At the interface of materials and objects in peripheral vision

by

Shaiyan Keshvari

B.S. in Physics and Astronomy
B.A. in Cognitive Sciences
Rice University, 2010

Submitted to the Department of Brain and Cognitive Sciences in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

August 2016 [September 2016]

© Massachusetts Institute of Technology 2016. All rights reserved.
At the interface of materials and objects in peripheral vision

by

Shaiyan Keshvari

Submitted to the Department of Brain and Cognitive Sciences on August 12th, 2016 in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Abstract

Humans are able to simultaneously perceive the world as discrete, distinct "objects", as well as regions of statistical regularity, or "textures". This is evident in the way we describe our perceptual world. A street is made up of concrete and asphalt "stuff", while the people and dogs walking on it are the "things" that make use of it. Both of these types of representation, however, are derived from the same sensory input, and thus there must exist transformations that map one to the other. A complete model of perception must account for these transformations. I study the representations that lie at the interface of object and texture perception in vision, focusing on utilizing the intrinsically impaired perception in the periphery to disambiguate the predictions of different models. I find that many seemingly separate perceptual phenomena in crowding can be better understood as different aspects of a single underlying model. I extend this to the study of material perception, and find that considering images of materials as visual textures can explain human’s ability to recognize materials in the periphery. Furthermore, I examine how the limitations of peripheral vision affects the perception of visual designs, namely webpages.

Thesis Supervisor: Ruth Rosenholtz
Title: Senior Research Scientist, Department of Brain and Cognitive Sciences
Table of Contents

**Introduction**........................................................................................................................10

**Pooling of continuous features provides a unifying account of crowding** ......................13

2.1 Abstract .......................................................................................................................... 14

2.2 Introduction .................................................................................................................. 14

2.3 Generating predictions from the Texture Tiling Model (TTM) ........................................ 17

2.4 Experiments ................................................................................................................... 18

2.4.1 Can a simple substitution model account for crowding? ........................................ 18

2.4.1.1 (Freeman et al., 2012)’s experiments ......................................................................18

2.4.1.2 Stimuli and procedure for generating TTM predictions for (Freeman et al., 2012)’s stimuli .................................................................................................................18

2.4.1.3 Results of (Freeman et al., 2012) replication ..................................................... 19

2.4.2 Do different visual features crowd in concert or independently? ............................ 21

2.4.2.1 (Greenwood et al., 2012)’s experiment .................................................................21

2.4.2.2 Stimuli and procedure for generating TTM predictions for (Greenwood et al., 2012)’s stimuli .............................................................................................................22

2.4.2.3 Results of (Greenwood et al., 2012) replication ................................................ 23

2.4.3 How does crowding affect feature integration? ........................................................ 24

2.4.3.1 (Põder & Wagemans, 2007)’s experiment ...............................................................25

2.4.3.2 Stimuli and procedure for generating TTM predictions for (Põder & Wagemans, 2007)’s stimuli ........................................................................................................ 25

2.4.3.3 Results of (Põder & Wagemans, 2007) replication .............................................. 25

2.5 Discussion ..................................................................................................................... 27

**Material perception, texture, and peripheral vision** ...................................................... 30

3.1 Abstract ......................................................................................................................... 31

3.2 Introduction ................................................................................................................... 31

3.3 Baseline and texture experiment methods .................................................................... 34

3.3.1 Selecting material images for stimuli .........................................................................34

3.3.2 Procedure for synthesis of material textures ............................................................ 35

3.3.3 Observers .................................................................................................................. 36

3.3.4 Experimental apparatus ......................................................................................... 36

3.3.5 Experimental procedure ......................................................................................... 37
Dedicated to my family and friends.

You are all I see, foveally and peripherally.
Acknowledgments

Ruth, your guidance and friendship has been indispensable to me. When I first saw your bubbly personality and sharp wit at interviews, I knew that yours was an offer I couldn’t turn down. You’ve been the main driver and supporter of my research at MIT, and I wouldn’t have had it any other way. Not only have you inspired me to think critically but positively about research, I’ve learned from you that researchers can be people too. I hope I can pick up even a little bit of the natural charm and confidence that is uniquely yours.

To the Rosenholtz Lab, you’ve been my second home since I arrived at MIT. I’ve learned about so much from all of you, from peripheral vision and material classification to cuttlefish camouflage. Much of this work, especially Ali, is directly or indirectly also yours. I also cherish the opportunity I’ve had to share lunch and office space with Ted’s Lab, with whom I’ve had wonderful times. The McDermott lab from its very beginning has been a fun and welcoming place. I’ve made some great friends there, and had an unparalleled opportunity to learn all about hearing. Thank you especially John-Elmer Canfield for all the work you do making MIT a great place to be.

Maarten, in addition to introducing me to the field of material perception, you showed me what true scientific passion is; I haven’t met another person that can get excited by simply looking at pictures of pieces of fabric.

Josh (McDermott), thank you for introducing me to the world of auditory perception and taking a vested interest in my development, from allowing me assist with your class, teaching me in classes, helping me get a fellowship, and advising my research. I really appreciate the opportunities you’ve given me and I know they have made a huge impact on my life and career.

Ted, I’ll just repeat what I’ve heard others say about you: He’s an amazing person. I won’t forget all you’ve shared with us, from stories youthful hijinks to expert lessons on presenting research clearly. I also appreciate knowing you in a non-academic context, when you invite us over for dinner and creative activities.

Josh (Tenenbaum), you and your lab have been an inspiration to me. I think you are all way too brilliant and kind to be real. I doubt I’ll ever be around a group of people as special as you all.

Prior to MIT, I developed my passion for cognitive science and neuroscience by We Ji Ma and Ronald van den Berg, who continue to be friends and an inspiration for me as I continue my research and career. I look forward to hopefully working again with you in the future.

Friends (in alphabetical order): Affi, Alex, Emily, Galen, George, James, Jiayong, Julian, Karen, Kim, Max, Omer, Praneeth, Zenna (and many more): thank you for the conversations, academic and otherwise, all the nights, food, trips. You are the people that I came to MIT to meet. I’m grateful for the opportunity to call you guys my friends, and you will be the people I will continue to cherish in the future.

Mike Hart, you know why you’re in here, for better or for worse. Adam, Nathan, Sunny: thanks for making Boston a little bit more like Texas.
Random Hall and Nina, I’m so grateful for the opportunity to spend four years learning from you in many more ways than a standard PhD program could offer. I’ve made many good friends, like Phil, Ben, Alison, Nora, and Erons, all of whom I look forward to spending time with as a graduate.

Thirsty Ear Pub, thanks for the opportunity to tackle a project completely different from science, and make friends along the way.

To Sophie, you were worth the wait, so glad my bad jokes finally worked.

To my family, I cannot possibly adequately put in words how important you’ve been during this intellectually and emotionally challenging part of my life. So, I’ll keep it simple: this thesis is for you.
Chapter 1

Introduction
When perceiving the world, our perceptual faculties simultaneously represent both detailed information and general statistical information about the incoming stimuli. We can hear the detailed rhythm of a popular drum solo playing on the car radio, while simultaneously being aware of the background sounds associated with driving on a highway. All at once, we can see the exact outline of a sailboat, with its triangular sail and rounded hull, while noticing the general size of the waves in the bay it is sailing through. Many models have been developed to understand how visual objects (DiCarlo, Zoccolan, & Rust, 2012) are parsed and identified by the brain. On the other hand, models that consider peripheral vision as texture representation (Balas, Nakano, & Rosenholtz, 2009; Freeman & Simoncelli, 2011; Keshvari & Rosenholtz, 2016; Rosenholtz, Huang, & Ehinger, 2012) have also gained ground recently. This divided progress is reflected in computer vision. The state of the art in computer vision algorithms concentrate on determining the identity of objects in an image (Krizhevsky, Sutskever, & Hinton, 2012) or segmenting the objects from an image (Arbeláez, Maire, Fowlkes, & Malik, 2011).

Progress in connecting the levels of representational abstraction has been slow. This is partly because statistical models are only recently becoming computationally tractable and precise, with new techniques for synthesis gaining traction. Foundational work has already been done in trying to understand the limitations in visual object perception as a consequence of statistical representations in the periphery (Balas et al., 2009; Rosenholtz, Huang, & Ehinger, 2012; Rosenholtz, Huang, Raj, Balas, & Ilie, 2012; Rosenholtz, 2011a). We continue this line of research, specifying the limitations of statistical models while exploring the possibilities of applying it to vision in general.

In particular, we approach the subject through three studies. First, in the spirit of (Balas et al., 2009), we apply a statistical model of vision to crowding, a behavioral phenomenon that is traditionally thought of as a result of difficulty with maintaining accurate object-like representations in the periphery. By testing whether a single statistical model can explain the main effects in three crowding experiments with drastically different stimuli, we will help reveal the representation that is relevant for peripheral vision.

Second, we test whether foveal material perception can be explained using a statistical model, and compare the results to peripheral material perception. Humans have the remarkable capability to rapidly detect and identify materials in a visual scene. This ability is crucial to navigating the visual world. For example, the ability to identify frozen regions of the ground is vital to winter trekking, and can lead to painful results if unused. Our understanding of material perception has developed greatly over the last few decades and informed the design of better material recognition algorithms (Fleming, 2014b; Sharan, Liu, Rosenholtz, & Adelson, 2013). While there is a detailed history of using texture to model material perception, most attempts have been unsuccessful (Tomita, Shirai, & Tsuji, 1982), suggesting that texture cannot be a complete model of material perception. We investigate whether a modern statistical model might work better. We also examine material perception in the periphery, as the visual periphery has recently been modeled successfully with statistical texture models.

Third, we investigate peripheral vision in a more applied context, namely visual design. Web pages are particularly interesting examples of design because they can combine elements of natural

---

3 As a technical note, anywhere when we mention “texture” or “mongrel”, we refer (unless otherwise noted) to single pooling region Texture Tiling Model (Balas et al., 2009) syntheses, which is equivalent to Portilla-Simoncelli texture syntheses (Portilla & Simoncelli, 2000).
images (photographs), text, and graphical or style elements (buttons, menus, etc.). The appeal and usability of a webpage depend on how accessible it is visually, which is driven by how well a user can extract information from it at a glance, specifically a single eye fixation. Given a single eye fixation, most of what a user sees on a webpage will fall in the periphery. For this reason, we determined how well observers can assess aspects of a webpage, such as its semantic category, the location of its menu, and the presence of an ad, in the periphery.

These three projects bring us closer to understanding what limitations of peripheral vision are, whether object and material perception can be understood as consequences of forced statistical representations, and how the limitations affect perception of designs.
Pooling of continuous features provides a unifying account of crowding

Adapted from Keshvari & Rosenholtz, 2016, Journal of Vision
2.1 Abstract

Visual crowding refers to phenomena in which the perception of a peripheral target is strongly affected by nearby flankers. Observers often report seeing the stimuli as "jumbled up", or otherwise confuse the target with the flankers. Theories of visual crowding contend over which aspect of the stimulus gets "confused" in peripheral vision. Attempts to test these theories have led to seemingly conflicting results, with some experiments suggesting that the mechanism underlying crowding operates on unbound features like color, or orientation (Parkes, Lund, Angelucci, Solomon, & Morgan, 2001), while others suggest it "jumbles up" more complex features, or even objects like letters (Korte, 1923).

Many of these theories operate on discrete "things" - features of the display items, such as the orientation of each line, or the identity of each item. By contrast, here we examine the predictions of the Texture Tiling Model (TTM), which operates on continuous feature measurements (Balas et al., 2009), or "stuff". We show that the main effects of three studies from the crowding literature are consistent with the predictions of TTM. Importantly, TTM is fundamentally different from the existing models; it measures complex statistics "stuff" on image pixels, and does not explicitly encode any segmented information about the "things" in the image. This suggests that many of the stimulus-specific curiosities surrounding crowding are the inherent result of the informativeness of a rich set of image statistics, rather than an amalgam of thing-based phenomena, for the particular tasks.

2.2 Introduction

Colloquially, peripheral vision refers to vision outside of the center of gaze, i.e. outside the central 1-2 degrees of visual angle known as the fovea\(^2\). Peripheral vision is where you are not looking, i.e. not pointing your eyes. Despite it being markedly worse than central vision, peripheral vision is a key player in many natural tasks, including gaze guidance, object recognition, scene perception, and navigation (Strasburger, Rentschler, & Jüttner, 2011; Whitney & Levi, 2011).

Peripheral vision has limited acuity and color perception compared with central vision (Anstis, 1998; Strasburger et al., 2011). Whereas these deficits are well studied and significant, they only explain a small fraction of the difficulty viewers have when performing peripheral tasks. For example, the modest decrease in acuity with eccentricity – distance to the point of fixation – would imply that one could read an entire page of 12 pt text at reading distance without moving one's eyes (Anstis, 1998). The crucial limit to peripheral processing is not lack of acuity, but rather visual crowding.

Crowding refers to empirical phenomena in which observers have difficulty performing a peripheral task, such as identifying or discriminating a target, when that target is surrounded by nearby flankers or is otherwise too complex or cluttered. Traditionally, crowding has often been demonstrated and studied with recognition of a target letter flanked by other letters, but crowding is fundamental to peripheral vision, and not specific to letter stimuli (Pelli & Tillman, 2008). A crowded stimulus often subjectively looks "mixed up", such that the details and exact locations of the stimulus' visual features become difficult to discern. (Lettvin, 1976) described the subjective experience as one in which the stimulus "only seems to have a 'statistical' existence... The loss of spatial order does not involve angles or any other property that we commonly discuss..." This line of thinking has developed into conceptualization of crowding as "forced texture perception."

What, however, does "forced texture perception" mean? Researchers have answered this question in different ways, analogous to different dominant approaches to modeling texture perception (Rosenholtz, 2014). Whether attempting to predict texture segmentation, discrimination, or representation, most models of texture perception fall into two broad classes: models that describe

\(^{2}\text{Also referred to as the foveola, or rod-free fovea}\)
texture in terms of the features of discrete, pre-segmented texture elements ("things"), and models that
describe texture by a rich set of image statistics ("stuff"). As an example of the "things" class of models,
some theories of texture segmentation represent texture by a set of discrete features, called textons.
Whether segmentation occurs depends on whether two abutting textures contain the same number
of attributes, like vertical lines, endstops, closed curves, and arrow junctions (Julesz & Bergen, 1983;
Julesz, 1981; Pomerantz & Cragin, 2014). On the other hand, filter-nonlinearity-filter models of texture
segmentation (Landy & Bergen, 1991; Malik & Perona, 1990; Rosenholtz, 2000), as well as texture
representations based on image statistics (Heeger & Bergen, 1995; Portilla & Simoncelli, 2000) fall into
the class of stuff models of texture perception.

If we characterize crowding models in this fashion, several prominent classes of models are
thing models, operating on discrete features. In averaging models, the visual system is presumed to
measure, say, the orientation of each individual item, and encode only a few summary statistics such as
the mean and variance (Greenwood, Bex, & Dakin, 2012; Parkes et al., 2001). On the other hand,
substitution models hypothesize that peripheral vision detects discrete things – for example
texton-like features or letter identities – but loses or jumbles their location information (Pelli,
Cavanagh, Desimone, Tjan, & Treisman, 2007; Strasburger et al., 2011; van den Berg, Johnson,
Martinez Antón, Schepers, & Cornelissen, 2012; Whitney & Levi, 2011). A key question for both of
these classes of discrete-feature models is what are the “textons”, the “atoms” of crowding? In other
words, what things are averaged? What things are detected before the loss of position information?
Are the atoms unbound features, such as the orientation of each display item? Or bound features such as
conjunctions of orientation and color? Or object identity, like letters?

The other approach to modeling crowding (and texture perception) essentially says that
the primitives of crowding are not “atoms” at all, but rather stuff (Adelson, 2001). This other class of
models operates on continuous features, such as the outputs of a cascade of filtering operations and
nonlinearities. For example, a number of researchers have attempted to reason about the implications
of a simple pooling model which averages continuous feature measurements over a pooling region.
Saarela, Sayim, Westheimer, & Herzog, 2009). However, this simple pooling model, conceived of as pooling
within at most a handful of feature bands, has been disproven by a number of experiments (Kooi,
Malania, Herzog, & Westheimer, 2007; Manassi et al., 2012, 2013; Nandy & Tjan, 2012; Sayim,
Westheimer, & Herzog, 2010; van den Berg, Roerdink, & Cornelissen, 2007), and serves more as a straw
man than as a real contender to model crowding. In contrast, our Texture Tiling Model (TTM) represents
its inputs with a high-dimensional set of local image statistics (stuff), known to be good for capturing
texture appearance (Balas, 2006; Portilla & Simoncelli, 2000). This model measures correlations of
the magnitude of responses of oriented V1-like wavelets across differences in orientation, neighboring
positions, and scale, and phase correlation across scale, as well as the marginal distribution of luminance
and luminance autocorrelation (Rosenholtz, Huang, & Ehinger, 2012; Rosenholtz, Huang, Raj, et al.,
2012). A similar model, measuring the same set of statistics, has been put forward to describe early
visual cortex (Freeman & Simoncelli, 2011; Freeman, Ziemba, Heeger, Simoncelli, & Movshon, 2013).
We have previously shown that this model can predict the results of a number of crowding experiments
(Balas et al., 2009), among other phenomena (Rosenholtz, Huang, Raj, et al., 2012).

