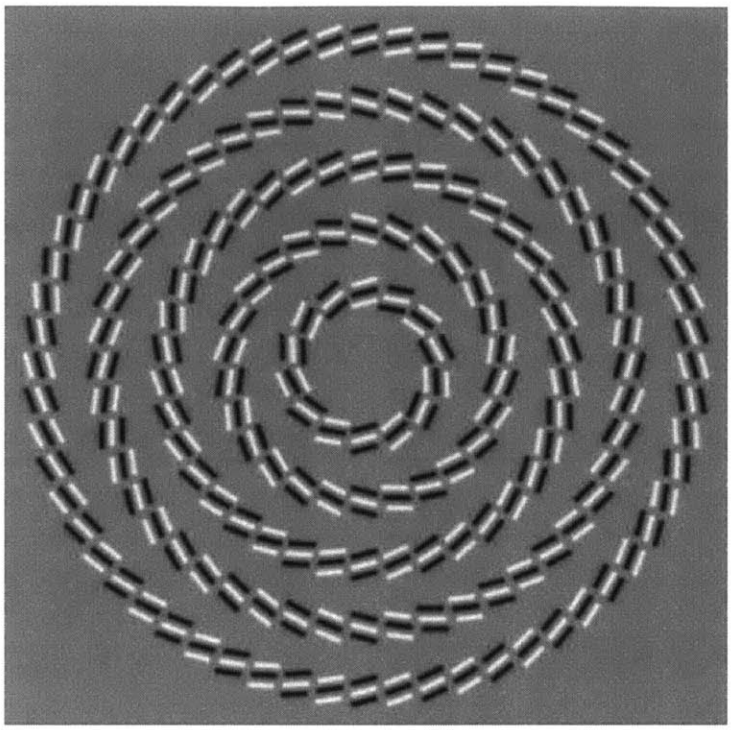


Figure 6-18: Filling in the middle of the "square" with the alternate polarity of the sides roughly visualizes the oriented filter responses (as appropriately rotated). Speculatively, the middle line corresponds to "illusory" line segments that are aligned along the squares' tilt on each ring. These give the impression of longer tilted line segments along each "ring".

to-normal were dropped from the study. Subjects on the Mechanical Turk website were allowed to participate in either one or both sets of images.

Intertwining Set: Thirty subjects participated, ages 19 to 64.

Spiraling Set: Thirty subjects, participated, ages 20 to 63.

Method

The methodology is identical to that of Experiment 1, except different images are used. Please refer to the section on square tilts in Appendix A for all the images used in these sets. In the intertwining set, 11 subjects were dropped for not meeting the standards as mentioned in Experiment 1, and in the spiraling set, 6 subjects were dropped.

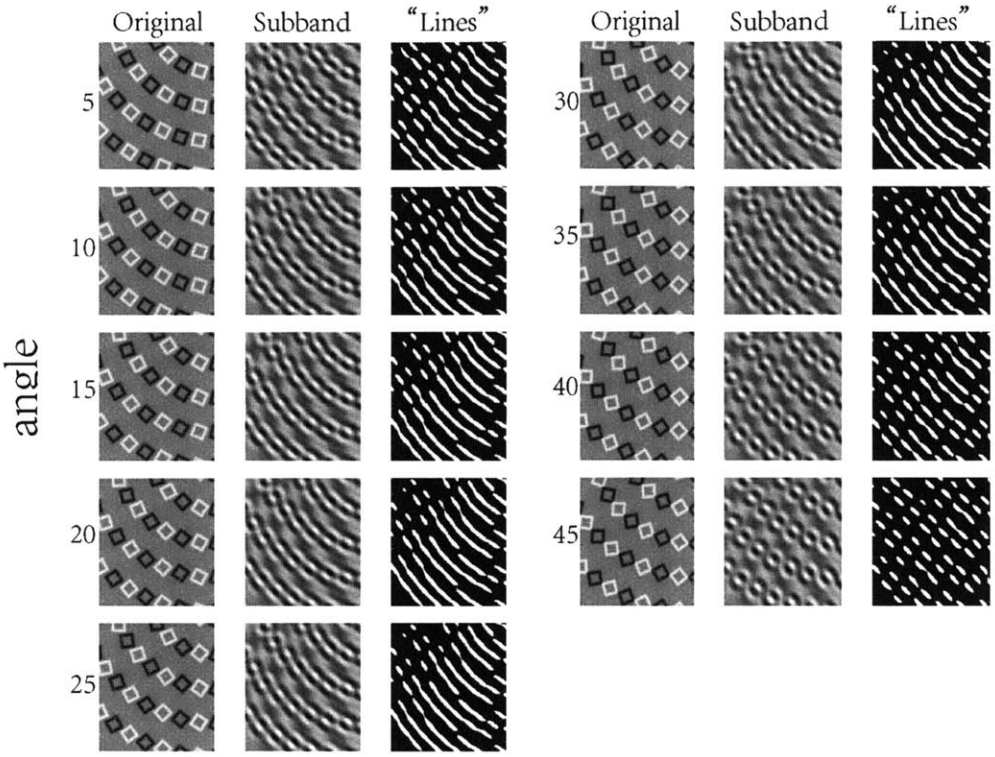


Figure 6-19: Patches taken from applying various tilts to the squares of the spiraling illusion. From left to right: original patch, oriented subband (first derivative), and "bumps" or local maxima (i.e., thresholded second derivative)

Results and Discussion

The results are shown in Figure 6-21. As in Experiment 1, all possible pairs from the respective image set were presented to subjects and the relative illusory strength was computed similarly. The spiraling tilts varied slightly from the intertwining tilts in perceived illusory strength. In both conditions, the perceived illusory strength at least qualitatively match the graph of line length vs tilt amount.

6.7 Visualizing the statistics of the illusions

We have demonstrated that various aspects of the illusion are predicted by a statistical understanding of peripheral vision. It would be desirable to train a machine classifier to use the statistics from the image to predict its percept. To achieve this, one needs a dataset with lots of examples of each type of percept. However, we only have a

small dataset of artificially generated stimuli that differ only slightly from each other so there is very little variation. Training a classifier with such a dataset will cause it to overtrain (i.e, incorrectly generalize that one can classify the percepts based on the artificial differences between these images).

Because training a robust machine classifier is not currently an option available to us, we turn to visualizing the statistics once more. Can we use the techniques from Chapter 3 to visualize the ambiguities that arise from the original images and quantify the result in a meaningful way?

Figure 6-22 visualizes the available statistical information if one were fixating in the middle of the image. The synthesis seems to create contours between rings in inconsistent directions. While the synthesis captures some of the aspects of the percept of the intertwining circles illusion, it fails at matching the regularity of the illusory percept. We speculate that some of the "chaos" is likely some synthesis artifacts, and that some form of regularity priors may prefer the ambiguities to be interpreted as a regular pattern instead of these somewhat chaotic contours all over the visual field that the synthesis produced.

Figure 6-23 visualizes statistics from the spiraling illusion. Notice that the synthesis creates contours between rings in a consistent clockwise direction and hallucinates line segment cues connecting the rings.

And finally, in Figure 6-24, the statistics of the white squares image are less ambiguous than in the intertwining and spiraling case, but the model indicates ambiguities about the tilts of the squares throughout the image. The model does not predict any contours between rings, and except for some noise in the image, essentially shows a set of slightly noisy concentric circles.

To better understand these visualizations, we attempt to quantify the differences between them by analyzing their orientation profiles. We do this by extracting the "rings" of stuff by finding non background pixels (i.e. pixels that are not gray), then we apply a Gaussian filter with σ approximately equal to the width of the ring to fill in any gaps. This captures the entire concentric rings structure. Figures 6-25, 6-26, and 6-27 show the extracted rings from the three original illusions and their

corresponding syntheses.

Then, we warp these images to $(radius, \theta)$ coordinates to effectively "linearize" the rings. If the rings were perfectly concentric, this operation will warp each ring to a line with a constant *radius* (x-axis) value. The linearized rings are shown in the first row of Figures 6-28 and 6-29. Notice that in the non-synthesized images, the rings are essentially perfectly concentric, which is why the linearized rings have a constant *radius*.

The orientation profile of the linearized rings are computed by plotting the orientations with the most votes (top 0.1 percent) as computed by using the Hough transform for line detection, as shown in the bottom rows of the same figures. There is little to no variance in the orientation profile of the linearized original images, and only little variance in the synthesized white squares image. There is a distinct non-zero tilt in the spiraling synthesized image, and there is higher variance in the intertwining image.

These orientation profiles are consistent with qualitative assessments of the percepts of the illusions. The white-squares synthesis elicited a percept that there was little or no illusion, and the analysis shows little or no deviation from concentric rings. The analysis of the spiraling illusion synthesis shows evidence of a tilt in the linearized rings, indicative of a spiral present. And finally, the analysis of the intertwining contours illusion synthesis shows more randomly oriented curves, consistent with the percept of intertwining contours. This analysis shows at least a proof of concept that the visualization of the statistics captures many qualities of the percept of these illusions. With more examples of images, we would be better able to quantify the variances present in this method, and will eventually allow us to use this method to predict the percept of an image.

6.7.1 Spirals vs Circles

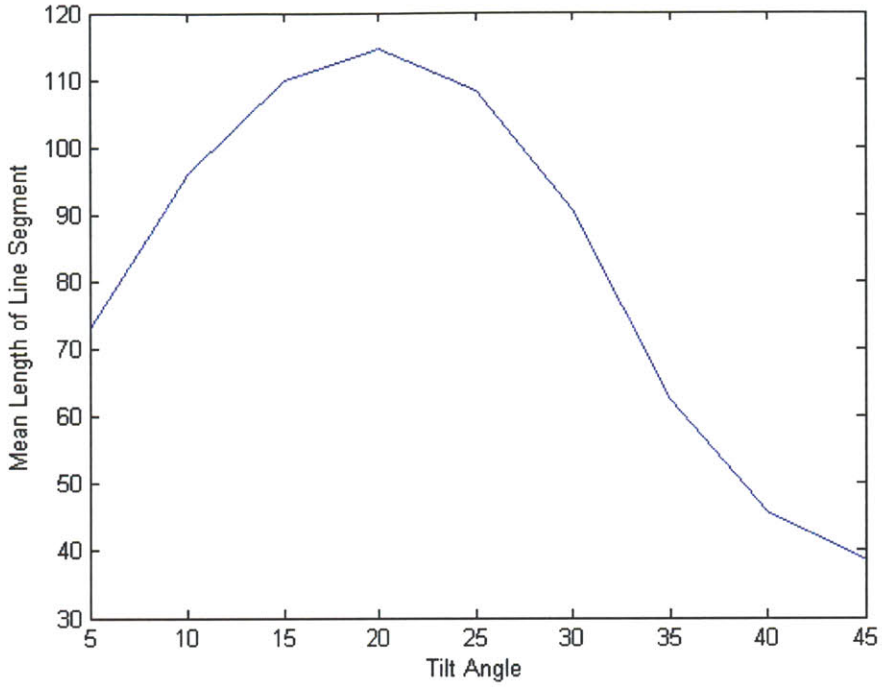
It should also be noted that spirals and circles are extremely similar. Figure 6-30 illustrates this point with a demonstration. It is difficult to notice that the spiraling image is on the left when only allowed to stare at the red dot. The spiraling illusion

presented in this chapter seems to be more than simply a spiral, but more like multiple spirals or a vortex. The analysis in section 6.7 does not capture this qualitative aspect, but we argue that the syntheses do seem to exhibit qualities that loosely resemble it by visualizing very subtle lines between rings. Quantifying and further testing that concept is left for future research.

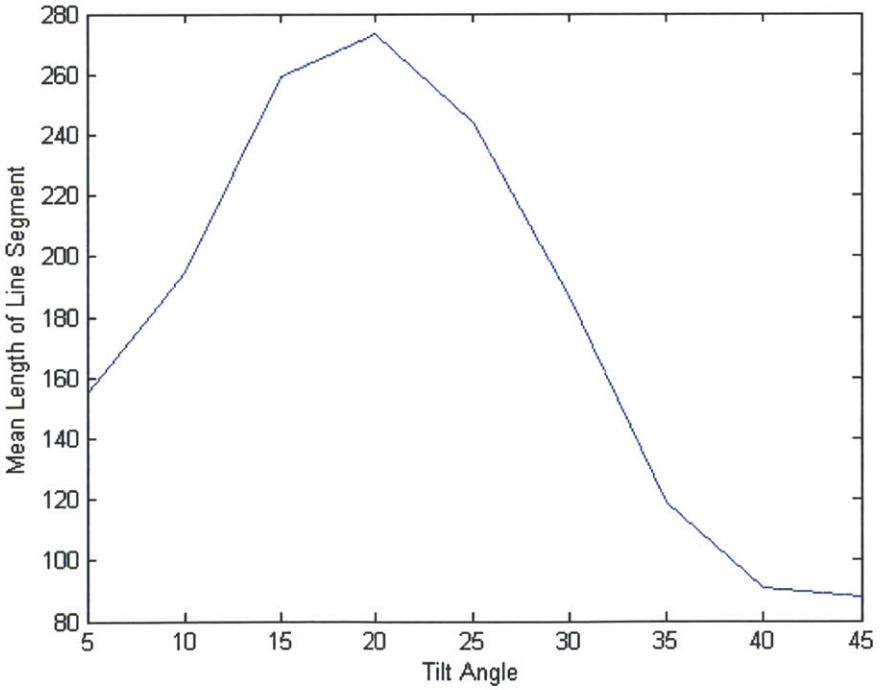
6.8 Conclusion

In this chapter, we showed that a statistical understanding of peripheral vision correctly predicts how polarity and widths of squares affect the perception of the Pinna-Gregory illusions. In addition, we show that visualizing the statistics in the spiraling and intertwining illusion shows many qualities that correspond to the percept of those illusions.

In future, it is important to test the model on a number other parameters and to lesion the model to see which parts of the model are necessary in capturing this illusion. The robustness of this algorithm must also be tested to see how variable the results of synthesis from this model is.

