

UC Berkeley

UC Berkeley Electronic Theses and Dissertations

Title

Invariance in Human Visual Perception

Permalink

<https://escholarship.org/uc/item/46d969f8>

Author

Nandakumar, Chetan

Publication Date

2011

Peer reviewed|Thesis/dissertation

Invariance in Human Visual Perception

by

Chetan Nandakumar

A dissertation submitted for the degree of

Doctor of Philosophy

in

Vision Science

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:

Professor Jitendra Malik, Chair

Professor Bruno A. Olshausen

Professor Stephen E. Palmer

Spring 2011

Invariance in Human Visual Perception

Copyright © 2011

by

Chetan Nandakumar

Abstract

Invariance in Human Visual Perception

by

Chetan Nandakumar

Doctor of Philosophy in Vision Science

University of California, Berkeley

Professor Jitendra Malik, Chair

This dissertation explores invariance in human visual perception via three unique studies. In the first two studies we probe the visual system to see how robust it is to impoverished stimuli. These investigations not only offer limits on perceptual abilities, but also offer key insights into the mechanisms underlying vision.

The first study explores rapid category detection, as discovered by Thorpe, Fize, and Marlot (1996). This study demonstrated that the human visual system can detect object categories in natural images in as little as 150 ms. To gain insight into this phenomenon and to determine its relevance to naturally occurring conditions, we degrade the stimulus set along a wide variety of image dimensions and investigate the effects on perception. We discover that rapid category detection in humans is quite robust to naturally occurring degradations and is mediated by a non-linear interaction of visual features.

This investigation into degradation is followed by our second study where we explore the limits of 3D shape perception. The shape-from-texture and shape-from-shading perspectives would motivate that 3-D perception vanishes once low-level cues are disrupted. Is this the case in human vision? Or can top-down influences salvage the percept? In this study, we explore this question by employing a gauge-figure paradigm similar to that used by Koenderink et al (1992). Subjects were presented degraded natural images and instructed to make local assessments of slant and tilt at various locations thereby quantifying their internal 3-D percept. Analysis of subjects' responses reveals recognition to be a significant influence thereby allowing subjects to perceive 3-D shape at high levels of degradation. Specifically, we identify the medium-blur condition, images approximately 32 pixels on a side, to be the limit for accurate 3-D shape perception. In addition, we find that degradation affects the perceived slant of point-estimates making images look flatter as degradation increases.

These 2 studies, in conjunction with previous work, point to 32-pixel color images as a rough threshold for a rich perceptual experience. The first study demonstrates that rapid recognition breaks down at around this point, and the second shows that it is also the limit

to reliably perceive 3-D shape. Presumably many perceptual abilities are tied together at this level - shape, recognition, etc, so that when one percept is lost, other percepts break as well.

In the third study, we explore how invariant properties of the natural world drive perceptual coding mechanisms in the brain. Specifically, we explore how the statistics of object regions in natural images motivate a sensitivity to hue by the perceptual system. To investigate this question, we compute the coding advantage of using hue angle to encode color inside real-world object regions. For this analysis, we use natural image datasets which provide pre-segmented object regions and surfaces.

Acknowledgements

First off, I'd like to thank my parents for their undying support, love and encouragement throughout my life. They instilled in me a deep desire for learning and served as amazing role models, both as people and as researchers.

I thank Jitendra Malik, my advisor, for guiding my degree and training me as a researcher. I've always been amazed at his ability to tune in to the essence of a problem and translate this understanding into ground-breaking research.

I'd like to thank Bruno Olshausen and the Redwood Center for serving as my second academic home. The culture and conversation was truly an example of science at its best and I'll always cherish the experience of spending time in this community.

The friendship and support of my fellow graduate students has also been incredibly valuable. This group includes Jon Barron, Subhransu Maji, David Bressler, Chunhui Gu, Pablo Arbalaez, Alex Berg, Ashley Eden, Pierre Garrigue, Charles Cadieu, Jascha Sohl-Dickstein, and Jimmy Wang. I thank them for engaging in lively discussions about science, engineering, society and life.

I'm in deep gratitude to new extended family who have supported me in countless ways over the past year to finish this degree – Lili Lambert, Adil Kassam, Mikki Willis, Nadia Salamanca, Steve Ehlinger, Paul Sparks, Kevin Carpenter, Darakshan Farber, Kevin Jones, Jeska Summerfield, Autumn Graca, James Graca, Jay Pullano, Laura Fredrickson, Gary Tomchuk, Aaron Farber, and Adam Bailey.

Contents

1	Introduction	1
2	Understanding Rapid Category Detection via Multiply Degraded Images	4
2.1	Introduction	4
2.2	Methods	5
2.3	Results	10
2.4	Discussion	10
3	How little do we need for 3-D shape perception?	17
3.1	Introduction	17
3.2	Methods	20
3.3	Results	24
3.4	Discussion	27
4	Natural Image Statistics underlying Color Perception	37
4.1	Introduction	37
4.2	Methods	38
4.3	Discussion	43
5	Contributions	46

Chapter 1

Introduction

Visual perception is the remarkable ability of the human brain to translate the matrix of light meeting the eye into a rich percept representative of the external world. One of the perceptual system's key challenges is remaining robust to an ever-changing and dynamic environment. Objects and scenes routinely undergo a host of transformations that radically change the two-dimensional projection of light entering the retina. Example transformations include changes to pose, viewpoint, lighting, and scale, and the visual system must remain invariant to these shifts in order to sensibly interpret the physical world.

In the scientific community, there has been a variety of studies exploring invariance in visual perception. These studies have investigated what the visual system remains invariant to and also what mechanisms or representational systems underly this ability. Studies such as Tarr and Pinker (1989), Rock, Di Vita, and Barbeito (1981), and Yin (1969) investigate rotation invariance in humans. In physiology, Brainard (2004) investigates color constancy and Ito, Tamura, Fujita, and Tanaka (1995) investigate size and position invariance. Many computational models have also been proposed to tackle different invariances. Examples include Olshausen, Anderson, and Van Essen (1993) who propose a computational model that is scale and shift invariant, and Lowe (1999) who proposes the SIFT descriptor, a local descriptor robust to affine transformations and lighting variation.

This dissertation explores the topic of invariance via three unique studies. In the first two studies, we explore the perceptual effects of degradation. The term degradation refers to the reduction of information content in a visual stimulus, and we explore the degree to which visual abilities remain invariant in the face of this transformation. By measuring the perceptual effects of degraded stimuli, we not only probe the limits of human perception, but we also gain key insights into the mechanisms driving it. In the third section, we explore how invariant properties of object regions drive perceptual coding mechanisms of color.

1.0.1 Rapid Category Detection

The first study explores the topic of rapid category detection in the context of degraded stimuli. Rapid category detection is the ability of the brain to detect the presence of a object category at very rapid time-scales. This phenomenon was first discovered by Thorpe et al. (1996) who demonstrate that subjects are capable of detecting the presence of an animal in a natural image in 150ms.

Yet the question remains as to how this phenomenon bears on real-world vision since natural stimuli is routinely degraded in various ways. Objects in the periphery and at a distance are degraded spatially and objects that fall under shadow are degraded with regards to luminance depth. How does rapid category detection cope under these transformations? Previous studies of degraded stimuli have explored certain select degradations. For instance, Mooney (1957) used 2-tone face images as stimuli to study visual development and demonstrated developmental differences in face-recognition ability. Harmon and Julesz (1973) demonstrated image recognizability in a severely degraded image of Abraham Lincoln, and Bachmann (1991) took this result further by demonstrating that face recognition is tolerant of extreme degradation. Exploring imagery beyond just faces, Torralba (2009) demonstrated the recognizability of tiny natural images.

In this study, we investigate how rapid category detection fares under a wide palette of different degradations in the image plane. This palette also includes multiple degradations where a given image is degraded along more than one dimension.

1.0.2 3D shape perception

The second study explores degradation's influence on 3D shape perception. 3D shape perception has been extensively studied over the years. Key work includes Mingolla and Todd (1986) where subjects made point-wise estimates of local slant and tilt and Ramachandran (1988), who demonstrated that the human visual system has a top-down lighting bias in its perception of 3-D shape. Koenderink et al. (1992) made a significant contribution by pioneering the gauge-figure paradigm - an intuitive interface to quantitatively measure perceived estimates of slant and tilt.

A few studies have also explored 3D shape perception in degraded environments. However, previous work in this space has focused on artificial stimuli. For instance, Christou et al. (1996) used a gauge-figure paradigm to measure how subjects perceive simple computer-generated 3-D objects in different renderings: grayscale, silhouette, and line-drawing. And Cole et al. (2009) demonstrated that line drawings of computer-rendered objects are sufficient for accurate 3-D perceptions.

How do these findings cope in the context of naturalistic stimuli? In addition, what mechanisms mediate this ability? Is it purely a bottom-up phenomenon or are top-down influences also important? This study explores these questions.

1.0.3 Perceptual Coding of Color

In the third study, we explore how invariant properties of the natural world drive perceptual coding mechanisms in the brain. When light hits the retina, it undergoes a series of transformations. After activating long, medium and short wavelength photoreceptors in the retina, the signal travels up the optic nerve to the LGN. Due to the high redundancy of activations by the photoreceptors, biology has learned to decorrelate this input for purposes of efficient coding. Principal components analysis is a mathematical transformation that closely approximates this decorrelation of photoreceptor activation in the visual system (Buchsbaum and Gottschalk 1983, Zaidi 1997, Ruderman et al., 1998). Further studies have found sensitivity to dimensions off the 2 color-opponent axes found in the LGN. Furthermore, Kiper et al. (1997) finds that most color sensitive cells in V2 are responsive to the angle between the stimulus color and the neuron's preferred color.

If decorrelation motivates the shift from coding color in terms of photoreceptor activation to color-opponent axes, what motivates the shift evidenced in V1 and V2? In this study, we explore the hypothesis that invariant properties of real-world object regions offer coding advantages that motivate this shift.

Chapter 2

Understanding Rapid Category Detection via Multiply Degraded Images

Joint work with: Professor Jitendra Malik

2.1 Introduction

Thorpe, Fize, and Marlot's (1996) study demonstrated the robustness of the visual system at extremely short time scales. Using an ERP setup, the authors flashed images of natural scenes, where only a portion of the image set contained animals. Subjects were instructed to make a go/no-go response of whether an animal was present in the image. Analyzing the ERP signals between animal-present and animal-absent trials, a significant difference was found 150 ms after stimulus onset. Although there is controversy around this finding (Johnson and Olshausen, 2003; Van Rullen and Thorpe, 2001), it is clear that category detection can occur at very rapid time scales.

The question motivating our study is the following: Can one do rapid category detection when the image is degraded? By the term degradation, we are referring to various ways of reducing information content in the image. Over the years, selected studies have investigated the perceptual effects of degrading images along a singular image dimension. Harmon and Julesz (1973) and Bachmann (1991) demonstrate that with respect to spatial resolution, only 18x18 pixels per face are sufficient for robust recognition, and these findings have been extended to the domains of faces, objects, and scenes by Torralba, Fergus, and Freeman (2008) and Torralba and Sinha (2001). Along the dimension of luminance depth, Mooney faces are a classic demonstration of visual processing working in extreme cases of luminance depth degradation (Mooney, 1957). Robustness to degradation falls under the larger umbrella of invariances. Invariance is a percept's tolerance to different transformations such as scaling, lighting, or rotation. There's been extensive study of invariance in psychology

and computer science. Studies such as Tarr and Pinker (1989), Rock, Di Vita, and Barbeito (1981), and Yin (1969) investigate rotation invariance in humans. In physiology, Brainard (2004) investigates color constancy and Ito, Tamura, Fujita, and Tanaka (1995) investigates size and position invariance. Many computational models have also been proposed to tackle different invariances such as those of Olshausen, Anderson, and Van Essen (1993), who propose a model that is scale and shift invariant, and those of Lowe (1999), who proposes the SIFT descriptor, a local descriptor robust to affine transformations and lighting variation.

Why do we think that rapid category detection maybe adversely affected by image degradation? One line of reasoning stems from the SpikeNet computational model proposed by Thorpe (2002). Here the author attempts to build a computational model capable of mirroring the rapid categorization effect originally demonstrated by Thorpe et al. (1996). This model centers on rank order coding - average local contrast determines the timing of neuronal firing. This perspective suggests that degraded images, such as blurred images, would have a delayed response since average local contrast is reduced. This link between reduced contrast and a delayed signal is also echoed in the explanation of the Pulfrich Pendulum phenomenon. A second line of reasoning, which suggests a delay for degraded stimuli, is that impoverished stimuli rely on top-down influences as argued by Cavanagh (1991). This might suggest that feedback mechanisms, and a corresponding increase in processing time would be required for degraded images.

In this study, we have crafted an experiment combining rapid stimulus presentation with degraded stimuli. We use a stimulus set of degraded images where each condition is degraded along a singular image dimension or pair of image dimensions. Through the use of degraded images, we are able to remove image information in a principled manner and by measuring the corresponding effects on perception, we also gain key insights into the mechanisms driving rapid category detection.

With the findings from above, we conducted another experiment to understand how well the latest object recognition algorithms compare to human recognition abilities. It has been argued (Serre, Oliva, and Poggio, 2007) that the current state of the art in computer vision can model the rapid category effect from Thorpe et al. (1996). By running such models with degraded images as input, we further test the robustness of this claim.

2.2 Methods

2.2.1 Experiment 1

2.2.1.1 Subjects

Twenty-one volunteers with normal or corrected-to-normal vision performed a 2AFC visual discrimination task. The experimental procedures were approved by the UC Berkeley ethical committee.