In this study, we examine the results from three different studies, each of which tested
hypotheses about the supposed atoms of crowding – the discrete features upon which the mechanisms
of crowding are presumed to act. One study, on letter crowding (Figure 2.1A), indicates that the atoms
are not individual letters (Freeman, Chakravarthi, & Pelli, 2012). A study with “T”-like stimuli (Figure
2.1B) suggests that the position and orientation of the cross bar are in some sense bound (Greenwood et al., 2012). A study with Gabor stimuli (Figure 2.1C) indicates imperfect binding between color, orientation, and spatial frequency (Pöder & Wagemans, 2007). These experiments and their analyses are quite complex; given the differences between the stimuli, it is not obvious which discrete-feature model of crowding could explain the full set of results. On the other hand, a strength of image processing-based, continuous-feature models is their applicability to arbitrary stimuli. We use TTM as our candidate continuous feature model, since it has shown promise on a number of crowding results (Balas et al., 2009; Rosenholtz, Huang, & Ehinger, 2012). We test whether TTM can account for the key results in all three studies. If so, this provides evidence in favor of continuous-feature pooling models of crowding, suggesting that perhaps the primitives of crowding are not atoms after all.

Testing this hypothesis is fairly complex. First, we must make TTM predictions for the previous experiments, essentially running our model through each experiment as a “subject”. We describe the methodology for doing this, which involves using human observers, below. We then compare these TTM predictions to the original behavioral results. This alone, however, is insufficient. In each of the three papers described above, the authors fit computational models of various degrees of complexity in order to draw conclusions about the likely atoms of crowding; the authors’ conclusions did not follow merely from eyeballing the data, nor from a few simple tests of significance. Here we ask, if TTM were a “subject” in the experiment, would the authors have come to the same conclusions by fitting its “data” (i.e. its predictions) as they did fitting their own experimental data? Taking this approach allows us to have a better sense of whether our model predicts key features of the data. Note that we are agnostic about both the details of the previous computational models, and about the conclusions drawn from those results. We simply ask whether a crowding mechanism based on pooling of a rich set of continuous features could have led to the previously observed data and the resulting conclusions.

The purpose of this study necessitates that we present a lot of information about the original experiments, the original models, and our attempts to test a unified account of these results. We have done our best to organize the study to make it as clear as possible. First, we describe our standard methodology for generating predictions from TTM. Then, each original study appears in its own section, with four parts, to (1) present the motivation of the original study, (2) describe the experimental methodology, (3) compare our predictions to the original data, and (4) present the outcome of applying
the computational modeling in the original study to the “data” of our model “subject.” Each section has a figure with sample stimuli and relevant results. Finally, we discuss the broader consequences of our findings.

2.3 Generating predictions from the Texture Tiling Model (TTM)

To gather predictions of the Texture Tiling Model, we use the same methodology as in our previous work (Balas et al., 2009; Rosenholtz, Huang, & Ehinger, 2012; Rosenholtz, Huang, Raj, et al., 2012; Rosenholtz, 2011a; Zhang, Huang, Yigit-Elliott, & Rosenholtz, 2015). In short: the model represents a stimulus image as vectors of image statistics computed over localized, overlapping sections (“pooling regions”) of the image. These measured statistics are then fed into an algorithm that generates a new image that matches the original image’s statistics, but is otherwise random. This methodology allows us to literally see the information encoded by the given set of statistics. The set of images that can be synthesized from a given image (with the same statistics but generated from different random seeds), form the equivalence class of the model. We call these synthesized stimuli “mongrels” (Balas et al., 2009).

The full version of TTM measures statistics over a number of pooling regions that overlap and tile the entire visual field. It takes hours to synthesize each image. Here we run the local version of the model, which extracts statistics from only a single pooling region. This reduces to Portilla & Simoncelli (P-S) texture analysis/synthesis ((Portilla & Simoncelli, 2000), available at http://www.cns.nyu.edu/~lcv/texture/), applied to stimuli we have blurred to account for reduced peripheral acuity. For stimuli such as those here, which lack larger scale grouping structure and are fairly homogeneous, syntheses of a single pooling region typically appear quite similar to those from the full version of the model.

For color stimuli, we use the color texture synthesis algorithm available at (http://www.cns.nyu.edu/pub/eero/colorTextureSynth.zip). This algorithm runs principle components analysis on the original color image, and measures statistics on each of the three PCA channels. It synthesizes a new image by iteratively applying P-S texture synthesis to each of the three color bands, enforcing, on each iteration, the correlations between bands corresponding to the eigenvectors of the PCA. For grayscale images, this reduces to ordinary P-S.

We generate mongrels from image statistics measured on the stimuli in the three original studies. We generate many of these mongrels per condition, and ask observers to perform essentially the same task as in the original study. The observers can view the mongrels for an unlimited time and without restriction of eye movements. Before each experiment, subjects were shown examples of the original stimuli along with their mongrels, and told colloquially how the process of creating the mongrel can mix up locations, features, and legibility of the originals. This procedure attempts to minimize higher level factors in order to test the effect of low-level encoding on complex symbol recognition tasks. See (Geisler & Chou, 1995) and (Zhang et al., 2015), for more discussion. Observer performance on the mongrel task gives us the predictions of TTM for a single pooling region.

The P-S texture analysis/synthesis method assumes that the image wraps around, meaning the top is adjacent to the bottom, and the left adjacent to the right (i.e. periodic boundary conditions). Essentially, it assumes that the image is on a torus. This leads to mongrels in which the “stuff” from the center of the original patch (e.g. the stimulus letters) can be centered on or near one of the edges, rather than the center of the image. This makes doing tasks like “report what letter appeared in the center of the original patch” somewhat tricky for naïve subjects, even when told about the wraparound. As the “center” of the mongrel is completely arbitrary – it is simply a matter of what starting point one
uses when unwrapping the torus – we correct for the wraparound by toroidally shifting each mongrel to move the center of mass of the foreground pixels to the center of the image.

2.4 Experiments

2.4.1 Can a simple substitution model account for crowding?

Substitution is undeniably a phenomenon in crowding, in which a subject, asked to report the identity of a target, instead reports one of its flankers. This phenomenon might suggest a mechanism of crowding, which measures features for target and flankers, but either does not encode their location or encodes it in a noisy way (Chung & Legge, 2009; Strasburger & Malania, 2008; van den Berg et al., 2012). This loss of location information would predict substitution of flankers for the target, though it is not the only mechanism that will do so (see (Freeman et al., 2012) and Discussion).

Is crowding due to substitution operating at the level of entire objects like letters – i.e. does the mechanism of crowding identify complete letters, but lose their location? Or does crowding operate on a different level of representation? (Freeman et al., 2012) examined the explanatory power of a simple substitution model at predicting responses in a crowded letter-identification experiment. In simple substitution, the response to a crowded stimulus is determined by independently sampling either the target or a flanker. This means that on a given trial, the response is either: the target, a letter similar to the target, a flanker, or a letter similar to that flanker. Importantly, because simple substitution is only a function of a single letter, any interaction between target and flankers, such as their similarity, does not directly influence the responses.

2.4.1.1 (Freeman et al., 2012)’s experiments

(Freeman et al., 2012) presented subjects with triplets of letters which were viewed peripherally (Figures 2.2A and Figure 2.2B). The subject’s task was to report the identity of the central letter, the “target”, disregarding the flankers on either side. (Freeman et al., 2012) based their conclusions largely on the results in their experiments 1 and 2, and so we address those two experiments here. In experiment 1, letter triplets were chosen such that one flanker was similar to the target and the other was dissimilar. Care was taken in choosing letter triplets to avoid introducing target-flanker correlations. This is necessary partially because observers more easily identify letters at the ends of a crowded array (Bouma, 1970). If target-flanker correlations were allowed, a subject who identified only the flankers could exploit the correlations to guess the target. A stimulus set was created for each subject such that each letter was used as a target, a similar flanker, and a dissimilar flanker. In experiment 2, an alphabet of two groups of three letters each (LTI, MNW) was used for all subjects. Within each group, the letters are similar, but between groups, they are dissimilar. The flankers were identical to each other in the second experiment, and both the target and flanker identities were chosen randomly on each trial.

(Freeman et al., 2012) proposed that a key prediction of simple substitution is that subjects will report a similar flanker just as often as a dissimilar flanker. They find that this null hypothesis does not hold, and conclude that simple substitution cannot account for letter-identification performance under crowding. A recent paper (Hanus & Vul, 2013), however, argues that simple substitution with letter confusions can account for increased similar flanker reports based their own modeling of the impact of individual letter confusability. We remain largely agnostic to the specifics of the models discussed in those two studies, and focus on the phenomena of similar vs. dissimilar flanker reports (see also (Bernard & Chung, 2011)).

2.4.1.2 Stimuli and procedure for generating TTM predictions for (Freeman et al., 2012)’s stimuli

We tested 5 subjects in our reproduction of experiment 1 from (Freeman et al., 2012) Four were experienced observers in other mongrel experiments, and all were naïve as to the design of the
experiments. (Three were lab members, and thus may not have been completely naïve as to the purpose of the experiments.) We tested 4 experienced psychophysical observers in experiment 2. Two of these observers were naïve to the purposes of the experiment, one of who also participated in experiment 1. Of the two remaining observers, one was an author.

Using P-S synthesis, we generated mongrels of the letter triplet stimuli in (Freeman et al., 2012) (Figures 2.2A and Figure 2.2B). Prior to their first main experiment, (Freeman et al., 2012) measured the confusion matrix for each observer between uncrowded peripheral letters with added noise. The confusion matrix was used to design a set of stimuli in experiment 1 for each observer, such that each letter in the alphabet was associated with a similar and dissimilar letter, resulting in a unique letter triplet of a target, a similar flanker, and a dissimilar flanker. Rather than carry out this procedure for each subject, we used the stimuli generated for their first observer, J.F., for all observers in the mongrel task. Given the success of observer-independent letter-confusion modeling (Watson & Fitzhugh, 1989), it is reasonable to assume that letter confusion matrices do not differ greatly in structure between observers, an assumption verified by (Hanus & Vul, 2013). We will show that we (similarly to (Hanus & Vul, 2013)) can nonetheless distinguish between simple substitution and pooling models. For each of the original 52 stimuli (each letter of the alphabet appears with equal frequency as the target, and the similar flanker can appear first or last in the triplet), we generated 10 unique mongrels, i.e. 10 members from the equivalence class of the original stimuli, according to TTM. This gave us a total of 520 trials per observer. Letter size and spacing were matched to the original experiment as closely as possible. Trial order was randomized for each subject.

To generate predictions for their second experiment, we generated letter triplets out of two groups of three letters, as in (Freeman et al., 2012). Enumerating all possible combinations produces 36 unique triplets. We generated 10 mongrels for each triplet. Stimulus order was randomized for each subject.

In both experiments, instead of instructing the subjects to name a single target letter in the center of the mongrel image, we required subjects to enter three letters, and used the second letter as the response. We did this to facilitate performing possible additional analyses of the responses — e.g. how errors in flanker reports relate to target errors; as such analyses are not relevant to the present study we do not present them here. For experiment 1, subjects were informed that in the original patch all letters occurred with equal frequency in all three locations, and that each mongrel represented a triplet of three unique letters. For experiment 2, they were told the 6 possible letters (examples of which were shown on the screen at all times) and told that the flankers would always be identical. Any trials in which subjects did not report identical flankers were discarded (1.39% of trials). Response time was unlimited and subjects received no feedback.

2.4.1.3 Results of (Freeman et al., 2012) replication

2.4.1.3.1 Data analysis

In experiment 1, trials in which subjects mistakenly reported too many letters, too few letters, or repeated letters were discarded (0.62% of all trials). Subjects reported the correct letter on 24.9 ± 1.1% (M ± SE) of trials, which is well above chance (1/26, or 3.85%), p < 1E-10. This is noteworthy because unlike simple substitution, the TTM model does not explicitly encode information about letter identity; pooled image statistics are sufficient to support identification of letters. Subjects reported the similar flanker on 13.8 ± 0.4% of trials and the dissimilar flanker on 8.4 ± 1.6% of trials (Figure 2.2C). The two types of flankers are not reported equally (p < 1E-5, two-sided permutation test), consistent with what (Freeman et al., 2012) observed in their first experiment. This result was also significant for four of the five subjects individually (p < 0.035, p = 0.12 for the fifth, two sided permutation test).
In experiment 2, subjects reported flankers more often when they were similar (26.8 ± 3.8%) to the target than when they were dissimilar (14.4 ± 2.1%) (Figure 2.2D). This trend is highly significant (p < 1E-5, two-sided permutation test) This difference in flanker reports is qualitatively similar to the results from (Freeman et al., 2012), who found that flankers were reported 60 ± 6% of the time on similar-flanker trials and 24 ± 1% on dissimilar-flanker trials. Subjects in the mongrel experiment had higher overall reports of absent letters than in the (Freeman et al., 2012) study. In short, we find that in both experiments 1 and 2, similar flankers are reported more often than to be expected by a pure substitution model, and thus our data support the notion that simple substitution is not a complete account of crowding.

**Figure 2.2** (A) Example stimulus from experiment 1 in (Freeman et al., 2012), with two mongrels to its right used in our Experiment 1A. Mongrels are generated by iteratively enforcing image statistics on a random noise seed. Subjects were asked to report what three letters they believed the mongrel was generated from, free-viewing with no time limit. Notice that the mongrels show effects of substitution (swapping of letter positions), as well as complex interactions between letters. In agreement with (Freeman et al., 2012), the flanker (“A”) more similar to the target (“X”) tends to be reproduced more faithfully in the mongrels than the dissimilar (“S”) flanker. (B) Original and mongrels for (Freeman et al., 2012)’s experiment 2 (mongrels in our Experiment 1B). Interestingly, the letter “I” seems to appear in the mongrels, despite being absent in the original. (C) Subject reports of the similar flanker (first bar) compared to the dissimilar flanker (second bar), averaged across subjects, for both (left) (Freeman et al., 2012) and (right) our Experiment 1A. Notice that in both studies, subjects reported the similar flanker significantly more often than the dissimilar (asterisk indicates p < 1E-5, two-sided permutation test). (D) How often subjects mistakenly reported one of the identical flankers when they were similar (first bar) to or dissimilar (second bar) to the target, for both (left) (Freeman et al., 2012)’s experiment 2 and (right) our Experiment 1B.

2.4.1.3.2 Applying (Freeman et al., 2012)’s models to TTM predictions

(Freeman et al., 2012) interpreted their first experiment as showing that simple flanker-substitution (where responses are based on, but not necessarily identical to, a flanker chosen at random) cannot account for letter-similarity effects in crowding, which is consistent with our results. Moreover, using their mixture modeling technique, we determined that flanker substitution could at
most account for 72% of responses in our mongrel experiment, compared to the 55% found by (Freeman et al. 2012).

Similarly, using the mixture model from their experiment 2, we found simple substitution to account for at most 84% of trials. These results, taken together, suggest that substitution cannot be a full account of crowding. While our percentages attributable to substitution are higher than that found in the original study, it is important to keep in mind that the mixture model maximizes the role of substitution in fitting the data; only trials in which it is impossible for substitution to explain are attributed to pooling. We still reach the same conclusion about simple substitution; in other words, if our model was a subject in their experiment, they would have drawn qualitatively the same conclusions (if quantitatively different) as they did when looking at actual behavioral data.

2.4.2 Do different visual features crowd in concert or independently?

If the atoms of crowding are not whole letters, as discussed in the previous subsection, what aspects of the stimuli are encoded prior to crowding, i.e. loss of location information or computation of summary statistics? (Greenwood et al., 2012) designed a set of experiments to determine which visual processing stage is responsible for crowding. One natural question to ask is whether different features of a stimulus, say position and orientation, are treated independently or “bound” together when crowding occurs. In a substitution-style model, in which some amount of recognition occurs (bars vs. corners vs. shapes and so on) prior to loss of position information, more binding implies that crowding operates at a later stage. Similarly, some models suggest that features of discrete items are identified and then averaged, producing crowding. The complexity of the feature measured and then averaged may suggest something about the stage at which crowding occurs. Thus, for both substitution and averaging classes of models, how much binding occurs gets at the important question of where in visual processing the mechanisms of crowding operate. One should note, however, that for models measuring image statistics rather than features of discrete items, the connection between amount of binding and stage of visual processing is not as straightforward. As we will see in our mongrels, a rich set of image statistics can capture a fair amount of information about feature binding even when that information is not explicitly encoded, much as the statistics capture some information about letter identity without explicit encoding of said identity.