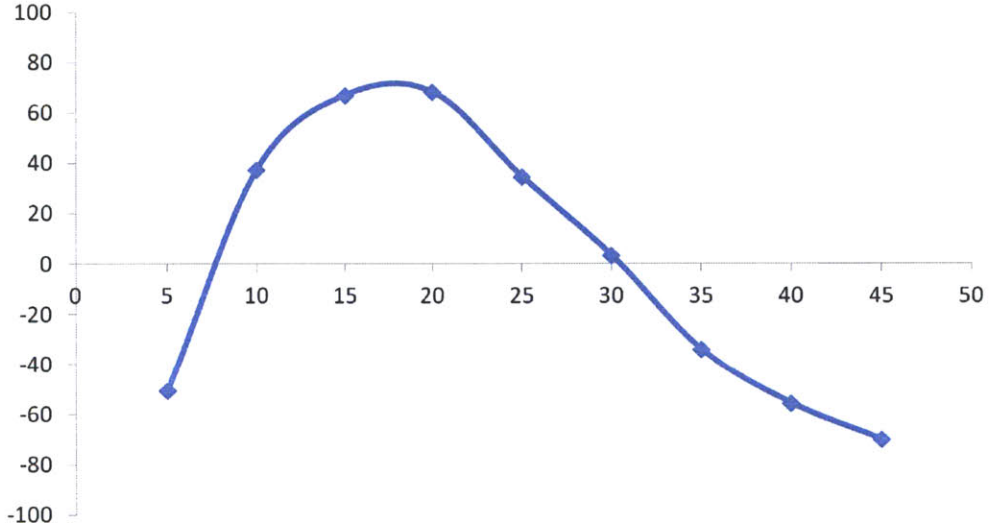


(a) Spiraling Tilts

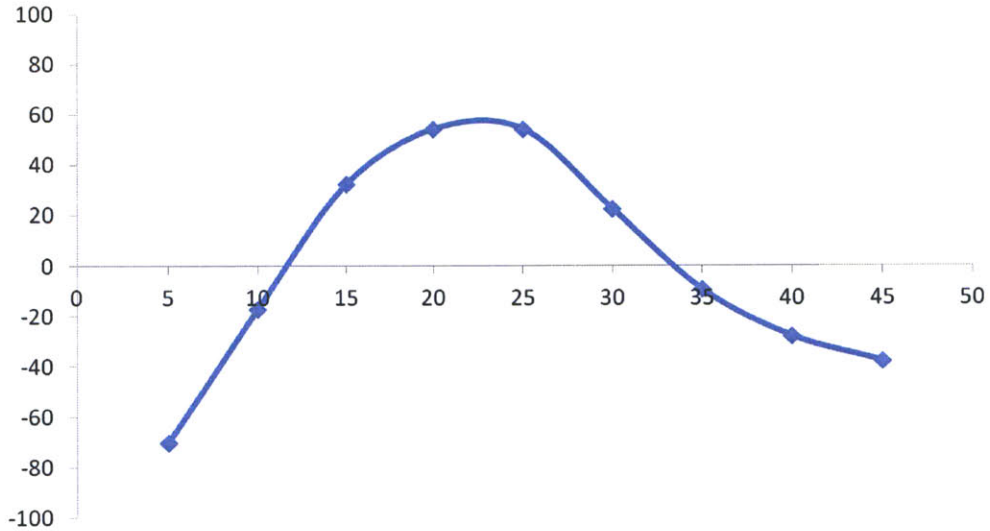


(b) Intertwining Tilts

Figure 6-20: Mean Line Length vs Tilt Angle



(a) Spiraling Tilts



(b) Intertwining Tilts

Figure 6-21: Relative Illusory Strength vs Tilt Angle

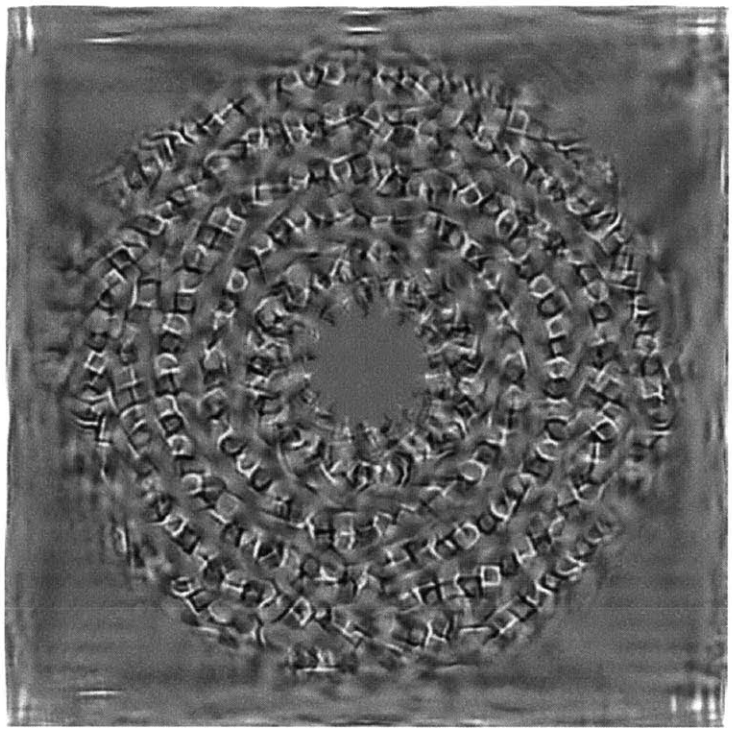


Figure 6-22: Visualization of the statistics in the intertwining illusion

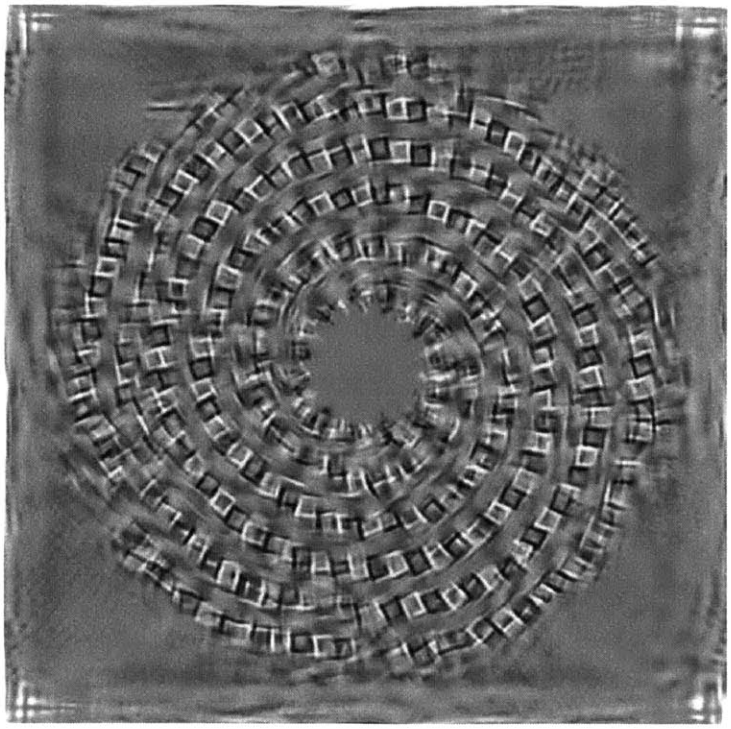


Figure 6-23: Visualization of the statistics in the spiraling illusion

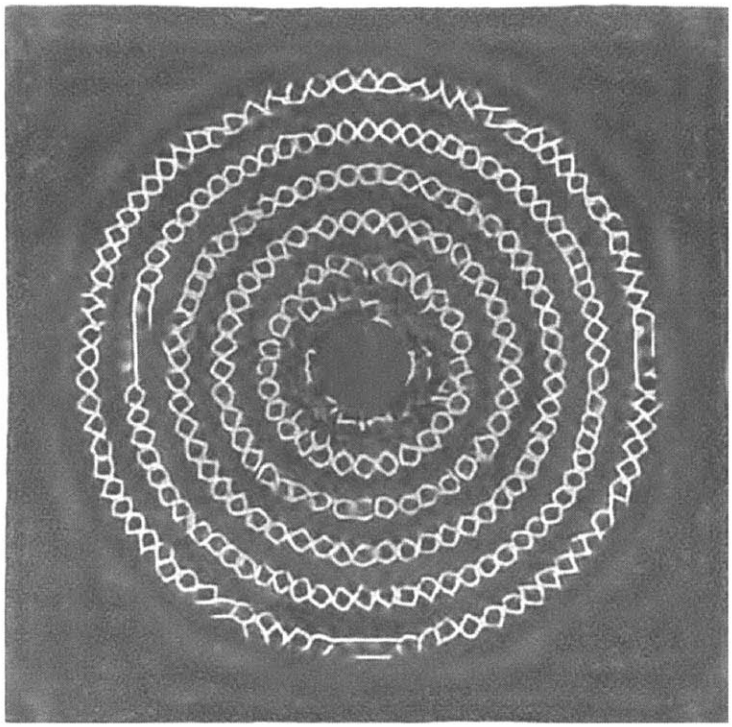
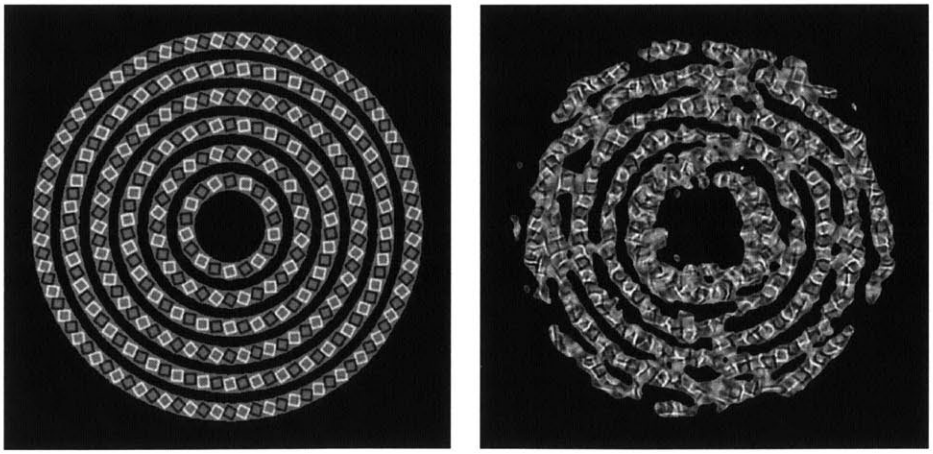
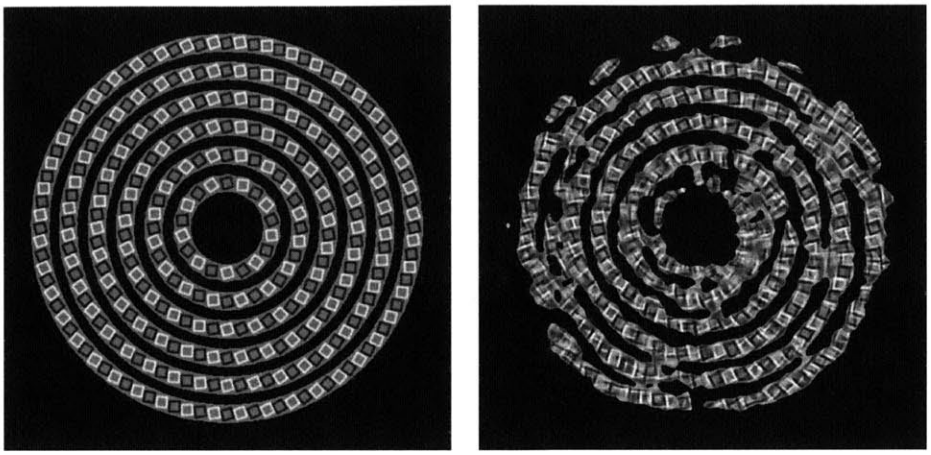


Figure 6-24: Visualization of the statistics in the white squares stimuli



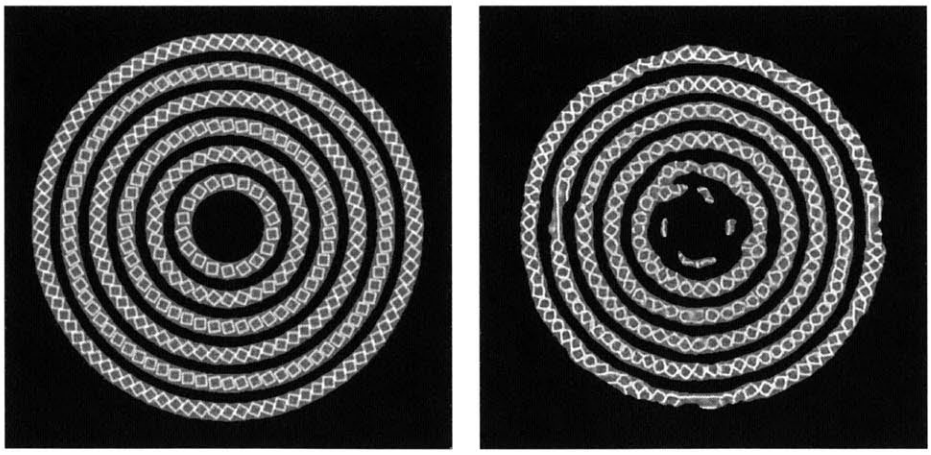
(a) Original Image (b) Synthesis Image

Figure 6-25: Rings extracted from the intertwining image: original and synthesized



(a) Original Image (b) Synthesis Image

Figure 6-26: Rings extracted from the spiraling image: original and synthesized



(a) Original Image (b) Synthesis Image

Figure 6-27: Rings extracted from the white squares image: original and synthesized

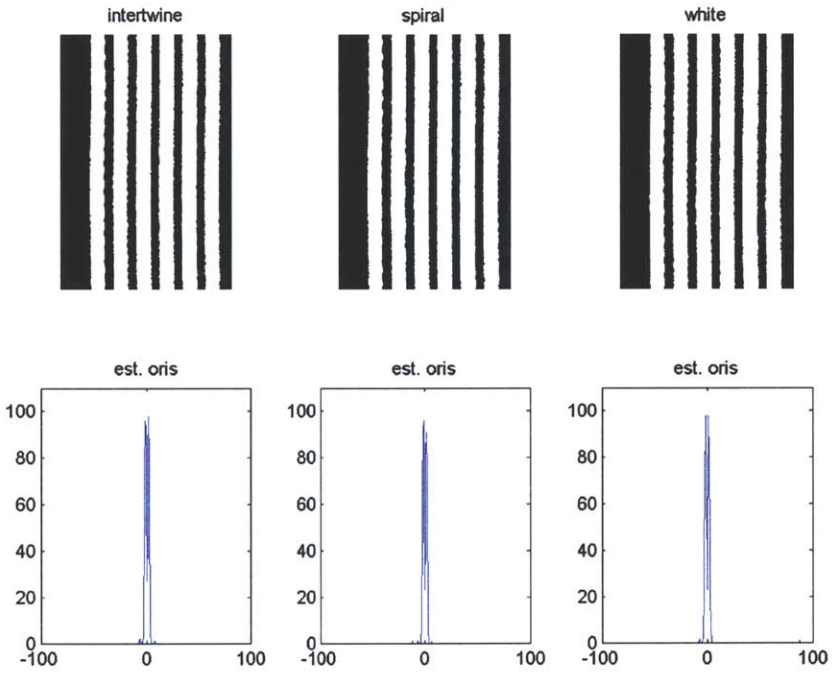


Figure 6-28: Orientation Profile of "Linearized" Illusory Images. Because these images are actually composed of concentric circles, their resulting orientation profiles are of lines with constant *radius* (x-axis)

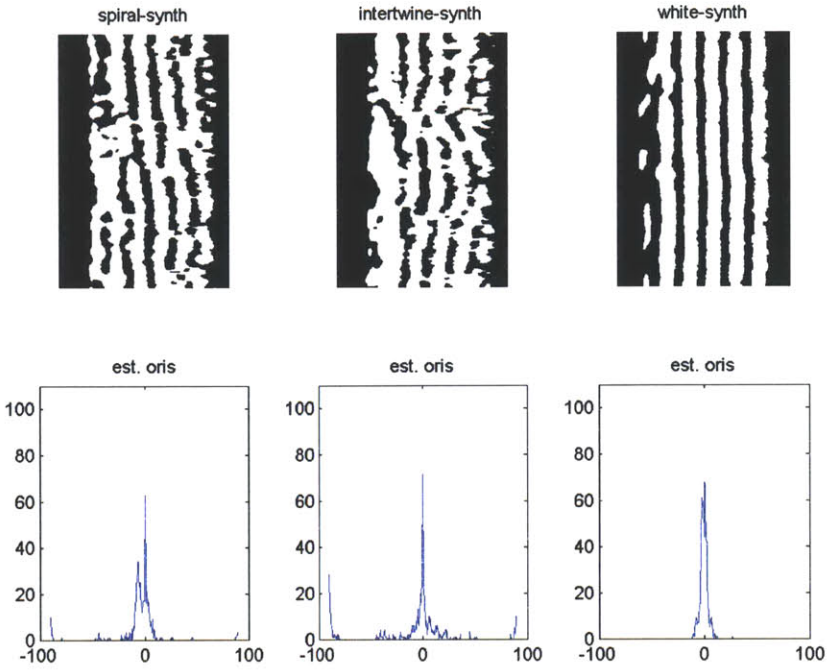


Figure 6-29: Orientation Profile of "Linearized" Visualizations of Statistics from illusory Images. These synthesized images exhibit some properties of the percept from their respective original images. The orientation profile of the white-squares synthesis essentially resembles concentric circles, as per 6-28, while that of the spiraling and intertwining syntheses produce orientation profiles that are consistent with spiraling or multiple oriented curves.

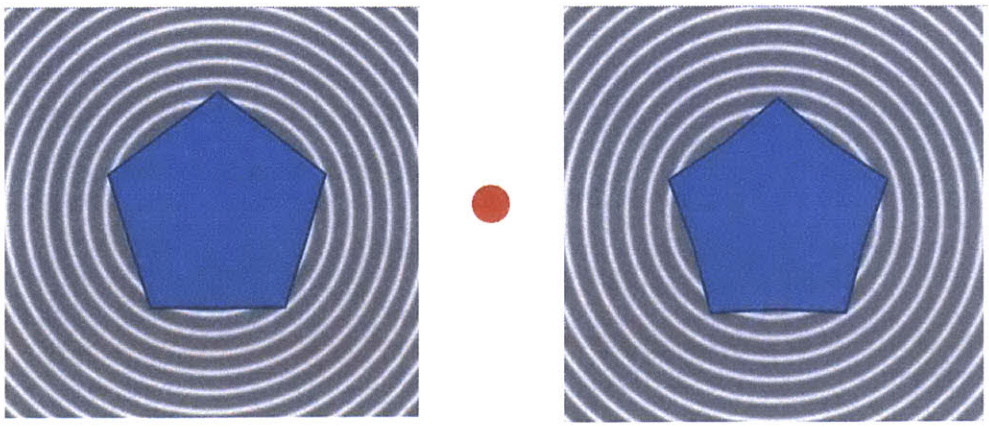


Figure 6-30: Stare at the red dot. It is difficult to classify which image is actually the spiral. They share highly similar visual statistics.

Chapter 7

Object Substitution Masking

7.1 Outline

In this chapter, Object Substitution Masking (OSM) and its relation to grouping is examined. Object substitution masking describes a form of masking where a sparse, non-overlapping, and temporally trailing mask impairs the perception of an object when attention is distributed over a large region.

We investigated whether different types of groupings can affect masking strength. We find that collinear grouping produced less masking, but containment grouping produced more. This result is in contradiction of a theory about OSM which predicts that grouping should increase masking strength. Our results suggests that there is a complex relationship between grouping and OSM.

We speculate a potential link between the statistical model of the periphery to OSM, and suggest a future line of research to investigate its merit.

All the work presented in this chapter is research I conducted by myself under the primary supervision of Ruth Rosenholtz and occasional discussions with Benjamin Balas.

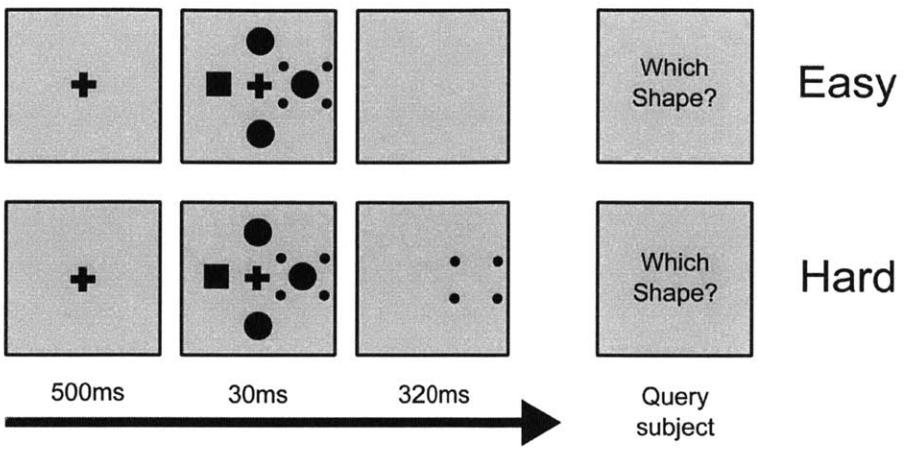


Figure 7-1: Object Substitution Masking

7.2 OSM

Object Substitution Masking refers to the phenomena in which perception of an object is impaired by a non-overlapping, temporally trailing mask. Figure 7-1 illustrates the phenomena. Subjects are asked to identify what shape was surrounded by the four dot mask. In the case when the object disappears at the same time as the four dots, subjects are easily able to report the identity of the indicated object. Intriguingly, if the four dots remain visible after the object has disappeared, identifying the object is very difficult, and some subjects report that the object was never there [12].

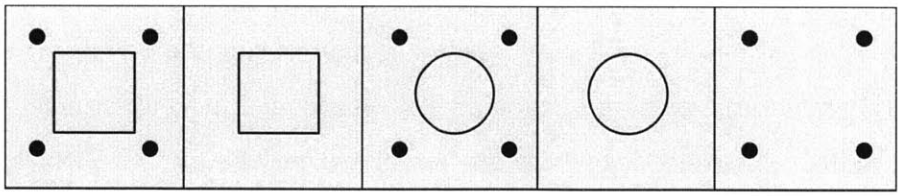


Figure 7-2: Each box represents a different hypothesis. The stimuli on display activate various hypotheses about what object is present at a given location. These hypotheses have temporal inertia in order to be robust to noise, and their strengths slowly degrade in time. [28]

Enns and Di Lollo identify a few conditions for producing this effect [12]. (1) Attention should be distributed over a large region by, for example, having many objects on screen. (2) a sparse mask. (3) the mask remains on screen after the objects are no longer visible. The timing of each stage in the standard OSM experiment affects

the amount of masking observed in a complex manner [15].

Expanding on their earlier work, Di Lollo, Enns and Rensink later proposed a theory for what is going on. They suggested that the stimulus activated a number of possible hypotheses in parallel. Figure 7-2 shows an example of possible hypotheses of the stimuli. Each hypothesis' activation strength depends on how well the appearance of the stimuli matches up with hypothesis' expectation of its appearance [28].

When the objects disappear, the activation signals slowly degrade over time. In the simultaneous offset case, where everything disappears at the same time, all hypotheses' strengths degrade over time, and so it is possible to still pick out the maximum of the competing hypotheses. Performance is still good because people can remember what was there.

However, in the delayed offset situation, all hypotheses' activation signals degrade except for the hypothesis that only a mask was present at that location all along. That hypothesis of only the mask being present does not decay due to longer exposure of the mask. Subjects' performance in this case deteriorates. Many report only seeing the mask and cannot recall the original shape.

Later work on OSM suggests other rules for how it works. Moore and Lleras show that when mask and target are grouped by color or motion, the resulting masking effect is stronger. They hypothesize that when the mask and target are within the same "object file", the trailing mask "overwrites" the file so only the mask is easily remembered. When mask and target are not grouped, there are two "object files", so the mask only overwrites itself [32].

If this hypothesis about how OSM is affected by the interaction of grouping and "object files" is true, then all types of groupings are equivalent and will affect OSM in the same manner. Masking should occur for many kinds of groupings, and not only color and motion. However, it is unclear whether this observation that grouping leads to more masking holds for all types of grouping. Further, there is reason to suspect that different types of groupings will behave differently. In our statistical model, not all types of groupings are alike. Grouping by containment, for example, is more visually complex then grouping by collinearity. Our statistical model represents

collinear groupings more faithfully than it does containment groupings. We investigated how collinearity and containment grouping between target and mask affects masking strength.

7.3 Experiment 1: Collinearity Grouping

Twelve subjects participated in this experiment after giving informed written consent. They received monetary compensation for participation. All subjects reported normal or corrected-to-normal vision.

7.3.1 Method

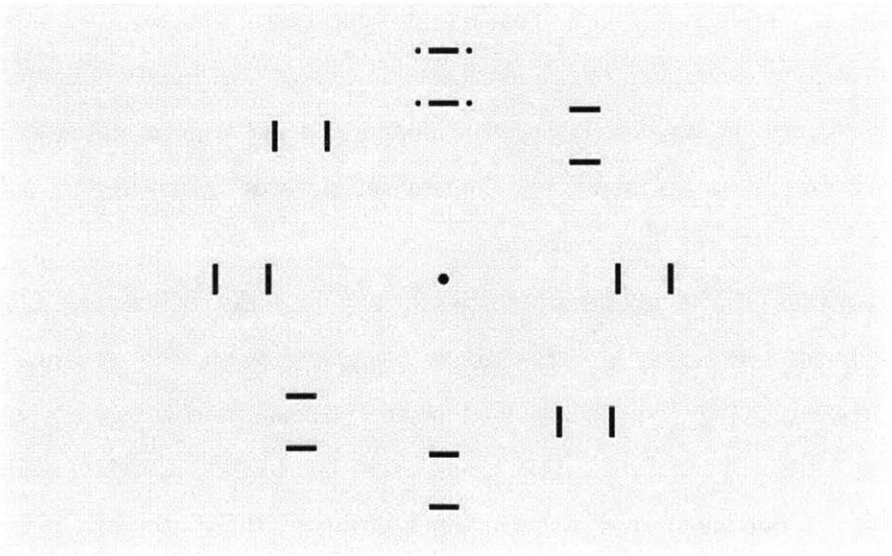


Figure 7-3: A trial where the mask was collinear with the target.