2.2.1.2 Experimental Setup

The experimental setup is modeled after Kirchner and Thorpe’s (2006) 2AFC animal/non-animal detection task. Subjects were seated in a dimly lit room with their heads mounted in a chin rest. Gray scale images were presented to subjects on a CRT monitor placed 60 cm away from the subject. Stimuli were centered laterally 10 degrees from fixation and subtended approximately 6 degrees of visual angle. We use an Arrington eye tracker to track subject’s eye movements; the accompanying software package presented the stimuli and appropriately created log files of the subjects’ responses. These log files were later analyzed using MATLAB scripts. The stimuli were selected from the COREL collection of natural images. The stimulus set was divided into an equal number of targets and distracters. The target images each contained an animal at an arbitrary pose and location, and the distracter images were natural scenes such as landscapes or forest scenes and did not contain any animals.

2.2.1.3 Protocol

In each trial, subjects were flashed a fixation cross for 2.5 s, followed by a blank gray screen for 200 ms, and then flashed a pair of images in the right and left hemi-fields for a duration of 30 ms. Each pair contained one target and one distracter image. The target location was equiprobable in the left and right hemi-fields. After stimulus presentation, two fixation crosses were presented for 1 s at ± 6 degrees. Subjects were directed to make an eye movement to the cross on the same side as the animal and were measured for performance in different conditions. Each subject was presented with the Full or control condition and then a subset of the variable conditions. Please see Figure 2.1 for a visualization of the experimental protocol.

2.2.1.4 Stimulus Conditions

The first set of conditions contains images degraded to different degrees along one of the following dimensions: spatial resolution, luminance depth, inversion, and reverse contrast. To understand how these different visual dimensions combine to influence perception, the above degradations are paired to create multiply degraded images.

Each dimension of degradation was chosen with a specific intention. Spatial resolution degradation is a naturally occurring degradation with objects at a distance and objects in the periphery, and luminance depth degradation occurs in low-light conditions. The inversion and reverse contrast conditions, although not naturally occurring conditions, offer insight into visual cognition. In the context of face recognition, both transformations have severely impacted recognition performance and thereby offered insight into the underlying representations (Liu, Collin, Burton, and Chaurdhuri, 1999; Yin, 1969). Inversion probes the role of global features as global features are warped in an inverted image while local features

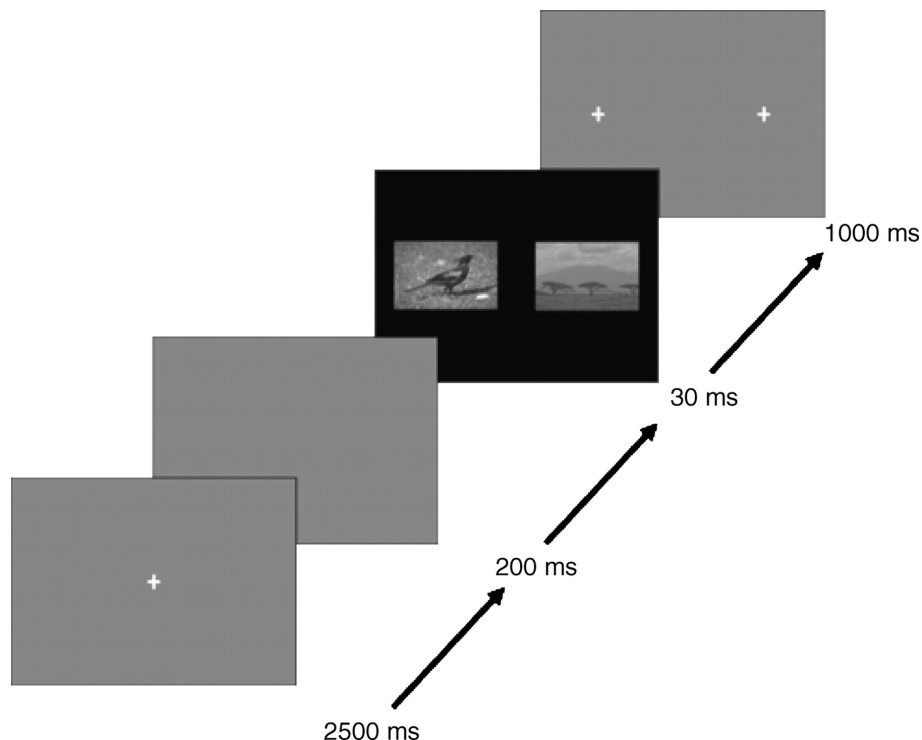


Figure 2.1: Visualization of the stimulus presentation protocol. On each trial, a fixation cross is displayed, followed by a blank gray screen, then the target/distracter pair, and finally a screen with two fixation crosses. The subject is instructed to saccade to the fixation cross on the same side as the target.

(eg. texture) are not affected. The reverse contrast condition probes the importance of edge polarity since each edge switches its polarity under this transformation.

Each condition consists of 50 trials, and between four and seven subjects participated in each condition. Example images from all conditions are shown in Figure 2.2. The baseline condition, labeled Full, contains 8-bit gray scale images of animals and natural scenes.

Along the dimension of spatial resolution, there are three degrees of degradation. The first, labeled Blur100, is created by resizing the original image, sized at 512 x 768 pixels, to 66 x 100 pixels. This miniaturized image is then resized back to its original size. The Blur50 and Blur25 conditions are created in an analogous manner. This manner of blurring allows us to compute an upper limit of the information contained after degradation. For instance, in the Blur100 condition, we have $66 \times 100 = 6600$ dimensions where the value of each dimension is specified by 8 bits.

Along the dimension of luminance depth, there are two degrees of degradation. The

first, labeled 4Tone, is created by discretizing the original 8-bit image into a 2-bit image. In other words, the original image, containing 256 levels of gray, is now reduced to only four luminance levels (black, white, and two levels of gray). Likewise, in the 2Tone condition, each image pixel is either black or white. This first set of conditions also includes the Inverse condition where each image is vertically inverted and Reverse where the contrast of each image is reversed.

In the second set of conditions, the above degradations are combined in pairs. In the first such condition, spatial resolution degradation is paired with luminance depth degradation: The Blur100 condition combines with the 4Tone condition to create the Blur100 + 4Tone condition, and Blur50 combines with 4Tone to create the Blur50 + 4Tone condition. In the next set of conditions, Inversion is paired separately with Blur100 and 4Tone to create the Inversion + Blur100 and Inversion + 4Tone. Lastly, Reverse is paired in a similar manner to create the Reverse + Blur100 and Reverse + 4Tone conditions.

2.2.1.5 Response recording and detection

Subjects' eye movements are tracked using an Arrington Eye Tracker. This unit consists of a single infrared camera setup with a 30-Hz sampling speed. From the eye tracking signal, the first movement to the left or right side of the screen is extracted and taken to be the binary response for that given trial.

2.2.1.6 Statistical Analysis

A repeated measures ANOVA was conducted to determine significance of a given condition against the baseline condition, Full. This test has one independent factor and one dependent factor. The independent factor is the experimental condition, which has 2 levels - the control condition and the experimental condition. This factor served as the repeated measure in the experiment since a given subject participated in both the control and experimental conditions. The dependent factor in this analysis is the accuracy of the subject in the task. The number of subjects participating in each condition varied between 4 and 7.

2.2.2 Experiment 2

To understand how object recognition algorithms cope with degraded images, we test two algorithms: the spatial pyramid kernel (Lazebnik, Schmid, and Ponce, 2006) and the GIST descriptor (Oliva and Torralba, 2006). The spatial pyramid kernel serves as a leading approach to object recognition as it has performed well on Caltech 101, one of the key benchmarks in the computer vision community. It works by pooling orientation features across different spatial scales into a feature vector, and this feature vector is then fed through a support vector machine classifier. The classifier uses the training set to find a separating

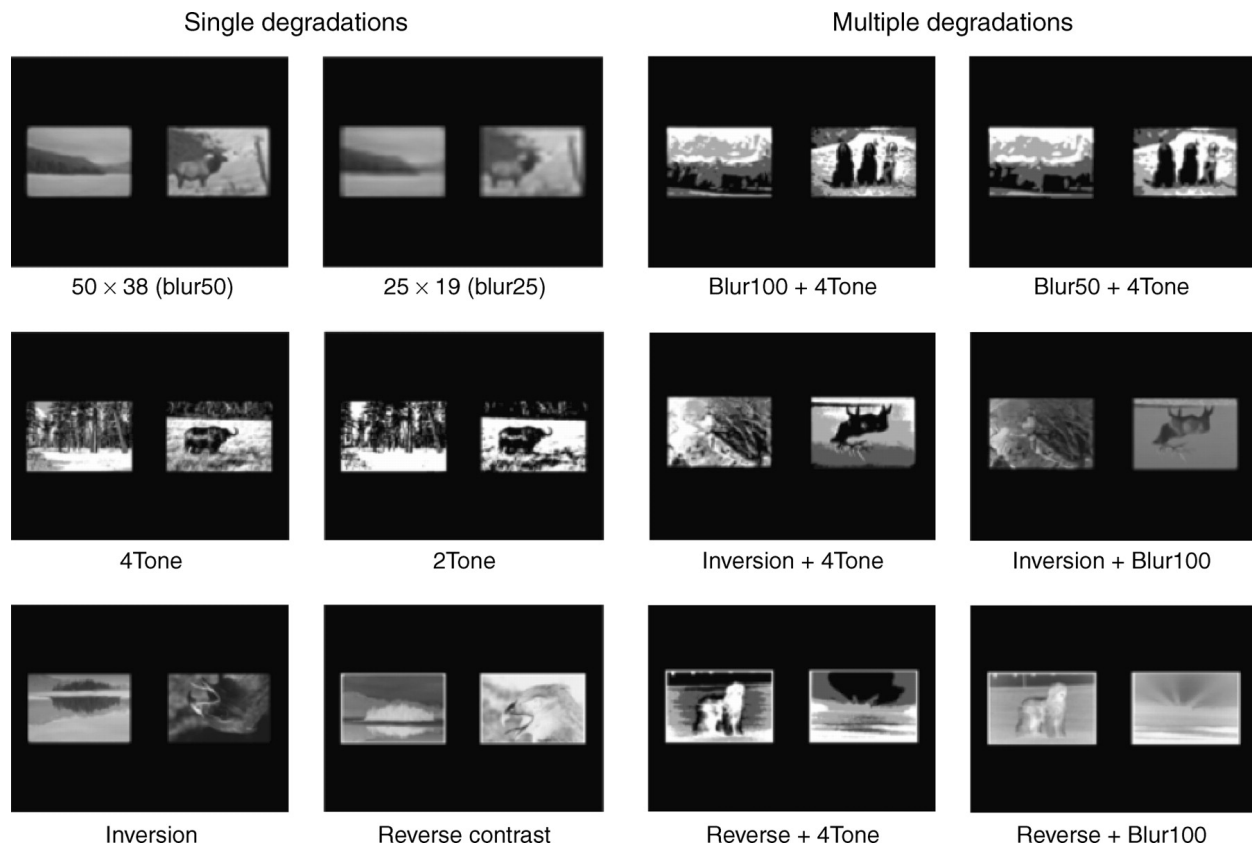


Figure 2.2: Examples of stimuli from select conditions. The conditions displayed in the left column contain images degraded along a singular condition, and the conditions displayed in the right column contain images degraded along pairs of conditions.

hyperplane between the positive and the negative examples in the feature space. A given test point is classified in reference to this hyperplane - the side it falls on is the class and the distance from the hyperplane is the confidence.

While the spatial pyramid kernel was designed for object recognition, the GIST descriptor was designed to capture the gestalt of an image for scene classification. The GIST descriptor was used to label an arbitrary scene as a beach scene, office scene, etc. The GIST descriptor also uses orientation energy but pools it in a different way. It tiles up the image into blocks and in each block, it sums the energy in different orientation and spatial frequency bands. This feature vector is then fed through a support vector machine classifier. In this experiment, the GIST descriptor was tailored to only use the low spatial frequencies of the image.

We trained each classifier with 200 animal and 200 non-animal images. These images

were drawn from the same set used in the human experiments above and each was tagged with its respective category. We tested the performance of the classifiers under different conditions, and to create results comparable to the human results from Experiment 1, we simulated trials where each trial contained a randomly paired animal and non-animal image. A correct trial is one where the confidence value assigned to the animal image is greater than that of the non-animal image.

As a baseline, we tested the classifier with 200 full resolution (the Full condition) images, where 100 were of animal images and 100 were non-animal ones. We then tested the classifiers in all of the degraded conditions.

2.3 Results

For Experiment 1, the histogram of reaction times across all conditions is shown in Figure 2.3. Mean reaction time is at 275 ms. Mean accuracy results along with error bars and p-value are displayed for each condition in Figure 2.4, where chance performance is at 50% accuracy. Significance ($p < .05$) from the baseline condition, labeled Full, was found in the following conditions: Blur25, Blur100 + 4Tone, Blur50 + 4Tone, and Inversion + 4Tone. Further analysis looking for a link between reaction time and accuracy is displayed in Figure 2.5 and Figure 2.6. In Figure 2.5, we plot speed vs. accuracy across all subjects and conditions, and in Figure 2.6, we look at the relationship between speed and accuracy in each different condition. We do not find a relationship between speed and accuracy in either analysis.

For Experiment 2, we print the results of the spatial pyramid kernel and GIST descriptor for all conditions in Figure 2.7 and the graph the results in Figure 2.8. We observe that neither the GIST descriptor nor spatial pyramid kernel match human performance (Figures 2.5 and 2.6).

2.4 Discussion

We can draw several implications from this study for models of visual category recognition. Our experiments show that cues used for rapid visual categorization are robust under significant spatial and luminance depth degradation, and it thereby puts marked constraints on models of rapid visual category recognition. Unlike faces, we also see that this effect is tolerant to both inversion and reversal of edge polarity.