(Greenwood et al., 2012) fit two models to their data. In both, crowding (i.e. averaging) mechanisms either operate on a given trial or they do not, in a way that depends probabilistically on the stimulus. In one model, this “gating” occurred for position and orientation features independently. In the other, whether or not orientation crowded also gated whether position crowded. They found that the latter model in which one gate controlled crowding for both features better fit the data. This result led them to the conclusion that binding of features occurs prior to the mechanisms of crowding. Here we compare the fits of their two models to our mongrel data, and ask whether TTM would have led to the same conclusion.

2.4.2.1 (Greenwood et al., 2012)’s experiment

(Greenwood et al., 2012) presented triplets of cross-like symbols in subjects’ periphery in their experiment 4 (Figure 2.3A and Figure 2.3B). On each trial, one cross-like target was flanked by two other cross-like symbols, and subjects were to report the relative position and orientation of the target crossbar (above or below the midline? tilted clockwise or counterclockwise?). In some conditions, the orientation of the target and flanker were close to zero, thus inducing “strong crowding” (± 10°) (Figure 2.3A). In other conditions, it was very different from zero, thus inducing only “weak crowding” (± 40°) (Figure 2.3B). The strength of crowding as a function of orientation difference between target and flanker was determined separately in their experiment 1. By changing the level of orientation crowding
independently of position crowding, Greenwood et al. (2012) argue that they can test whether crowding in one dimension induces it in the other, or whether they "crowd" independently. (In what follows, strong and weak crowding refer to experimental conditions, not performance.)

![Figure 2.3](image)

Figure 2.3: (A) (left) An example stimulus from Greenwood et al. (2012) displaying "strong crowding" and a difference in both position and orientation between target and flankers (both-differ condition). Orientations of the target and flankers (target at 5°, flankers at -10°) are relatively close to zero. The correct response to the orientation and position (relative to the vertical bar’s midpoint) of the center crossbar in this case would be “clockwise and below”. (right) Two of the corresponding mongrels. Notice that the orientation of the center object becomes ambiguous in the left mongrel, and swaps with a flanker in the right mongrel. (B) (left) Example from the “weak crowding” condition for which both position and orientation differ. Correct response to these stimuli would also be “clockwise and below”. The orientation difference between target and flankers is large here (target at 35°, flankers at -40°). (C) – (F) The average data over four subjects (bars) along with averaged model fits (circles and triangles). Each subplot corresponds to a particular relationship between target and flankers, just as in (Greenwood et al., 2012). The different colors are different types of responses the subjects made. Within each plot, the bars are grouped according to the orientation of the target. Notice that subjects tend to make position errors when the target and flankers differ in position (D). They make orientation errors only when orientation is different (E) by a small amount (±10°, strong crowding) but not when orientation differs greatly (±40°, weak crowding). In the case where both position and orientation differ (F), subjects tend to make errors in both features under strong crowding. When crowding is weak, though, subjects get both features correct most often. The model fits in the last plot (F) show that the independent model (solid triangles) predicts more position errors than the data show. The joint model (white circles), however, correctly predicts a greater number of correct responses than position errors in this case.

2.4.2.2 Stimuli and procedure for generating TTM predictions for Greenwood et al. (2012)'s stimuli

We tested 4 subjects, all experienced psychophysical observers. All subjects had prior experience viewing mongrels. One was an author and one was naïve to both the design and purpose of the experiments. The other two were lab members, and naïve as to the design but perhaps not the broad purpose of the experiment. Because this task was quite unnatural and difficult, we used observers experienced with mongrels in order to focus on difficulty due only to information loss by the model. For each of the 32 original stimuli in experiment 4 from Greenwood et al., 2012, we generated 20 synthetic
mongrel images (Figure 2.3A and Figure 2.3B for example), for a total of 640 trials per subject. Subjects in the mongrel experiment were asked to do their best to infer from the appearance of the mongrel the orientation and position of the target crossbar in the original stimulus (from which the mongrel was derived). Subjects were shown examples of original stimuli, told how those original stimuli were generated, e.g. that the irrelevant “flankers” were always identical to each other. They were also told that when stimuli are synthesized, the objects can switch positions, flip, etc. Subjects had unlimited viewing time to respond as to the orientation and position of the target crossbar by selecting one of four possible responses. To match the conditions in (Greenwood et al., 2012), trials were split into blocks of 320 trials each. Each block had trials with either strong (difficult) or weak (easier) crowding. The order of blocks was counterbalanced between subjects, and subjects received no feedback.

2.4.2.3 Results of (Greenwood et al., 2012) replication

2.4.2.3.1 Data Analysis

As in (Greenwood et al., 2012), we separate the data into four categories which correspond to the relationship between the target and flankers: both-match, position-differs, orientation-differs, and both-differ (Figure 2.3C through F). In each of these categories, when crowding occurs (leading to weighted averaging a la (Greenwood et al., 2012)), the observer will be more likely to report the value of a flanker’s features. According to this model, in the both-match condition, averaging target and flanker responses would still lead to a correct response; e.g. the average of three positions above the midline is above the midline. We find that our subjects are most likely to be correct in both features, regardless of crowding level (Figure 2.3C). In the position-differs case, the predominant responses in our data are position errors (also regardless of crowding level), although not much more frequent than correct responses. This suggests that while position judgments were crowded, if “gating” occurred a la (Greenwood et al., 2012), it would operate on around or slightly less than 50% of trials (Figure 2.3C). In the orientation-differs case, the results are more complex. In the strong crowding sub-case, observers most commonly responded with an orientation error and to a lesser extent (but not significantly less) they responded correctly, suggesting 50% or more of trials were crowded in orientation. In the weak crowding case, however, orientation errors dropped to nearly zero, and the most common response was to be correct (Figure 2.3D). This is not surprising, as (Greenwood et al., 2012) designed the weak crowding condition with large tilts to preclude orientation errors.

Importantly, according to (Greenwood et al., 2012), the key set of trials consists of those in which both the position and orientation differ between target and flankers (both-differ, Figure 2.3F). In the case of strong crowding, their subjects most commonly made mistakes in both position and orientation. When crowding is “released” in the weak crowding condition, however, the dominant response in their experiment was to be correct in both position and orientation. This effect is also prominent in our results.

To be specific: when crowding is “released” by increasing the difference in orientation between target and flankers, subjects have fewer orientation and position errors. This can be shown by computing the proportion of correct responses expected if releasing crowding only affected orientation judgments, and comparing it to the actual proportion of correct responses. If releasing crowding only affected orientation judgments, the proportion of correct responses in the weak crowding case should be equal to the number of correct responses with strong crowding, plus the number of orientation errors with strong crowding, minus the number of orientation errors with weak crowding. If the actual number is higher, then releasing crowding in orientation also releases crowding in position, supporting (Greenwood et al., 2012)’s “joint” model of crowding. If the expected and actual proportions are the same, then an independent model of crowding, where position and orientation don’t interact, cannot be ruled out. We find that in three of four subjects individually, the proportion of correct responses is 22.5
3.8% higher than expected if weak orientation crowding only affected orientation judgments ($p < 0.025$, two-sided permutation test for each subject), and higher (6.25%) but not significant for the fourth subject ($p = 0.4581$, two-sided permutation test for one subject). This is the same trend observed by Greenwood et al. (2012).

2.4.2.3.2 Applying (Greenwood et al., 2012)’s models to TTM predictions

In order to interpret their results, Greenwood et al. (2012) fit two models to their data. The models were similar in that both assume (1) noisy encoding of position and orientation, (2) a mechanism that “gates” orientation crowding on or off with a probability that depends on how close the target and flanker orientations are, and (3) the effect of crowding is modeled as a mechanism that takes a weighted average of the target and flanker features. The models differ in that one model (the “independent model”) has a separate parameter for the probability gating on or off position crowding, while the other (the “joint model”) posits that crowding in position happens if and only if crowding in orientation happens on a given trial. For a schematic of these models, see Figure 6A in Greenwood et al. (2012).

We fit these two models to our mongrel data and compare the fits. Because we had difficulty converging to a good fit using the models and data from Greenwood et al. (2012), we made one minor change in both of the models. Specifically, we used the thresholds determined from their experiment 3 to set the standard deviation of position noise in their model (their equation A4). This was done to keep parameter estimates at intermediate values and help the error minimization converge more quickly. Importantly, this component of the modeling is identical for both models, and occurs before the “binding” step, so it should not affect the ability to distinguish the models.

We fit both models to each of our subjects individually and found that for each subject the independent model, i.e. the model in which crowding happens independently for orientation and position, fit worse than the joint model. The root mean squared error (RMSE) of the independent model was 1.71 ± 1.1 percentage points worse than the joint model. The joint model fits the mongrel data better for each subject individually and overall. Compare this to the fits in Greenwood et al. (2012), where they find RMS errors of 6.29 for the independent model, and 3.78 for the joint model, a difference of 2.51 ± .0012. The joint model also fits their data better. Given the difference between the fits of the independent and joint models, our data also supports the conclusion that orientation and position are bound prior to crowding.

2.4.3 How does crowding affect feature integration?

(Pöder & Wagemans, 2007) also study to what degree feature binding survives crowding. They ask whether fully bound items are mixed up by crowding (high level substitution), whether the “mixing” happens in each feature band separately, or something in between. By analyzing feature errors and fitting models to the subject responses, they conclude that the level at which crowding occurs is neither extreme, but includes partially bound features. For example, the authors found a correlation between the number of flankers with a particular feature value and the number of times that feature value was erroneously reported. This correlation, however, was not perfect, suggesting some binding of the target features. Additionally, they argue that a Feature Integration Theory (FIT)-style model in which features bind probabilistically within a spotlight of attention fits the data very well, despite having only two free parameters.

We used the method described by (Pöder & Wagemans, 2007) to generate stimuli, which we then used to generate mongrels. Subjects viewed the mongrels and did the same task as in the original
experiment. As in the previous sections, we use the same analysis and modeling on our data to test whether we would arrive at the same conclusions about feature binding in crowding.

2.4.3.1 (Pöder & Wagemans, 2007)'s experiment

Using Gabor patches as stimuli, (Pöder & Wagemans, 2007) tested the effect of flankers on an identification task in which the observer must report three features of the target. Each Gabor patch had three orthogonal properties: color (red or green), orientation (vertical or horizontal), and spatial frequency (high or low). The target was selected randomly from the 8 possibilities (2 x 2 x 2). The target was flanked by 2, 4, or 6 Gabors, randomly chosen on each trial. The flankers' features were randomly chosen such that no flanker was identical to the target. The flankers were placed around the target, with their centers regularly spaced on a notional circle of radius 0.8 deg (Figure 2.4A and Figure 2.4B). On each trial, the array of Gabors was briefly displayed peripherally and subjects indicated the target identity by selecting one of 8 possible responses.

2.4.3.2 Stimuli and procedure for generating TTM predictions for (Pöder & Wagemans, 2007)'s stimuli

Four subjects participated in our experiment. One was an author, two had significant experience with mongrels, and the fourth had a small amount of experience with mongrels. All subjects except the author were naive to the purposes of the experiment. All properties of the stimuli were chosen to match those of the original study. The subjects were instructed to respond as to what they thought were likely the features of the central target in the original stimulus. Subjects responded by choosing one Gabor patch of the 8 possibilities displayed at the bottom of the screen and received feedback on each trial.

2.4.3.3 Results of (Pöder & Wagemans, 2007) replication

2.4.3.3.1 Data Analysis

In their analysis, (Pöder & Wagemans, 2007) organize the data in several different ways to make sense of the phenomena. First, they note that for all subjects, performance decreases with increasing number of distractors. They also plot the predictions of a random selection model, where the subject randomly reports one of the Gabors in a given stimulus (essentially losing all position information). This random selection model predicts the same trend but severely underestimates actual performance. Our data (for individual subjects and on average) also show both this decline in performance with increasing number of distractors and better performance than predicted by random selection. This means that while TTM's representation loses information about the stimulus, it does preserve some spatial configuration information. One can directly observe this in the mongrels (Figure 2.4A).

Another important finding reported by (Pöder & Wagemans, 2007) is that subjects’ probability of responding with a particular distractor is proportional to the prevalence of that particular distractor in a given display. They attempt to discriminate between two mechanisms that might drive this performance: random selection (mentioned previously), or misbinding, where the subject reports a combination of features randomly chosen from those present in the stimulus. They show that the probability of reporting a flanker is proportional to how prevalent that flanker is in a particular stimulus. However, the proportionality is less than 1, indicating that at least some responses are not due to random selection. Likewise, they provide evidence for misbinding by showing that increasing the number of times a particular feature value appears among the flankers increases the likelihood of it being reported. For example, having more green flankers in the display leads to a higher probability of subjects reporting a green target. We find strong evidence that TTM produces both of these effects (Figure 2.4C and Figure 2.4D).
2.4.3.3.2 Applying (Pöder & Wagemans, 2007)’s models to TTM predictions

(Pöder & Wagemans, 2007) present several computational models of their experiment. The one they find most compelling is a FIT-style model. Briefly, this model asserts that responses on a given trial are a result of a doubly-stochastic process: First, a “center of attention” is chosen randomly from a 2D normal distribution centered at the stimulus center (the target). The variance of this normal distribution – essentially the error in centering attention on the target – is a free parameter of the model. Second, the probability of choosing features from a given display item falls off with distance from the center of attention, according to another 2D normal distribution whose variance is a second free parameter of the model. Pöder & Wagemans (2007) simulated this model and found the values of the free parameters that best fit the data. Specifically, they fit the distribution of responses as a function of the number of flankers, the number of feature-errors (how many features differed between the response and the target), and whether the response was present in the flankers (see Figure 7 in Pöder & Wagemans, 2007). Organizing the data in this way, we compare the results of our study and theirs in Figure 2.4E.
They find that this model fits well compared to other models tested (Pöder & Wagemans, 2007). They find the best fitting variance parameters for both normal distributions to be 0.32 degrees visual angle. We also fit their model to our data, and found it to fit well, with the optimal parameters being effectively 0.48 and 0.2 degrees visual angle in the original crowding experiment (Pöder & Wagemans, 2007). The model fits to their and our data are shown in Figure 2.4F.

2.5 Discussion

The Texture Tiling Model (TTM) of visual processing, in which a rich set of image statistics are measured over sparse pooling regions that tile the visual field, accounts for a variety of crowding phenomena (Balas et al., 2009; Rosenholtz, Huang, & Ehinger, 2012; Rosenholtz, Huang, Raj, et al., 2012; Zhang et al., 2015). Importantly, because the model operates on images rather than hand-labeled, experiment-specific features, the model is flexible enough to make predictions for arbitrary stimuli. We have shown here that three reasonably different crowding phenomena can be captured by this one model.

First, (Freeman et al., 2012) find that when viewing letter arrays in the periphery, performance cannot be described by independent position and identity confusion; a more complex mechanism (pooling) is required to account for crowding. We also find our model predicts that their crowding results cannot be completely explained by simple substitution.

Second, when we compare the predictions of TTM to those of (Greenwood et al., 2012)'s model, we find that both result in the so called “release of crowding” for position under conditions of weak crowding for orientation; orientation and position seem to be in some sense “bound” by peripheral vision. It is important to note that TTM predicts this release from crowding phenomenon, even though it has no explicit “gating” mechanism that turns crowding mechanisms on or off depending upon the stimulus. This suggests that given the information encoded and lost by peripheral vision, some stimuli inherently lead to more or less difficult recognition of the target.

Finally, in (Pöder & Wagemans, 2007), subjects attempted to identify peripheral objects in which target and flankers varied independently along three orthogonal feature dimensions. Responses seemed to be a mix of substitution of whole flankers (full binding), and substitution of partially-bound features (illusory conjunctions). They found that a parsimonious explanation of this phenomenon was a FIT-style model with a soft-edged and semi-randomly placed attentional window. The TTM also predicts a mixture of whole flanker and illusory conjunction reports; and importantly generates data which are well approximated by the same FIT-style model.