Stimuli were presented on a 40 cm x 28 cm monitor, with subjects seated 75 cm away in a dark room. We ran our experiments in MATLAB, using the Psychophysics Toolbox [5]. Subjects were presented with a ring of eight items. The ring had a radius of 9 degrees v.a., and each item was 1.4 degrees v.a. by 1.4 degrees v.a.. The target was cued by a four-dot mask. Subjects had to report the orientation of the target. The target could either be a pair of horizontal or vertical lines. These lines were either collinear with the mask as in Figure 7-3 or not collinear as in Figure 7-4

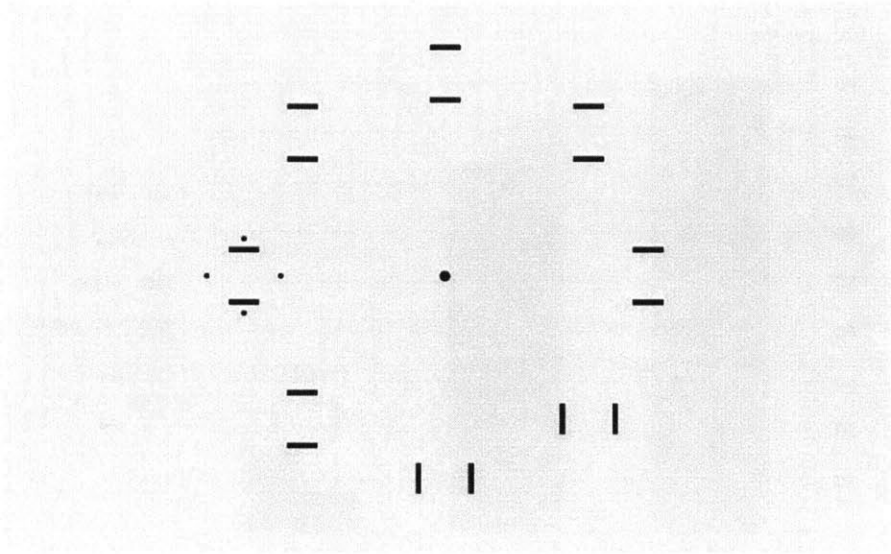


Figure 7-4: A trial where the mask was not collinear with the target

by rotating the mask by 45 degrees. Subjects finished 256 trials with factors Target Horizontal, Target Vertical X Mask Collinear, Mask Non-Collinear X Delayed offset, Simultaneous offset) equally and randomly distributed in the trials. Target location was randomized.

The order and timing of the events are as follows: Subjects view a fixation cross for 500ms, followed by presentation of the ring of 8 objects with a four dot mask around one of the randomly selected objects for 30 ms, after which either everything disappears (the simultaneous offset case) or only the four dot mask remains for 320 ms (delayed offset case).

7.3.2 Results

We look at the masking effect (percent correct simultaneous delay - percent correct delayed offset), a commonly used measure of the amount of masking in OSM tasks [32]. Collinearity significantly ($p = .0002$) relieved masking, with a masking effect of 8.07 percent (93.53 percent minus 83.46 percent) for the collinear mask, and 17.45 percent (91.01 percent minus 73.57 percent) for the diamond mask.

Collinearity is often a strong grouping cue, and so when target and mask were arranged collinearly, they formed a stronger group than when target and mask were

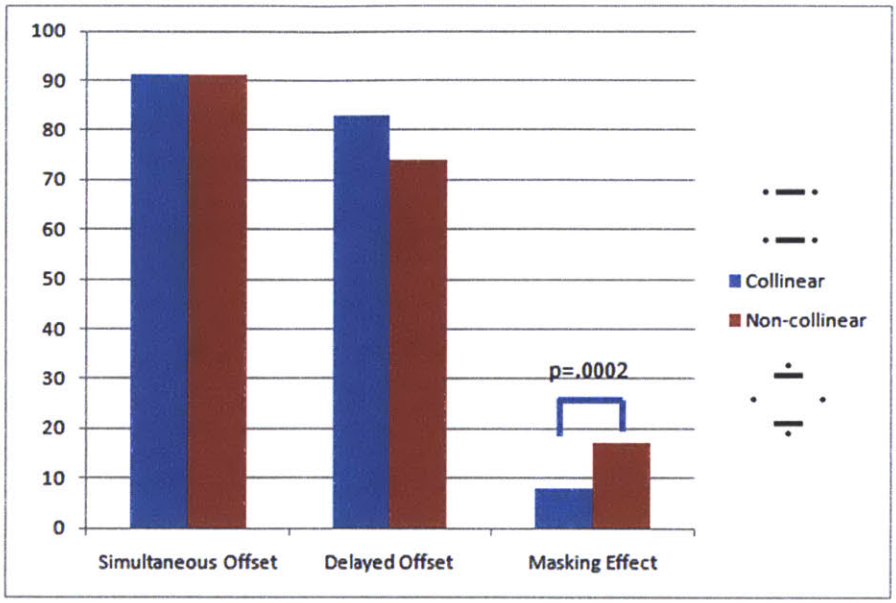


Figure 7-5: OSM impairs the non-collinear grouping more than the collinear grouped stimuli. In this case, grouping produced less masking.

not arranged collinearly. We find that when the items were grouped collinearly, the masking effect was relieved compared to when the items were not grouped. This result contradicts the prediction made by the "object file" account that grouping should always produce stronger masking effects.

7.4 Experiment 2: Containment Grouping

7.4.1 Subjects

Twelve subjects participated in this experiment after giving informed written consent. They received monetary compensation for participation. All subjects reported normal or corrected-to-normal vision.

7.4.2 Method

Like in Experiment 1, stimuli were presented on a 40 cm x 28 cm monitor, with subjects seated 75 cm away in a dark room. We ran our experiments in MATLAB, using the Psychophysics Toolbox [5]. Subjects were presented with a ring of eight

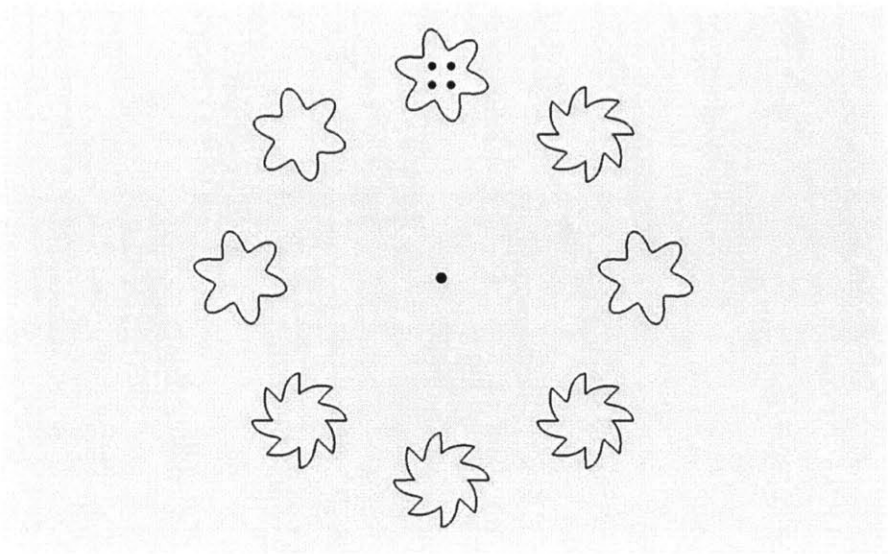


Figure 7-6: A trial where the mask was inside the target.

items. The ring had a radius of 9 degrees v.a., and each item was 1.4 degrees v.a. by 1.4 degrees v.a.. The target was cued by a four-dot mask that could either appear inside or outside a target item. Subjects had to report whether target was wavy or spiky, in cases where the mask was inside (Figure 7-6)or outside (Figure 7-7) the shape. Subjects completed 256 trials with factors Target Wavy, Target Spiky X Mask Inside, Mask Outside X Delayed offset, Simultaneous offset) equally and randomly distributed in the trials. Target location was randomized.

The order and timing of the events are the same as in Experiment 1. Subjects view a fixation cross for 500ms, followed by presentation of the ring of 8 objects with a four dot mask around one of the randomly selected objects for 30 ms, after which either everything disappears (the simultaneous offset case) or only the four dot mask remains for 320 ms (delayed offset case).

7.4.3 Results

We find that masking effect was significantly (p=.0419) stronger when the mask was inside, 18.36percent (85.29 percent minus 66.93 percent), compared to outside, 11.85 percent (72.40 percent minus 60.55 percent).

Grouping is stronger when an object encircles another object, as opposed to when

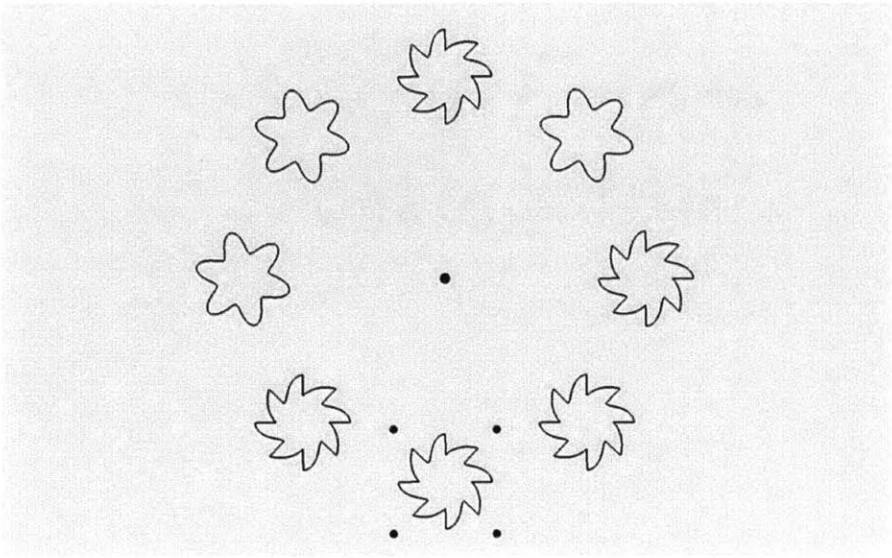


Figure 7-7: A trial where the mask was outside the target

a sparse set of dots are located outside another object. We found that in the strong grouping condition (when the shape encircled or contained the mask), the masking effect was stronger. In this case, grouping did produce a stronger OSM masking effect.

7.5 Discussion

Interestingly, grouping by collinearity produced less masking effect, while grouping by containment produced a stronger masking effect. This suggests that the strength of grouping between target and mask does not directly predict the magnitude of the masking effect. We speculate that there could be an explanation of this effect, based on a space-time version of the peripheral model we have been presenting in this thesis. There are many other temporal peripheral illusions that some have attributed to temporal crowding [46].

In addition, Holcombe has listed several instances where the speed at which stimuli changes affects what information one is able to report [18]. For example, when green dots move left then change color to red and move right, the rate at which the stimuli is presented affects performance in identifying which color was associated with left or right motion, yet people are able report that they perceived left and right motion

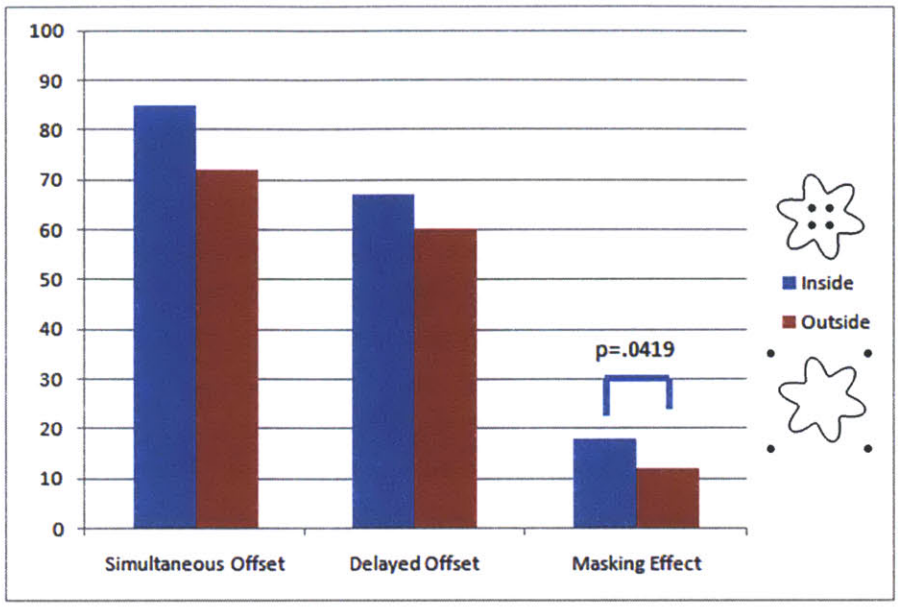


Figure 7-8: OSM impairs performance more when the four dots are inside the target item, indicating that containment grouping produce more masking.

as well as seeing green and red dots. It was only the conjunction of direction of motion as well as the color that proved difficult to report, which seem analogous to the difficulty in reporting color and orientation conjunctions that was discussed in Chapter 3. This suggests a link to temporal crowding.

McDermott and Simoncelli have used a temporal version of the Portilla-Simoncelli statistics to represent audio perception [31]. However, it remains to be seen whether the statistical encoding presented in this thesis can be generalized to include space and time. One simple method of extending the model towards that goal is to consider placing ellipsoidal pooling regions that pool visual information over space and time, and whose size increases both with distance from fovea, and with time in a three dimensional block of video.

OSM affects peripherally displayed stimuli more than foveally displayed items [12]. That observation seems to indicate that OSM is a promising area for research to see whether the ambiguities in peripheral representations also underlie some aspects of OSM. With a space-time version of the model, we could test this hypothesis. If space-time statistics underlie OSM, the model should have more ambiguities in representing

a video clip of object+mask followed by mask, compared to that of a video clip of object+mask followed by nothing. In this situation, the model must also be able to explain why the timing matters greatly in producing this phenomena.

One observation that may explain why "collinearity" relieved masking is that grouping by collinearity is special in that it is easily represented by correlations, while containment is not. It is possible that collinearity is inherently more easily represented by statistics, while containment is not. Perhaps the difficulty in representing a stimulus also plays a role in object substitution masking.

7.6 Conclusion

In conclusion, we have shown that an "object file" account of Object Substitution Masking is complicated by the results presented in this chapter. Grouping by collinearity produced less masking, but grouping by containment produced more masking effect.

Chapter 8

Applications and Conclusions

8.1 Outline

In this chapter, we apply the peripheral vision model to gain insight into user interface design, perception of mazes, and classic visual cognition tasks. We suggest some lines of future work for applying the statistical model of the periphery developed in this thesis. We argue that the model can help inform designers about how to modify a user interface design in order to improve user experience. We also show some work in progress regarding maze perception and classic visual cognition puzzles.

We discuss the relationship of the work presented in this thesis to some computer vision methods used to represent objects and scenes with local feature descriptors. The pros and cons of using this statistical model for general computer vision algorithms are evaluated. Some ideas for future work are presented as well.

Finally, we conclude with a summary of the contributions made in this thesis.

The preliminary work presented in this chapter is the result of applying the work on visualizing statistical information that was presented in Chapter 3. This work was done in collaboration with Ruth Rosenholtz and Benjamin Balas.

8.2 Efficient User Interactions

At any given instant, much of a display appears in a user's peripheral vision. Based on the information available in the periphery, the user moves their eyes, scanning the display for items of interest, and piecing together a coherent view of the display. Much of this processing happens unconsciously.

We argue that an understanding of what types of information the periphery is capable of representing well can help design better information visualizations and user interfaces by making important information comprehensible at a glance. The statistical encoding has some implications for what type of information is available to a user in the periphery.

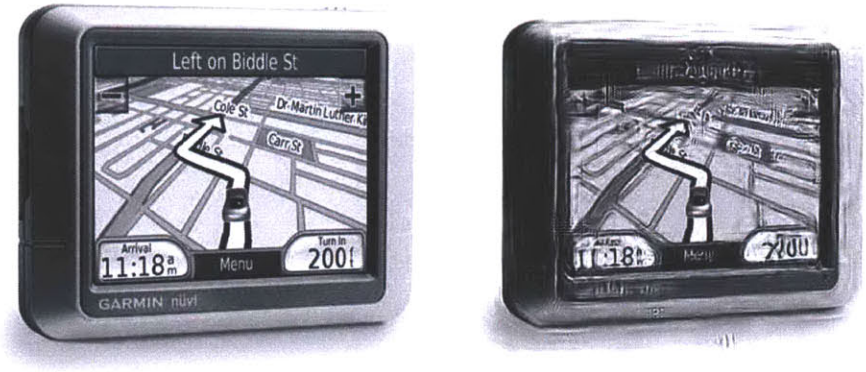
Some of these implications may be expressed in general rules of thumb. For instance, users are able to perceive some low-level idea of what shapes are present, but details on position and phase information are likely to be incorrect. Text is difficult to read in the periphery, and designers should not expect users to be able to read peripheral text in general. More cognitive effort is required to comprehend visually complicated parts of a display.

Beyond those rules of thumb, we can visualize the information available in the periphery, as discussed in Chapter 3, to answer specific questions a display. For example, we might want to know whether a driver will be able to tell whether he needs to turn left very shortly with only one look at the GPS. What information can a user comprehend in one glance at a map?

8.2.1 Analysis of some designs

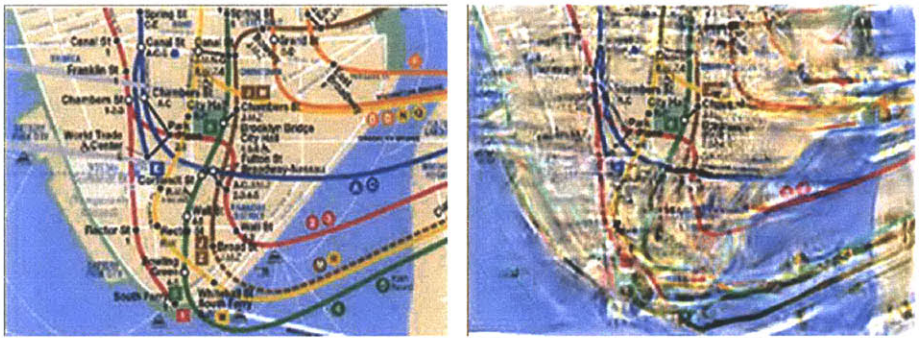
Figure 8-1 shows an image of (a) a GPS unit, and (b) the information contained in the periphery should one be fixating on the car. Most of the text is not readable. The general layout of streets seems preserved in the statistics, though some details are incorrect. In this case, it seems that the model predicts that people will be able to tell that a turn is coming up.

What can a designer expect someone to notice while looking at a location on a



(a) Original (b) Synthesized

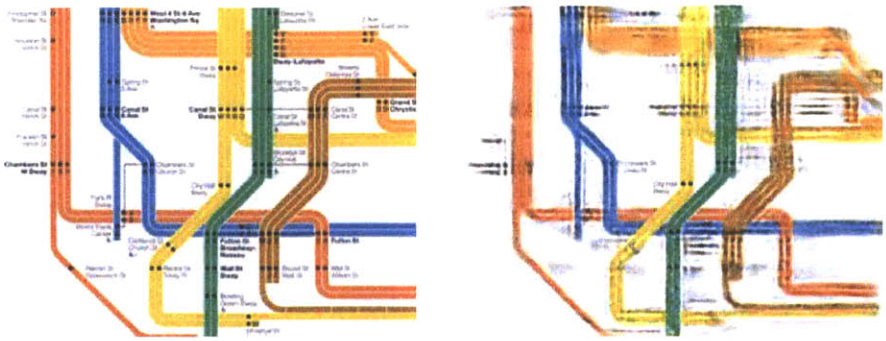
Figure 8-1: (a) What can people tell about the GPS display? (b) Visualization of information available in the periphery, fixating on the car.