We also note the non-linear interaction among the different visual cues. A striking example is in the Blur100 and 4Tone conditions. Individually, neither is significantly different from baseline but by combining the degradations into the Blur100 + 4Tone condition, we find a significant drop in human performance. We observe the same phenomenon with the Inversion + 4Tone condition (Figure 2.4). This finding leads to conclusions in direct contrast

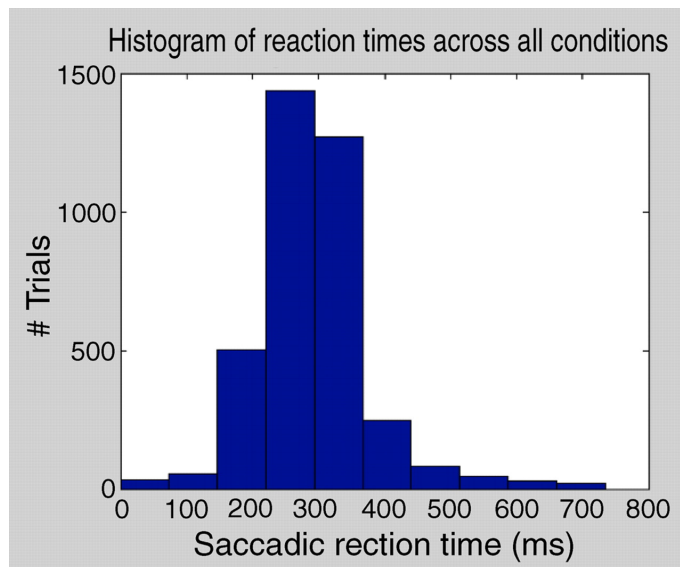


Figure 2.3: Histogram of saccadic reaction times. Data is pooled across all conditions and all subjects.

to claims made by Guyonneau, Kirchner, and Thorpe (2006). In this study, the authors rotated the target and distracter images to different degrees. Since this transformation did not impact performance, the authors concluded that rapid category detection could not depend on the global distribution of orientations within the image. Since we see significance with the Inversion + 4Tone condition, we conclude that both local features and global features are used in rapid category detection.

In addition, this study extends the initial work done by Thorpe et al. (1996) to naturally occurring conditions such as blurring and luminance depth degradation. In such degradations, the image information is not nearly as pristine as in Thorpe’s original experiment. This study demonstrates that rapid category recognition is not merely a laboratory effect but is a process that can be at play in real-world settings.

Lastly, many groups are constructing computational models of object recognition, such as Frome, Singer, Sha, and Malik (2007), Lazebnik et al. (2006), and Serre et al. (2007). Since human subjects are quite robust to degradations, we suggest that robustness to degraded input be a critical measure for all such models.

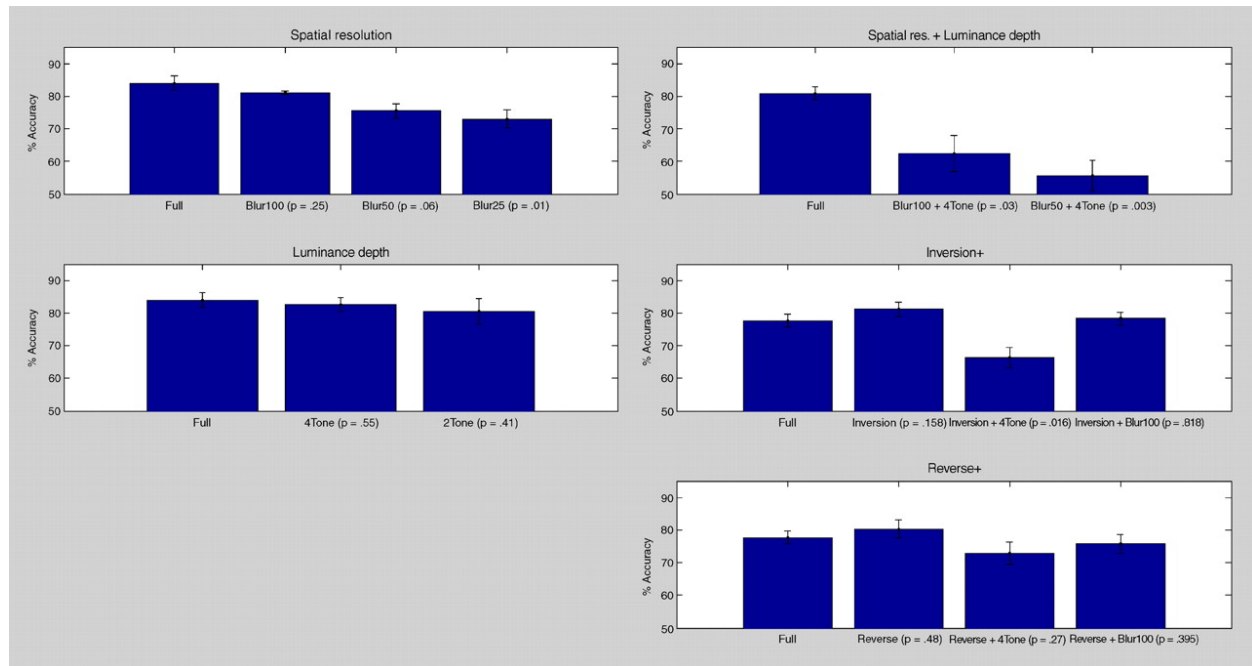


Figure 2.4: Mean accuracy results along with error bars displayed for each condition. Between 4 and 7 subjects were used in each condition, and a repeated measures ANOVA was used to determine significance. Significance from the baseline condition, labeled Full, was found in the following conditions: Blur25, Blur100 + 4Tone, Blur50 + 4Tone, and Inversion + 4Tone. The p-values are printed alongside in the condition labels in the plots.

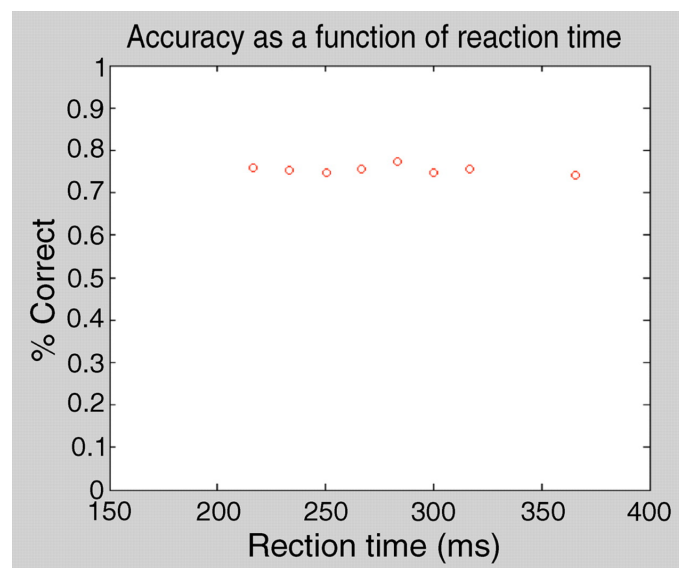


Figure 2.5: Plot of speed vs. accuracy across subjects and conditions. This plot uses adaptive binning so that each bin contains the same number of points.

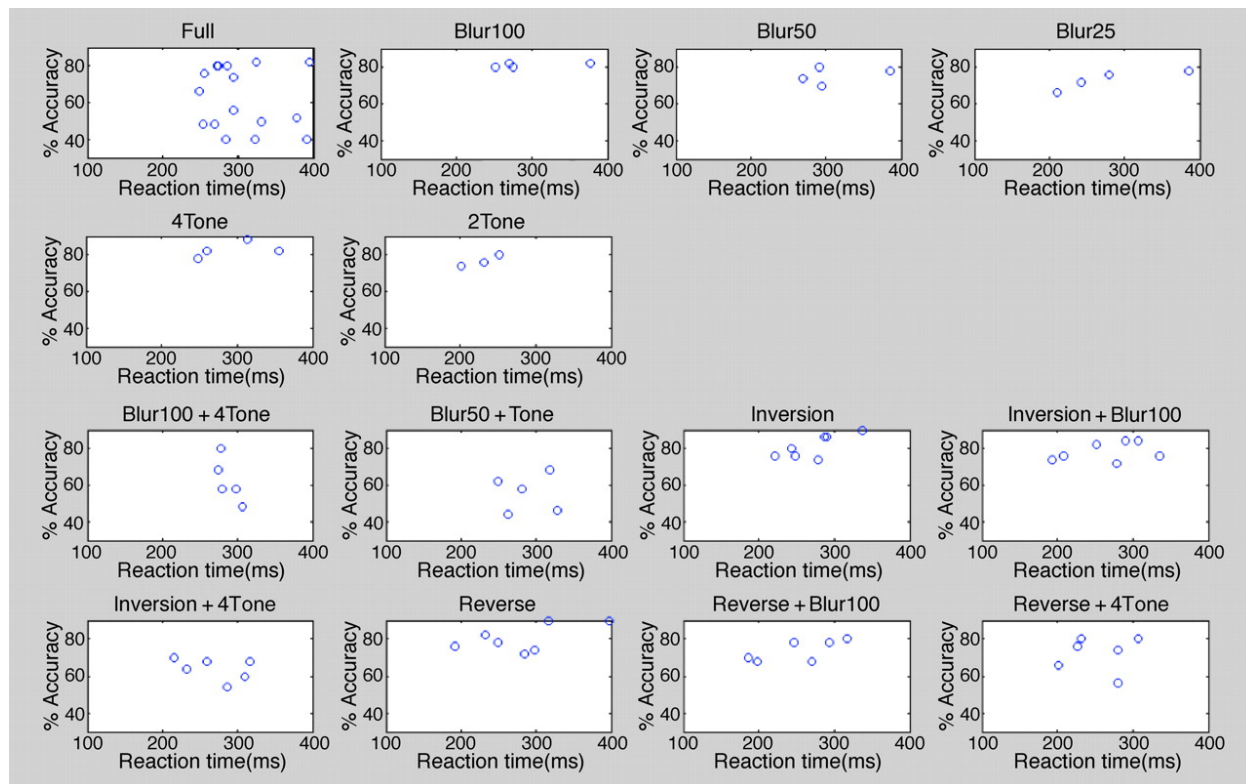


Figure 2.6: Plot of speed vs. accuracy on a per-condition basis. Each blue dot corresponds to the mean reaction time and accuracy for a given subject in the named condition. Please note that since Full was a common condition across all subjects, its plot above is a consolidated plot containing the responses from all subjects.

Condition	Human performance (% correct)	Spatial pyramid kernel (% correct)	GIST descriptor (% correct)
Full	84	73	84
Blur100	81	56	72
Blur50	75	59	55
Blur25	73	63	46
4Tone	82	51	75
2Tone	80	57	74
Blur100 + 4Tone	62.4	50	58
Blur50 + 4Tone	55.6	50	52
Inversion	81.1	57	65
Inversion + 4Tone	66.2	51	64
Inversion + Blur100	78.2	57	61
Reverse	80.2	60	76
Reverse + 4Tone	72.8	52	75
Reverse + Blur100	75.7	67	56

Figure 2.7: Performance of the spatial pyramid kernel in different conditions.

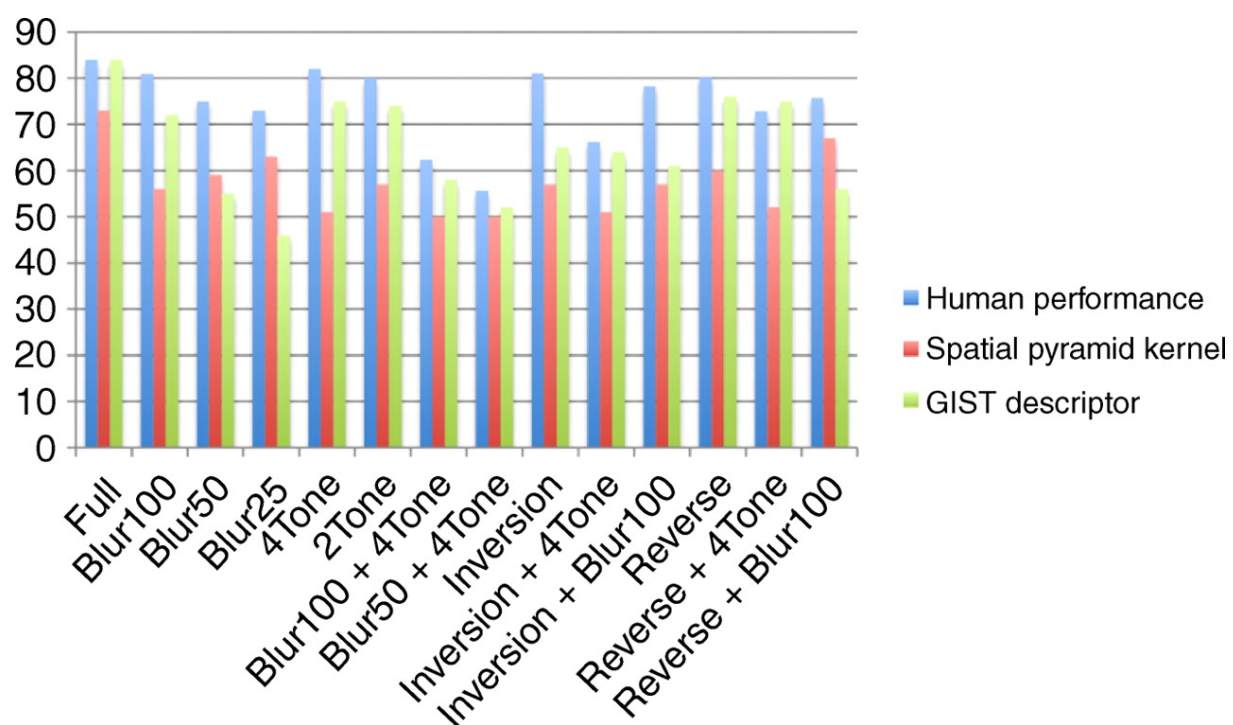


Figure 2.8: Plot of the data from Figure 2.7.