Predictions made by the TTM, overall, are thus consistent with a wide range of results from a diverse set of experiments. The qualitative success of the model is encouraging. This success likely derives in part from it belonging to a class of models that includes a cascade of oriented bandpass filtering, nonlinearities, and pooling over sizeable but sparse regions in order to compute a large number of summary statistics. The specific image statistics hypothesized by the model may not be exactly right, as also suggested by the present work. Though the qualitative results are good, quantitatively, our model predictions did not always agree with the data. For example, in the letter-identification task (originally from (Freeman et al., 2012)) subjects viewing mongrels in our study reported significantly more absent letters than in the original study (Figure 2.2C, the decrease in “similar flanker” responses in our data as compared to (Freeman et al., 2012)) is attributable to absent-letter responses). The number of reports of absent letters might decrease with the use of the full version of TTM, which gathers additional information from multiple pooling regions. However, in addition, the model statistics may need further refinement. We would have been surprised if the image statistics in (Portilla & Simoncelli,
developed to capture texture appearance, turned out to be precisely the statistics necessary to predict the phenomena of visual crowding.

One powerful feature of TTM is the ability to visualize the equivalence classes of the model, as represented by the mongrels. Mongrels can be used to gain intuitions about peripheral vision, without having to run an entire experiment. These intuitions facilitate the development of new experiments and theories. For example, we have observed that triplets of identical letters (e.g. “A A A”) were better preserved in mongrels than triplets of non-identical letters (e.g. “A B A”) (Figure 2.5A vs Figure 2.2A and Figure 2.2B, respectively). The hypothesized encoding better represents identical letters, likely leading to easier recognition in such cases. This is in addition to any decision-making effects, in which identical letters also have an advantage because reporting a flanker also gives the correct answer. An interesting effect that becomes apparent when viewing mongrels of letter triplets is that letter order also makes a difference; mongrels suggest that “G N W,” for example, may result in a qualitatively different percept than “W N G”. This makes sense, as the local correlations measured by the model will be different in the

Figure 2.5 (A) Three mongrels of letter triplet stimuli from (Freeman et al., 2012) in which the letters in the original stimulus were identical (III, TTT, and WWW). Notice that identical letter triplets tend to be better represented by the mongrels than when the letters are different (Figure 2.2A and Figure 2.2B), although they aren’t perfect (in the WWW case for example). The improvement in representation comes from the fact that identical letter triplets form a more coherent visual texture, irrespective of the letter identity itself, and textures in general are better represented by summary statistics. (B) Example stimulus that is similar to (Greenwood et al., 2012)’s stimuli, and two mongrels to its right. Notice that in the mongrels, orientation and position are bound; the tilted bar intersects the vertical bar in each object at the same place as the original, even though objects can become shuffled around. (C) Plot of all relevant data from the original three studies (horizontal axis) vs. TTM predictions, i.e. the data in our three experiments (vertical axis). The data points for (Freeman et al., 2012) derive from Figure 2.2C and Figure 2.2D. The data points from (Greenwood et al., 2012) are those in Figure 2.3C through Figure 2.3F. The data points from Pöder & Wagemans (2007) are copied from Figure 2.4E. A viable model of crowding must be able to generate a plot like this one, where predictions can be generated for arbitrary stimuli in the form of images.
two cases. Future experiments might determine whether or not letter order actually makes a difference in letter triplets viewed peripherally.

The mongrels we generated of the stimuli in (Pöder & Wagemans, 2007) resulted in illusory conjunctions mainly between color and other features. In other words, spatial frequency and orientation of Gabors were more “tightly bound” to each other than color was to either of the other features. A follow up experiment could more carefully examine the nature of the illusory conjunctions in the crowded periphery. (Rosenholtz, Huang, Raj, et al., 2012) have previously demonstrated that TTM predicts orientation-color illusory conjunctions underlying difficult conjunction search tasks.

An interesting observation from the mongrels is that the locations of individual objects are often jumbled up such that the target does not necessarily appear in the center of the array. In other words, the model predicts substitution-like phenomena without a substitution-like mechanism. At least for some simple stimuli, the image statistics measured by TTM are sufficient to encode the correct binding between features of the same object. This means that if the original stimulus was the letter triplet “A B A”, the mongrels would often contain two “A”s and a “B”. If the subject knows that the stimulus will contain two of the same letter and one different, they can bring to bear this knowledge and pick the target by choosing the letter that is different. Rather than a crowded letter identification task, as the experimenter may have intended, the subject may actually be doing an oddball identification task. By better understanding the representation stage, as visualized by the mongrels, we can better understand the role of decision-making in performing tasks under conditions of crowding. Experimenters must not only model the encoding losses due to crowding, but also take into consideration the subject’s prior knowledge of the task.

This study sets a new benchmark for crowding research, by attempting to predict a range of experimental results with a single model (Figure 2.5C). Looking at Figure 2.5C, a standard linear regression model results in a best-fit line with slope $0.608 \pm 0.057$, intercept $0.07 \pm 0.02$, and adjusted $R^2 = 0.665$. (Putting aside, for purposes of discussion, that this standard fitting procedure is not strictly correct for these data. Some of the data points represent percentages of trials with a given response, which must add up to 100%; this introduces unavoidable correlations between the values within the same study. Furthermore, the percentage data theoretically violate assumptions of homoscedasticity [uniform variance].) If we assume TTM is a complete model of peripheral encoding, without any fitted parameters, then the line fit to the data should have slope 1 and intercept 0. Enforcing this results in $R^2 = 0.39$, with no free parameters. One could add a single fitted parameter, e.g. a multiplicative constant on the mongrel predictions, in order to improve the fit, but the value of that parameter would not be easily interpretable in the context of understanding the model. There is clearly room for improving the model in future studies.

Nevertheless, TTM makes testable predictions without having to hand-tune it to the task-relevant stimulus features or augment it with ad hoc crowding mechanisms like grouping or gating. This is a direct result of the “visual texture” representation it uses, which is image-computable and operates on “stuff” rather than pre-segmented and processed “things” like items and bars. Given the parallels between models of texture processing and models of crowding, it is worth noting that stuff models of texture perception have been more successful than thing models (Rosenholtz, 2014). Models based on image statistics make testable predictions for a far broader range of input stimuli, and have performed well across a range of stimuli and tasks (Heeger & Bergen, 1995; Malik & Perona, 1990; Portilla & Simoncelli, 2000; Rosenholtz, Huang, Raj, et al., 2012; Zhang et al., 2015). In order to be viable, future models of crowding must, like TTM, be image-computable, and able to make predictions on the wide range of stimuli behavioral researchers can throw at them.
Chapter 3

Material perception, texture, and peripheral vision

Work done in collaboration with Maarten Wijntjes
3.1 Abstract
Humans can rapidly detect and identify materials such as a hardwood floor or a smooth leather jacket, even within a complex visual scene. Given a single image, one can easily identify the underlying "stuff", even though a given material can have highly variable appearance; fabric can be stretched or sewn into arbitrary shapes and printed with a wide variety of patterns, and plastic can be used to make imitation fruits. What cues does the visual system use to determine material identity? Prior research shows that simple "texture" features such as the power spectrum capture some information about material properties and identity. Few studies, however, have tested richer, biologically motivated models of texture. To determine whether texture is a sufficient cue for material identity, we compared the results of two experiments. To first get baseline performance, human observers did a 6-way identity classification task with windowed grayscale images of materials and unlimited viewing time. We compared performance in this experiment to performance in the same task done with synthetic textures generated from each original material image using the Portilla-Simoncelli algorithm (Portilla & Simoncelli, 2000). These synthetic textures retain statistical information about the original material images, but are otherwise random. We found that observers' performance in the texture task was well below baseline, suggesting that texture alone is not enough to support material perception. Interestingly, recent studies have hinted that peripheral vision might employ a more statistical, texture-like representation. Might texture properties better capture performance at material judgments when the latter judgments were made in the periphery? We find that performance classifying materials in the periphery is highly correlated with performance in the texture experiment, even when considering individual stimuli. This suggests that peripheral, rather than foveal, material classification can be modeled by a rich visual texture model.

3.2 Introduction
From a single image, humans can extract rich, structured information about the properties of a material. Looking at the first photograph in Figure 3.1, it is immediately obvious that the pertinent material is water. One also perceives the scene as dynamic; droplets and waves suggest how the water got to be there and where it might go next. It is perhaps also apparent that all of the action is taking place inside a steel sink.

Figure 3.1 Example material images from the MIT-Flickr Materials Database (Sharan et al. 2014). The images came from the water, stone, and foliage categories, respectively. Notice the large range in viewpoint, illumination, scale, and context.
Perceiving material qualities allows for rich scene descriptions, and for disambiguating between objects of the same category ("the wooden plate"). However, material perception also allows us to safely navigate the world, interact with objects, and infer physical processes. Being able to reliably and quickly tell whether a patch of road is wet or dry is vital for driving and biking, for instance. When grasping soft-serve ice cream, we naturally reach for the solid waffle cone, rather than the creamy filling. Mistaking the material properties of the ice cream could lead to a mess. Furthermore, looking at an image of squash, we can tell whether it is raw or cooked, solid or pureed, fresh or rotten.

A central goal of research on the perception of materials, or the "stuff" that makes up the world (Adelson, 2001), is to discover what image cues and computations allow humans to determine material properties. (Fleming, 2014a) argues that models of material perception have fallen into two broad categories. First are inverse optics models, which postulate that the visual system has a generative model of images it uses in reverse to infer the properties that caused the image (Pizlo, 2001). According to such theories, for example, our ability to visually estimate the thickness of fabric might come from an implicit knowledge of the physical mechanics of fabric and light reflection, the prior distribution of relevant parameters (the range of possible fabric thicknesses and light sources etc.), and how they interact to form the images that fall on the retina. Another class of models posits instead that the visual system relies on ad-hoc regularities in the retinal image, or "statistical appearance", to infer material qualities. For example, observers seem to use the skewness of an image’s luminance histogram as a cue to the glossiness of the material (Motoyoshi, Nishida, Sharan, & Adelson, 2007). While skewness is clearly only a single cue to glossiness, in many cases it is diagnostic of intrinsic material properties. Both approaches, inverse-optics and statistical appearance, are certainly good models of material perception, perhaps one being more appropriate than the other (or a combination of the two) for certain tasks and stimuli. Here, we study to what extent a particular statistical appearance model, an image-processing based model visual texture, can account for material perception.

Visual textures, or textures for short, are loosely defined as images or regions of images with homogenous appearance. A pile of leaves, a marble countertop, and a wavy ocean all give rise to texture-like images. Textures have been studied for understanding perceptual organization (Rosenholtz & Wagemans, 2014; Rosenholtz, 2000), peripheral object recognition (Balas et al., 2009), and most relevantly for this study, material perception (Balas, 2015; Sharan, Rosenholtz, & Adelson, 2014). There is a long history of modeling in texture, and models fall into two general classes: object, or “texton” based models, and image-processing based models. The former posits that the basic elements of texture are individual features, such as angles, endstops, intersections, etc., and that the more different two textures are in their textons, the more perceptually different they will be. The latter family of models has gained ground recently for applications in computer graphics (Gatys, Ecker, & Bethge, 2015; Heeger & Bergen, 1995; Portilla & Simoncelli, 2000), and promises to be a more general approach because it operates directly on the image itself, rather than using hand-labelled features. The present study considers a specific texture model, known as Portilla-Simoncelli (Portilla & Simoncelli, 2000), that represents a texture using a high-dimensional set of image statistics that are thought to be relevant for early human vision (Freeman & Simoncelli, 2011; Freeman et al., 2013) and peripheral vision (Balas et al., 2009; Keshvari & Rosenholtz, 2016; Raj & Rosenholtz, 2010; Rosenholtz, Huang, & Ehinger, 2012; Rosenholtz, Huang, Raj, et al., 2012). To our knowledge, there has not been a systematic study of the appropriateness of Portilla-Simoncelli texture statistics for material perception.
There is, however, extensive prior work on the usefulness of texture models for material classification by humans and computers (Gibson, 1986). It is worth explicitly noting here that material perception and texture perception are not identical, as many studies attest (Sharan et al., 2013; Tomita et al., 1982). Nevertheless, we introduce what is known about texture for material perception. (Sharan et al., 2014) found that the information encoded by a non-parametric texture model was not sufficient to capture material categorization performance. Specifically, they generated textures from material images using the patch-based synthesis method in Efros and Freeman (Efros & Freeman, 2001). Observers viewed these textures for one second, and classified them into one of 9 categories. The observers were significantly less able to classify these textures than the original material images. There were significant differences in percent correct between the 9 different material categories; baseline material categorization performance however was too close to ceiling to correlate to texture performance.

On the other hand, in a detailed study on machine classification of materials, (Sharan et al., 2013) found that small and large-scale texture-like descriptors were essential for good classification performance. Specifically, they compute jet and SIFT (Lowe, 1999) features (in addition to measures of color distribution and material shape) on material images and bilaterally filtered versions (for small-scale information). They trained a classifier to map the feature set from each image into a material category. They found that SIFT was the single best feature for classification, although other features were quite important overall. This suggests that such features are diagnostic of material category. In a pertinent extension, they measured the ability of their algorithm and humans to classify textures generated from material images. That is, they scrambled material images using Efros and Freeman (Efros & Freeman, 2001) texture synthesis, and presented those to their algorithm and human observers. They found that these texture images were not only harder to classify for both the algorithm and humans, but nearly equally difficult for both algorithm and human. This suggests potential similarities in encoding.

Humans are clearly able to extract a much richer representation of a material than simply its identity; classification, however, depends on many material properties and is less dependent on the subjective preferences or abilities of observers than material judgments like gloss estimation or aesthetic appeal. Classification is also arguably an important task for vision, and can inform the estimation of other properties. For example, knowing an object is made of glass reveals something about its material qualities like gloss, roughness, and color (Sharan et al., 2014). Moreover, it is straightforward to compare classification performance between different experimental conditions to get insights into perceptual abilities. This is reflected by a growing body of work on material classification by humans.

(Sharan et al., 2014; Sharan, 2009) found that humans can classify images of materials reliably and quickly, despite large variations in color, scale, and context. The images were taken from the MIT-Flickr database (Sharan et al., 2013). Critically, they found that this ability does not depend on a single particular cue, such as color, shape, or small-scale texture; it is rather a more basic and holistic ability. Using the same material image database (Fleming, Wiebel, & Gegenfurtner, 2013) collected a set of subjective ratings, as well as a set of subjective ratings for material classes. They found that the ratings for individual images correlated well with the ratings of their classes, and that k-means clustering could correctly classify materials with 90% accuracy using only the averaged subjective ratings. (Wiebel, Valsecchi, & Gegenfurtner, 2013) measured the time course of material perception compared to object recognition in a rapid, masked classification task. By changing contrast and luminance values, they were
able to normalize images of both materials and objects to have equal detection rates for a given presentation time. This allowed them to find out whether material classification was faster or slower than object classification for a given level of performance. The researchers found that with these normalized images, observers could classify materials with the same accuracy as objects, but only at longer presentation times. Furthermore, it turned out that observers relied on color information more for material than object classification. These studies suggest that by studying the properties of images of materials, we can better understand how humans perceive materials in general.

Our study’s purpose is twofold. First we test whether a state-of-the-art model of texture, Portilla-Simoncelli texture statistics, can fully capture what aspects of an image cue its material category. This can be understood as an important extension of the texture work done by (Sharan et al., 2014), and a rigorous assessment of the extent to which texture supports material perception. Second, we test the ability of observers to classify material images shown peripherally. Since Portilla-Simoncelli texture statistics are the backbone of a couple of powerful models of peripheral vision (Balas et al., 2009; Freeman & Simoncelli, 2011), it is necessary to compare texture performance to peripheral performance in material perception. Previous studies (Balas, 2015; Sharan et al., 2014) have made a foray in this direction, but our work introduces essential control conditions. Namely, we control for color, and use more material categories. We also use a gaze-contingent display to enforce fixation while allowing unlimited viewing times. These innovations enable a fair direction comparison of peripheral to texture perception of materials.

3.3 Baseline and texture experiment methods
3.3.1 Selecting material images for stimuli
We used images from the MIT-Flickr Materials Database (Sharan et al., 2013). The MIT-Flickr database has several advantages for our purposes over other material image databases. Prior work has shown that people are quite good at categorizing the images in the database (Sharan et al., 2014), despite the large variability in color, pose, scale, semantic content, and illumination within and between material categories. Furthermore, since the images are highly variable, whatever image cues the observers use to do the task are more likely generalize to real world perception. Prior material image datasets are usually too restricted in viewpoint, illumination, color, content, etc. to be useful for probing general material perception ability. The materials in CURF (Varma & Zisserman, 2009), for example, are laid out flat, illuminated identically, and photographed from a single viewpoint and distance. The goal of controlling the images so closely was to facilitate the estimation of the Bidirectional Reflectance Distribution Function (BRDF) for each material; this can only be done in restricted settings. Using such a restricted set might erroneously lead to the conclusion that texture is sufficient for material classification. For example, a texture model might not be able to capture the essential qualities of glass if different lighting angles are used for different exemplars.