(a) Original (b) Synthesized

Figure 8-2: (a) The New York city subway map (b) Visualization of information available in the periphery while fixating on "city hall".

map? Figure 8-2 shows (a) the typical map of the New York subway system, and (b) the information in the periphery if the model fixates on "city hall". The large land mass' shape is roughly preserved in the visualization, but many details about where lines connect and travel get messed up. When using a more stylized version of the subway map in Figure 8-3 (a), the visualization of peripheral information in (b), also fixated on "city hall", shows that most of the lines are preserved well. In both styles, text is mostly unreadable in the periphery.



(a) Original (b) Synthesized

Figure 8-3: (a) Stylized New York city subway map (b) Visualization of information available in the periphery while fixating on "city hall".

8.2.2 Future Work

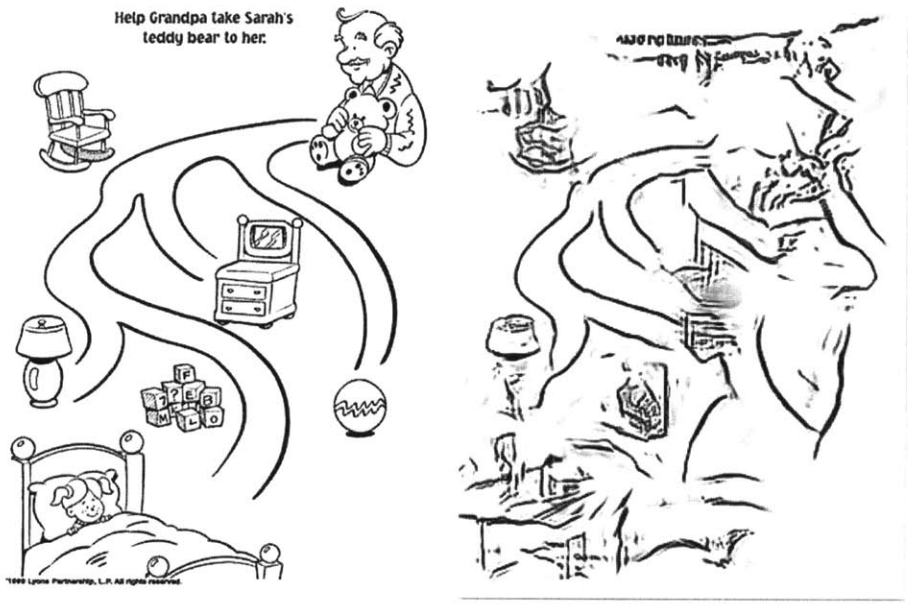
These syntheses indicate potential for a fully developed tool to aid visual designers. Studies are required to test its efficacy in aiding designers, and improvements in the algorithm's efficiency is needed for it to be useful to designers who will not want to wait many hours to see the results of the algorithm.

8.3 Mazes

Once again, we repeat the theme of this thesis. If the peripheral visual system allows ambiguities on complex stimuli, we should expect to see performance limited by the information contained in the periphery. Two particular situations we examine here are how easily one can solve a maze, or whether one can determine if two dots are on the same line or not.

In Figure 8-4, a relatively simple maze is shown. The information in the periphery, shown in (b), essentially exhibits a clear path from the start to the exit with little ambiguity. This probably means that one can solve the maze without many additional fixations, and so the maze is easy.

In Figure 8-5, the maze is more complex. The visualization of the information in the periphery shows many ambiguities arise just a short distance away from the fixation point. This then requires one to make many fixations to find out where a



(a) Original (b) Synthesized

Figure 8-4: (a) This maze is trivial to solve (b) Visualizing the statistics shows that one can easily find a path from the start to the end without needing to move the fixation

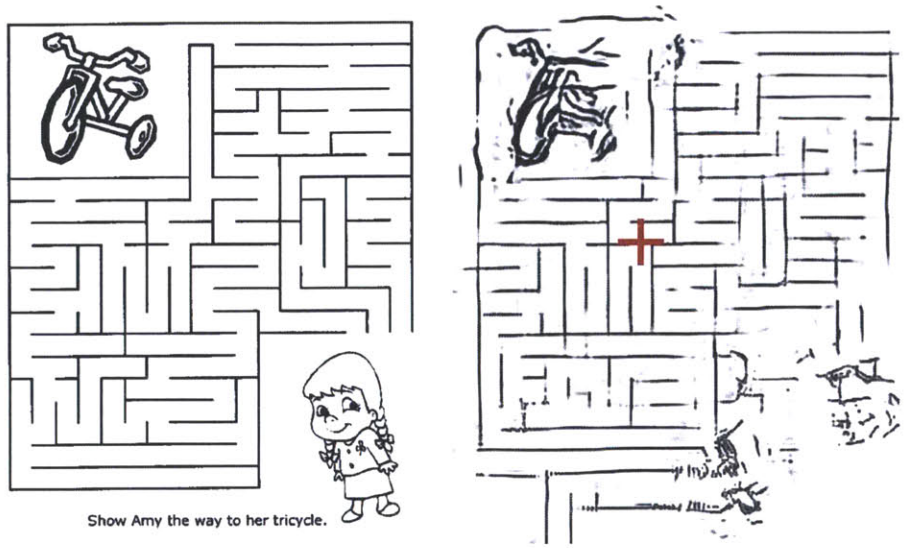
path leads, making this a more difficult puzzle to solve.

8.3.1 Future Work

This is some preliminary evidence that the peripheral information limitations underlie how difficult some maze tasks are. Additional work is being conducted by Benjamin Balas, exploring how the statistics of mazes are affected by changing the thickness of the walls and why the layout of the pooling regions predict why stretching the image in one direction will hamper performance, but enlarging the image while maintaining its aspect ratio will not cause much difference in performance.

8.4 Summary of Contributions

This thesis extends previous work on a statistical model of the peripheral visual system. If the periphery loses the information this model suggests, then we should see evidence of it in tasks that use peripheral vision.



(a) Original (b) Synthesized

Figure 8-5: (a) This maze is more difficult to solve (b) Visualizing the statistics shows that one needs to make more fixations to figure out where a path leads.

This thesis shows that many classic visual search results can be explained by the ambiguities in representing peripheral stimuli. Beyond showing the correlation of search results and the discriminability of the statistical representations [44], this thesis also proposes a quantitative model of visual search that tries to estimate the average number of fixations needed to find a target. We show that it is possible to construct such a model to fit human performance in visual search.

Next, this thesis shows that the Pinna-Gregory illusion is predicted by the ambiguities in the statistical representation of peripheral vision. In particular we show how item width and polarity affect the perceived illusory strength of several modifications of the basic illusion. We also show that a visualization of the information contained in the periphery exhibits many qualities present in the illusory percepts.

The thesis also describes work in showing that a phenomenon known as Object Substitution Masking (OSM) has different effects when grouping by collinearity than when grouping by containment. This result is in contradiction of some work in the field which predicts that OSM should behave in the same way to all kinds of grouping. We also suggest a future line of research in which a spatio-temporal version of the peripheral model might underlie OSM as well.