Chapter 3

How little do we need for 3-D shape perception?

Joint work with: Professor Antonio Torralba, Professor Jitendra Malik

3.1 Introduction

When monocularly looking at a natural image, humans have a vivid experience of 3-D shape. Koenderink et al. (1992) quantitatively measured this subjective percept through pioneering the gauge-figure paradigm, an intuitive interface to measure perceived estimates of local slant and tilt.

Yet what features mediate humans' ability to perceive 3-D shape from monocular images? Over the years there has been a variety of key studies offering greater insight into the features and mechanisms underlying 3-D shape perception. These studies cleave into two camps: the bottom-up perspective that focuses on texture, shading, and contour; and the top-down one that focuses on cues such as recognition and 3-D primitives.

With regards to contour, work dates back to Attneave (1954) who suggested that the regions of high curvature along an object's contour are perceptually more informative and Guzman (1969) who attempted to infer the 3-D structure of a block world scene from the edge information alone. Barrow and Tenenbaum (1981) argued that since shading is so variable in natural images, contour serves as a more robust cue. Certain works investigated the particulars of how contour translates into a 3-D interpretation. Richards et al. (1987) identified mathematical properties of smooth surfaces that drive perception of 3-D shape from 2-D contour; Koenderink (1984) illustrated how the curvature of a contour can be used to infer intrinsic properties of the surface geometry. For instance, the convexity of the contour corresponds to a convex patch of the surface and a concavity to a saddle-shaped patch. Mamassian and Landy (1998) explored Bayesian models for extracting 3-D volume from a

2-D contour representation. In the experimental literature, Mamassian et al. (1996) used the gauge-figure paradigm to demonstrate that humans can use contour alone to accurately perceive the shape of a selected artificial stimulus. Christou et al. (1996) investigated subjective perception of slant and tilt in a block world environment and demonstrated that shading and texture offer little information beyond contour alone in 3-D perception. Norman and Raines (2002) asked subjects to make ordinal depth judgments of a variety of stimuli (both natural and artificial) and demonstrated the power of the silhouette on 3-D perception. Panis et al. (2008) investigated what areas of a contour are most informative for object identification by presenting fragmented outlines of selected stimuli to subjects and probing them for their recognition abilities.

In shape from texture, work dates back to Gibson (1950) who introduced the idea of texture serving as a cue for 3-D shape. Cutting and Millard (1984), Todd and Akerstrom (1987), and Rosenholtz and Malik (1997) explored what image features mediate the texture gradient in human vision. Creating computational models of shape from texture, Purdy (1958) specified an algorithm to compute the slant of an isotropic texture on a planar surface and Garding (1992) extended these results to curved surfaces. Malik and Rosenholtz (1997) designed an algorithm that estimates the affine transforms between nearby image patches in order to estimate 3-D shape parameters, and Fleming et al. (2004) discovered how to recover 3-D shape from specular distortions. Todd et al. (2004) demonstrated the limitations of the current models of texture perception on human perception, and Todd and Thaler (2010) developed an algorithm to recover the shape of a continuously curved surface with anisotropic textures.

In the shape-from-shading approach, 3-D shape is attributed to shading variation in the image. In the human vision community, key work includes Mingolla and Todd (1986) where subjects made point-wise estimates of local slant and tilt and Ramachandran (1988), who demonstrated that the human visual system has a top-down lighting bias in its perception of 3-D shape. Kleffner and Ramachandran (1992), in a series of experiments, went further to conclude that the shape-from-shading computation happens early in the visual system, prior to perceptual grouping, motion perception, and vestibular correction. Gerardin et al. (2007) further proved that there is a left-lit bias in addition to a top-down one. Erens et al. (1993a, 1993b) explored how subjects categorize shaded quadric surfaces in different illumination conditions, and discovered that shading is a weak cue for 3-D shape categorization when used in isolation from other visual cues. Lastly, the computational approach to the shape-from-shading problem was pioneered by Horn and colleagues (Horn 1970; Ikeuchi and Horn 1981; Horn and Brooks 1989).

Top-down influences on 3-D shape have been much less studied. But noted studies include Biederman (1987), who posited that human vision breaks down objects into an assembly of 3-D primitives called geons. This work builds upon the work of generalized cylinders by Binford (1971) and Marr (1982). Other work in this camp is Peterson (1994) who demonstrated that humans employ object-level knowledge in figure-ground segregation. Works in Bayesian cue

combination, such as those of Mamassian and Landy (1998) and Hillis et al. (2002), offer models that can combine information from top-down and bottom-up sources to create a percept informed by both sources of information.

Yet, what happens when substantial information is missing from the stimulus? Since degradation is a routine transformation for objects in the periphery and objects at a distance, this question bears on real-world vision. The effects of degraded input on recognition have been investigated in several studies. Mooney (1957) used 2-tone face images as stimuli to study visual development and demonstrated developmental differences in face-recognition ability. Harmon and Julesz (1973) demonstrated image recognizability in a severely degraded image of Abraham Lincoln, and Bachmann (1991) took this result further by demonstrating that face recognition is tolerant of extreme degradation, both spatially and temporally. Exploring imagery beyond just faces, Torralba (2009) demonstrated the recognizability of tiny natural images, and Nandakumar and Malik (2009) demonstrated that rapid recognition of object category survives with degraded input.

In addition to recognition, a few studies have been devoted to 3-D shape perception in degraded environments. Christou et al. (1996) used a gauge-figure paradigm to measure how subjects perceive simple computer-generated 3-D objects in different renderings: grayscale, silhouette, and line-drawing. They found that the line-drawing version of the image is sufficient for a full-rendering and that shading adds little extra. Strengthening this conclusion is the study of Cole et al. (2009) who demonstrated that line drawings of computer-rendered objects are sufficient for accurate 3-D perceptions. With regard to spatial blurring, Jaa-Aro and Kjell Dahl (1997) investigated how image resolution relates to depth perception, Gerardin et al. (2007) demonstrated a preference for globally convex shapes in conditions of high blur, and Norman et al. (2010) investigated 3-D shape discrimination under retinal blur. They asked subjects to discriminate artificially created 3-D shapes under different degrees of retinal blur and found that 3-D shape discrimination is robust to degradation.

In this study, we are interested in studying 3-D shape perception in monocular natural images. Previous studies with natural images demonstrated that recognition copes incredibly well with degradation. Although a tremendous amount of information is missing, humans still have a rich visual experience - they can recognize objects, the surrounding context, and even do so rapidly. Yet how much more can be perceived with such impoverished input? Specifically, is 3-D shape perception also possible at low resolution? Without a bottom-up signal, can top-down influences salvage the percept? In two experiments, we degraded monocular natural images and measured the corresponding effects on 3-D shape perception.

3.2 Methods

3.2.1 Subjects

Twenty-one subjects with normal or corrected-to-normal vision performed the experimental task. The experimental procedures were approved by the Committee for Protection of Human Subjects (CPHS) at UC Berkeley.

3.2.2 Experimental setup

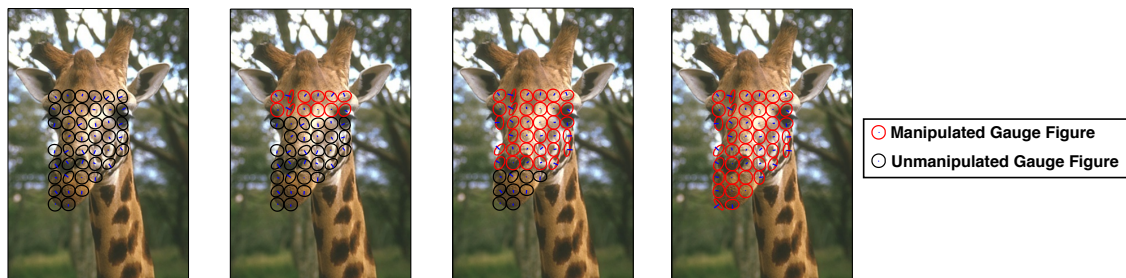
The experimental setup was modeled after Koenderink et al.'s (1992) gauge-figure task. Subjects were seated at a desk and performed the task at a computer with a flat-screen LCD monitor. Subjects sat approximately 60 cm from the monitor and made responses using a combination of the mouse and the keyboard. The stimuli were selected from the Berkeley Segmentation Dataset (Martin et al. 2001), which is, in turn, a subset of the COREL collection. The Berkeley Segmentation Dataset contains human-marked object boundaries for each image in the dataset.

3.2.3 Experimental procedure

On each trial, a natural image was presented and a portion of the image was overlaid with gauge figures. The region overlaid with gauge figures was pre-specified by a mask created for each image. This mask highlights a key object or key objects in the image and is created in advance by the experimenter. A gauge figure is a disk the subject can manipulate with the mouse and is intended to measure the subject's local perception of slant and tilt at the corresponding point in the image.

The collection of gauge-figure estimates offers an empirical measurement of the subject's perception of shape. The subject is instructed to manipulate the gauge figure with the mouse so that the surface normal of the gauge figure is aligned with the perceived surface normal at the corresponding point in the image. The interface is designed so that the movements of the mouse quickly and intuitively move the gauge figure in 3-D. As seen in Figure 3.1, all gauge figures are present simultaneously, thereby allowing subjects to revisit a given gauge figure, if so desired. The gauge figures are initially set to random orientations, and the subject is required to manipulate all gauge figures before he/she is permitted to move to the next image. Once a gauge figure is manipulated, it turns red, thereby giving a visual marker for which figures have been manipulated. In addition to manipulating the gauge figure, the subject is also instructed to describe the perceived content of the image in a textbox. This response allows us to look for a link between recognition and 3-D shape assessment.

Subjects were first given a demonstration of the task by the experimenter. They were shown how to use the interface to manipulate the gauge figures and then given the opportu-



[h!]

Figure 3.1: For each image presented, the subject responds with his/her assessment of 3-D structure. 3-D perception is measured via the gauge- figure paradigm of Koenderink et al. (1992) - each image is overlaid with gauge figures and the subject adjusts each gauge figure so that its surface normal is aligned with the perceived surface normal in the image. The gauge figures are randomly oriented at the outset and turn red after being manipulated by the user.

nity to get comfortable with the interface on a demonstration trial. If the subject had any questions, the experimenter was there to answer. Subjects were then directed to execute the experiment on their own.

In both experiments, we used a stimulus set of natural images spanning a wide variety of content: natural scenes, man-made objects, faces, animals, and so on. There was also a great variety in the pose and scale of the foreground object. This variety required that any effects found were general in nature and not tied to a specific domain.

3.2.4 Experiment 1

In the first experiment, there were 32 different images, each degraded to 4 different levels: 'full resolution' (256 pixels), 'small blur' (64 pixels), 'medium blur' (32 pixels), and 'large blur' (16 pixels). In this notation, the number of pixels corresponds to the miniaturized length of the shorter dimension, and the longer dimension is appropriately resized as per the aspect ratio of the image. After miniaturization, the image is then enlarged back to its original size. This process was executed with a Matlab script that miniaturized the original image to its reduced size using a binomial filter and then rescaled this thumbnail back to its original size with the same filter. Please see Figures 3.2 and 3.3 for example stimuli from the experiment.

To ensure that all stimuli got the same number of assessments and that no subject saw the same image in different conditions, the stimulus bank (32 images x 4 conditions =128

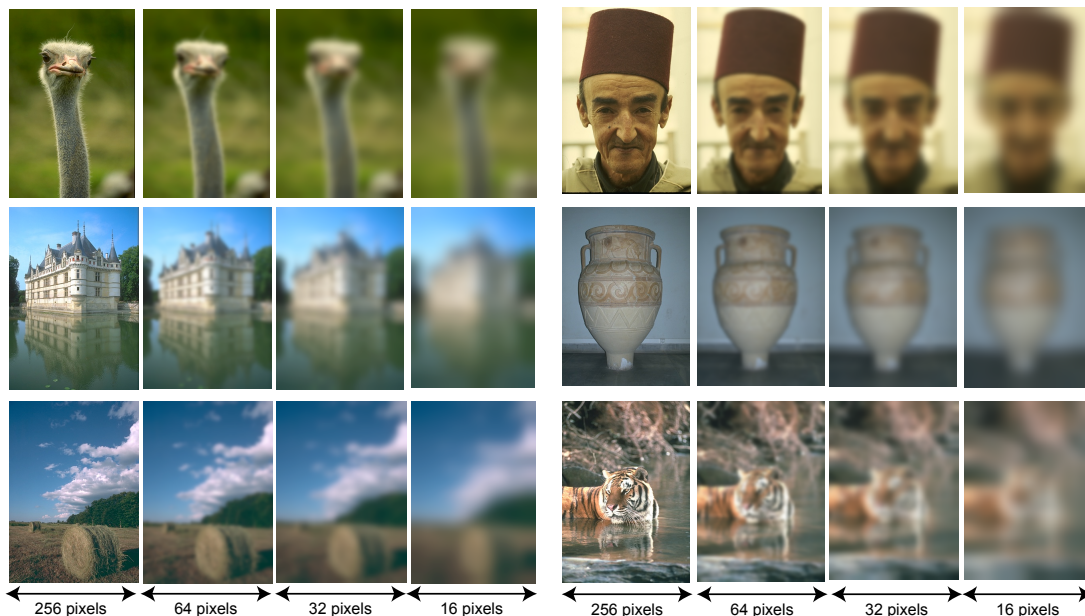


Figure 3.2: Example stimuli from Experiment 1. Each image is degraded to 4 different levels: full resolution (256 pixels), small blur (64 pixels), medium blur (32 pixels), and large blur (16 pixels).

stimuli) was split into 4 groups where each group contained an equal number of stimuli from each of the 4 levels of blur. Each subject was assigned to one of the groups so that he/she was presented only with that subset of the stimulus bank. This subset of stimuli was presented to the subject in a random order. In this experiment we used twelve subjects; therefore there were three subjective assessments for each stimulus.