The MIT-Flickr database contains a total of 900 images, with 9 material categories and 100 images in each category. We only used 6 of the categories (stone, water, wood, fabric, foliage, and leather; leaving out metal, plastic, and glass) to facilitate an easier 6- rather than 9-alternative forced choice task. For each category, the database contains 50 “close-up” and 50 “object-level” images. The object-level images usually contain more background (of a different material than the relevant one), which we thought would introduce extra variability uninteresting for our purposes. Therefore, we left those out of this study, resulting in 300 total material images. When generating synthetic textures, we
chose to convert all images to grayscale and windowed them, for reasons described in Procedure for synthesis of material textures. In order to maintain parity between the baseline and texture task, we also used similarly grayscaled and windowed material images for the baseline experiment. This helped ensure that differences in performance between baseline and texture conditions are driven by the texture representation. See Figure 3.2 for some example stimuli.

3.3.2 Procedure for synthesis of material textures

Textures are generated using the basic Portilla-Simoncelli synthesis algorithm (Portilla & Simoncelli, 2000). We first converted the images to grayscale by converting to CIELab space and keeping only the luminance channel. This was done for two reasons: first, observers are nearly as good in grayscale as with color (Sharan et al., 2014), so it is not essential for material perception. Second, there is not a color version of the texture model which is widely accepted and tested with respect to human vision. The grayscale Portilla-Simoncelli model has been tested extensively with a variety of stimuli and tasks (Balas et al., 2009; Keshvari & Rosenholtz, 2016; Rosenholtz, Huang, Raj, et al., 2012). Then, for each material image, we compute P-S statistics (with default parameters), and use the P-S synthesis algorithm along with a random noise seed to generate a synthetic version of the same size. We run the algorithm for many iterations (150) to ensure convergence. This procedure results in 300 synthetic textures, one for each original material image.
The algorithm assumes that the image wraps around top-to-bottom and side-to-side, i.e. it assumes the original and synthesized images lie on a torus. This means that the edges of the image are non-intuitively structured, and it is cleaner to leave them out by windowing (mentioned in Selecting material images for stimuli). Note that the windowing is done after synthesis (textures are computed over the whole, un-windowed material image; the synthesis produces a full-sized image which is then windowed). We used a circular window of 2 deg visual angle in radius, with a smooth Gaussian fall off of standard deviation of 0.5 deg. Figure 3.3 illustrates some materials and their texture counterparts.

3.3.3 Observers
16 observers participated in experiments. 5 observers did the baseline experiment, and 11 did the texture task (these latter observers also did the peripheral task in a separate block as discussed in the Methods for peripheral experiment). All observers were naive as to the purpose of the experiment, and all had normal or corrected-to-normal vision. Observers in the baseline experiment were paid $10 for about 30 minutes of experiment, the rest were paid $15 for about an hour of experiment.

3.3.4 Experimental apparatus
Stimuli were presented using Psychtoolbox 3 (Kleiner et al., 2007) and MATLAB on a CRT monitor with a mid-gray background. Observers used a chinrest in all conditions. Observers responded by using a mouse to click on one of 6 circles, each labeled with one of the categories.

Figure 3.3 Example materials along with their synthetic texture versions. The left image in each column is the original, and the right column is a sample texture.
3.3.5 Experimental procedure

For all experiments, observers had unlimited viewing and response time. Observers were first oriented to the task by the experimenter and shown example stimuli along with category labels. The experimenter also informed them that all categories occurred with equal frequency, and to do the best possible without spending too much time on a particular trial. Observers received feedback on the first 25 trials. Each observer saw all 300 unique stimuli, in random order.

![Image](image_url)

*Figure 3.4 Procedure for a single trial in the baseline experiment. If the trial contains feedback, it is displayed immediately after clicking. The procedure for the texture experiment is identical, except displaying a texture image instead of a material.*

For each trial, the procedure is as follows: the stimulus (material image) appears in the center of the screen after a 1 second central fixation (fixation circle remains on throughout each trial). The observer then has unlimited time to push spacebar to end the presentation. The stimulus is then removed and the decision screen appears. On the decision screen, the name of each of material category is shown in a circle of radius 2 deg, at 8 deg eccentricity (evenly spaced in a notional circle centered at the screen center). The observer then moves the mouse to the desired choice, and clicks to make a response, and the decision screen is removed. Each observer received 25 training trials at the beginning. Training trial had visual feedback as to correctness of the response (a change in the color of the central fixation for 0.5 seconds). During the main experiment, observers received no feedback, with the fixation remaining white for 0.5 seconds. The next trial then begins immediately. The observer receives an untimed break every 75 trials. A schematic of the procedure is illustrated in Figure 3.4.

3.3.6 Baseline task

For the baseline task, observers viewed all 300 windowed material images presented foveally, and did the classification task as described in Experimental procedure. It is important to get a baseline measure of performance for several reasons. If we are to examine texture as a cue for material category, we need to know how well observers can tell category with all cues present, i.e. the original materials. To our knowledge, there has not been a study of untimed, grayscale material recognition with the MIT-Flickr database using the subset of images we choose here. Importantly, it is not obvious that observers will be perfect at this task. The images come from a wide range of 3D-shapes, object identities, surface reflectances, physical scales, and illuminations, even within a category. Our later experiments compare
performance in this baseline condition to performance under degraded viewing conditions. If observers are less able to categorize materials with the textures than the baseline, this would imply that texture is not a sufficient cue for category. Put alternately, it would imply that the information lost by converting a baseline material to a Portilla-Simoncelli texture, e.g. shape or large-scale layout information, is necessary for robust material classification. If on the other hand texture classification performance is close to or identical with baseline performance, we cannot rule out texture as sufficient for material classification.

3.3.7 Texture task

For the texture task, observers foveally viewed windowed synthetic textures generated using the procedure described in Procedure for synthesis of material textures. Specifically, the observers did the same classification task as described in Experimental procedure, except viewing textures. The experimenter gave a colloquial explanation of how the textures are generated from the original materials, including that parts of the image might be translated, swapped, and mixed together with respect to the original material image. Importantly, they were instructed to respond to what material category the texture was generated from, rather than what material the texture itself might look like. This distinction encourages the observer to do as well as possible with the available information and take into account their intuitions about the texture-generation process. In this way, the observer is essentially acting as a classifier on the texture statistics (Balas et al., 2009).

![Confusion matrices](image)

**Figure 3.5** Confusion matrices for the baseline and texture experiments. The row indicates the true material, and the column indicates the response. Darker shades indicate a higher prevalence of the response. The color bar on the right provides a reference for exact values. Notice that most of the mass is along the diagonal in the baseline case, indicating high performance. In the texture case, the diagonal is weaker, and errors are widely distributed.

3.3.8 Baseline and texture experiment results

We analyzed the results in several ways. First, we discuss the results from each experiment separately, and then compare them. Unless otherwise noted, statistical significance values are computed using a 2-sided random-permutation test.
3.3.8.1 Baseline experiment results

5 observers completed the baseline experiment, each doing 300 trials. The average performance (proportion correct) over all 6 categories and all subjects was 0.882 (chance is 1/6 = 0.167). Each observer performed well above chance in each category (p < 2E-5). This is in line with previous work, where (Sharan et al., 2014) found that observers’ performance was 0.866 (chance is 1/9 = 0.11) for 9-way grayscale material classification. A confusion matrix of responses is shown in Figure 3.5A. There does not seem to be significant off-diagonal structure in the confusion matrix, although we do not present a formal measure.

3.3.8.2 Texture experiment results

10 observers completed the texture experiment, each doing 300 trials. The same subjects did the peripheral experiment in a different block. Average performance over subjects and categories was 0.40. Each subject performed above chance (p < 2E-5) averaging over all categories. Each subject was also above chance within each category (using a p < 0.05 criterion), with a few exceptions: two subjects were not above chance at classifying fabric, three observers were not above chance with foliage, and one observer was not above chance with leather. The confusion matrix is shown in Figure 3.5B.

3.3.8.3 Comparison of baseline and texture results

We directly compare the elements in the confusion matrix from one experiment to the other. We find that baseline performance is not well predicted by performance using the texture images. Namely, performance is too close to ceiling in the baseline case, precluding an interesting comparison to texture performance (Figure 3.6).

3.4 Interim discussion

We can draw several conclusions from the baseline and texture experiments. First, it is clear that humans are excellent at categorizing grayscale images of materials. Performance is near ceiling for each subject in each category, and the overall confusion matrix does not seem to contain any interesting off-diagonal structure. Second, performance in the texture task is much lower than ceiling. Although all subjects are performing above chance, for certain categories, there are subjects that do not perform above chance. There are also no obvious category confusions that are consistent between subjects, although some confusions do seem more likely than others. For example, fabric was often classified as leather, and foliage was rarely perceived as wood. Comparing baseline to texture performance, it is clear that texture statistics are not sufficient to convey material category, performance is simply too high in the baseline case for there to be significant correlations between misclassification rates between experiments. It is important to note however that it was not obvious a priori that this particular texture model would be insufficient.

Interestingly, however, there is a large range of performance in the texture case. Some materials are more accurate than others when viewed as textures. Water textures, for example, are on average more easily classified than foliage textures. Furthermore, there are significant performance differences for individual texture stimuli within a category; for some textures, all of the subjects correctly classified
them, while for others, no subjects were correct. If foveal material perception is too accurate to explain such variation in performance, might there be another aspect vision that does?

![Graph](image)

Figure 3.6 Comparison of elements in the texture confusion matrix (Fig. 5A) to elements in the baseline confusion matrix (Fig. 5B). The cluster of points on the right are the diagonal elements, the proportions of correct responses. Notice that while baseline is not well predicted by texture, texture performance is above chance (1/6).

3.5 The relevance of peripheral vision

One place to look is peripheral vision. Humans can identify materials peripherally, for example when noticing wet leaves on a road while driving and looking ahead. This plays into rapid scene categorization, where observers quickly classify a natural scene. Research on rapid scene categorization argues that this ability depends on low-level image cues (Greene & Oliva, 2009a, 2009b; Oliva & Torralba, 2006), rather than by identification of objects in the scene. This finding is relevant to the present study for two reasons: First, due to the rapid presentation, most of a scene can only be seen peripherally. Second, in natural scenes, most of the image consists of regions of various materials rather than individual objects. A waterfall scene, for example, might have water running along the middle with foliage and stone on the sides; there may not even be easily individuated plants or rocks.

Furthermore, there is an established body of work modeling peripheral vision as texture perception. This line of research suggests that a statistical representation that pools information over large regions of the visual field, namely visual texture, captures the information available to peripheral vision. Visual texture has made successful predictions for many peripheral vision phenomena, such as crowding (Balas et al., 2009; Keshvari & Rosenholtz, 2016), visual search (Alexander, Schmidt, & Zelinsky, 2014; Rosenholtz, Huang, Raj, et al., 2012; Zhang et al., 2015), and scene perception (Freeman & Simoncelli, 2011; Oliva & Torralba, 2006). One study has even made a cursory examination of peripheral vision...
material perception (Balas, 2015). For these reasons, we explicitly test peripheral material classification, and compare it to results from both baseline and texture material classification.

3.5.1 Methods for peripheral experiment

In the peripheral task, observers viewed windowed grayscale material images (not textures) at 10 deg eccentricity. The images were randomly shown to the left or right of fixation. We used the Eyelink 2000 (SRI Labs) for eyetracking, along with the standard built-in calibration procedure. The image was only on while the observer was within 2 deg of the central fixation; if the gaze fell outside of the central 2 deg, the image was removed until the gaze returned. Eyetracking was not used during the decision stage. Thus, the peripheral condition was identical to the baseline condition except that the material images could only be viewed peripherally.

3.5.2 Peripheral experiment results

11 observers completed the peripheral experiment, each doing 300 trials, in the same session but different block as the texture experiment. Average performance over subjects and trials was 0.44. As in the texture experiment, all subjects performed well above chance when averaging over all categories ($p < 2E^{-5}$). For each subject and category, performance was above chance ($p < 0.05$) except for one subject for fabric, one for foliage, and one for leather. As before, the confusion matrix of responses is shown in Figure 3.7.

responded material

```
<table>
<thead>
<tr>
<th></th>
<th>fabric</th>
<th>foliage</th>
<th>leather</th>
<th>stone</th>
<th>water</th>
<th>wood</th>
</tr>
</thead>
<tbody>
<tr>
<td>fabric</td>
<td>1.0</td>
<td>0.2</td>
<td>0.4</td>
<td>0.8</td>
<td>0.6</td>
<td>0.4</td>
</tr>
<tr>
<td>foliage</td>
<td>0.2</td>
<td>1.0</td>
<td>0.6</td>
<td>0.4</td>
<td>0.8</td>
<td>0.2</td>
</tr>
<tr>
<td>leather</td>
<td>0.4</td>
<td>0.6</td>
<td>1.0</td>
<td>0.2</td>
<td>0.4</td>
<td>0.6</td>
</tr>
<tr>
<td>stone</td>
<td>0.8</td>
<td>0.4</td>
<td>0.2</td>
<td>1.0</td>
<td>0.8</td>
<td>0.4</td>
</tr>
<tr>
<td>water</td>
<td>0.6</td>
<td>0.8</td>
<td>0.4</td>
<td>0.2</td>
<td>1.0</td>
<td>0.8</td>
</tr>
<tr>
<td>wood</td>
<td>0.4</td>
<td>0.2</td>
<td>0.6</td>
<td>0.4</td>
<td>0.8</td>
<td>1.0</td>
</tr>
</tbody>
</table>
```

Figure 3.7 Confusion matrix for peripheral viewing experiment. Similar to the confusion matrix for texture, most of the responses fall along the diagonal (indicating correct classifications), and there is a large spread in the errors (off diagonal elements).

3.5.2.1 Comparisons between peripheral results and other results

3.5.2.1.1 Confusion matrix comparison

Comparing the confusion matrices between the baseline and peripheral, we find that baseline performance is also not well predicted by peripheral; this is unsurprising given the large discrepancy between overall performance. Comparing the elements in the matrices between texture and peripheral experiments, however, we find a strong correlation (Figure 3.8A). Note that since these "data points" do not obey the assumptions made by regression (independent errors, homoscedasticity, and weak
exogeneity), the $R^2$ value (computed assuming a fit of slope one and intercept zero) is not strictly appropriate. Nevertheless, the strong correlation intuitively indicates not only that the “correct” responses are similar between texture and peripheral conditions, but so are the confusions (off-diagonal elements).

3.5.2.1.2 Comparison of precision between material categories

One way to measure classification ability for different categories is to measure the precision, or the true positive rate divided by the sum of true positive and false positive rates, of each category and for each subject. Using precision instead of proportion correct better accounts for potential biases for reporting particular categories. Looking at Figure 3.8B, we see that different materials have different levels of precision. Furthermore, we can see that the precision values tend to lie on the diagonal, suggesting that difficulty for different materials is matched between peripheral and texture conditions. We also computed the covariance of precision values for each material in the texture vs peripheral space, and found that most of the variance points along the diagonal direction. This suggests that even within a material category, texture and peripheral performance are correlated across subjects.

3.5.2.1.3 Comparing performance for individual stimuli between peripheral and texture experiments

Each subject in the texture and peripheral experiments saw each stimulus in both conditions. We can thus compute for each stimulus the proportion of subjects that correctly classified the stimulus in each task, and compare the two. In Figure 3.9, we directly compare the two tasks (a small random jitter is added to each point to avoid multiple points being occluded). Stimuli that fall along the diagonal are similarly difficult between the two conditions. The off-diagonal stimuli indicate where the texture model fails to capture peripheral appearance; stimuli in the upper left quadrant are easier to classify with textures than in the periphery, while stimuli in the lower right quadrant are more difficult. In the

Figure 3.8 (A) Comparison of elements in the texture confusion matrix (Fig. 5A) to elements in the peripheral confusion matrix (Fig. 7). Notice that not only correct classifications (the values above about 0.3 in both dimensions) fall close to the diagonal; off diagonal elements (confusions) do as well. (B) Comparison of precision values between texture and peripheral experiments. Each point represents average precision for a subject and category. Each ellipse indicates the $\alpha = 1$ equal-probability curves of a 2D Gaussian fitted to the mean precision values of each material. Notice the ellipses are oriented diagonally up and right, suggesting that within a category, subjects’ classification ability is correlated between tasks.

42
discussion we comment on what stimulus properties might cause a material image to be easier with
texture vs peripheral and vice versa.

Figure 3.9 Comparison of proportion of subjects correct with textures vs in the periphery for each stimulus.
Each point corresponds to a stimulus, and the color corresponds to category. The points are jittered slightly
to avoid overlap, since there are only 12 possible values (there were 11 observers). As in Fig. 8B, ellipses are
equal-probability curves corresponding to the stimuli of different materials.