Appendix A

Images Used In Mechanical Turk Experiment

A.1 Gold Standard Dataset

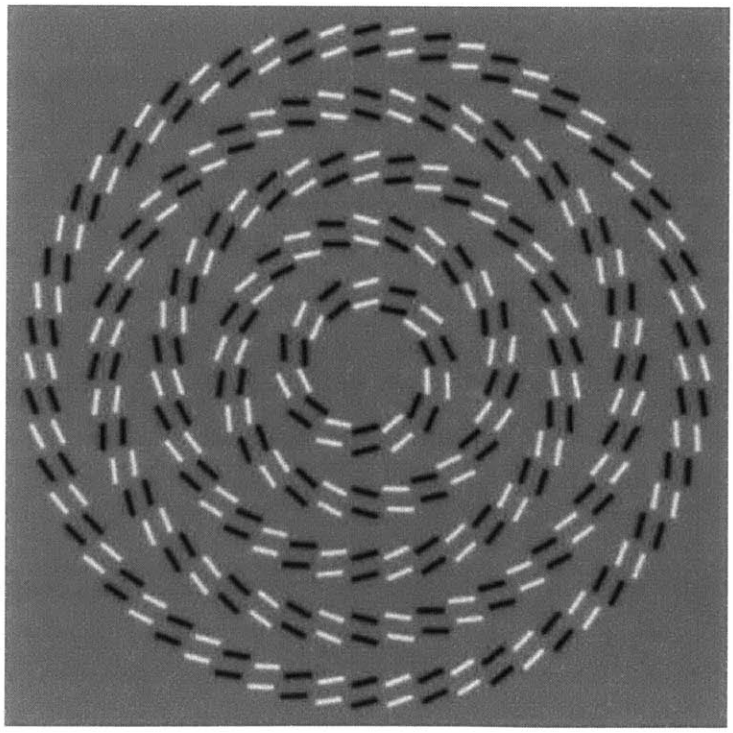


Figure A-1: Two-Lines Intertwining Illusion

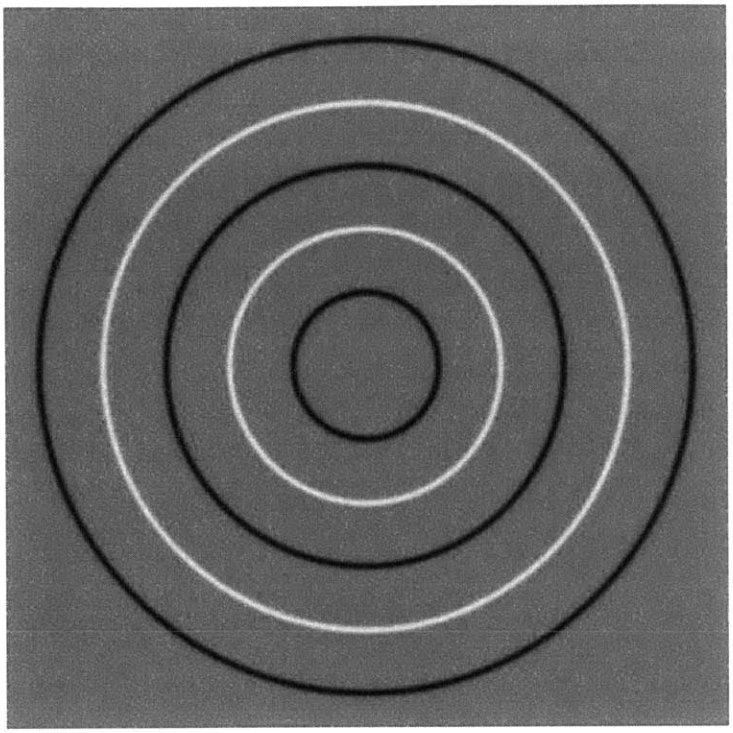


Figure A-2: Concentric Circles Alternating Polarity of Rings

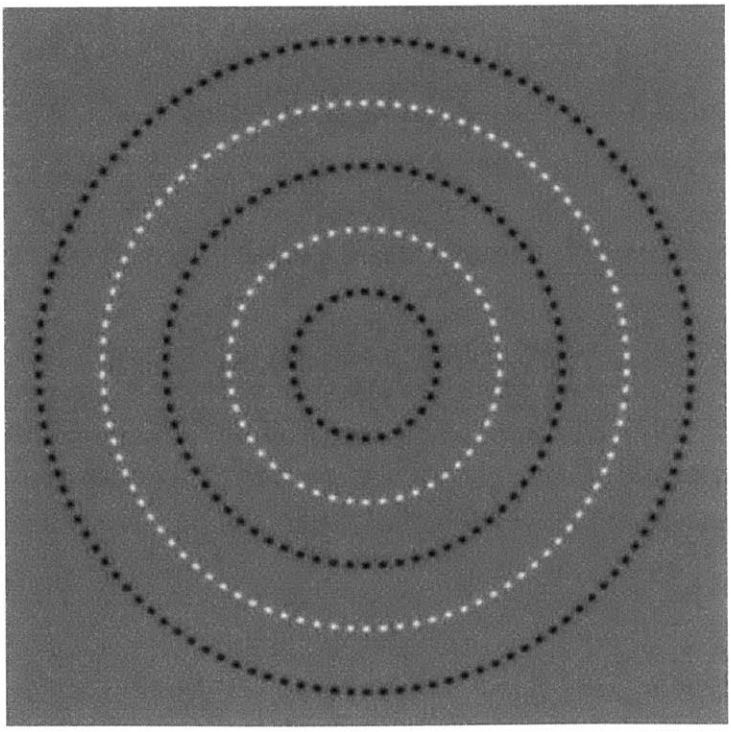


Figure A-3: Dots in Concentric Circles Alternating Polarity of Rings

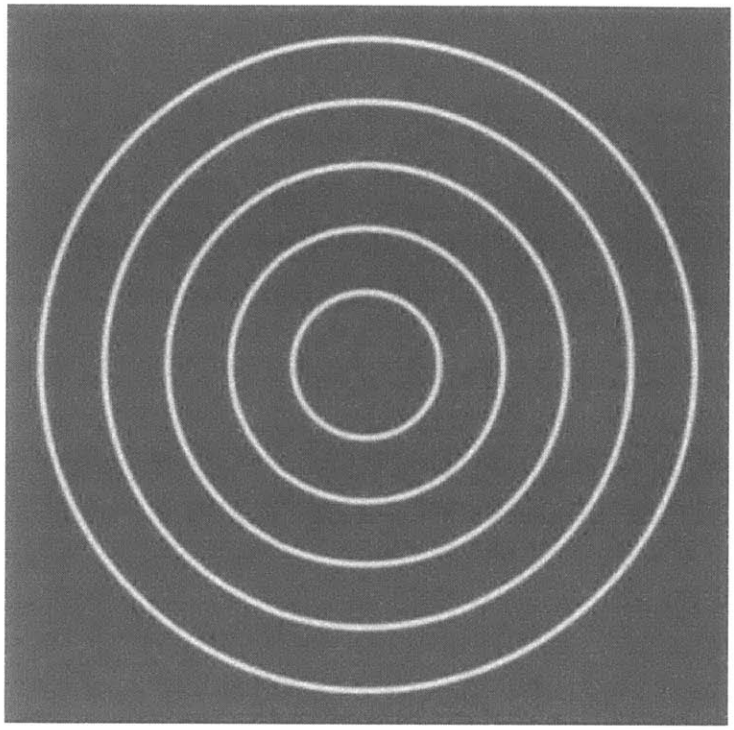


Figure A-4: Concentric White Circles

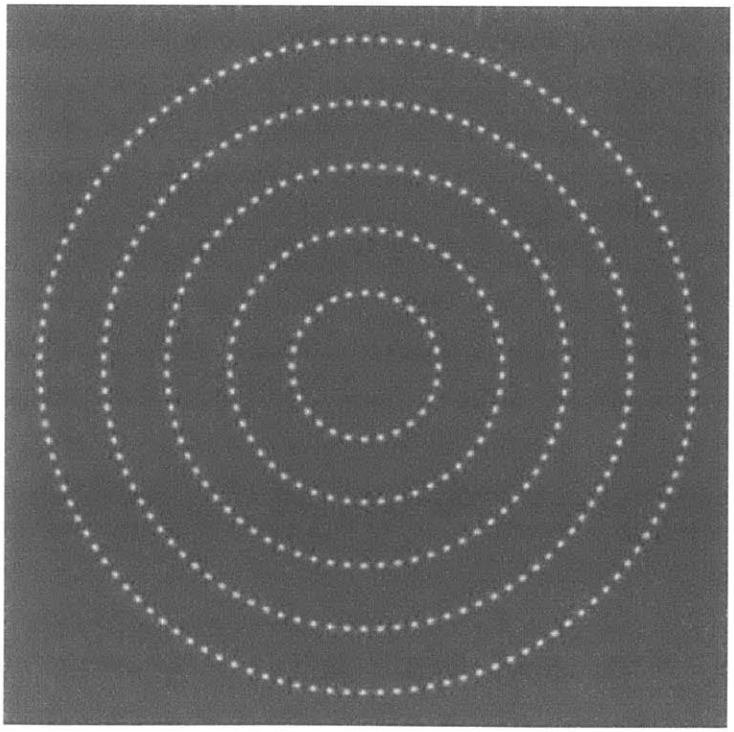


Figure A-5: Concentric White Dots in Circles

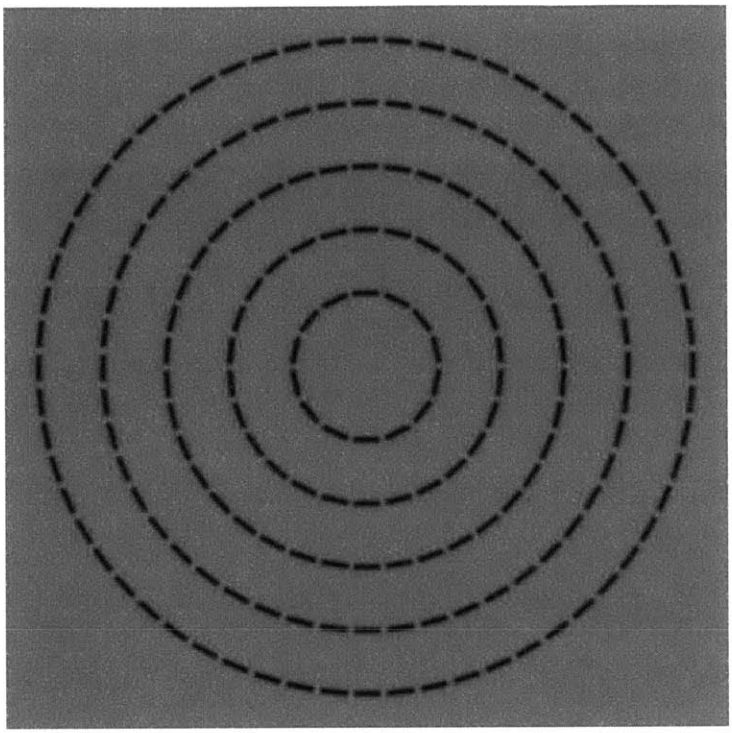


Figure A-6: Concentric Circles With Black Lines

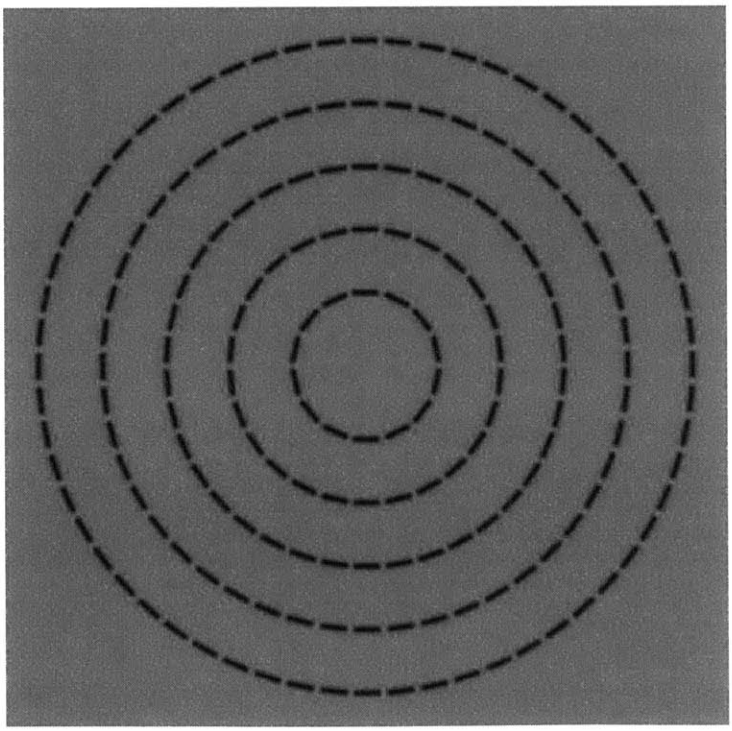


Figure A-7: Concentric Circles With White Lines

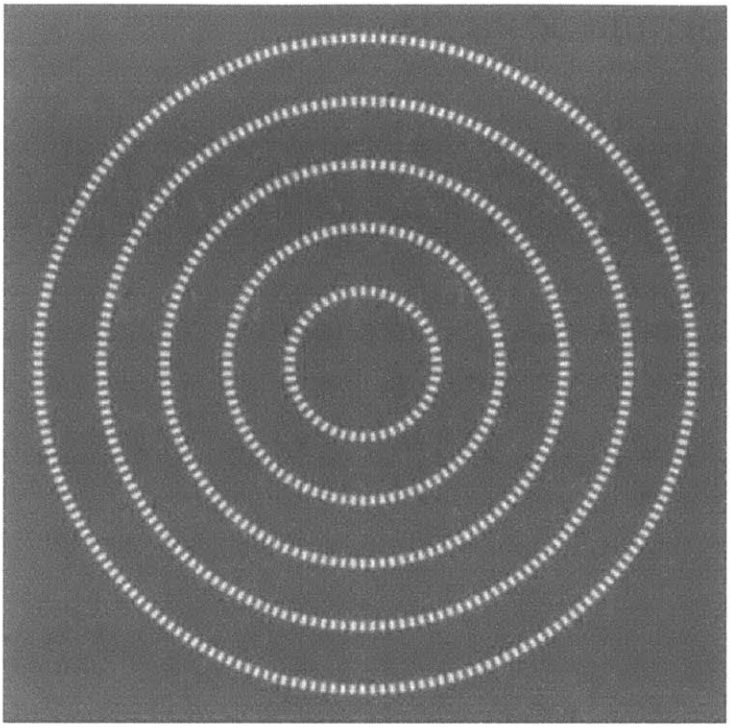
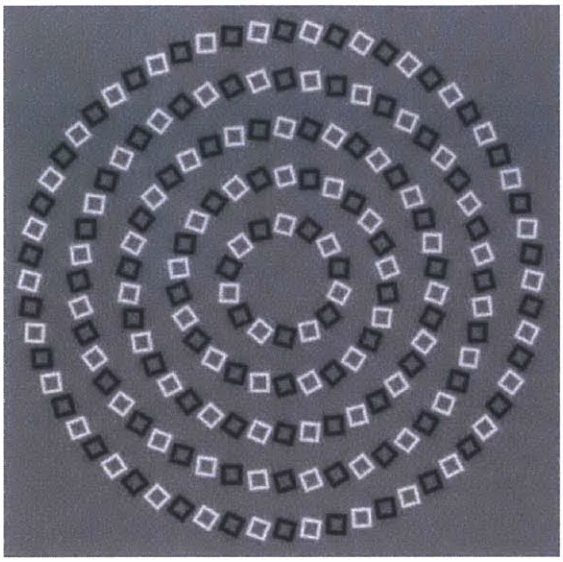
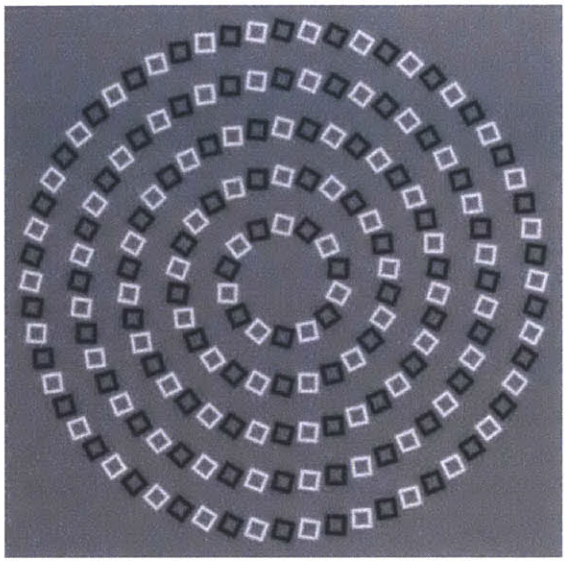


Figure A-8: Concentric Circles With White Lines 2

A.2 Color and Polarity

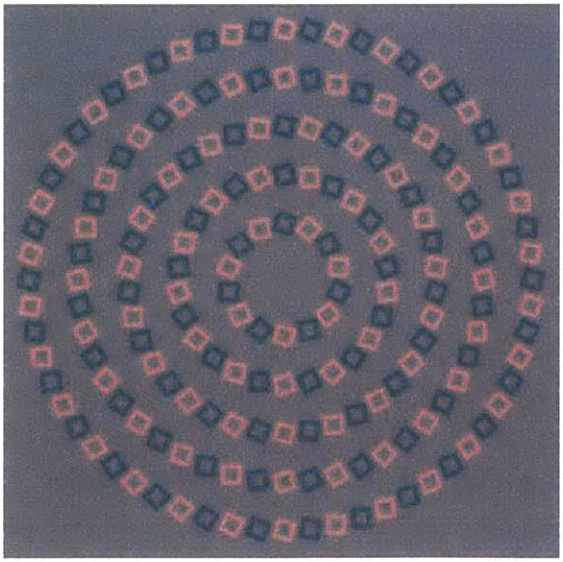


(a) Intertwining Illusion

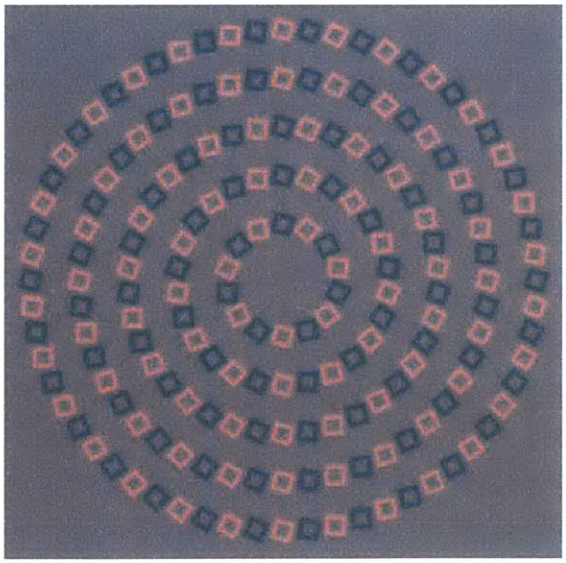


(b) Spiraling Illusion

Figure A-9: Unmodified Illusions

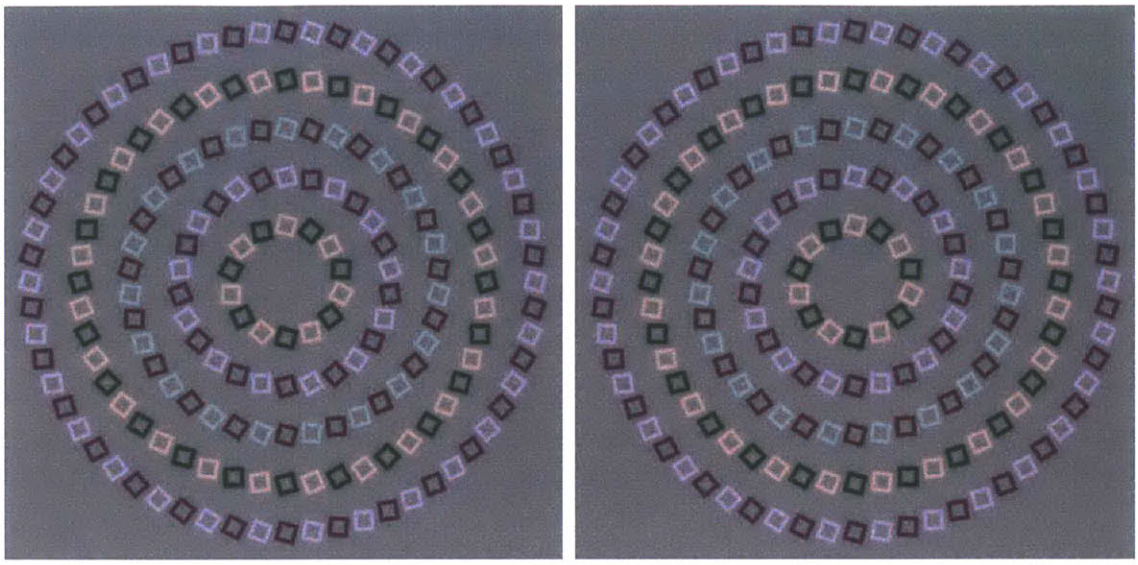


(a) Intertwining Illusion



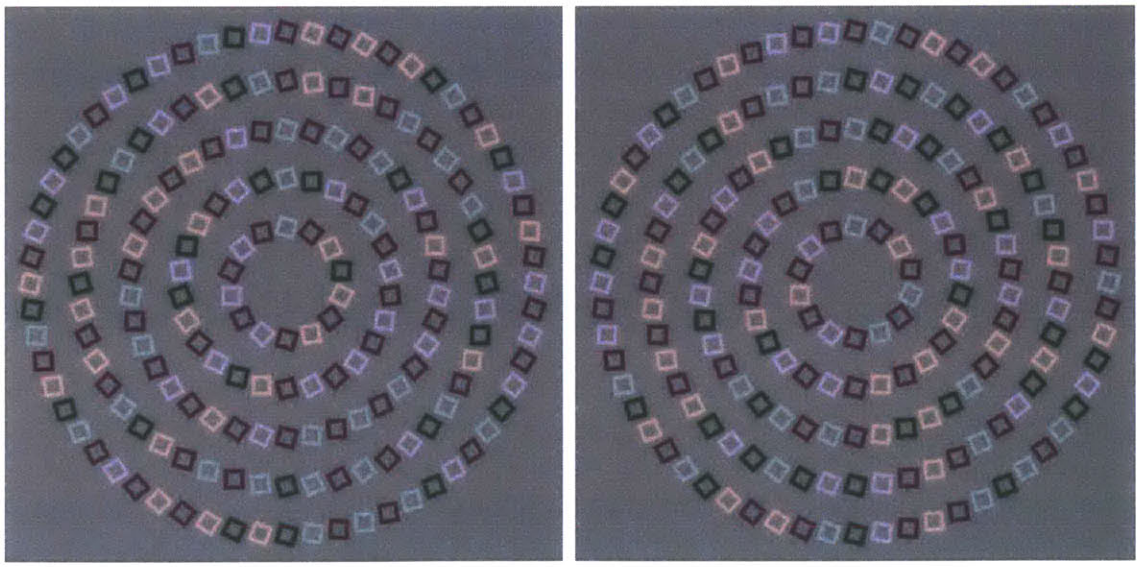
(b) Spiraling Illusion

Figure A-10: Alternating Polarity in Color



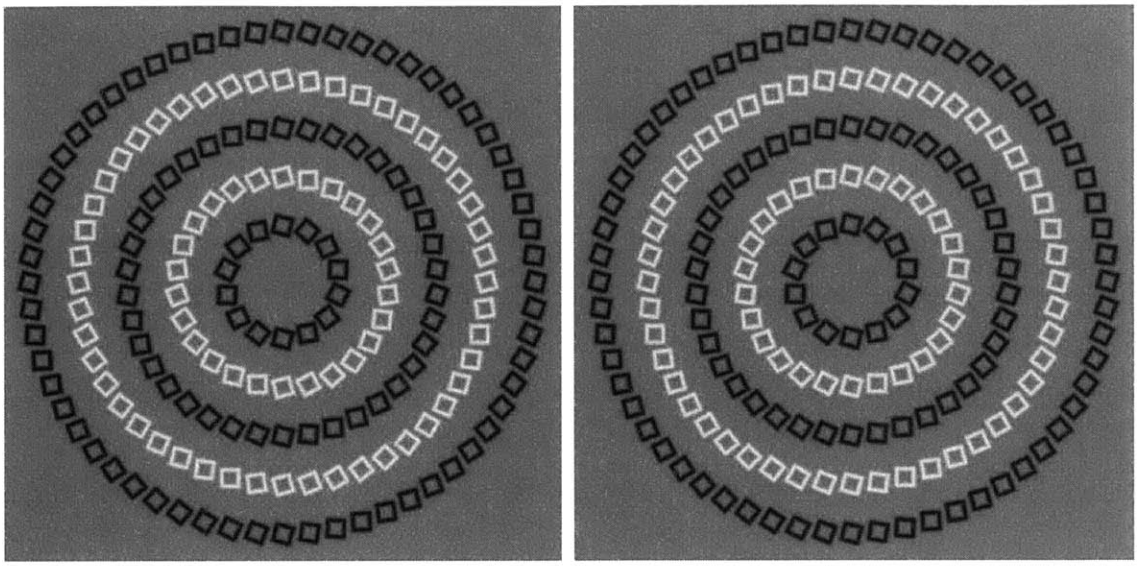
(a) Intertwining Illusion (b) Spiraling Illusion

Figure A-11: Alternating Polarity with Multiple Colors



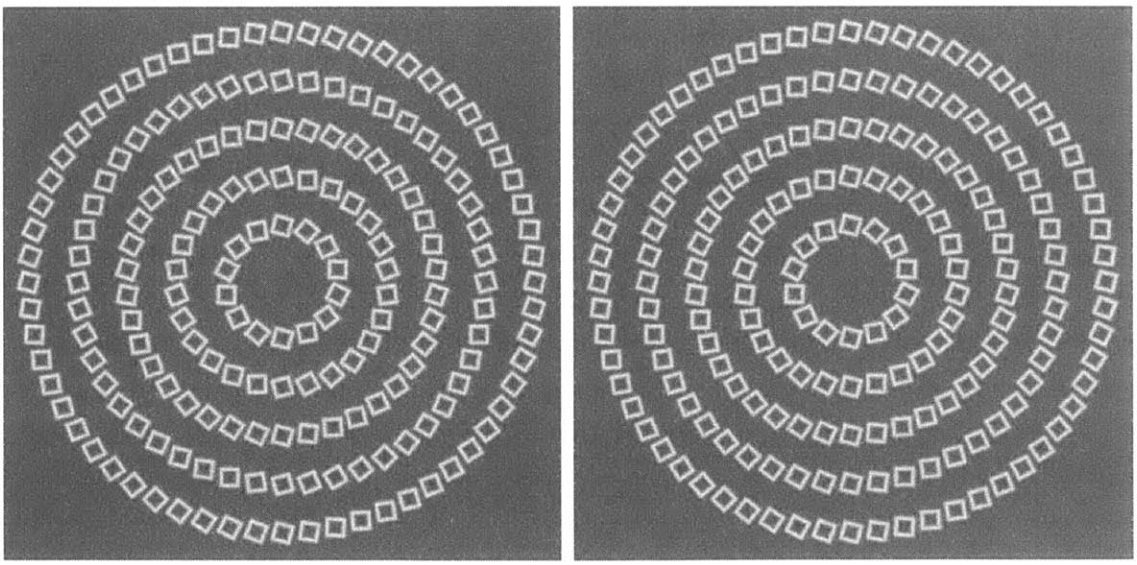
(a) Intertwining Illusion (b) Spiraling Illusion

Figure A-12: Alternating Polarity with Multiple Colors, Randomized Slightly



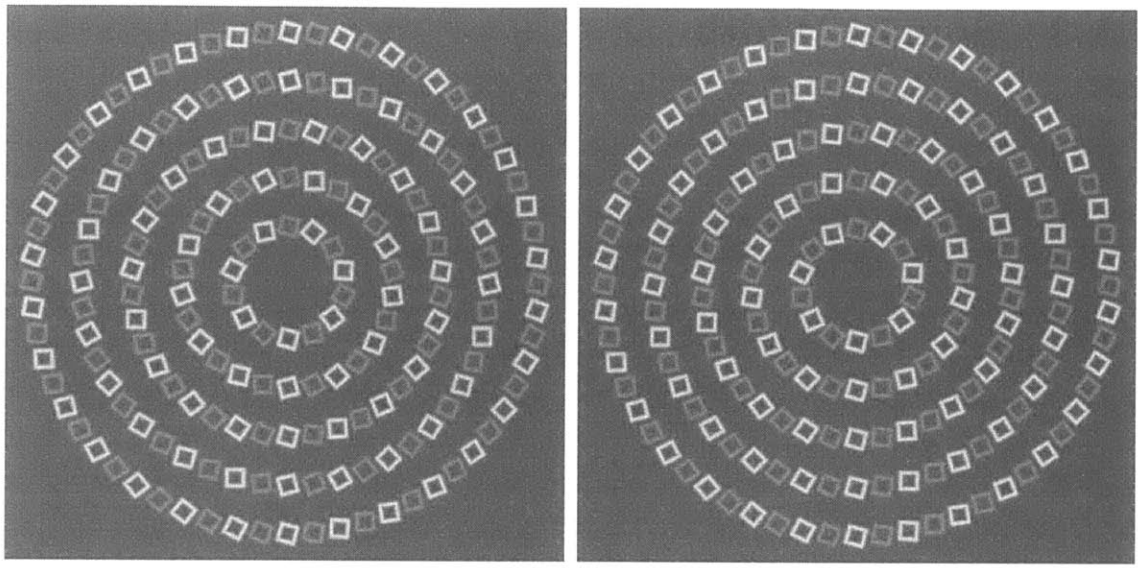
(a) Intertwining Illusion (b) Spiraling Illusion

Figure A-13: Alternating Polarity of Rings



(a) Intertwining Illusion (b) Spiraling Illusion

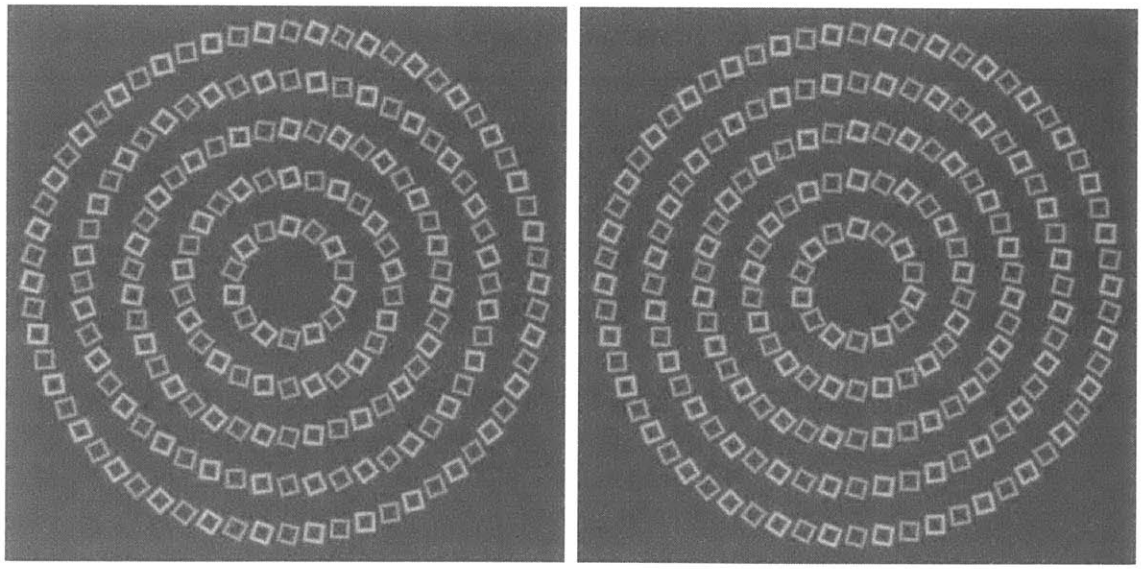
Figure A-14: Positive Polarity in One Tone