3.2.5 Experiment 2

Experiment 2 used a similar gauge-figure setup as Experiment 1 (the only differences being that there was no text input field and the gauge figures were oriented to be frontoparallel at the outset). In Experiment 2, the key difference was in the stimulus set presented to the subjects. The stimulus bank consisted of 15 different images in 2 conditions: ‘undegraded’ and ‘region-averaged’. For the region-averaged condition, we used the human-marked region masks provided by the Berkeley Segmentation Dataset. Each mask contained a tiling of the image into closed regions. We averaged the color in each such region to create the stimuli in the region-averaged condition. Please see Figure 3.4. Please note that region-averaging is a

Chapter 3. How little do we need for 3-D shape perception?

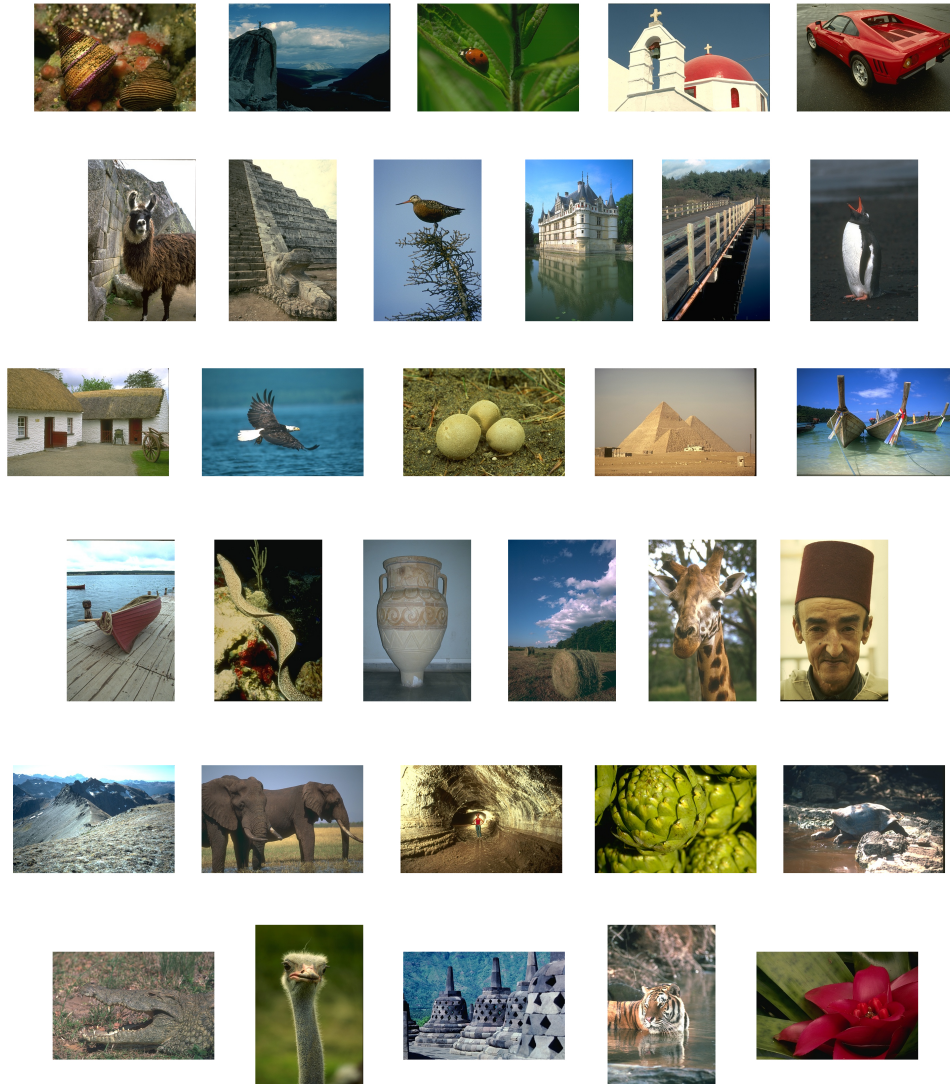


Figure 3.3: All stimuli from Experiment 1 in full resolution.

transformation not present in nature. Although it is a laboratory condition, it allows us to measure the relative contribution of contour independently of texture and shading.

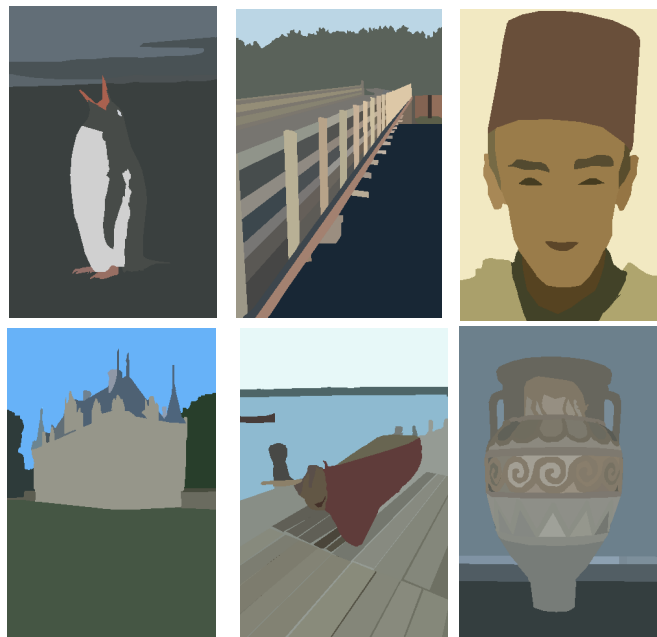


Figure 3.4: Example stimuli from Experiment 2. Each image is first segmented by a human observer into closed regions. Each region is then replaced with its average color creating this bank of ‘region-averaged’ images.

3.3 Results

3.3.1 Qualitative Analysis

Before entering into a quantitative analysis of the data, let us take a qualitative look at subjects’ responses. Please consult Figure 3.5 for a visualization of selected trials. We display the surface normals marked by the subject along with the corresponding 3-D volume. This 3-D volume is obtained by using a thin-plate approximation. The depth is written as a linear combination of thin-plate radial basis functions (to reduce the noise, we set the number of basis functions to be half the number of locations probed with the gauge figure). To obtain the coefficients of each basis function, we constrain the derivatives of the shape to match the average normals given by the observers. Please note that this 3-D volume has been created for visualization purposes only and is not used in any subsequent analysis. With the top three images, we find the general shape to remain robust across degradation and a significant flattening effect at high levels of blur. However, there are notable differences between images. For instance in the last image of the cave, the image is completely misinterpreted at the

highest level of degradation instead of being interpreted as a concave structure, the cave is interpreted as convex. In Figure 3.6 we showcase examples of two correctly and two incorrectly recognized trials in the medium-blur (32 pixels) condition to offer greater insight into how subjects responded to the stimuli.

In Figure 3.7 we showcase subjects' responses for selected images. Scatter plots are displayed for the three levels of degradation. In each plot, the horizontal axis is the average response across subjects at full resolution (256 pixels) and the vertical axis is the mean of all responses across subjects for a given level of blur. Clustering around the 45 degree diagonal signifies that the two quantities are closely aligned. With most images, we find that the error increases with increasing levels of degradation. However, there are exceptions to the rule - the image of the three balls and the image of the pyramids fare quite well even after substantial degradation. We presume this is because their shapes are simple and easily recognizable.

3.3.2 Flattening of blurred scenes

To quantitatively analyze the data, we built a measure of angular error. Angular error measures the quality of a given response and we analyzed how this error changes with increased levels of blur. There are four versions of each image in the stimulus bank; each version is the image at a different level of degradation. Each stimulus in the bank contains responses from three unique subjects.

For a given image, we created a best-case measure by computing a point-wise average of the assessments of two subjects for the image in the full-resolution (256 pixels) condition. In other words, this best-case measure is a vector where each entry is the average response by two subjects in the full-resolution (256 pixels) condition of the orientation for the corresponding gauge figure. Angular error takes a subject's assessment for a given image at any level of degradation and first computes a point-wise difference between the orientation estimate at each point and the corresponding orientation estimate in the best-case. It then averages these differences to obtain a single number that captures the angular error for that subject's response. Please note that in the full-resolution (256 pixels) condition we compute angular error between the best-case and the third subject's response to obtain a top-level measure of performance.

By averaging the angular error across all the stimuli at a particular level of blur, we arrive at the plot in Figure 3.8. To compute a worst-case measure of performance we looked at an array of randomly oriented normal vectors and measured the average angular offset between this array and the frontoparallel plane. We plot this as 'chance' in Figure 3.8. As expected, we see that subjects perform far better than chance and that angular error increases as degradation increases. Performing the same analysis with the region-averaged images in Experiment 2, we arrive at an angular error of 19 degrees, which is close to that of the full-resolution (256 pixels) condition of Experiment 1. However, what underlies this

increase in angular error? We analyzed the selective distributions of slant and tilt to uncover what was driving the increasing error. We display these distributions in Figure 3.9.

We investigated slant and tilt distributions for different conditions using a paired-comparison test. It works as follows. For each image, we compute the mean of the responses across subjects for the given condition and compare between conditions. A '1' is reported in the entries where 'condition A' is flatter than 'condition B'. A '0' is reported for the opposite. Performing this for each image in our set, we generate a binary vector. We then use the binomial distribution to find the probability of the number of 1s in the vector. Please consult Figure 3.12 for these values. The only statistically significant comparisons ($p < .05$) are between full resolution (256 pixels) and large blur (16 pixels), and between small blur (64 pixels) and large blur (16 pixels). Therefore, images look flatter at high levels of degradation. This makes sense, since image information is being taken away in degraded input, making the image more akin to a flat 2-D plane.

Repeating the same analysis with the tilt distributions, we did not find any significant difference between any pair of conditions as demonstrated in Figure 3.13. With Experiment 2, we found no significant difference between the distributions of slant or the distributions of tilt. Please see Figure 3.10.

3.3.3 The influence of recognition

To probe the role of recognition in subjects' ability to ascertain 3-D shape, we analyzed the text input entries of what subjects perceived in each image. This was accomplished by having a naive observer score all the responses as correct or incorrect. The observer used a custom-built interface that presented each stimulus in its full-resolution (256 pixels) form alongside the subject's description. The observer marked whether or not the subject correctly recognized the content of the image. Using these responses we could then divide the subjects' responses into correctly identified trials labeled 'recognition positive' and incorrectly identified trials labeled 'recognition negative'.

We plot the angular error for recognition-positive trials and recognition-negative trials across different levels of blur in Figure 3.11 and detail the parameters of the t-test used in Figure 3.14. We arrive at a p-value of 0.035 between recognition-positive and recognition-negative trials in the medium-blur (32 pixels) condition making the two distributions significantly different. This analysis demonstrates that memory and past experience play a significant role in 3-D shape perception when an image is sufficiently degraded.

At higher levels of resolution, we did not find an impact of recognition performance on 3-D shape perception. This finding makes sense as higher-resolution images contain sufficient low-level information to mediate 3-D perception. In very low levels of resolution, namely the large-blur (16 pixels) condition, we found that low-level information was so disrupted that 3-D perception was impaired irrespective of recognition.

In addition, we compared the average angular error in the full-resolution (256 pixels)

condition against the average angular error of the recognition-positive trials at different levels of blur, and display the results in Figure 3.15. We found that there is no significant deviation between full-resolution (256 pixels) and both the small-blur (64 pixels) and the medium-blur (32 pixels) conditions. It is only in the large-blur (16 pixels) condition that we found a significant difference, implying that recognition allows one to adequately perceive 3-D shape up to the medium-blur (32 pixels) limit. Lastly, we investigated the flattening effect for the recognition-positive trials and display the results in Figure 3.11 and Figure 3.16. We found that blurred input is perceived as flatter regardless of recognition performance.

3.3.4 The role of contour

Lastly, we demonstrated the role of good contour in 3-D shape perception of natural images. In the region-averaged images, subjects could recognize all images, yet there was absolutely no shading or texture information present in the images. In this setup, we arrived at an angular error of 19 degrees. We compared this estimate with the recognition-positive trials from Experiment 1. In Experiment 1, shading, texture, and contour are disrupted as the image is progressively blurred. In the large-blur (16 pixels) condition, all three are sufficiently impoverished, yet we found that when just contour was preserved in Experiment 2, the error rate was on par with that of the small-blur (64 pixels) condition. Therefore the visual system can make up for lack of shading and texture in the context of good contour and recognition. This finding builds upon those of Mamassian et al. (1996), Norman and Raines (2002), and Cole et al. (2009). In the first study, the authors used a gauge-figure task to measure the subjective percept of an artificial stimulus - a donut with no shading or texture information, leaving only the contour present. Subjects' performance was not hindered in this condition. A similar conclusion was reached by Cole et al. (2009) and Norman and Raines (2002). They both found that shape perception is robust to contour representations of artificial stimuli. Our study extends these results to natural images.