3.6 Discussion

Materials are ubiquitous in natural visual experience, and humans are remarkably good at
identifying materials. This ability is robust to large changes in viewpoint, illumination, scale, color, and
subclass. We found that a texture statistics support this ability to some extent, but are not sufficient to
explain foveal material classification. It is well known that perception is very detailed in the fovea; this
runs in contrast with a statistical model, which recodes an image with a smaller number of translation-

invariant features. While some examples of materials are well-captured by this representation (for
example, see the two top right images of wood and their textures in Figure 3.3), the majority of material
examples contain information that is not retained by texture statistics. This finding is in line with
previous studies, which found a nonparametric model of texture (Sharan et al., 2013, 2014) to be subpar
for representing material category to humans and algorithms.
Not all materials images are represented with equal accuracy by the texture model, however. Some materials, such as water, seem to be inherently more texture-like than others. This is due to regular statistical properties inherent in images of water: waves, caustics, and distortions that occur regularly throughout an image at many different scales. Some materials, like foliage, are less likely to exhibit such regularity. Foliage can contain small, texture-like repetitive structures like leaves or bark, but can also be defined more by shape, such as a single leaf or flower. In these cases, a texture model is doomed to fail; it purposefully tosses out large-scale shape information, information which happens to be necessary for material identification. This draws an interesting parallel with sound textures. Sound textures come from repetitive, independent events (such as raindrops hitting a surface or voices in a big party), or continuous and simple sources (such wind blowing through trees or a waterfall). Such sounds are well modeled using sets of statistics that are similar to the one here (McDermott & Simoncelli, 2011). Importantly, when part of a sound violates the texture-like quality, like a single bird call near a waterfall, the statistical model fails to describe human hearing ability (McDermott, Schemitsch, & Simoncelli, 2013).

![Figure 3.10 Example materials stimuli with corresponding textures. (A) Materials for which more observers are correct with textures than in the periphery (upper-left region of Fig. 9). (B) Materials for which more observers are correct in the periphery than with textures (bottom-right region of Fig. 9) The differences between these materials is useful for spotting the shortcomings of texture as a model of peripheral vision.](image)

The most compelling finding from the present study is the fairly close correlation between performance with textures and peripheral viewing. Not only is performance per category similar between experiments, confusions between categories are similar. For instance, fabric was mistaken as leather 25% of the time with textures, and 20% of the time in the periphery (correct classification of fabric was 33% and 31%, respectively). Under foveal viewing conditions, observers made this confusion only 6% of the time, and correct classification was 88%. While it is not clear whether this confusion is ultimately due to visual or semantic similarity between the categories, the process of viewing a stimulus as texture or peripherally arguably degrades the available information to a similar level.

By showing each material image both peripherally and as a texture to each subject, we are able to visualize per stimulus where the statistical model and peripheral vision differ. We can glean some intuitions by considering stimulus-texture pairs in cases where more subjects were correct with textures
than peripherally, and vice versa. In the case where texture performance is better than peripheral performance (Figure 3.10A), there seem to be shape cues which are not diagnostic (or even misleading) of material category. The statue of a child in the bottom right image is made of wood; statues can be made of several different materials, including stone. If the periphery is relying on shape information, this particular image’s shape will not be useful for the task. Similarly, for the football in the middle right, the overall curved shape does not necessarily hint at leather; only by “stripping off” the surface material with texture is the leathery look obvious. We can also consider cases where peripheral vision is better. Looking at Figure 3.10B, the water droplet in the top left can easily be identified as water by the precise locations of the highlights and edges. In the texture version, all the positions are scrambled, resulting in an ambiguous image. Similar logic applies to the fabric example in the middle right. The vertical folds give away the fabric nature of the material; these long-range structures are washed out in the statistical representation.

One conclusion that we can draw from this study is that a full model of peripheral vision must represent these shape cues with a higher fidelity. This may be done through explicitly adding shape information to the model, as implemented for machine classification (Sharan et al., 2013) or for visual search (Alexander et al., 2014). It might also be achieved by using multiple, overlapping regions where texture statistics are computed. Extensive and promising work has been done in this domain using the Texture Tiling Model. Future research should address whether the long-range correlations afforded by models like the TTM are sufficient to explain peripheral texture perception.
Chapter 4

Web page gist: what can you see in one fixation?

Work done in collaboration with Ali Jahanian (lead author) and Ruth Rosenholtz
4.1 Abstract

Considerable work has been done to determine what is visible at a glance in a natural scene or an artificial display. Web pages are complex, designed interfaces, usually made of a mixture of several images, organized text, and graphical elements. An important question to ask is, what can we see at a glance on a webpage? From a basic vision science view, the answer advances our understanding of human visual capabilities. Practically, the answer informs our understanding of user experience, as it reveals not only what tasks the user can accurately perform within the first few hundred milliseconds, but more broadly the general usability of the page. Here we test what aspects of a web page are visible in short durations: how well can observers categorize a webpage, detect ads, and localize the menu? Borrowing from scene perception terminology, we are attempting to determine what information is in the gist of a webpage by determining what tasks the representation supports. We find that users, fixating at the center of a webpage shown for only 120 milliseconds, are well above chance at classifying web pages into one of our ten chosen categories. Furthermore, we find evidence that in doing so they can make use of readable text. Our results also show that users can reliably localize the menu bar and detect ads. The latter is more difficult, however, perhaps because advertisers aim to camouflage ads. We discuss the implications of our findings for vision science and human-computer interaction.

4.1 Introduction

A fundamental constraint on performance in any visual task is the extent of information available at a glance. In the context of navigating a web page, for instance, if a user immediately sees evidence of a menu in the periphery, she need not search for it; she can simply direct her gaze to it. If she notices an ad, she can choose to click on it or ignore it. If she can quickly determine the category of web page, it allows her to more easily decide if it will contain information she needs. At the other extreme, if the user cannot get much information out of a page at a glance, she may be forced to read a significant amount of the text, a slow and perhaps frustrating process. In the extreme, the user may decide to leave
the page to find a new page that is more visually comprehensible. While web pages certainly contain pertinent information beyond what is available at a glance, if well designed they can efficiently display cues for planning subsequent eye fixations and allow a viewer to easily piece together a holistic understanding of the page.

Precisely, by “at a glance”, we mean the duration of one single eye fixation, prior to starting any saccadic movements. During this time, our visual system encodes information and plans for eye movements towards where to look next. This duration can last between 100 and 300 milliseconds (ms) (Harris, Hainline, Abramov, Lemerise, & Camenzuli, 1988; Pieters & Wedel, 2012; Wedel & Pieters, 2000), while typical fixations fall in the range of 200-250 ms (Rayner & Castelhano, 2007). A considerable amount of information about a stimulus, such as a scene or display, is available in a single fixation. This summary information has been termed the “gist” of the stimulus. Colloquially, the gist is the sentence one might use to describe that stimulus. Often this is operationalized as “the perceived contents of a scene given a certain amount of viewing time” (Fei-Fei, Iyer, Koch, & Perona, 2007), often in a single fixation of the eyes (Fei-Fei et al., 2007; Oliva, 2005). Studying this gist representation informs us about human vision and practically speaks to understanding user experience, as it determines not only what tasks the user can accurately perform within the first few hundred milliseconds, but more generally how usable a design is.

Studying at a glance vision has proven valuable in constraining models of human perception and attention (Oliva & Torralba, 2006). Experiments in this field have been done mostly with the goal of uncovering the basic mechanisms of vision. Rarely, however, have researchers in human vision extended this work to complex designed stimuli like web pages, from which we might not only learn about human vision, but also gain insight into design for usability. It is worth noting that getting the gist of a natural scene is easy perhaps because the human visual system is biologically optimized for it; getting the gist of an artificial design is not guaranteed to be easy or natural. To this end, here we examine the strengths and limits of human vision as applied to current design practice. We choose web pages as our stimuli because they have a great deal of variation and richness, usually containing a mixture of several images along with organized text and graphical elements. This confounding of design, prior knowledge, and human perception is a strength rather than a limitation of the present study; because we examine performance with screenshots of real web pages, our findings are relevant to actual implemented designs. Our interest lies specifically with how well designers convey the message of a web page at a glance.

Researchers in the HCI (Human-Computer Interaction) community have begun to study perception of web pages at a glance. This work, however, has focused mostly on rapid subjective judgments. For instance, recent studies have suggested that users form subjective impressions of the appeal, trustworthiness, and usability of a web page in the first 50 ms of viewing (Lindgaard, Dudek, Sen, Sumegi, & Noonan, 2011) and respond consistently when shown the same stimulus later. Furthermore, quick impressions of visual appeal correlate well with subjective judgments upon longer viewing (Tractinsky, Cokhavi, Kirschenbaum, & Sharfi, 2006). While this is clearly valuable, visual appeal is not the only issue in understanding at a glance perception of web pages. It is also possible that the cues that support aesthetics judgments are not the same as those used for semantic judgments. Recent studies attempt to understand correlation between aesthetics judgements and low-pass spatial frequencies (as a perceptual feature) of web pages when the stimulus is exposed up to 1sec (Thielisch & Hirschfeld,
Although low-pass frequencies play a role in getting the gist (Oliva & Torralba, 2001), there must be other vision cues that contribute to this.

Here we investigate the contents of the gist of a web page. In other words, we ask what can be seen a brief, single fixation of a web page. While there are many relevant ways to ask this question, here we examine only a few that we find particularly pertinent. First, we ask whether users can rapidly ascertain the category of a web page, and whether categorization ability depends on reading text on the web page. This is important because designers’ choices can make web pages inherently more or less distinct from the average category, affecting how they are perceived. Surprisingly, in the HCI community, the common belief is a user cannot get much semantic information, like the category of a web page or any readable text, given short presentation times less than 500 ms (e.g. see (Lindgaard et al., 2011)). Next, we ask how well a user can locate a menu. By making a menu more or less obvious, possibly for aesthetic reasons, a designer can greatly affect the navigability of a page. Finally, we determine whether ads are detectable at a glance. Ads are often designed to be difficult to distinguish from other page elements; users may decide to avoid web pages with excessive or distracting ads.

We display web pages for 120 ms, in order to limit observers to a single fixation. Display times of this magnitude are typical for such studies in the field of human vision (Fei-Fei et al., 2007). In Experiment 1A, we ask whether users can accurately assess the category of a web page, i.e. whether it is a blog, a news page, etc. Do they manage, in a brief glance, to read any of the text, and does that inform their categorization, or does it derive from other web page elements (Experiment 1B)? Can they quickly tell ads from other page elements (Experiment 2)? Can they quickly understand the layout, for instance locating the navigation menu (Experiment 3)? To what extent do they require that an observer be able to read words in the web page, which may be challenging due to the limits of peripheral vision? What observers can see within a single fixation depends heavily upon the strengths and limitations of peripheral vision, in particular, and peripheral vision is highly sensitive to clutter and to the perceptual organization of the stimulus (Pelli & Tillman, 2008).

4.2 Related work

The question of what one can see in a single fixation is particularly interesting, both because it asks what we can perceive, given the strengths and limitations of peripheral vision, and because the information available in a single fixation guides later eye movements, serving as an important determinant visual task difficulty eye movements are allowed (Rosenholtz, 2011b).

Research on human vision has shown that observers can at a glance identify the scene category (e.g. beach vs. forest, indoor vs. outdoor), and how much room there is to navigate through the depicted scene (Greene & Oliva, 2009a; Rousselet, Joubert, & Fabre-Thorpe, 2005). Observers can identify whether given objects are present, such as an animal (Kirchner & Thorpe, 2006; Li, VanRullen, Koch, & Perona, 2002; Thorpe, Fize, & Marlot, 1996), vehicle (VanRullen & Thorpe, 2001), or a human face (Crouzet, Kirchner, & Thorpe, 2010). Furthermore, experiments in which observers freely report whatever they perceived in the scene, i.e. the gist contents, have revealed the richness of the perception of lower and mid-level properties, such as the colors and textures present(Fei-Fei et al., 2007).

While this experimental paradigm has been highly fruitful for understanding human vision, it has rarely been applied to understand perception of real-world designs such as web pages. In a seminal
work, (Lindgaard, Fernandes, Dudek, & Brown, 2006) studied the perceived visual appeal of web pages after a brief exposure, compared to longer viewing times. They examined a 500 ms presentation time for 50 web pages. Participants rated each page twice, in two phases. They found a high correlation between the participants’ mean ratings in the two phases, suggesting that participants’ judgments are consistent across time. When the researchers decreased the presentation time to 50 ms, the correlation between the mean ratings of the two phases still was high; participants make consistent visual appeal judgments even after only 50 ms of viewing. Later, (Lindgaard et al., 2011) demonstrated that users also make consistent subjective ratings on trust and usability of web pages in 50 ms. While the results provide valuable knowledge in understanding first impressions, the study focuses on the subjective aspect of perception. In contrast to subjective judgments, we ask more objective questions about perception of short exposures. Objective questions allow us to more accurately probe the contents of the resulting gist. Furthermore, this study assumes the common misconception that not enough semantic information is available in the first impression to form a holistic picture, or too few elements (e.g. readable text) from the web page can be perceived. That is, they argue that fast exposure marginalizes the effect of content (semantics and topic of the page) in subjective judgements. The current study explicitly investigates this assumption, and to preview the results, provides objective evidence otherwise.

(Thielsch & Hirschfeld, 2010, 2012) study the relation between spatial frequencies of web pages and aesthetics judgements. (Thielsch & Hirschfeld, 2010) show low frequency (LF), high frequency (HF), and unfiltered (UF) versions of web pages to participants. They find high correlation between LF and aesthetics, and HF and usability. Their stimuli are chosen from 10 categories such as e-commerce, corporate, etc. In their work, they borrowed the idea from previous work in vision (e.g. (Oliva & Torralba, 2007)) that LF is a cue for extracting the gist of the visual stimulus. In other words, without really aiming for studying gist of web pages, they take LF as a proxy of gist. This finding is interesting and relevant to our purposes, and hints at what might be common between perception of natural scenes and designs. LF, however, is clearly not the only stimulus cue that affects the gist of a web page.

Another relevant work studies people’s perceptions of ads on web pages (Roth, Schmutz, Pauwels, Bargas-Avila, & Opwis, 2010) and on magazines (Pieters & Wedel, 2012). (Roth et al., 2010) find evidence of consensus between participants about where ads and menus may appear on a web page (in three categories of “news portal”, “online shop”, and “company”). However, they did not study any rapid exposure stimuli. On the other hand, (Pieters & Wedel, 2012) study viewers’ gist when viewing ads in magazines. They find that people can distinguish between ads and editorial articles (selected from magazines), with high accuracy in one eye fixation (100ms in their experiments). People could also very accurately tell which kind of product (among four selected categories: cars, financial services, food, and skincare) a typical ad was about. In our work, we are interested in web pages, which often include various design elements and styles of representing information. In our study of ads, we ask a more basic question: Is it possible to recognize ads within a web page?

4.3 What are the web page categories?

As it is known that people can quickly get the basic-level category of a natural scene, we wanted to know if they can also quickly get the basic-level category of a web page. In order to cleanly ask whether people can do this categorization at a glance, we first needed people to agree on the categorization of our particular set of exemplars when given unlimited time. First, we needed to define
plausible categories. We picked category definitions to minimize overlap between categories. Then we needed to acquire a corpus of pages belonging to these categories, and confirm our ground truth categories. The procedure of defining the categories as well as collecting corresponding web pages is described below.

4.3.1 Picking the web page categories

We needed to define some plausible categories, at the level of “company web page,” “blog,” and “online shopping web page.” We conceived of categories in terms of what task a user might want to do with a web page. Users go to a news page to learn about current events, a tourism page to learn about destination’s facts, etc. It would be difficult, for instance, to allow both “sports” and “news” as categories, as many news web pages might report sports news, making the category ambiguous. Table 1 lists the resulting categories, and the definitions provided to experimental subjects.