(a) Intertwining Illusion

(b) Spiraling Illusion

Figure A-15: Positive Polarity in Two Tones

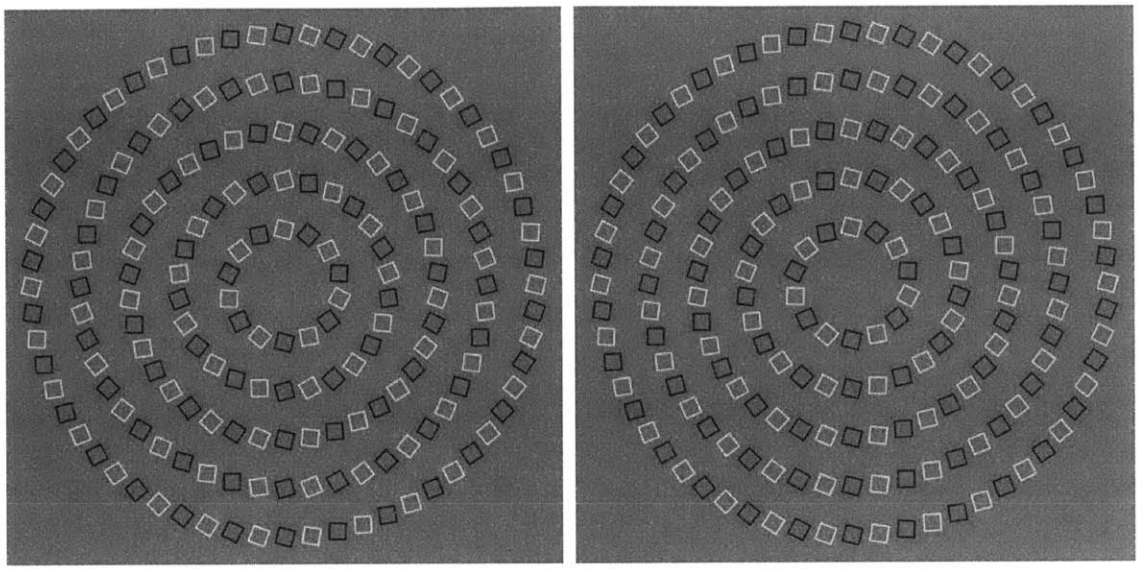


(a) Intertwining Illusion

(b) Spiraling Illusion

Figure A-16: Positive Polarity in Two Colors

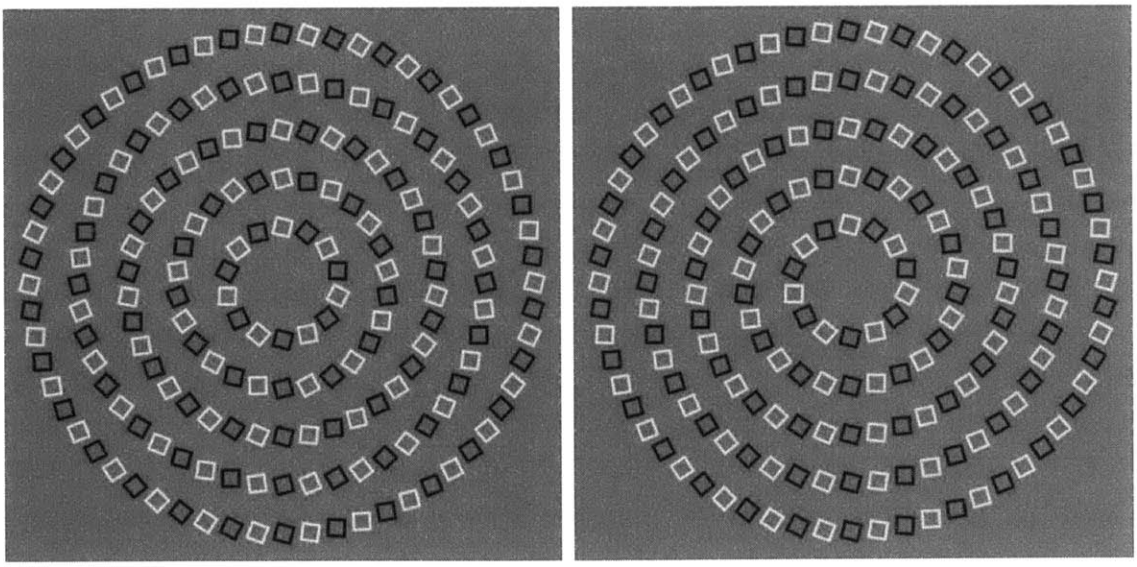
A.3 Square Width



(a) Intertwining Illusion

(b) Spiraling Illusion

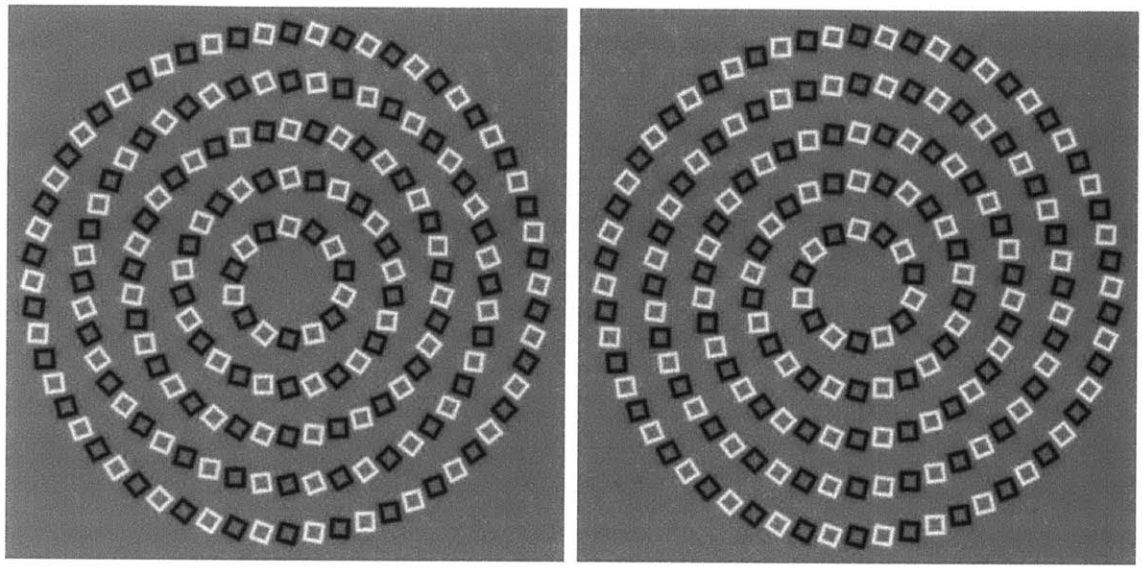
Figure A-17: Width 0.5



(a) Intertwining Illusion

(b) Spiraling Illusion

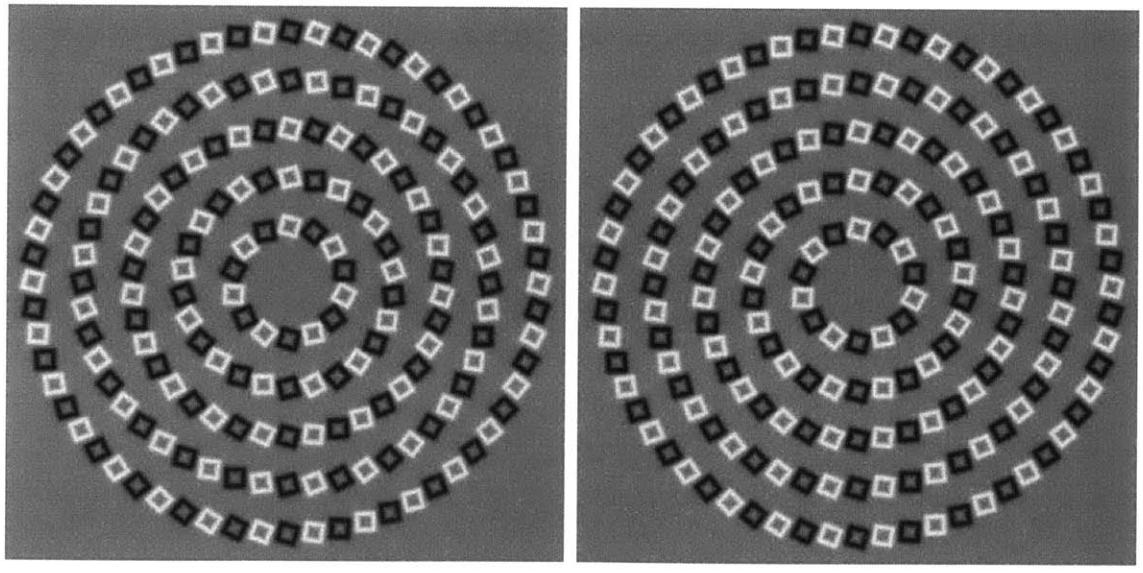
Figure A-18: Width 1.0



(a) Intertwining Illusion

(b) Spiraling Illusion

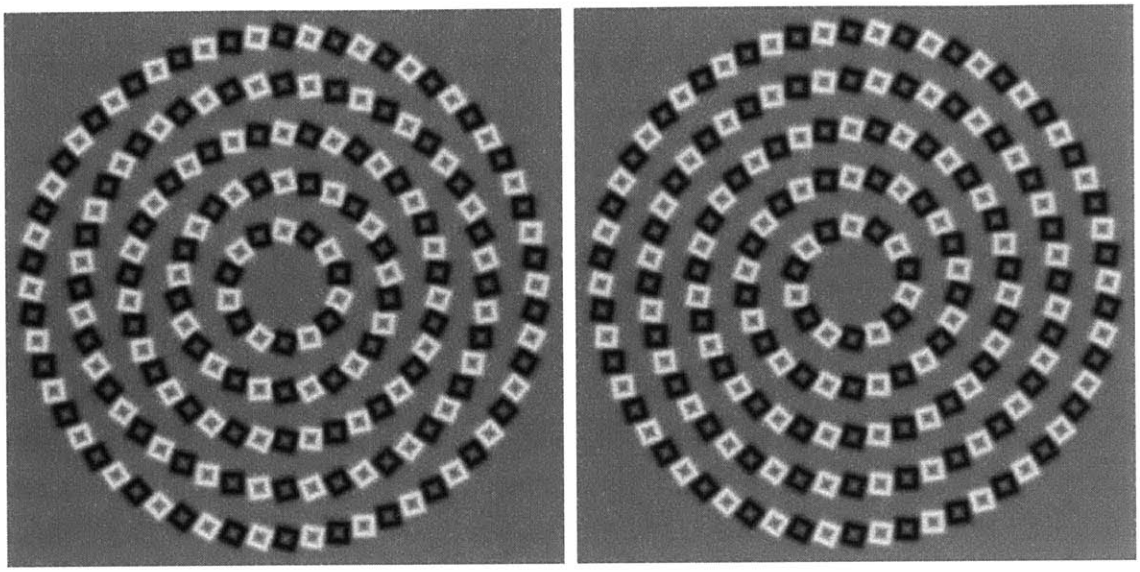
Figure A-19: Width 1.5



(a) Intertwining Illusion

(b) Spiraling Illusion

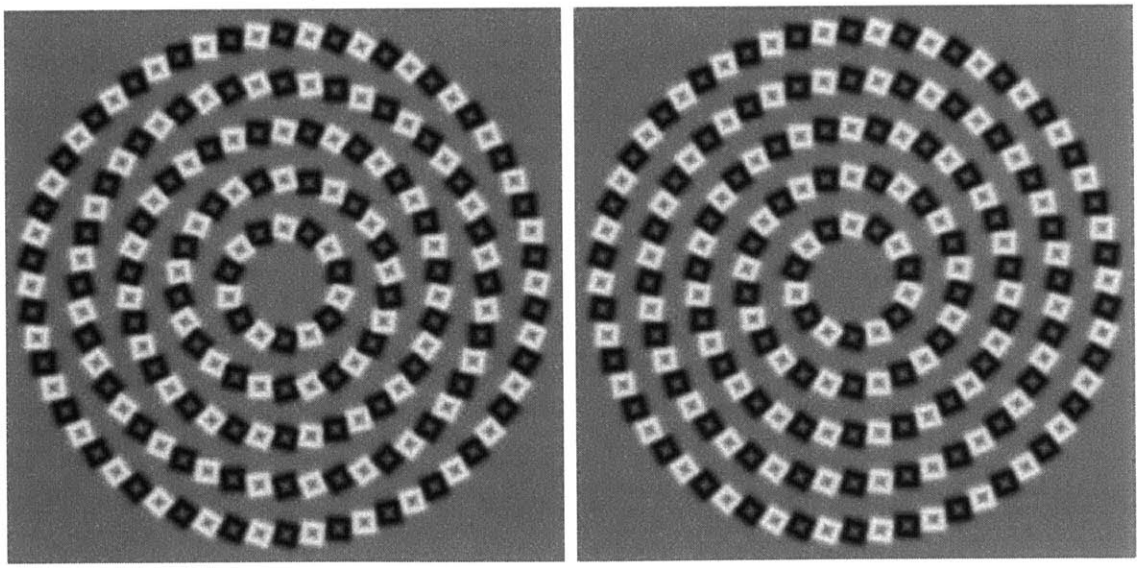
Figure A-20: Width 2.0



(a) Intertwining Illusion

(b) Spiraling Illusion

Figure A-21: Width 2.5

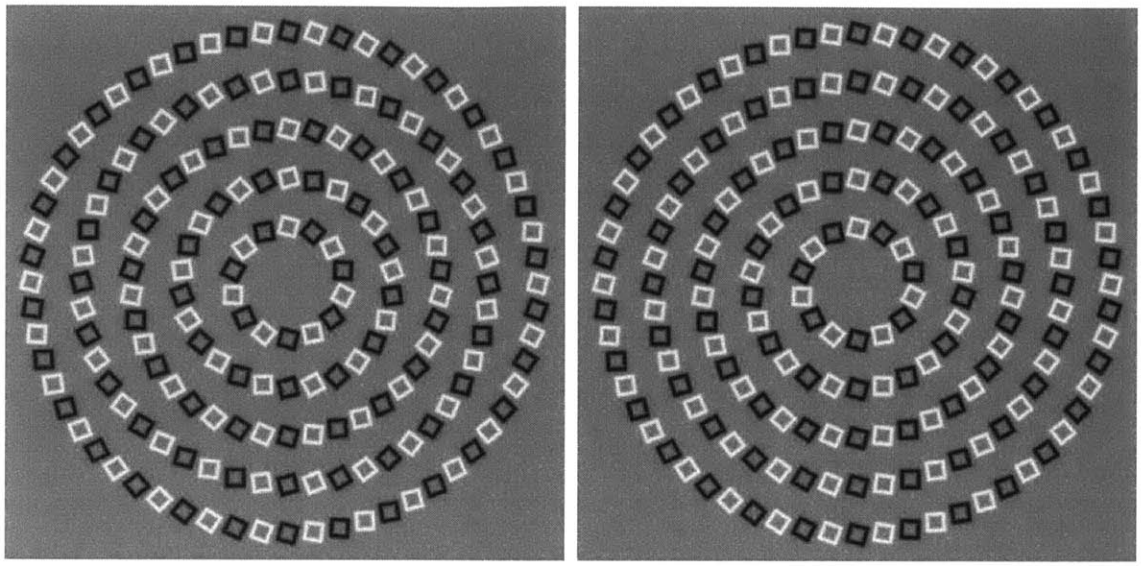


(a) Intertwining Illusion

(b) Spiraling Illusion

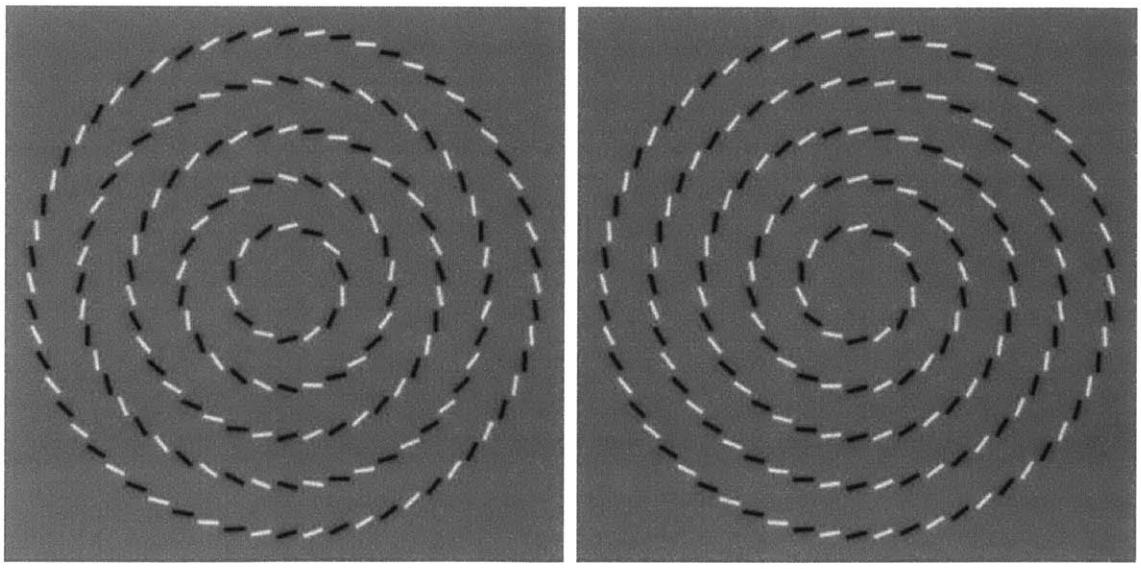
Figure A-22: Width 3.0

A.4 Shape of Elements



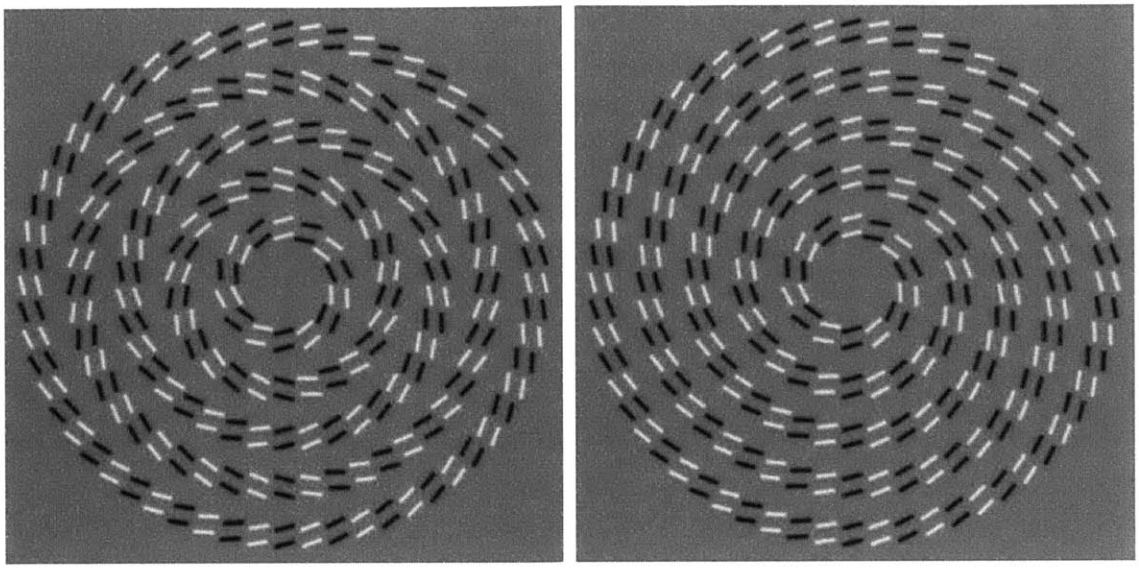
(a) Intertwining (b) Spiraling

Figure A-23: Squares (Unmodified Illusion)