3.4 Discussion

Our results provide quantitative evidence that human subjects perceive 3-D shape in impoverished natural images. This finding is in line with that of Norman et al. (2010) who demonstrated that 3-D shape discrimination survives under retinal blur. And, when bottom-up information alone is insufficient as in the medium-blur (32 pixels) condition, we demonstrate that top-down influences play a significant role. We propose that models of Bayesian cue combination, such as the one proposed by Hillis et al. (2002), offer mathematical models of how these different sources of information can be combined to create a singular percept. As images are progressively degraded, the perceptual error is captured by an increasing flattening effect that exists irrespective of recognition. Looking at these

results in conjunction with previous studies, we conclude that 32-pixel color images are a rough threshold for a rich perceptual experience. Previous studies (Nandakumar and Malik 2009; Torralba 2009) have demonstrated that recognition breaks down at around this point, and here we show that it is also the limit to reliably perceive 3-D shape. Presumably many perceptual abilities are tied together at this level - shape, recognition, etc, so that when one percept is lost, the other percepts break as well. In high-resolution images, however, it is possible to have one without the other - it is possible to perceive the 3-D shape without recognizing the content of the image as demonstrated in the small-blur (64 pixels) condition. Lastly we demonstrate that when even just one bottom-up cue is preserved, such as contour, top-down influences can combine to create an accurate 3-D percept.

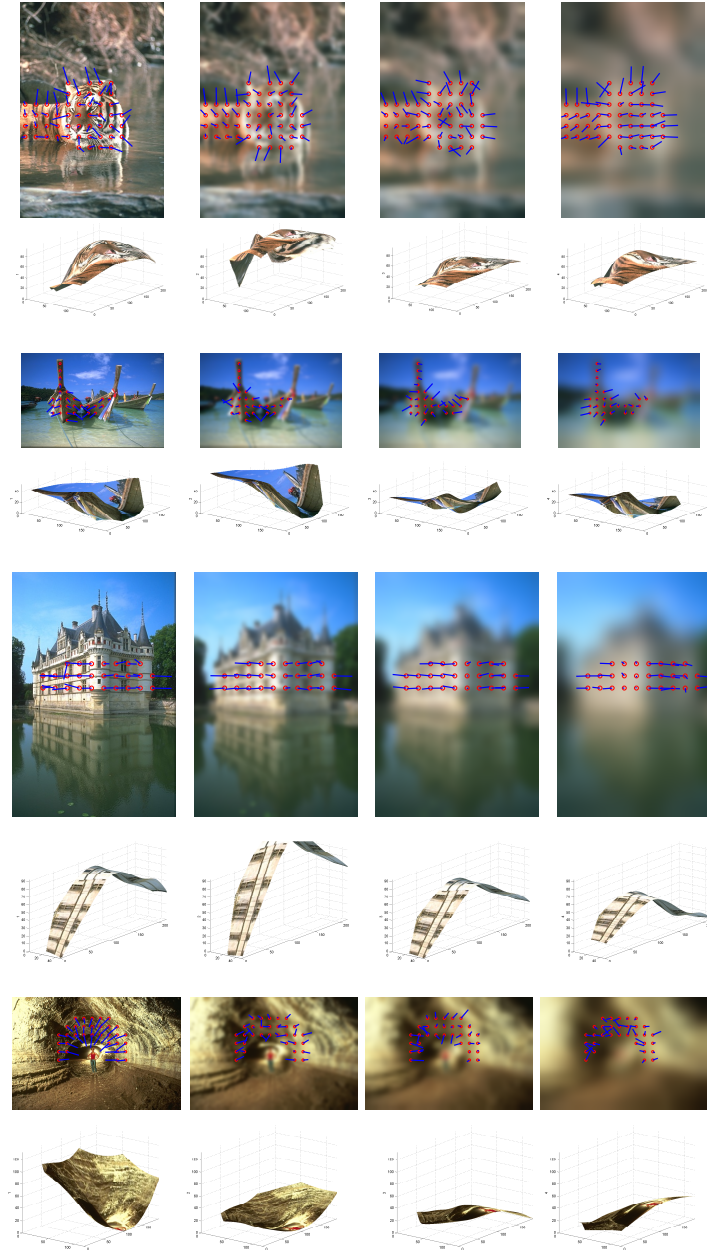
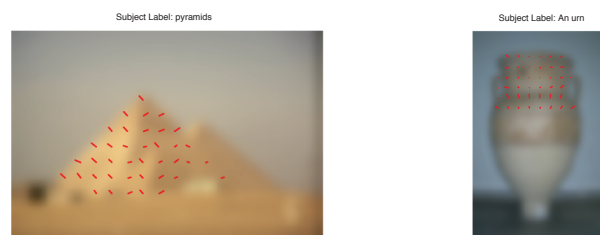


Figure 3.5: Visualization of subjects' responses for 4 images along with the corresponding 3-D reconstructions. For each image, we showcase the response from a representative subject at each level of degradation. Although there are global patterns across all images, we see that the character of the 3-D reconstruction varies from one image to the next. Some images are robust to high levels of degradation, whereas others, such as the image of the cave, suffer quite a bit under degradation. Please note that these reconstructions are not used for any subsequent analysis.

Examples of Correctly Recognized Trials in the Medium Blur (32 px) condition



Examples of Incorrectly Recognized Trials in the Medium Blur (32 px) condition

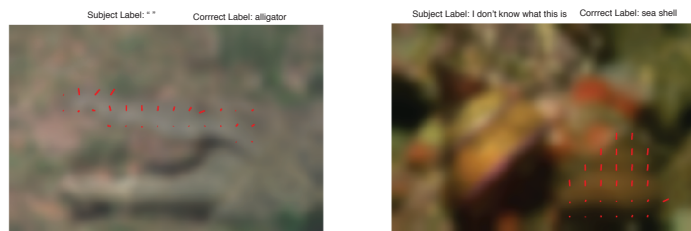


Figure 3.6: Examples of images recognized correctly (a) and incorrectly (b) in the medium blur (32 pixels) condition.

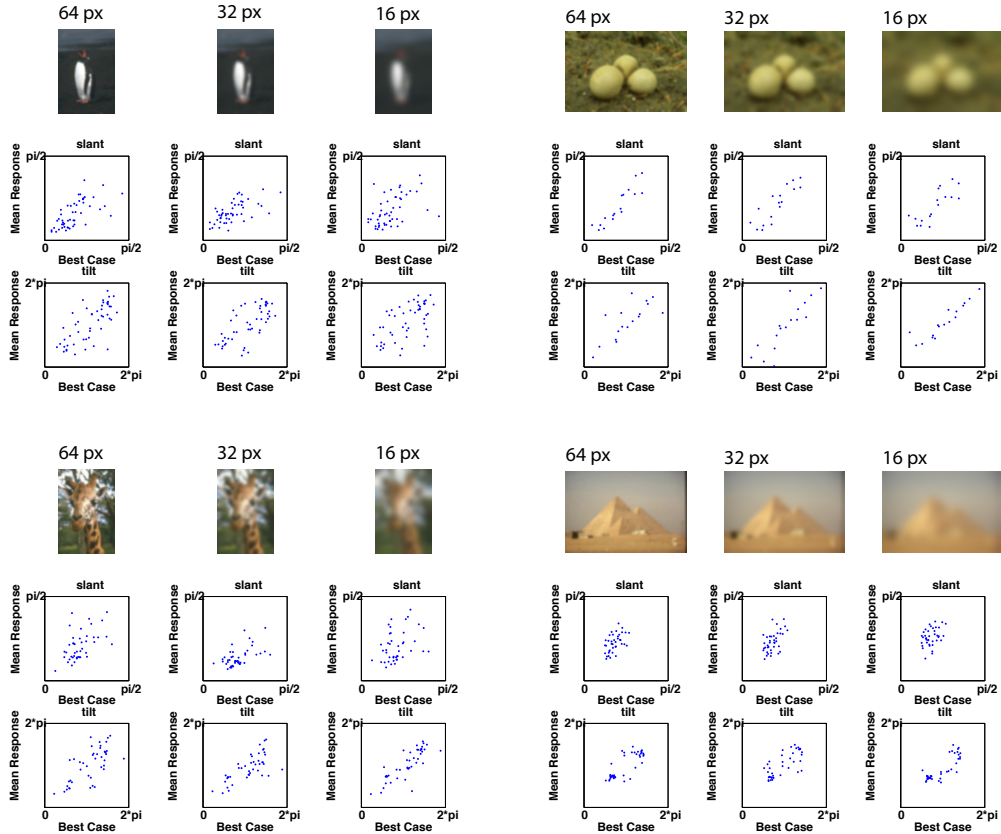


Figure 3.7: Visualization of subjects' responses for selected images. Scatter plots are displayed for the three levels of degradation. In each plot, the horizontal axis is the best-case, the average response at full resolution (256 pixels). The vertical axis is the mean of all responses for the given image in the given condition.

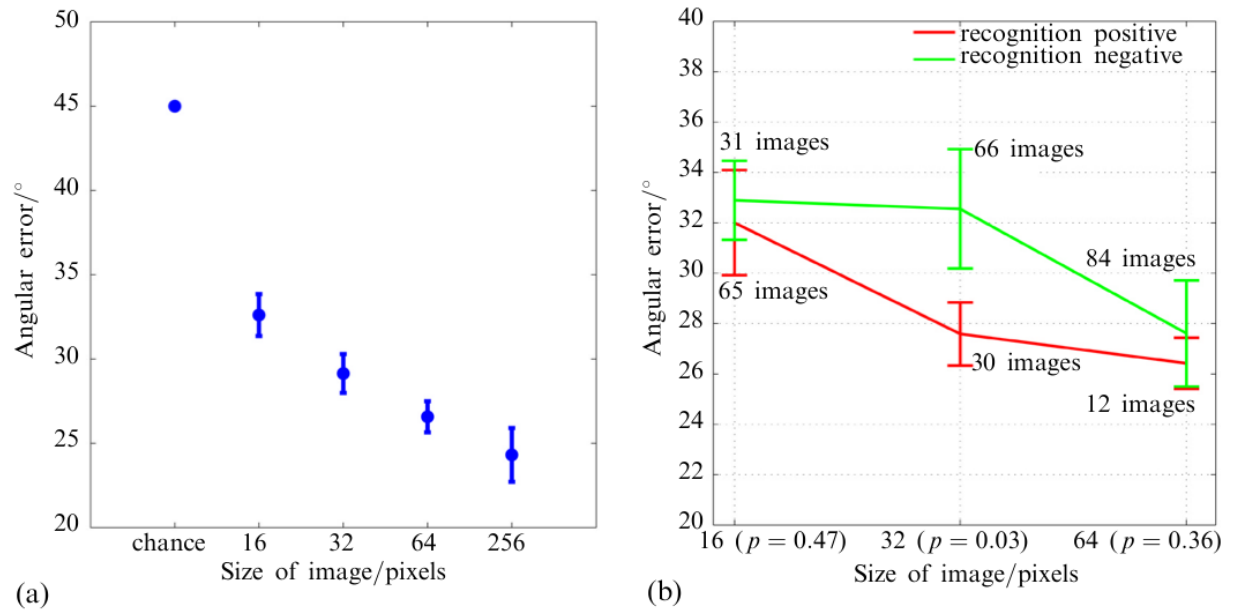


Figure 3.8: Angular error in Experiment 1 with standard error bars as the amount of information increases. Angular error is (a) averaged across all stimuli at a given resolution, (b) angular error divided between recognition positive (correctly identified) and recognition negative (incorrectly identified) trials, along with standard error bars. The number of images in each category is printed at each level of degradation.

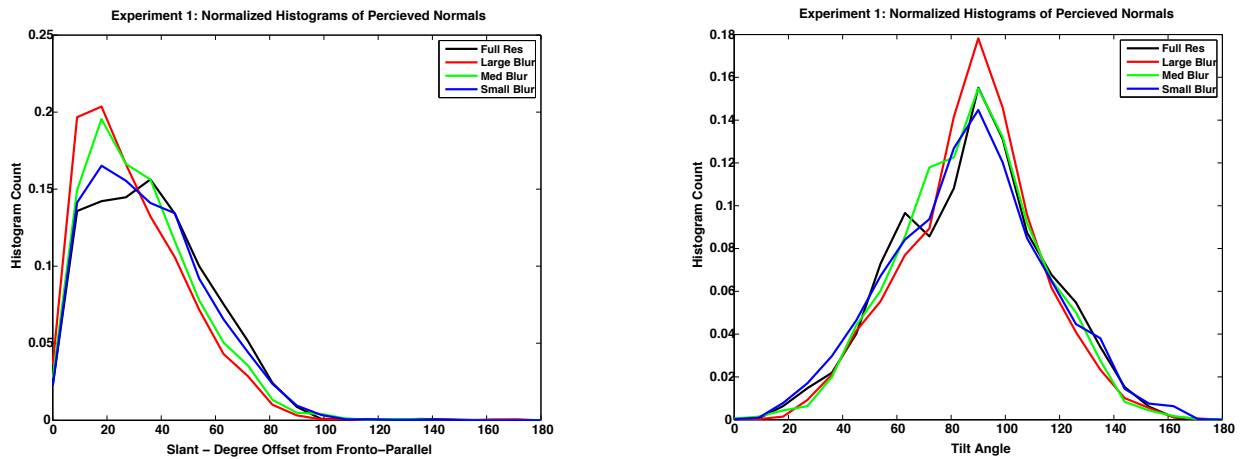


Figure 3.9: Experiment 1: normalized histograms of perceived normals. Distributions of slant and tilt at all levels of degradation.

Chapter 3. How little do we need for 3-D shape perception?