<table>
<thead>
<tr>
<th>Category</th>
<th># screenshots</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Art place</td>
<td>30</td>
<td>if you want to see art, e.g. exhibition, museum, galleries</td>
</tr>
<tr>
<td>2. Blog</td>
<td>32</td>
<td>has articles with titles and dates, it usually has opinion of a person with a form that you could write your comments</td>
</tr>
<tr>
<td>3. Company</td>
<td>33</td>
<td>if you want to learn about the services that it provides, e.g. consultation</td>
</tr>
<tr>
<td>4. Computer game</td>
<td>33</td>
<td>if you want to play a computer game either online or offline, or read the latest news about a computer game, buy or download a computer game</td>
</tr>
<tr>
<td>5. Helpline</td>
<td>32</td>
<td>if you have an emergency, e.g. need advice for your kid</td>
</tr>
<tr>
<td>6. News</td>
<td>34</td>
<td>if you want to read daily news</td>
</tr>
<tr>
<td>7. Online tutorials</td>
<td>28</td>
<td>e.g. learning a course, finding a tutorial for learning html coding</td>
</tr>
<tr>
<td>8. Shopping</td>
<td>34</td>
<td>if you want to by an online product or item, new or used</td>
</tr>
<tr>
<td>9. Society</td>
<td>37</td>
<td>unions, groups of people with same interest, e.g. if you are looking for a cultural club, or book reading club</td>
</tr>
<tr>
<td>10. Tourism</td>
<td>41</td>
<td>if you are looking for things to do in a destination, booking for a tour</td>
</tr>
</tbody>
</table>

Table 4.1 Selected web page categories and their descriptions. Each subject could consult this list at any time during the experiment.

4.3.2 Collecting web page screenshots

In order to automatically gather a candidate set of web pages belonging to the 10 categories, we first extracted the urls of various categories from the Open Directory Project (DMOZ.org). This is a common technique in url-based topic classification problem (Baykan, Henzinger, Marian, & Weber, 2011). We did not want to use a list of high-traffic urls (such as the one provided by Alexa.com), to avoid familiarity biases to the stimuli. We also did not use search engines to find urls, in order to reduce the potential biases caused by their algorithms in retrieving. We crawled OPD for our selected categories (left column of Table 4.1). Then, we randomly picked urls from those categories. The screenshots of the urls were captured in 1200 by 800 pixels using the webkit2png software package (https://github.com/adamm/python-webkit2png). This resulted in 1795 screenshots.

4.3.3 Getting ground truth categories for screenshots

By hand, we then culled exemplars that appeared, upon examination, to be ambiguous in category, written in a foreign language (not English), or not fully loaded at the time of screenshot. This
resulted in 714 screenshots. In order to more objectively assess whether the labeled category for each screenshot was correct, we had 5 participants categorize each screenshot into the 10 categories listed previously, given unlimited response time. (3 participants were male, 3 female, all native English speaker university students. One observer was left out because of apparent difficulty understanding the instructions.) A row of response buttons, one for each category, was displayed below the screenshots. Participants were encouraged not to spend more than several seconds on any particular web page. This was to avoid picking the category based on small or technical details in the text; we wanted to avoid overly subtle categorizations, such as deciding that a blog is a tourism web page because on careful reading it becomes apparent that many of the posts involve travel.

Each participant viewed 379 screenshots (we reduced the number from the original 714 to fit the experiment into a one-hour session.) Of these 379 screenshots, we found that for 283 (about 75%), all 5 observers agreed with our categorization, and for 334 (about 88%) at least 4 of 5 agreed. In the further analyses, we consider only web pages for which at least 4 of these 5 observers agreed with our initial categorization. This filtering left us with 334 trials for our experiments. The number of screenshots in each category is as follows. art place: 30, blog: 32, company: 33, computer-game: 33, helpline: 32, news: 34, online-tutorial: 28, shopping: 34, society: 37, tourism: 41.

4.4 General experimental procedure

In all of our experiments, observers viewed the stimuli on a Dell E2209W LCD monitor, with their eyes at a distance of approximately 55 cm from the center of the screen. Stimulus presentation and response collection was done using PsychToolbox-3 (Kleiner et al., 2007) in MATLAB.

In Experiments 1A and 1B, observers responded by using the mouse to click on-screen buttons, and received no feedback about performance. The order of presentation of the web page screenshots was randomized for each observer. All stimuli were displayed at the original 1200 by 800 pixels, and subtended an area of the screen approximately 33.75 cm by 22.5 cm (34 by 23 degrees visual angle). In other words, all web pages were displayed at the same size and resolution used to capture the screenshot. All of the text in the screenshots (except a few pixelated and uninformative bits of text, such as disclaimers or trademarks) was legible to an observer with normal or corrected-to-normal vision at the experimental viewing distance.

Prior to the experiment, the experimenter gave a verbal overview of each category. The participants were also instructed to consult with a list of all 10 categories and their short descriptions (see Table I), while making their decisions. This list was displayed on a separate screen. Participants were informed that they could take as many breaks as necessary and received an indication of their progress a few times during the experiment.

4.4.1 Recruiting participants

Performance will likely depend upon a user’s experience with web browsing, and with whether the user comes from the same culture as the designs studied (Indian news pages might look different from American ones). While it is an interesting question to ask how experience and culture influence perception, it is outside of the scope of the present study. We thus recruited participants from roughly the same age range and demographic: university (undergrad and grad) students with English as their first language. In total, we recruited 25 participants (average age = 23.56 years, standard deviation = 4.00 years, range of 18 to 35 years). The participants on average spent 3 hours per day surfing the
Internet. Each participant completed between 30 minutes and an hour of experiment, and received $15 USD as compensation for their time. All participants gave informed written consent prior to the experiment.

4.5 Experiment 1A: Can observers categorize web pages at a glance?

4.5.1 Procedure

Participants were told that they would be categorizing screenshots of web pages that were presented for very short time periods. They were asked to keep fixation at the center of the screen, where a cross appeared prior to each screenshot. When viewing natural scenes, observers are biased to look at the center, so enforcing central fixation for at a glance scene perception is a natural choice. For web pages, there is no clear bias, and typical fixation location depends on task. A prior study examined where people looked while surfing the Internet, and found that for information foraging tasks, subjects mainly looked at the center (center-center, left-center and left-center) of the web pages in the first second of fixation (Buscher, Cutrell, & Morris, 2009). One caveat is that in that study, a set of questions were given to the subjects (e.g. “Which of three given cars, Porsche, BMW, or Audi, has the best performance?”). In absence of a strong reason to use non-central fixations, we chose central fixation. Future work could examine web page perception based on typical fixations in a particular task.

![Figure 4.2 Schematic of the gist experiment. Participants were instructed to fixate on a cross in the center of the screen. A screenshot appeared for 120 ms, then was replaced with a response screen after a short blank screen. Participants had unlimited time to make their response, then the next trial began.](image)

We recruited 10 participants for this task (5 female). Participants first did a short training session of 30 trials to orient them to the experiment. The screenshots used in the training came from a separate set than the main experiment.

Each trial consisted of three steps. First, a fixation cross appeared on the screen for 900 ms, which we asked the participants to look at. The fixation was then immediately replaced with the web page screenshot. After 120 ms, the screenshot was removed. The response buttons then appeared about 30 milliseconds later on the screen, and remained until the participant made a response. The buttons disappeared immediately after the response was made and the fixation of the next trial.
appeared after 250 ms (see Figure 4.2) Each observer did a total of 379 trials, and we included 334 of them for the analysis (see Getting ground truth categories for screenshots).

4.5.2 Results

We aggregated the responses from all 10 participants and computed confusion matrices between all 10 categories. Overall, on average subjects were 47% correct. Each subject’s performance averaged over categories was well above chance ($p < 1E-5$; two-sided permutation test). Furthermore, performance in each category was also well above chance ($p < 1E-5$; two-sided permutation) when averaged across subjects.

What cues in the stimuli might support such high performance? Interestingly, several participants indicated to us that they could read words within the screenshots, and that they used this information in doing the categorization task. This was surprising given the common assumption in HCI that text is not readable at such short timescales (Lindgaard et al., 2006). We explored this possibility explicitly in Experiment 1B.

![Figure 4.3 Results of the gist experiment. Each row denotes the correct category of a given figure, and each column the responded category. The value in each cell is the fraction of trials which, for each true category (row label), the screenshot was identified as the category given by the column label. Values along the diagonal indicate the fraction of correct responses per category.](image)
4.6 Experiment 1B: Gist of scrambled web pages

Was some of the web page text at least partially readable, and might that have provided a categorization cue? It is plausible that observers could have read some of the text. According to research in Rapid Serial Visual Presentation (RSVP), humans can read at least 12 words per second (83.3 ms per word) when presented foveally (Potter, 1984). Peripherally, humans can also read words in such short times, provided that the font is large and spaced out enough (Latham & Whitaker, 1996). Otherwise, a phenomenon known as crowding strongly limits peripheral reading (Pelli & Tillman, 2008).

To determine whether participants could use any of the text on the page to do categorization, we asked a new set of participants to categorize web pages with text scrambled such that it was illegible (see Scrambling web page text). We recruited 10 participants for this task (5 female).

4.6.1 Scrambling web page text

In order to scramble the text, we used custom software developed by two undergraduate students (Kevin Kwok - kkwok@mit.edu and Guillermo Webster - gui@mit.edu). This software uses the Stroke Width Transform (Epshtein, Ofek, & Wexler, 2010) and connected components analysis to detect letters in an image, and vertically flips the bounding box for each letter upside-down. This results in an image where each letter of text is vertically flipped. While very accurate, this process did not catch all words, and occasionally mistakenly flipped regions of the screenshot that contained images and no text.

To remedy these shortcomings, an experimenter manually checked each scrambled screenshot, and inverted any text that was still readable. In order to do this, the experimenter used image-editing...
software to manually draw the minimal bounding box around each word, line, or paragraph (of uniform background), and vertically flip it. This was repeated for all readable (not already-inverted) text, including text in brand logos. Any significantly large image regions that were mistakenly flipped by the algorithm were unflipped by the experimenter as well, to avoid confusing the participants. See Figure 4.4 for examples of screenshots with scrambled text.

4.6.2 Procedure

The procedure was identical to that in Experiment 1A, except that all of the text in each screenshot was scrambled (except for the initial training, in which we used 15 scrambled and 15 unmodified screenshots).

4.6.3 Results

We aggregated the responses from all 10 participants, and computed confusion matrices between all 10 categories (Figure 4.5). As before, participants performed well (43% correct) on the categorization task. Specifically, each subject performed above chance ($p < 1E-5$; two-sided permutation test). Performance in each category, averaged over subjects, was also well above chance ($p < 1E-5$; two-sided permutation test), averaged across subjects.

Figure 4.5 Results of the gist experiment with scrambled text. As in the previous confusion matrix, the value in each cell is the percentage of trials for which, for each true category (row label), the screenshot was labeled as indicated by the column label.
Combining Experiments 1A and 1B, we ran a 10 (category) x 2 (scramble/no scramble) mixed model ANOVA. We found a significant main effect of scrambling the text \( (F(1,324) = 20.49, \ p < 1E-3) \); the categorization task is more difficult when the text is unreadable. This implies that observers in Experiment 1A were in fact inferring the category in part from readable text (although since performance in the scrambled condition was well above chance, it is clearly not the only cue). We also found a significant main effect of category \( (F(9,324) = 11.59, \ p < 1E-3) \). Some types of web pages are more difficult to categorize than others, both with and without readable text. Finally, we found a significant interaction between category and text scrambling \( (F(9,324) = 7.91, \ p < 1E-3) \), meaning that readable text provided a better cue for some categories than for others. To better visualize the effect of scrambling the text, we subtracted performance in the scrambled case from the unscrambled (see Figure 4.6).

![Confusion Matrix](image)

**Figure 4.6** The confusion matrix for Experiment 1A (gist of web pages) minus the confusion matrix for Experiment 1B (gist of scrambled web pages). Positive values (green) indicate that more responses were assigned to that category in the intact-text case than the scrambled case. Along the diagonal, positive values indicate that performance is better with intact text.

### 4.7 Experiment 2 - Does this web page have an ad?

In addition to basic-level category, the gist of a real world scene includes sufficient information to recognize a few objects (Fei-Fei et al., 2007; Oliva, 2005). Can observers also recognize elements of a web page at a glance? One tricky type of web page content is advertising. Web pages often display ads to generate revenue, and advertisers in many cases want users to click on these ads. As a result, designers of both web pages and ads take care in setting the location and style of an ad to best appeal to visitors, or possibly to trick them into clicking on the ad. Users, on the other hand, may desire not to click on irrelevant ads (or perhaps on any ad at all). In Experiment 2, we ask whether observers can
detect the presence of an ad within a web page at a glance. The same 10 participants in Experiment 1B completed this experiment in a separate session.

4.7.1 Procedure

We manually selected 50 screenshots with at least one visible ad, and 50 without. Note that these stimuli are selected from the original set of 714 screenshots, and have not been used in any of the previous experiments. To qualify as having an ad, the web page had to display the entire ad, it had to be fully loaded in the screenshot, and it could not link to the same web page as the current page. Thus, we would exclude an ad for a different product sold by the same shopping web page. We excluded a bank web page, for example, if it included an ad about its own banking service. The text on the screenshots was not scrambled. The screenshots were displayed with the same timing and steps as in Experiment 1A. The response screen on ad vs no-ad trials had two response buttons (one green for “ad”, one red for “no ad”), underneath the text “Did you see any ad(s)?”

Observers did this task randomly interleaved with the one in Experiment 3 (finding navigational menus). After each screenshot appeared, with equal probability the observer was asked to answer either about the presence of an ad, or where the menu was. The observer did not know which question would be asked until after the screenshot was shown. Observers were thus instructed beforehand to do their best to remember what they saw of each screenshot as the task was uncertain until afterwards. This task uncertainty limited how much spatial attention the participants could employ to focus on specific locations on the screenshot. For example, the observer who only paid attention to the top and left regions in order to do better on the menu-localization task (Experiment 3) would suffer in performance on the ad-detection task. This uncertainty also discouraged anticipatory eye movements.

4.7.2 Results

The results are shown in Figure 4.7. Each row indicates the ground truth, and each column represents the response. Thus, the top left value indicates correct detections of ads, the top right indicates missed ads, bottom left indicates false detections, and the bottom right indicates correct rejections. 8 of the 10 observers performed above chance ($p < 0.03$; two-sided permutation test), and the average performance over all subjects (64.7%) was also well above chance ($p < 1E-5$; two-sided permutation test). Thus, observers usually notice ads on a web page within a single glance. Since the observers responded “ad” and “no ad” with roughly equal frequency, it is unlikely that they were biased towards either response.

![Figure 4.7. Results of the ad-detection task, in terms of proportion response. Rows indicate the ground truth, while columns indicate the responses. Diagonal values indicate correct classifications.](image-url)
4.8 Experiment 3 - Where is the menu?

Other interesting questions concern whether a user can perceive the layout or organization of a web page at a glance: the columns or rows of text, alignment, grouping, and so on. The layout of a web page largely determines how effective it is. Menu bars are specifically defined by their perceptual organization: the alignment of a single line of words or phrases across the top of the page, or down the side. Sometimes the menu is segmented from the rest of the page by a colored box. They are used to represent cues of visual hierarchy, e.g., where to start, or where to find information. Designers take care to put in form elements, such as menus, to help the user easily navigate; a good design depends on users finding them quickly. We asked whether observers could tell at a glance where the menu bar was.

4.8.1 Procedure

For the menu localization task, we manually selected 100 screenshots from the original 714 screenshots, such that they are not shown in any other previous experiments. We selected 50 web pages that had a menu bar along the top (extending horizontally) of the page, and 50 that had a menu bar along the left (extending vertically). We excluded web pages that had menu bars in both locations.

Observers did this task interleaved with the one in Experiment 2, using the same procedure. The response screen for the menu localization task had two same-colored buttons, labeled “side” and “top”, beneath the text, “Was the menu on the left side or the top?”

4.8.2 Results

The results can be found in Figure 4.8. As before, each row indicates the ground truth, and each column represents the response. The diagonal elements indicate correct localizations of the menu bar, while the off-diagonal elements indicate errors. Participants were remarkably good at this task (84.6% correct on average), and each subject was significantly above chance ($p < 1E-5$; two-sided permutation test).

This suggests that a single fixation provides enough information to find the menu for many designs. The participants did not seem to be biased towards either location. It is perhaps not surprising that performance on this task exceeds that on the ad discrimination task, as users and designers share the goal of easy page navigation, whereas their goals may be different when it comes to identifying ads (e.g., designers may not want ads to be easy to identify).

<table>
<thead>
<tr>
<th>responded category</th>
<th>side</th>
<th>top</th>
</tr>
</thead>
<tbody>
<tr>
<td>side</td>
<td>0.82</td>
<td>0.18</td>
</tr>
<tr>
<td>top</td>
<td>0.15</td>
<td>0.85</td>
</tr>
</tbody>
</table>

*Figure 4.8 Results of menu-localization task, in proportion response. Rows indicate the ground truth, while columns indicate the responses. Diagonal values indicate correct classifications.*
4.9 Discussion

We have demonstrated that in a single 120 ms fixation, observers can quickly categorize a web page into one of ten common categories. Furthermore, it seems that observers are at least in part using text in the web pages to do the task. Future work should probe this further, and ask whether observers predominantly use text near the point of fixation, or whether they can read (or at least infer category from) bigger text that appears more peripherally. Clearly, however, the bulk of the performance comes from other cues, such as layout, presence and content of images, and so on. Formally investigating these cues is an important direction of future work. Furthermore, observers can discriminate pages that contain ads from those that do not, as well as localize elements of the layout, namely the menu bar. Ad detection is more difficult than menu localization, possibly resulting from designers’ different purposes for the two elements. Further research should pin down which menu or ad styles are more visible than others, and why.