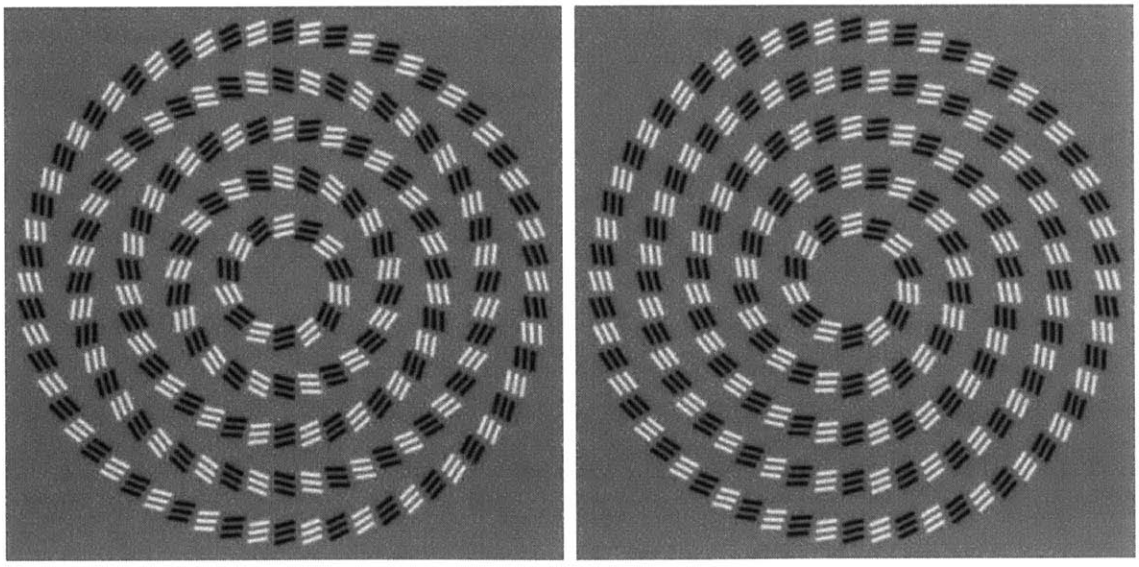
(a) Intertwining (b) Spiraling

Figure A-24: One Line



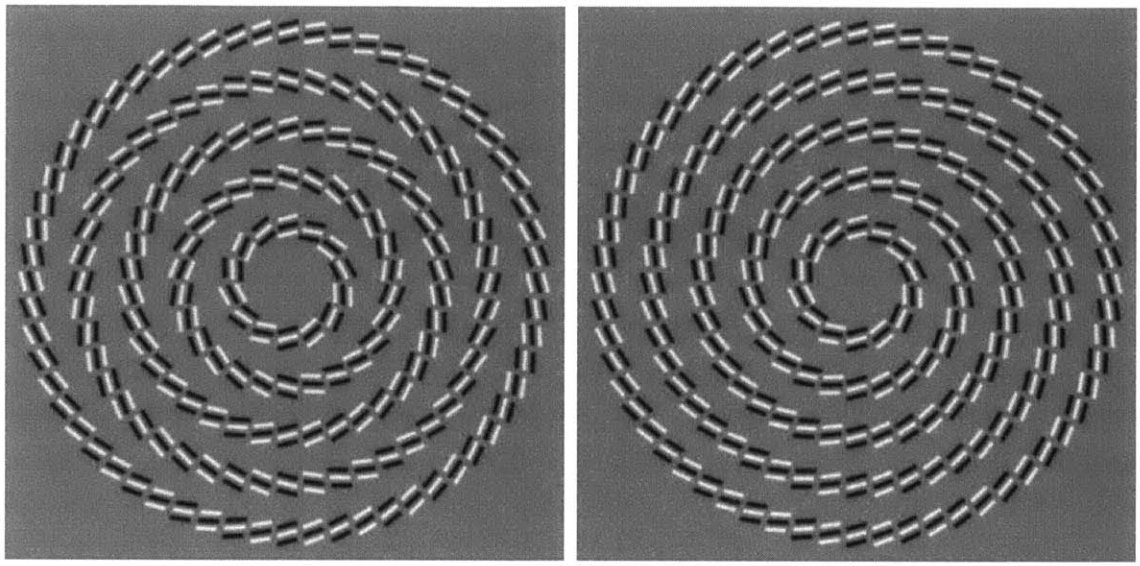
(a) Intertwining (b) Spiraling

Figure A-25: Two Lines



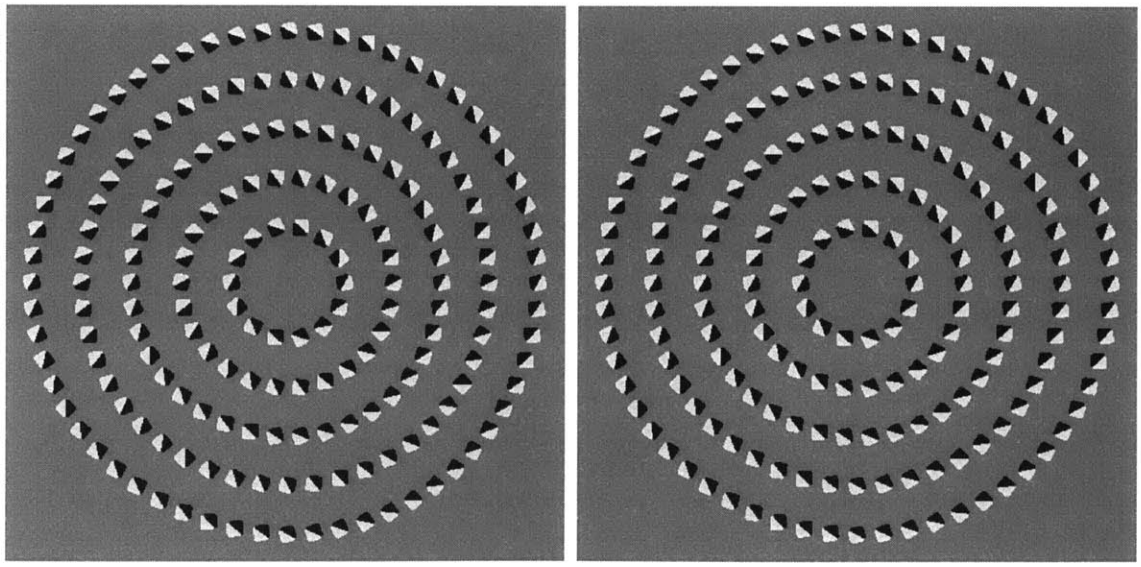
(a) Intertwining (b) Spiraling

Figure A-26: Three Lines



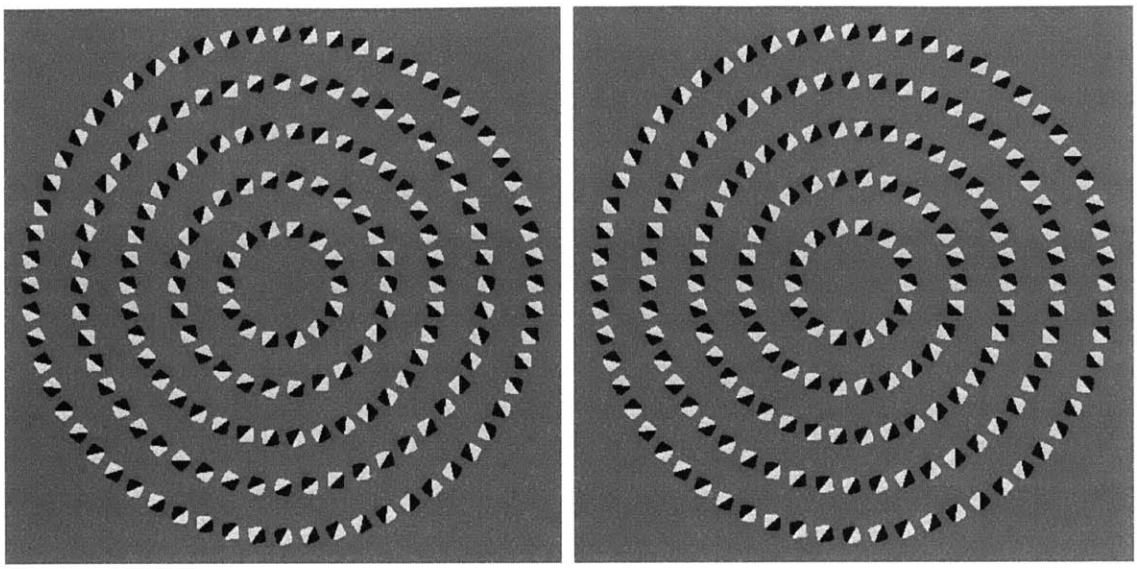
(a) Intertwining (b) Spiraling

Figure A-27: Three Lines, Middle Line Opposite Polarity



(a) Intertwining (b) Spiraling

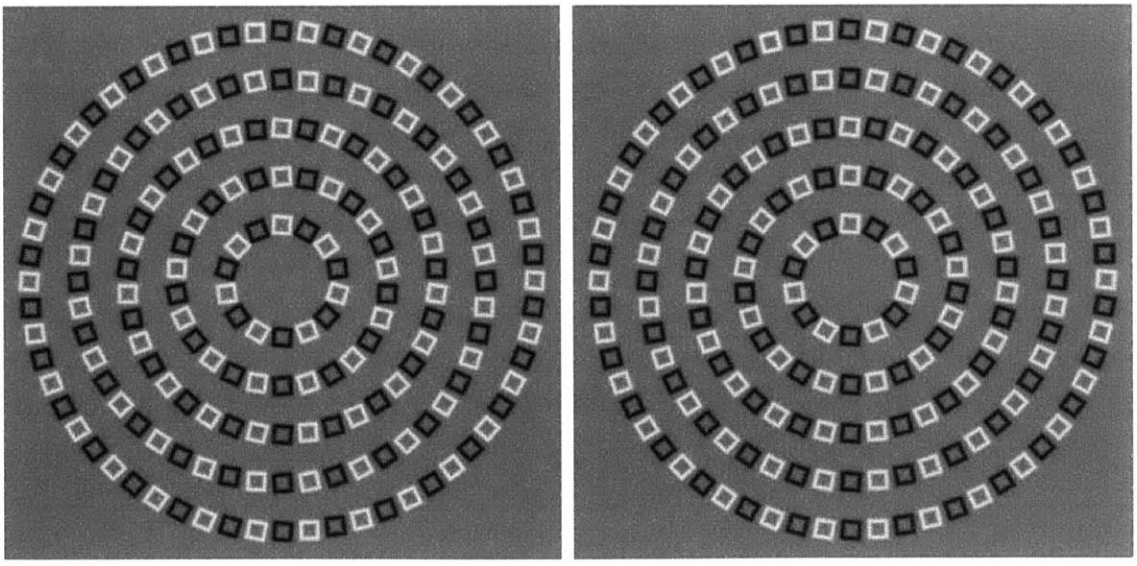
Figure A-28: Double Triangle 1



(a) Intertwining (b) Spiraling

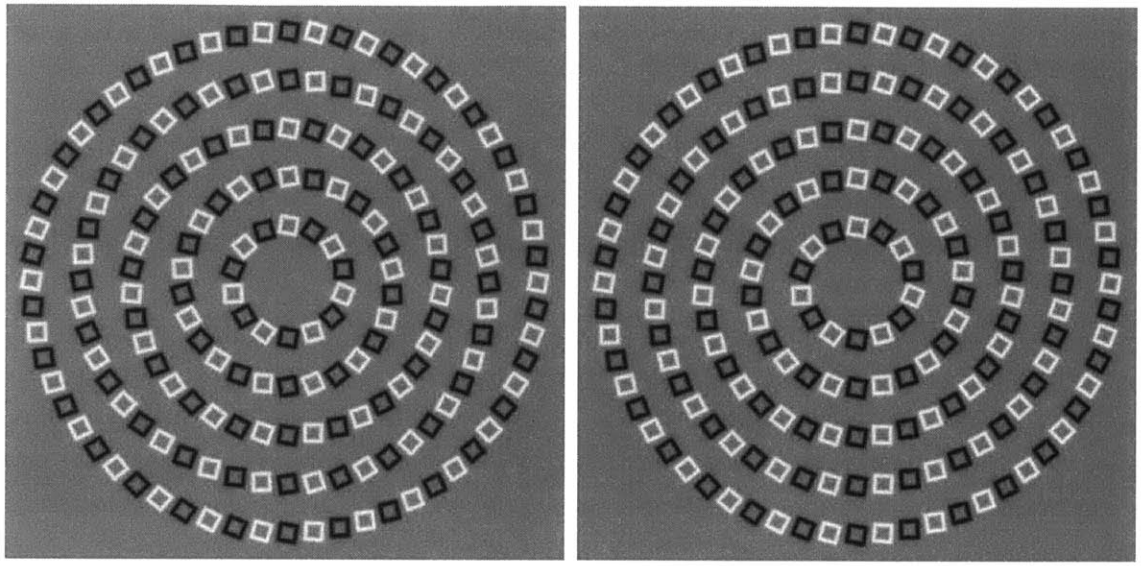
Figure A-29: Double Triangle 2

A.5 Tilt of Squares



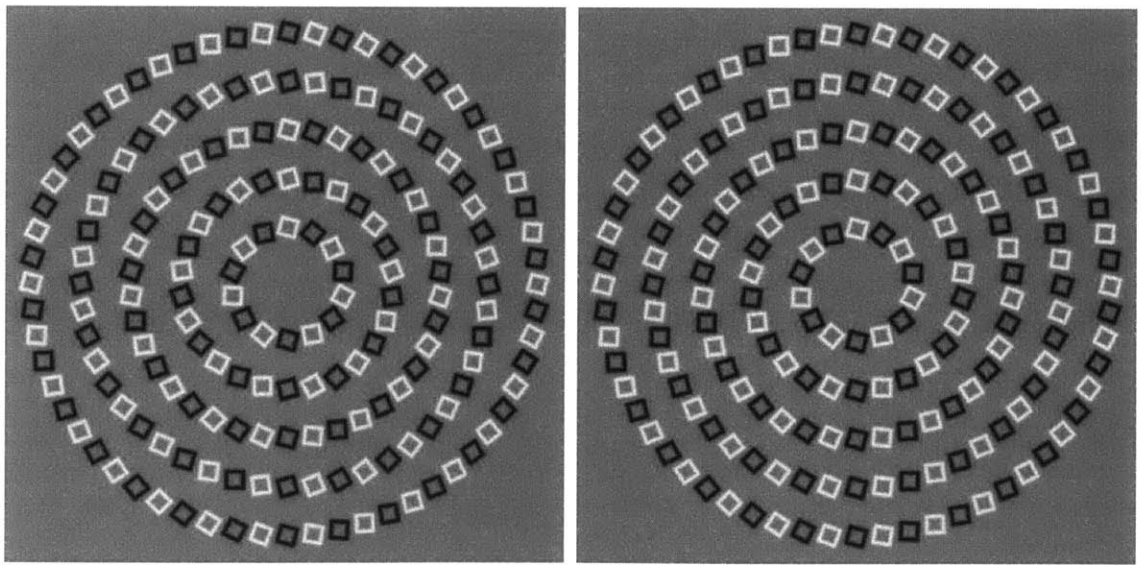
(a) Intertwining (b) Spiraling

Figure A-30: Tilted 5°



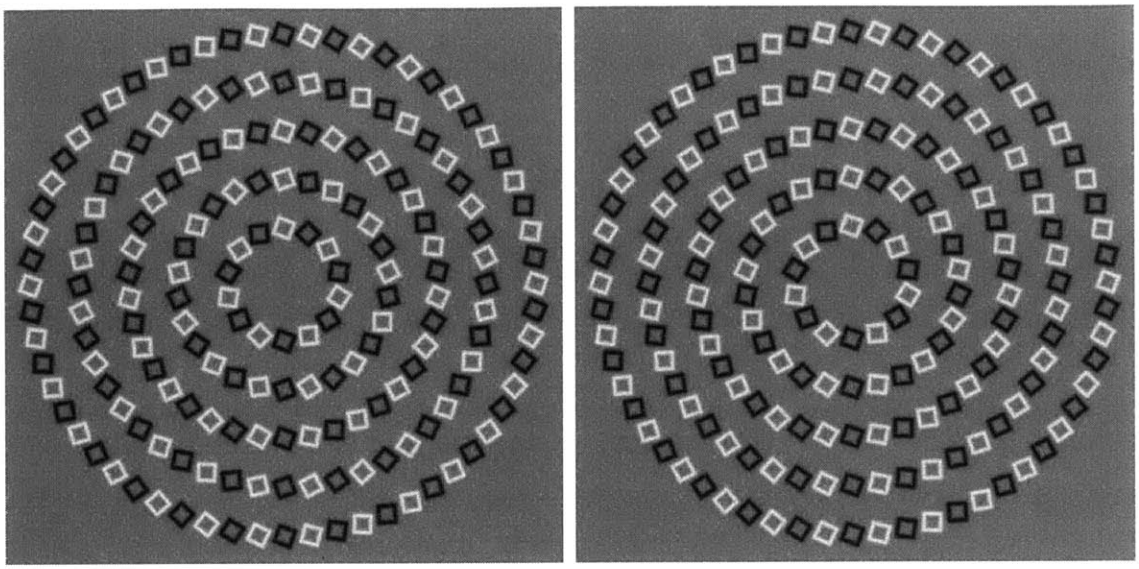
(a) Intertwining (b) Spiraling

Figure A-31: Tilted 10°



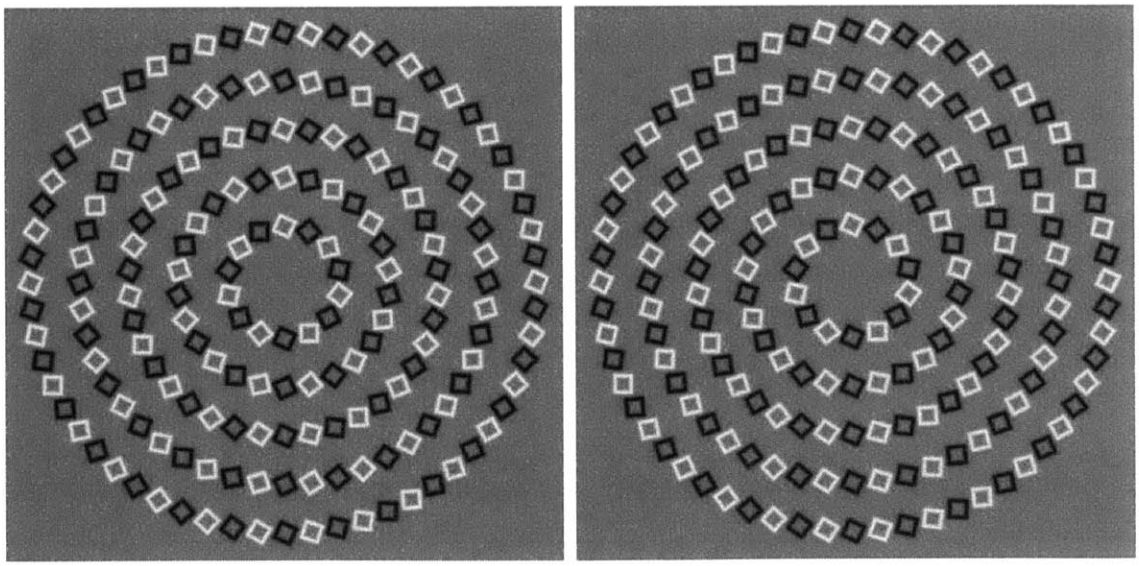
(a) Intertwining (b) Spiraling

Figure A-32: Tilted 15°



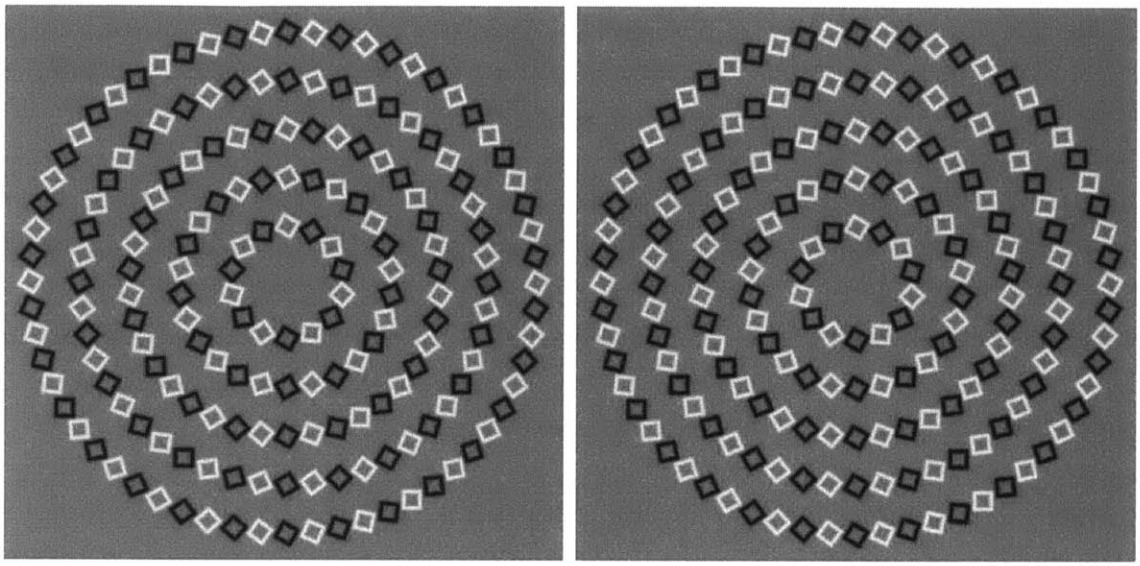
(a) Intertwining (b) Spiraling

Figure A-33: Tilted 20°



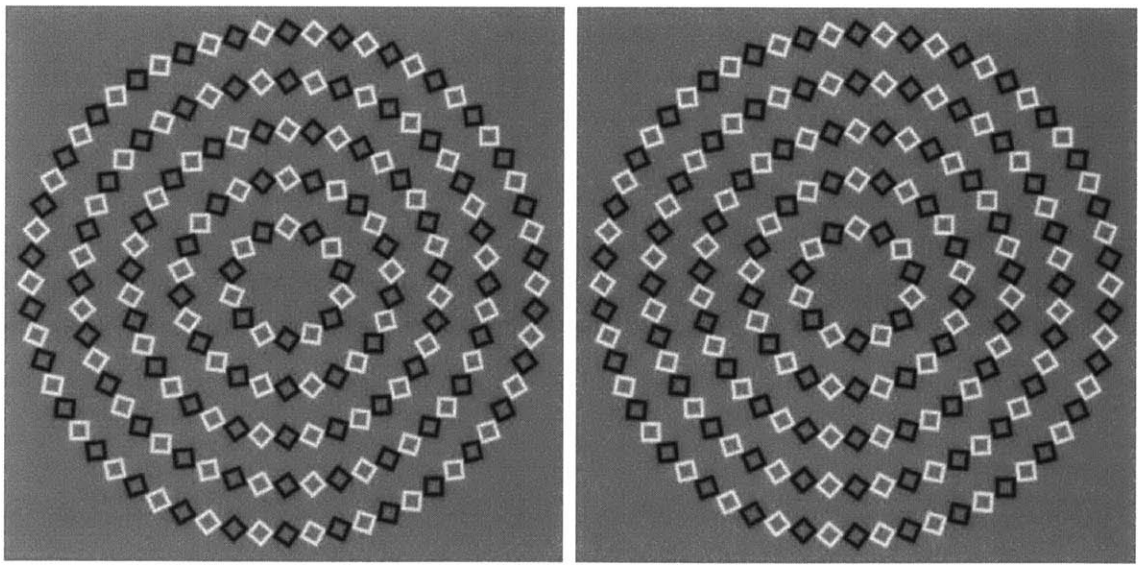
(a) Intertwining (b) Spiraling

Figure A-34: Tilted 25°



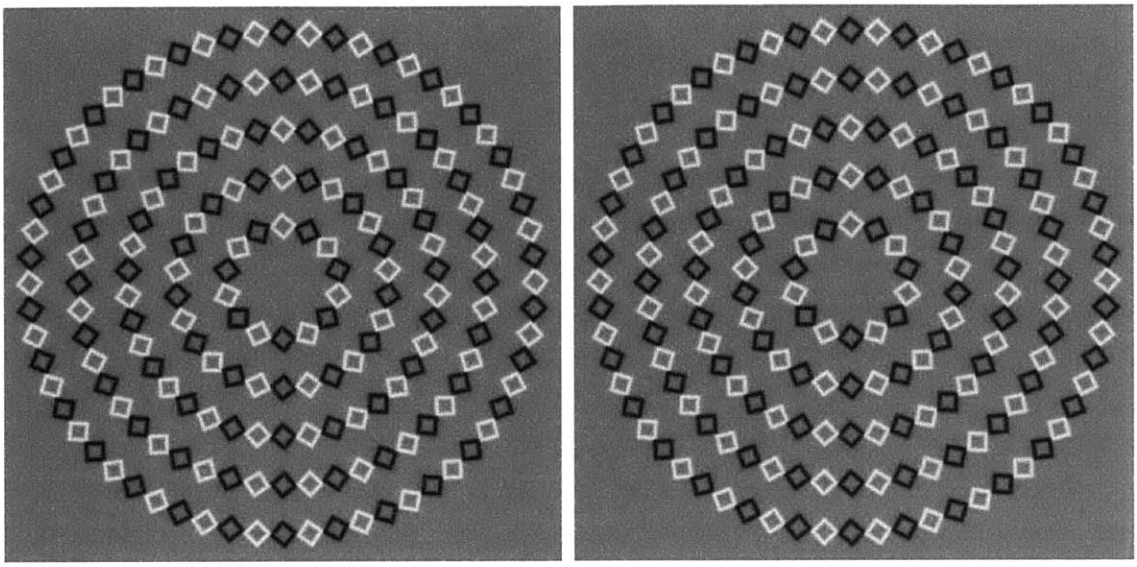
(a) Intertwining (b) Spiraling

Figure A-35: Tilted 30°



(a) Intertwining (b) Spiraling

Figure A-36: Tilted 35°



(a) Intertwining (b) Spiraling

Figure A-37: Tilted 40°

Appendix B

Illusory Patch Statistics

B.1 Patches from Illusion

Figures B-2 and B-4 show, respectively, visualizations of statistics taken from the images in Figures B-1 and B-3.

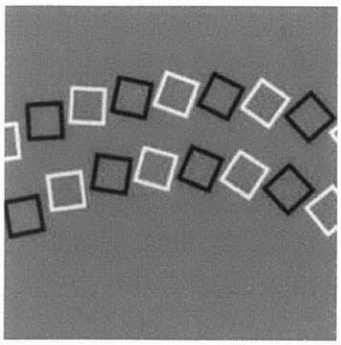


Figure B-1: Two-Lines Spiral Tilt: Black-White

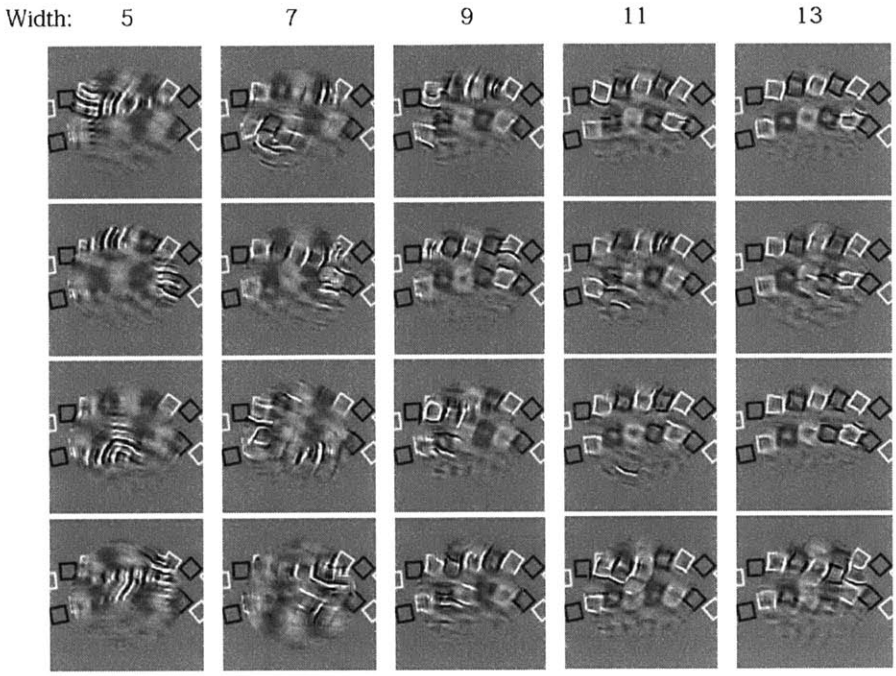


Figure B-2: Visualization of the statistics from Figure B-1. Each column corresponds to a different autocorrelation width window. Larger windows will collect more spatial information. The different rows correspond to different randomly generated seeds.

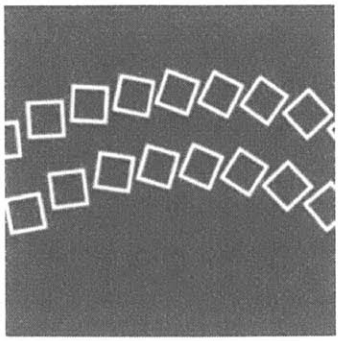


Figure B-3: Two-Lines Spiral Tilt: White

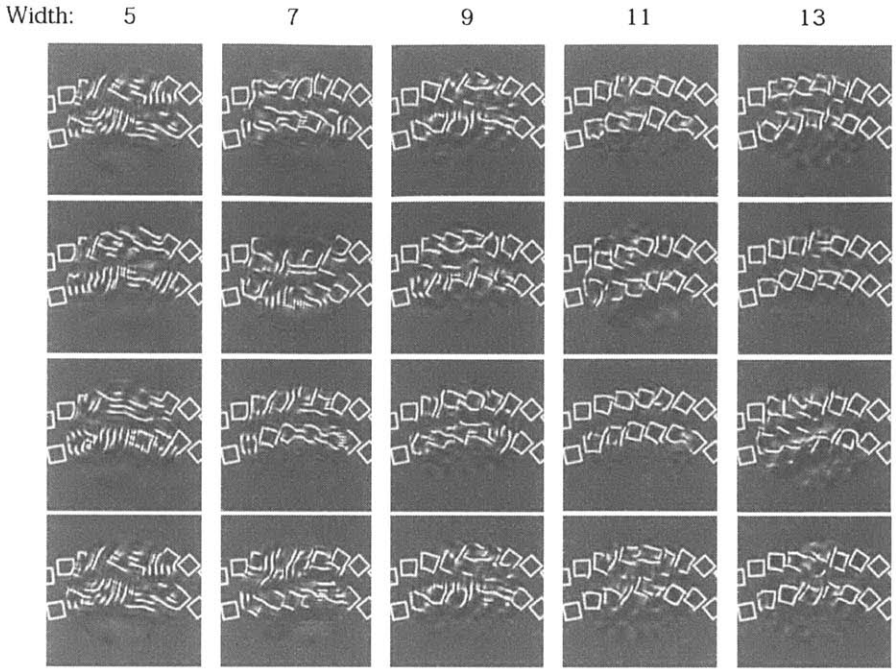


Figure B-4: Visualization of the statistics from the Figure B-3. Each column corresponds to a different autocorrelation width window. Larger windows will collect more spatial information. The different rows correspond to different randomly generated seeds.

Bibliography

[1] Stuart Anstis. Picturing peripheral acuity. *Perception*, 1998.

[2] Benjamin Balas. Texture synthesis and perception: using computational models to study texture representations in the human visual system. *Vision Research*, 2006.

[3] Benjamin Balas, Lisa Nakano, and Ruth Rosenholtz. A summary-statistic representation in peripheral vision explains visual crowding. *Journal of Vision*, 2009.

[4] H Bouma. Interaction effects in parafoveal letter recognition. *Nature*, 1970.

[5] D. Brainard. The psychophysics toolbox. *Spatial Vision*, 1997.

[6] Fermller C and Malm H. Uncertainty in visual processes predicts geometrical optical illusions. *Vision Research*, 2004.

[7] J. Duncan and G. W Humphreys. Visual search and stimulus similarity. *Psychological Review*, 1989.

[8] M. P. Eckstein, B. Drescher, and S. S. Shimozaki. Attentional cues in real scenes, saccadic targeting and bayesian priors. *Psychological Science*, 2006.

[9] Alexei A. Efros and William T. Freeman. Image quilting for texture synthesis and transfer. *Proceedings of SIGGRAPH 2001*, pages 341–346, August 2001.

[10] Alexei A. Efros and Thomas K. Leung. Texture synthesis by non-parametric sampling. In *IEEE International Conference on Computer Vision*, pages 1033–1038, Corfu, Greece, September 1999.

[11] J. Enns and R. Rensink. Sensitivity to three-dimensional orientation in visual search. *Psychological Science*, 1990.

[12] James T. Enns and Vincent Di Lollo. Object substitution: A new form of masking in unattended visual locations. *Psychological Science*, 8(2):135–139, 1997.

[13] J. M Findlay. Saccade target selection during visual search. *Vision Research*, 1997.

[14] Jeremy Freeman and Eero Simoncelli. Metamers of the ventral stream. *Nature Neuroscience*, 2011.

[15] Stephanie C. Goodhew, Troy A. W. Visser, Ottmar V. Lipp, and Paul E. Dux. Competing for consciousness: Prolonged mask exposure reduces object substitution masking. *Journal of experimental psychology*, 37(2):588, 2011.

[16] David J. Heeger and James R. Bergen. Pyramid-based texture analysis/synthesis. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, SIGGRAPH '95, pages 229–238, New York, NY, USA, 1995. ACM.

[17] Friedrich Heitger, Lukas Rosenthaler, Rdiger Von Der Heydt, Esther Peterhans, and Olaf Kbler. Simulation of neural contour mechanisms: from simple to end-stopped cells. *Vision Research*, 32(5):963 – 981, 1992.

[18] A Holcombe. Seeing slow and seeing fast: two limits on perception. *Trends in Cognitive Science*, 2009.

[19] T. Horowitz and J. M. Wolfe. Visual search has no memory. *Nature*, 1998.

[20] A. Hyvrinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley & Sons, 2001.

[21] Najemnik J and Geisler W. Optimal eye movement strategies in visual search. *Nature*, 2005.

[22] Ryota Kanai. "healing grid" annual best illusion of the year contest 2005. *Vision Sciences Society*, 2005.

[23] rni Kristjnsson. In search of remembrance: Evidence for memory in visual search. *Psychological Science*, 11(4):328–332, 2000.