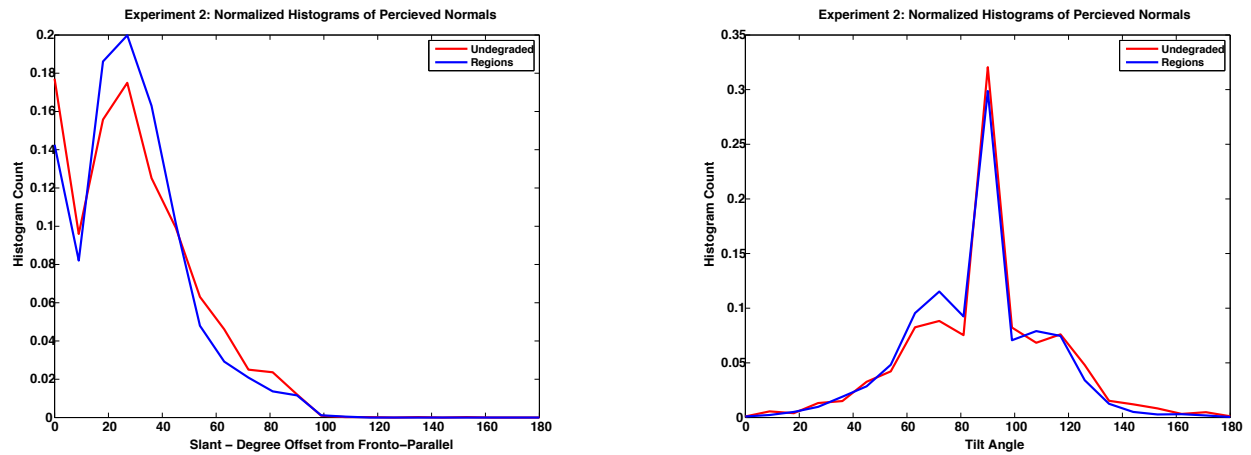


Figure 3.10: Distributions of slant and tilt from Experiment 2 (normalized histograms of perceived normals). There is no significant difference between subjects' responses in the undegraded and region-averaged conditions in either condition.

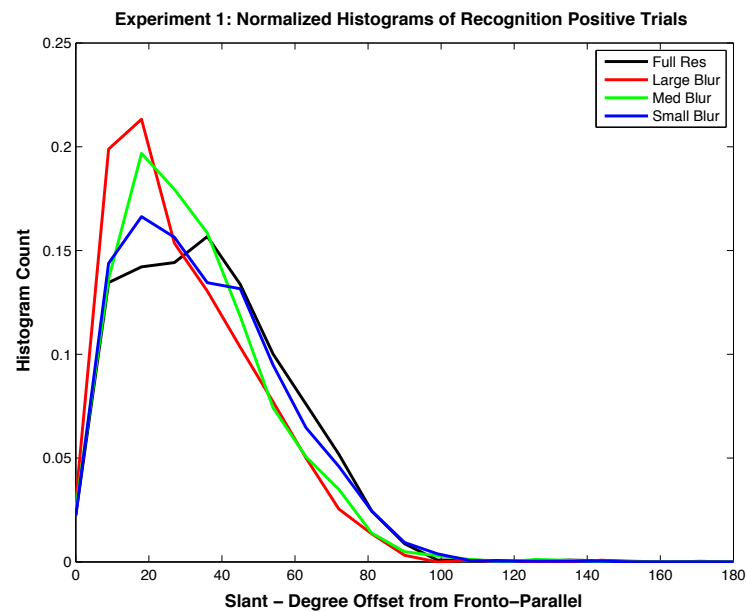


Figure 3.11: Experiment 1. Distribution of slant for recognition-positive trials (normalized histograms).

Condition	Full resolution (256 pixels)	Small blur (64 pixels)	Medium blur (32 pixels)	Large blur (16 pixels)
Full resolution	0	0.1399	0.0809	0.0008
Small blur		0	0.0809	0.0024
Medium blur			0	0.0526
Large blur				0

Figure 3.12: Using pairwise comparison testing between the slant distributions of different conditions, we display the p-value that response distributions for two conditions are the same. There is a significant difference between large-blur (16 pixels) and full-resolution (256 pixels) conditions and between large-blur (16 pixels) and small-blur (64 pixels) conditions.

Condition	Full resolution (256 pixels)	Small blur (64 pixels)	Medium blur (32 pixels)	Large blur (16 pixels)
Full resolution	0	0.1317	0.0809	0.1317
Small blur		0	0.1399	0.1399
Medium blur			0	0.1399
Large blur				0

Figure 3.13: Using pairwise comparison testing between tilt distributions, we display the p-value that response distributions for two conditions are the same. There are no significant differences between any pairs of conditions.

Image size	<i>t</i> -statistic	Degrees of freedom	<i>p</i> -value
Small blur (64 pixels)	0.5072	16.5562	0.47
Medium blur (32 pixels)	1.8533	45.9469	0.03
Large blur (16 pixels)	0.3408	63.8010	0.36

Figure 3.14: The *t*-statistic, degrees of freedom and *p*-value for each level of blur. These values correspond to the analysis visualized in figure 8b.

Condition	<i>t</i> -statistic	Degrees of freedom	<i>p</i> -value	Mean error rate/°
Small blur (64 pixels)	1.11	57.64	0.13	26.4
Medium blur (32 pixels)	1.61	68.67	0.06	27.6
Large blur (16 pixels)	2.92	56.57	0.002	32.0

Figure 3.15: Comparison of the average angular error in full resolution (256 pixels) with the average angular error of the recognition-positive trials in different conditions. Only the large-blur (16 pixels) condition is significantly worse than full-resolution (256 pixels) condition.

Condition	Full resolution (256 pixels)	Small blur (64 pixels)	Medium blur (32 pixels)	Large blur (16 pixels)
Full resolution	0	0.1399	0.03	0
Small blur		0	0.1317	0.0526
Medium blur			0	0.0065
Large blur				0

Figure 3.16: Using pairwise comparison testing between slant distributions of the recognition-positive trials, we display the p-value that response distributions for two conditions are the same.

Chapter 4

Natural Image Statistics underlying Color Perception

Joint work with: Professor Bruno Olshausen, Professor Stephen Palmer

4.1 Introduction

Human color perception is a multi-staged process that begins when light hits the retina. Light that has reflected off of objects and surfaces in the world activates Long (L), Medium (M) and Short (S) wavelength cone photoreceptors in the retina. These activations are translated into electrical voltages that travel via the retinal ganglion cells, through the optic nerve to the Lateral Geniculate Nucleus (LGN) in the thalamus and on up to the visual cortex. Due to the high redundancy of activations by the photoreceptors, biology has learned to decorrelate this input for purposes of efficient coding. Principal components analysis is a mathematical transformation that closely approximates this decorrelation of photoreceptor activation in the visual system (Buchsbaum and Gottschalk 1983, Zaidi 1997, Ruderman et al. 1998). Performing this analysis on photoreceptor activations, PCA transforms the dimensions of L, M, and S into the following dimensions: $L+M+S$, $L-M$, and $S-(L+M)$. Qualitatively the first dimension is luminance, the second is red-green color opponency and the third is blue-yellow color opponency. Validation of this scheme has come from both psychophysical and neurophysiological experiments. Krauskopf et al. (1982) conducted a psychophysical experiment that finds three unique color channels in human vision – luminance and 2 color-opponent directions. These psychophysical findings were corroborated in the neuroscience literature by Derrington et al. (1984) who found cells in the LGN that code along these color-opponent axes.

This transformation by the LGN maps the chromatic information of a pixel into a two dimensional plane parameterized by the red-green and blue-yellow color opponency dimen-

sions. Perceived hue is roughly encoded as the angular dimension in this plane and perceived saturation is roughly encoded as the distance of a given point from the achromatic origin.

Further studies have found sensitivity to dimensions off the 2 color-opponent axes found in the LGN. This sensitivity has been localized to V1 and V2 cortical areas. Lennie et al. (1990) and Cottaris and DeValois (1998) found that cells in V1 prefer colors that lie intermediate to the cardinal directions coded for in the LGN. Kiper et al. (1997) finds that most color sensitive cells in V2 are responsive to the angle between the stimulus color and the neuron's preferred color. Further evidence comes from Webster and Mollon (1994) who demonstrate that vision can selectively adapt to iso-hue lines in the cartesian plane, and that the perceived hue of other colors in the plane is a function of their angular separation with the adapted stimulus.

If decorrelation motivates the shift from coding color in terms of photoreceptor activation to color-opponent dimensions, what motivates the brain to be sensitive to iso-hue lines and to the angular dimension in the cartesian plane?

There has been great success in using natural image statistics to explain perceptual coding in the brain. Attneave (1954) and Barlow (1961) initially drew this link and recent years have seen a number of studies supporting this perspective. Olshausen and Field (1996) show that optimizing for sparseness and statistical independence in natural images yields basis functions very similar to the response properties of neurons in V1. Cormack et al. (2005) measured the distribution of binocular disparities occurring in forest environments and showed how this distribution corresponds to the disparity tuning of single neurons in MT of the macaque monkey (DeAngelis and Uka 2003). With respect to contrast response functions, there is a rough match between the contrast response functions of neurons in the early visual system and the distribution of contrasts that occur in natural environments (Brady and Field 2000, Clatworthy et al. 2003, Tadmor and Tolhurst 2000). With respect to color, Long et al. (2006) connect natural image statistics to psychophysical response curves of hue, saturation, and value. Lastly, Webster and Mollon (1997) and Webster et al. (2002) demonstrate how color perception adapts to the statistics of the natural environment.

4.2 Methods

In this study, we investigate how the color statistics of real-world images motivates a sensitivity to hue in human visual perception. Specifically, we explore how the statistics of color in object regions can be exploited for coding efficiencies by the visual system. We use 2 different datasets for this analysis, the Amsterdam Library of Object Images (ALOI) (Geusebroek et al., 2005) and the Berkeley Segmentation Dataset (BSDS) (Martin et al., 2001).



Figure 4.1: Selected images from The Amsterdam Library of Images (ALOI) dataset

4.2.1 The ALOI

The ALOI is a color image collection of 1000 small manmade objects. Each object is photographed against a black background in a canonical pose. Please see Figure 4.1. We begin by taking a qualitative look at the distributions of color in each object region. Since the images in this dataset are of isolated objects, the entire object in a given image is treated as a singular object region.

We begin by performing principal components analysis (PCA) against the entire set of RGB images in the ALOI. Since the first principal component largely encodes luminance and not chromaticity, we do not use it for this analysis. We mask out the object from the black background and map each pixel into the space of the second and third principal component (PC-2 and PC-3). We display these pixel clusters for an arbitrary selection of object images in Figure 4.2. We note a strong radial structure to the point clouds. In other words, we note that the angle of the first principal component of the point cloud is very aligned with the angle of its mean.

To get a snapshot of the entire dataset we display all the regions in PC-2, PC-3 space in Figure 4.3. Each region is denoted by its individual principal components plotted at the location of the region's mean. To clarify, PC-2 and PC-3 refer to the principal components of the entire dataset and inside this cartesian space, we display the principal components of each object region. In the plot, the length of each principal component is proportional to its eigenvalue, and we visually see a strong radial structure to this data just as we had with the individual objects.

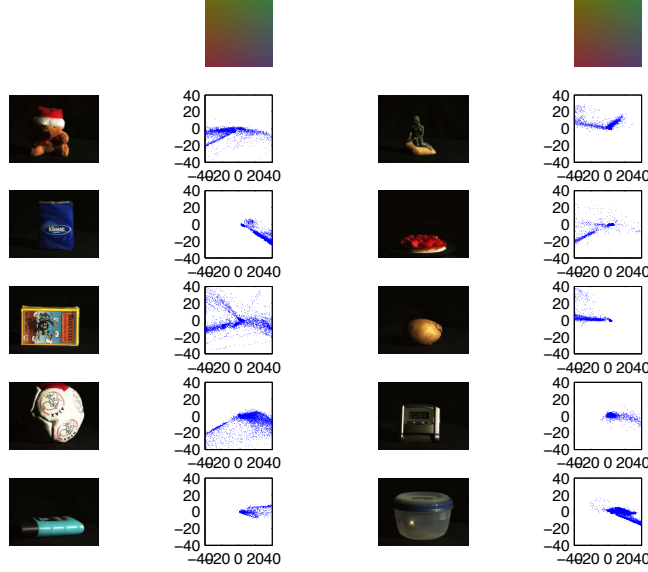


Figure 4.2: Select objects from the ALOI, and for each object, its colors are plotted in the PC-2, PC-3 plane. On the top row is a visualization of how PC-2, PC-3 plane maps out in colorspace.

4.2.1.1 Statistical Analysis

We investigate whether or not this structure can be exploited for a more efficient coding of color. Since there is a high correlation between the angle of the mean and the angle of the first principal component, it seems plausible that the coordinate space captured by the angle of the mean, Θ , and its orthogonal axis, $\Theta + \pi/2$, serves as an efficient coordinate space for the points in that region. We investigate this hypothesis for the images in the ALOI dataset.