That observers can extract this information from a glance at a web page is perhaps not at face value surprising; however, we emphasize that there has not been a quantitative study of quick perception of such a prevalent stimulus. As we become more enmeshed in the virtual world, more of our “natural vision” will be filled with artificial displays like web pages. We expect that research on at a glance vision of displays and designs will become more relevant and common in the near future.

These results on one hand speak to the capabilities of the human visual system. Designers could potentially make use of these abilities, for example to quickly display an RSVP stream of web search results. It informs design to know that users may be quickly reading text, identifying the category, and understanding the layout and some page elements. User performance also speaks to the quality of current design practice. Designers might wish to design for better categorization and layout performance, though perhaps the goals for ad design differ. Furthermore, these results speak to models of rapid assessment of visual appeal. Though recent work has attempted to predict early visual appeal ratings using low-level image features (Miniukovich & Angeli, 2015; Reinecke et al., 2013; Zheng, Chakraborty, Lin, & Rauschenberger, 2009), our work suggests that users may bring to bear higher-level semantic judgments, which they can also perform rapidly.

There are many other interesting questions one could ask about web page perception at a glance. What other elements can users quickly identify? How much layout information do users get at a glance? What design elements underlie the ability to accurately categorize a page, identify an ad, or find a menu? Our hope is, beyond presenting our specific experimental results, to bring the rich experimental paradigm of studying vision at a glance to the HCI community.
Chapter 5

Conclusion
5.1 General discussion

There is a striking divide between our models of perception. At one end, statistics-based models provide explanations about how the brain manages to compress the deluge of incoming perceptual stimulation into an overarching representation. At the other, object-based models postulate discrete representations that underlie our ability to interpret the information in a stimulus semantically. Both levels of abstraction play an important part of everyday perception, and only by explaining how they interact and drive each other can we have a rigorous understanding of perception. To this end, the experiments described here follow what the actual limitations of each approach are, while pointing in a direction that unifies them in a computationally rigorous fashion. We hope that these efforts will lead to a more general and predictive understanding of perceptual representations, particularly in peripheral vision.

Furthermore, beyond theoretical considerations, an improved understanding of peripheral vision can provide an abundance of practical benefits. Given how pervasive peripheral vision is, applications can be found in a wide variety of contexts. In the health sector, much work has been done to develop models of the phenomenology of peripheral vision loss (Hallum, Suaning, Taubman, & Lovell, 2005; Larson & Loschky, 2009; Wiececk, Pasquale, Fiser, Dakin, & Bex, 2012; Wood & Troutbeck, 1992). Many common disorders, such as glaucoma, eye strokes, retinitis pigmentosa, retinal detachment, and diabetes can lead to selective loss of peripheral vision, also called “tunnel vision” (Luo & Peli, 2004; Peli, Goldstein, Young, Trempe, & Buzney, 1991; Saunders, Russell, & Crabb, 2012; Turano, Geruschat, Baker, Stahl, & Shapiro, 2001). Tunnel vision causes severe difficulties in navigation, spatial judgment, and even posture (Zapf, Mei-Ying Boon, Lovell, & Suaning, 2015). Neural prosthetics or augmented reality systems may be able to help mitigate some of the difficulties caused by peripheral vision loss; it is important, however, that these approaches take into account what we know about peripheral vision (Luo & Peli, 2004; Peli et al., 1991; Zapf, Matteucci, Lovell, Zheng, & Suaning, 2014; Zapf et al., 2015). In contrast to tunnel vision, disorders like macular degeneration often lead to selective central vision loss, such that the patient has relatively intact peripheral but impaired foveal vision. Because these patients are forced to use only their peripheral vision, they have difficulty in high acuity vision (Kwon et al., 2012). Interestingly, many of such patients develop a preferred retinal locus (PRL) in the intact periphery, a new part of the retina where they direct their fixations to (Crossland, Culham, & Rubin, 2004; Kwon, Nandy, & Tjan, 2013; Seiple, Grant, & Szlyk, 2011; Tarita-Nistor, Gonzalez, Markowitz, Lillakas, & Steinbach, 2008; Van der Stigchel et al., 2013; Walsh & Liu, 2014). This “new” fovea gains some benefits over the preexisting baseline peripheral retina, but still has fundamental deficits (Bullimore, Bailey, & Wacker, 1991; Chung, Mansfield, & Legge, 1998; Elliott et al., 1997; Mäkelä, Näsänen, Rovamo, & Melmoth, 2001; Rubin & Turano, 1994). By using a modeling approach for peripheral vision, we might be able to understand the PRL in terms of existing model parameters, and possibly make predictions for therapies. Furthermore, by applying our modeling techniques, we can simulate peripheral or central vision loss to better understand the difficulties faced by impaired populations (Aguilar & Castet, 2011; Bernard, Anne-Catherine, & Eric, 2007; Janssen & Verghese, 2015; Kwon et al., 2013; Rubin, 2001; Varzori, Perez-Fornos, Safran, & Whatham, 2004; Walsh & Liu, 2014), possibly leading to new assistance tools and interfaces.

Peripheral vision plays a key role in the fields of HCI, user interaction, and design. Above the current work on at a glance perception of web pages, studying peripheral vision of maps, navigational instruments, and virtual displays is vital for designers and developers (Raj & Rosenholtz, 2010). It is well
known that visual clutter can impair the usability of a display. Peripheral vision is particularly sensitive to clutter, and using image-computable models like the one presented here can help designers make concrete predictions about their products (Raj & Rosenholtz, 2010; Rosenholtz, Dorai, & Freeman, 2011). Peripheral vision is also a major player in driving (Clay et al., 2005; Henderson, Gagnon, Bélanger, Tabone, & Collin, 2010). Current knowledge in driving research espouses oversimplified models of vision, namely models in which foveal and near-peripheral vision are collapsed into one notion called the “useful field of view”, or UFOV (Clay et al., 2005). New research, however, is attempting to apply the theories and models presented here to better understand what drivers can and cannot see both on the road and within the vehicle (Henderson et al., 2010). Finally, virtual and augmented reality (VR/AR) are poised to become parts of everyday life. Given that the vast majority of the visual field is in the periphery, it is crucial that the technology implement up-to-date knowledge about peripheral vision. In most current implementations of VR and AR, the field of view is quite limited, due to technological shortcomings and a general opinion that peripherally displayed information is not necessary for a good experience (Lin, Duh, Parker, Abi-Rached, & Furness, 2002). Our demonstration of the information available at a glance with webpages runs counter to this intuition. Our research predicts that successful technologies will focus more on making a realistic peripheral scene, and can inform how specific cues might be used to show information peripherally. Furthermore, because we have a better understanding of the limitations of peripheral vision, it is possible to use gaze-contingent rendering in VR/AR to only show at high fidelity the regions that are fixated at a point in time. By displaying the periphery in a compressed fashion, computation and communication costs can be reduced (Duchowski & Çöltekin, 2007; Guenter, Finch, Drucker, Tan, & Snyder, 2012). The particular scheme of compression, however, must be indistinguishable in peripheral vision; our modeling techniques can help guide the development of such schemes. Finally, with AR in particular, users see a digital image overlaid on the real world, and can navigate the world freely. An AR system should incorporate what is known about natural peripheral vision to create safe interactions, and avoid presenting or failing to present peripheral cues that distract or are necessary for navigation, respectively.

5.2 Convolutional neural networks and peripheral vision

Convolutional neural networks (CNNs) have recently stolen the show in computer vision applications (Krizhevsky et al., 2012). CNNs are the modern successor of standard feed-forward neural networks. In short, given an input image, CNNs alternate applying linearly rectified filter banks (convolved with the image), pooling layers (down-sampling), and fully connected layers, with a (typically) soft-max layer at the end with the same number of neurons as object classes; the activations in the final layer represent the likelihood of each object category being in the image. The filters and weight of connections between and within layers are optimized using gradient descent, typically to maximize object classification performance (Krizhevsky et al., 2012). CNNs bear a resemblance to a class of biology-inspired models of the visual ventral hierarchy (see (Riesenhuber & Poggio, 1999) for example) which consist of layers of cells, where each layer computes increasingly complex linear and nonlinear operations on the previous layer, and the connections between layers are tuned to maximize object recognition performance. Importantly, CNNs are typically trained with the goal of having the same sort of invariance to changes in size and position of an object as human vision. Both CNNs and hierarchical vision models end up learning what correlations between various features within and between spatial scales are useful for conveying the identity of an object. Inspired by the similarities between CNNs, biologically inspired models, and research on representations in visual cortex, a growing
avenue of research compares CNN models to brain activation during visual tasks (Cadieu et al., 2014; Güçlü & van Gerven, 2015; Khaligh-Razavi & Kriegeskorte, 2014).

Given the rise of CNNs as models of object recognition and of vision in general, a pertinent question to ask is whether they can be adapted to describe peripheral vision. Perhaps, since humans can identify objects peripherally (in the absence of strong crowding) under large changes in appearance (due to eye movements, etc.), the representations learned by CNNs trained on invariant object recognition could be pooled over large regions to model the pooling in peripheral vision. It is possible that this approach would match behavior better than pooling of handcrafted Portilla-Simoncelli statistics (Portilla & Simoncelli, 2000). This avenue of research is largely unexplored; studies that compare CNNs to human vision use foveal or parafoveal presentation in order to maximize performance. One way to get at this would be to generate “textures” using pooled CNN features, just as is done with Portilla-Simoncelli, in order to visualize the equivalence class of the stimulus representation.

Interestingly, a new approach for texture generation with CNNs might present a way to approach this idea. The new study (Gatys et al., 2015) uses the middle layers of a CNN trained on object detection to instead model visual textures. Specifically, the model measures the correlations between filter responses to a particular visual texture image in the convolutional layers of the CNN as a texture descriptor. The authors find that adjusting a noise image to share the same filter-response correlations as the original texture results in a synthetic image that retains the appearance of the texture, but is otherwise different. This procedure generates synthetic textures that seem to capture not only fine-scale information, but also more broad scale structure. The syntheses look quite convincing as samples from the original texture, on par with state of the art nonparametric resampling models and typically more convincing than Portilla-Simoncelli syntheses.

The success of the CNNs in capturing texture appearance is perhaps not entirely surprising given their original purpose. Specifically, they are trained to do invariant object recognition. This requires extracting features that are moderately invariant to transformations, such as translation, rotation, or rescaling, as object identity is largely independent of such properties. What is more important for object identity, arguably, is the co-occurrence between particular features; one can imagine that the image of a leaf is defined by a particular distribution of co-located hues, spatial frequencies, and oriented edges. Therefore, one should expect some part of the CNN to “learn” the co-activation of such features as a cue for object identity. The early convolution and pooling layers in particular are most likely to compute these features, since they measure activations over the whole image at various scales. By only enforcing correlations between these filtering and pooling activations, (Gatys et al., 2015)’s model effectively captures the defining stationarity of visual textures. Thus, the CNN can be thought of as a learned version of the handpicked texture statistics in parametric models (Heeger & Bergen, 1995; Portilla & Simoncelli, 2000). The main difference is that the Portilla-Simoncelli architecture consists of a set of intelligently handpicked features, chosen based on their relevance in natural image statistics and optimized to produce good texture syntheses. CNNs, on the other hand, learn the features (here in order to do object recognition); only the broad scale architecture is fixed. This line of reasoning is explicated in more detail in the original study (Gatys et al., 2015).

This new technique raises three important and related questions. First, how similar is the architecture of the CNN texture model to Portilla-Simoncelli (and therefore the Texture Tiling Model)? Both measure the outputs of convolution of oriented filters with the input image (although CNN filters
are not explicitly oriented, the training procedure does often produce Gabor-like filters), both analyze
the image at several scales, and both retain correlations between these measurements. Beyond that,
however, the models are quite different. Secondly, CNNs consist of many layers, each of which is
convolved with the image (or, for later layers, with pooled outputs of the preceding layer). On the other
hand, Portilla-Simoncelli can be approximated as a “two layer” model: the first “layer” is the multi-scale
convolution, and the next “layer” consists of the correlations computed on the convolutions of the
previous layer (akin to the computation of feature co-occurrences done by CNN pooling layers). There
are also more nuanced differences: CNNs do not explicitly apply filters over large image regions (in fact,
the filters are quite small), while Portilla-Simoncelli uses a multi-scale steerable pyramid. These multi-
scale features are vital for Portilla-Simoncelli synthesis. The hierarchical CNN, however, only gets large-
scale information through the filters learned in the later convolutional layers, which operate over the
spatially pooled responses of earlier layers. Would a CNN perform better at object recognition and
texture synthesis if it employed multi-scale features? A few studies suggest this might be the case
(Gong, Wang, Guo, & Lazebnik, 2014; He, Zhang, Ren, & Sun, 2014; Sermanet & LeCun, 2011; Yoo, Park,
Lee, & Kweon, n.d.), although there is clearly more room for work. Furthermore, Portilla-Simoncelli
explicitly measures summary statistics of the pixel histogram (1st - 4th moments) and local
autocorrelation. The CNN texture model (Gatys et al., 2015) only uses filter output correlations; perhaps
the large number of layers and neurons implicitly encodes such statistics. An important question is
whether the pooling done by TTM/Portilla-Simoncelli is done at the proper level of feature complexity, a
question that might be answered using CNNs (discussed in the next paragraph).

Second, how well does the CNN texture model mesh with intuitions about the biology and
behavior associated with peripheral vision? On a basic level, the complex features and correlations in
the CNN texture model are at least somewhat similar in flavor to Portilla-Simoncelli statistics, and might
thus be relevant. One important aspect of Portilla-Simoncelli is that information is pooled over the
whole image, and precise spatial details are lost. The CNN texture model also loses absolute position
information by only retaining correlations, but the loss of information is more modest; retaining
correlations from all convolutional layers preserves a large amount of relative position information.
Perhaps the level of features at which pooling occurs in the periphery falls at some point along the CNN
architecture; an interesting project would be to examine how well different subsets of the CNN model
predict peripheral vision (as in the development of (Portilla & Simoncelli, 2000) and “lesioning” studies
of (Balas, 2006)). Clearly, using the correlations from all convolutional layers in the CNN, as described by
(Gatys et al., 2015), provides too much information to match the losses of peripheral vision; observers
are likely to over-perform with foveally viewed synthetic images when all CNN correlations enforced as
compared to peripherally viewed original images. Using only the lowest layer results in syntheses that
look like spectrally-matched noise (Gatys et al., 2015), which would be too impoverished to account for
peripheral vision.

Furthermore, a key aspect of peripheral vision is that acuity decreases while pooling regions
grow in size with increasing eccentricity. CNNs do not have a resolution limit beyond the mild restriction
of using small filters, and do not pool information in an eccentricity-dependent way. Perhaps this could
be artificially built-in; it would be in contrast with the original goal of the CNN, however, of maximizing
object recognition performance. It might, however, help to decrease the amount of computation or size
of networks required to achieve good performance.
To answer some of the questions about modeling peripheral vision, we can take a similar approach as in Chapter 2. We could examine the textures generated from one or more layers of a CNN model, to see how well they can support a range of perceptual tasks. If this new “CNN-inspired” TTM model also does well at predicting peripheral vision, then that suggests that maybe the features that are important for object recognition are the ones employed by peripheral vision.

On the other hand, our work on peripheral vision raises questions about the next generation of object recognition models. If an important goal of vision is object recognition, one might imagine that the visual system should extract features that are invariant to the various translations caused by eye movements. It is possible that restricting a CNN’s representational complexity and giving it low-cost eye movements would lead it to learn to do object recognition with structures that are similar to TTM, such as foveation and pooling regions that grow with eccentricity. This raises an even deeper question: Why do we have foveated vision, where detailed information decreases with eccentricity, at all? One possibility is that foveation is a kind of optimal solution to vision (including tasks like object recognition) under limited neural resources and eye movements. If so, how does this lead to the particular kinds of features and statistics found to support peripheral vision? Perhaps the features are drawn both from statistics that efficiently describe our visual world, like the ones used by Portilla-Simoncelli, and from features that support ecologically important visual tasks, like invariant object recognition.
Chapter 6

References


http://doi.org/10.1023/A:1026553619983


74


Wiecek, E., Pasquale, L. R., Fiser, J., Dakin, S., & Bex, P. J. (2012). Effects of peripheral visual field loss on
http://doi.org/10.3389/fpsyg.2012.00472