[24] Vivek Kwatra, Arno Schodl, Irfan Essa, Greg Turk, and Aaron Bobick. Graphcut textures: Image and video synthesis using graph cuts. *ACM Transactions on Graphics, SIGGRAPH 2003*, 22(3):277–286, July 2003.

[25] Adam M. Larson and Lester C. Loschky. The contributions of central versus peripheral vision to scene gist recognition. *Journal of Vision*, 9(10), 2009.

[26] J Y Lettvin. On seeing sidelong. *The Sciences*, 1976.

[27] Dennis Levi. Crowding - an essential bottleneck for object recognition: a mini-review. *Vision Research*, 2008.

[28] Vincent Di Lollo, James Enns, and Ronald Rensink. Competition for consciousness among visual events: The psychophysics of reentrant visual processes. *Journal of Experimental Psychology: General*, 2000.

[29] Martelli M., Majaj N. J., and Pelli D. G. Are face processed like words a diagnostic test for recognition by parts. *Journal of Vision*, 2005.

[30] E Matin. Saccadic suppression: a review and an analysis. *Psychological Bulletin*, 1974.

[31] J McDermott and E Simoncelli. Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis. *Neuron*, 2011.

[32] Cathleen Moore and Alejandro Lleras. On the role of object representations in substitution masking. *Journal of Experimental Psychology: Human Perception and Performance*, 2005.

[33] M. Concetta Morrone, David C. Burr, and Donatella Spinelli. Discrimination of spatial phase in central and peripheral vision. *Vision Research*, 29(4):433 – 445, 1989.

[34] Aude Oliva and Antonio Torralba. Chapter 2 building the gist of a scene: the role of global image features in recognition. In S. Martinez-Conde, L.M. Martinez S.L. Macknik, J.-M. Alonso, and P.U. Tse, editors, *Visual Perception Fundamentals of Awareness: Multi-Sensory Integration and High-Order Perception*, volume 155, Part B of *Progress in Brain Research*, pages 23 – 36. Elsevier, 2006.

[35] J. Palmer, C. T. Ames, and D. T. Lindsey. Measuring the effect of attention on simple visual search. *Journal of Experimental Psychology: Human Perception and Performance*, 1993.

[36] J. Palmer, P. Verghese, and M Pavel. The psychophysics of visual search. *Vision Research*, 2000.

[37] L. Parkes, J. Lund, A Angelucci, J A Solomon, and M Morgal. Compulsory averaging of crowded orientation signals in human vision. *Nature Neuroscience*, 2001.

[38] Karl Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 1901.

[39] B. Pinna. Pinna illusion. *Scholarpedia*, 2008.

[40] B. Pinna and R. Gregory. Shifts of edges and deformations of patterns. *Perception*, 2002.

[41] Javier Portilla and Eero P. Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, 40:49–70, 2000. 10.1023/A:1026553619983.

[42] Mary Potter. Short-term conceptual memory for pictures. *Journal of Experimental Psychology: Human Learning*, 1976.

[43] RW Rodieck. *The First Step in Seeing.* Sinauer Associates, 1998.

[44] Ruth Rosenholtz, Jie Huang, Alvin Raj, Benjamin J. Balas, and Livia Ilie. A summary statistic representation in peripheral vision explains visual '. *Journal of Vision*, 12(4), 2012.

[45] B C Russell, A Torralba, K P Murphy, and W T Freeman. Labelme: A database and web-based tool for image annotation. *International Journal of Computer Vision*, 2008.

[46] Arthur G. Shapiro, Emily J. Knight, and Zhong-Lin Lu. A first- and second-order motion energy analysis of peripheral motion illusions leads to further evidence of feature blur in peripheral vision. *PLoS ONE*, 6(4):e18719, 04 2011.

[47] Lavanya Sharan. *The perception of material qualities in real-world images.* PhD thesis, Massachusetts Institute of Technology, September 2009.

[48] J Sun and P. Perona. Early computation of shape and reflectance in the visual system. *Nature*, 1996.

[49] S. Thorpe, D. Fize, and C. Marlot. Speed of processing in the human visual system. *Nature*, 1996.

[50] Alexander Toet and Dennis M. Levi. The two-dimensional shape of spatial interaction zones in the parafovea. *Vision Research*, 32(7):1349 – 1357, 1992.

[51] Antonio Torralba. Modeling global scene factors in attention. *Journal of Optical Society of America A (Special Issue on Bayesian and Statistical Approaches to Vision)*, 2003.

[52] A. Treisman and S. Gormican. Feature analysis in early vision: evidence from search asymmetries. *Psychological Review*, 1988.

[53] A. Treisman and J. Souther. Search asymmetry: A diagnostic for preattentive processing of separable features. *Journal of Experimental Psychology: General*, 1985.

[54] Anne M. Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97 – 136, 1980.

[55] R. van den Berg, J.B.T.M. Roerdink, and F.W. Cornelissen. On the generality of crowding: Visual crowding in size, saturation, and hue compared to orientation. *Journal of Vision*, 7(2), 2007.

[56] Rodieck R W. *The First Step in Seeing.* Sinauer Associates, 1998.

[57] A. H. Wertheim, I. T. C. Hooge, K. Krikke, and A.. Johnson. How important is lateral masking in visual search. *Experimental Brain Research*, 2006.

[58] J. M. Wolfe. Guided search 2.0: A revised model ofvisual search. *Psychonomic Bulletin and Review*, 1994.

[59] J. M. Wolfe. Asymmetries in visual search: An introduction. *Perception and Psychophysics*, 2001.

[60] J. M. Wolfe, K. R. Cave, and S. L Franzel. Guided search: An alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human Perception and Performance*, 1989.

[61] Zelinsky. A theory of eye movements during target acquisition. *Psychological Review*, 2008.