We compute the Shannon entropy, H , or bits per pixel, required to code the pixels in a given coordinate space via:

$$H_{tot} = H(axis_1) + H(axis_2) \quad (4.1)$$

$$H(axis_i) = (1/2) \ln\{2\pi e\sigma_i^2\} \quad (4.2)$$

where $axis_1$ and $axis_2$ are the two orthogonal axes for the given space and where σ_i^2 is the variance along that dimension. For the pixels in each object region, we compute the average H_{tot} under PC-2, PC-3 space and compare it to the average H_{tot} under $\Theta, \Theta + \pi/2$ space. To see if this difference is significant, we compute the difference in H_{tot} per region

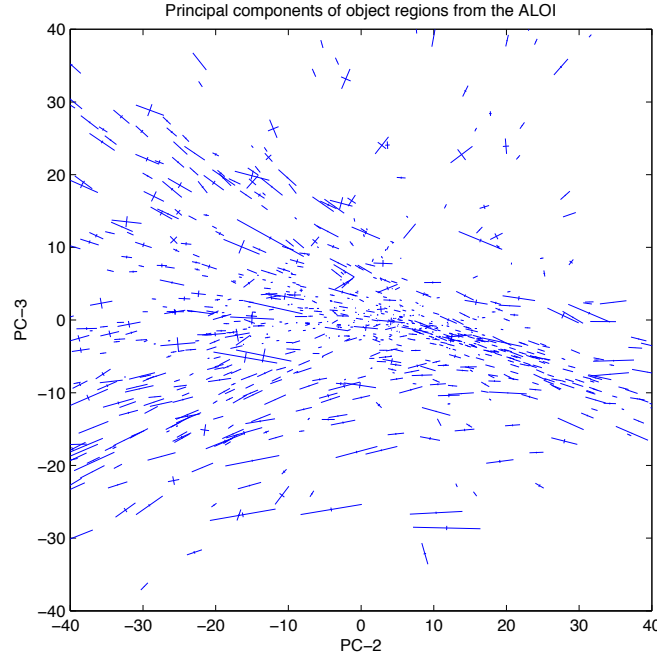


Figure 4.3: Principal components of object regions from the ALOI. The principal components for a given object region are plotted at its mean in PC-2, PC-3 space. Please note that PC-2 and PC-3 are the second and third principal components derived from computing PCA across all pixels in the dataset.

per pixel between the two spaces to create a difference vector, $\vec{v}$.

To construct $\vec{v}$, let $R_1, R_2, \dots, R_m$ be m number of regions. Let k_i represent the number of pixels in region R_i and $N = k_1 + k_2 + \dots + k_m$. Let $H_i = H_{\text{tot}, \text{PC}} - H_{\text{tot}, \Theta}$ for the given region R_i , and we denote the vector:

$$\vec{v} = (\underbrace{H_1, \dots, H_1}_{k_1 \text{ times}}, \underbrace{H_2, \dots, H_2}_{k_2 \text{ times}}, \dots, \underbrace{H_m, \dots, H_m}_{k_m \text{ times}}). \quad (4.3)$$

We then perform a t-test on $\vec{v}$ to see if this distribution is significantly different from the null hypothesis. We present the output of this analysis in Table 4.1.

We find a significant coding advantage to using $\Theta, \Theta + \pi/2$ as a coordinate space for each region. Since the ALOI is comprised of manmade objects, the question remains if these findings hold for natural settings. Furthermore, is this finding tied to object regions, or is it a local phenomenon?

Coord. Space	Objects
PC-2, PC-3	8.43
$\Theta, \Theta + \pi/2$	8.18
$p - value$	$1e - 5$

Table 4.1: The average H_{tot} (bits/pixel) in each coordinate space for objects in the ALOI. The p-value is the probability that these two numbers are drawn from the same distribution as described in the Statistical Analysis section.

4.2.2 The BSDS

To investigate the above questions, we used the Berkeley Segmentation Database (BSDS). The BSDS is a collection of natural images in which subjects were instructed to mark closed-loop regions corresponding to objects. It is organized by subject number, and for this analysis we use the set of images labeled by subject 1123 due to the high quality of his/her responses. This set includes 137 unique images and 3225 tagged object regions. Please see Figure 4.4 for select examples.

Since the dataset comprises a wide range of images, we find reasonable variety to the tagged object regions. One source of variability is that the regions can be from any scale in the image. For instance, an image of a face can be segmented as one big region or as multiple regions, where the eye, the nose and lips are all distinct regions. In addition, the collection of regions from a given image include both foreground and background regions. We perform the above analysis in two conditions. We first work with object regions in Condition 1, and then we work with random patches in Condition 2.

4.2.2.1 Condition 1

As we had done for the ALOI, we compute PCA on the entire dataset and then plot the first and second principal components of each object region on a unified plot. Please see Figure 4.5. Once again, we see a strong radial structure to this data. Computing the correlation coefficient between the angle of the mean and the angle of the first principal component across all object regions, we arrive at .21. However, we note that this correlation increases with distance from the origin. Please see Figure 4.6. Regions whose mean is distant from the origin are far more likely to be radially oriented. Even with one standard deviation away from the origin, the correlation coefficient increases to .41.

4.2.2.2 Condition 2

In this condition, we investigate to see if any coding advantage is tied to local statistics in the image plane as opposed to object-based statistics. Using the same set of images as

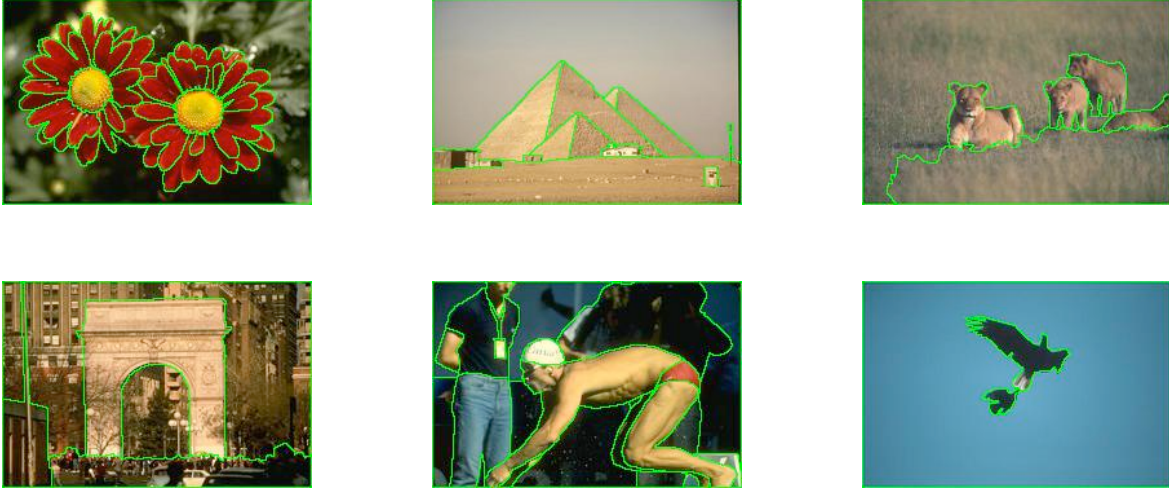


Figure 4.4: Example images from The Berkeley Segmentation Dataset (BSDS) dataset, with the human segmentations overlaid in green.

in Condition 1, we consider random patches as opposed to object regions. For each image in our collection, we create patches where each patch is a square drawn from the same size distribution as the original set of object regions. It is placed so that its center randomly lies on an edge of the human segmentation for the image. In this way, each patch straddles different object regions. This choice was made so that we insure that the set of random patches is distinct from the set of object regions.

We compute H_{tot} in Condition 1 and Condition 2 and display the results in Table 4.2. Because we find an increased radial structure as distance from the origin increases, we also consider regions whose mean is greater than 1 standard deviation from the origin as a separate category.

4.3 Discussion

With manmade objects, as represented by the ALOI, we find a significant coding advantage in the $\Theta, \Theta + \pi/2$ coordinate space as opposed to the PC-2, PC-3 coordinate space. Looking at a wider variety of naturally occurring images, as represented by the BSDS, we find a coding advantage in the $\Theta, \Theta + \pi/2$ coordinate space for object regions that are sufficiently distant from the origin. However, we do not find a coding advantage for object regions as a collective or for random patches.

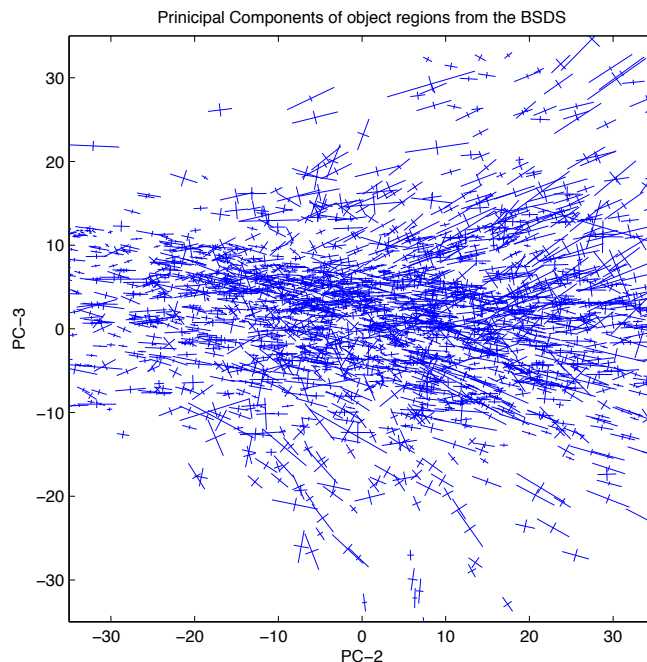


Figure 4.5: Principal components of object regions from the BSDS. The principal components for a given object region are plotted at its mean in the PC-2, PC-3 plane. Please note that PC-2 and PC-3 are the second and third principal components derived from computing PCA across all pixels in the dataset

Looking deeper at the data, we find that the area surrounding the origin in PC-2, PC-3 space is achromatic in character. Therefore a region centered in this area of the plane does not have a distinct hue, and the angle of its first principal component is only loosely correlated with the angle of its mean. This is shown explicitly in the first plot of Figure 4.6. Future work in this space can investigate how this model can be improved so that it is able to efficiently represent regions close to the origin while also being optimal for regions far away.

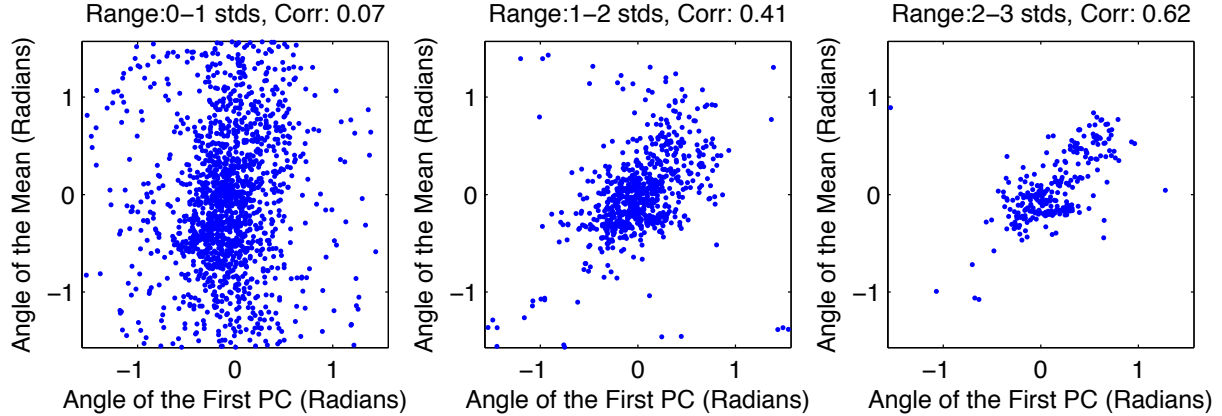


Figure 4.6: Scatter plots of the angle of the mean against the angle of the first PC for object regions in the BSDS. The three different plots correspond to object regions positioned at increasing distances from the origin in PC-2, PC-3 space.

Coord. Space	Objects	Objects>1 std	Random	Random>1 std
PC-2, PC-3	6.34	6.45	7.32	7.28
$\Theta, \Theta + \pi/2$	6.39	6.43	7.38	7.36
<i>p-value</i>	$1e-5$	$1e-5$	$1e-5$	$1e-5$

Table 4.2: The average H_{tot} (bits/pixel) in each coordinate space. We explore this statistic for object regions from the BSDS, for object regions whose mean is greater than one standard deviation from the origin, for random patches from the BSDS, and for random patches whose means is greater than 1 standard deviation from the origin. The p-value is the probability that $H_{tot,PC}$ and $H_{tot,\Theta}$ for a given condition are drawn from the same distribution. This analysis is described in the Statistical Analysis section.

Chapter 5

Contributions

The work in this dissertation has contributed to the scientific literature in a few key ways. With respect to rapid category detection, we extend the initial work done by Thorpe et al. (1996) to naturally occurring conditions such as blurring and luminance depth degradation. In such degradations, the image information is not nearly as pristine as in Thorpe’s original experiment. This work demonstrates that rapid category recognition is not merely a laboratory effect but is a process invariant to real-world degradations. In addition, this work also highlights the non-linear interaction between visual cues in rapid category detection.

With regards to 3D shape, we find that human 3-D shape perception is invariant to severe degradation in natural images. And, when bottom-up information alone is insufficient, we demonstrate that top-down influences play a significant role. As images are progressively degraded, the perceptual error is captured by an increasing flattening effect that exists irrespective of recognition. In addition, we demonstrate that when just one bottom-up cue is preserved, such as contour, top-down influences can combine to create an accurate 3-D percept.

These two studies, in conjunction with previous work, point to 32-pixel color images as a rough threshold for a rich, perceptual experience. The first study demonstrated that recognition breaks down at around this point, and the second shows that it is also the limit to reliably perceive 3-D shape. Presumably many perceptual abilities are tied together at this level - shape, recognition, etc, so that when one percept is lost, the other percepts break as well.

With respect to the perceptual coding of color, we find a link between the statistics of color in object regions and hue angle. Although this investigation is still in progress, future work may offer greater clarity on how these invariant properties of object regions explain the perceptual coding of color in the brain.

Chapter 6

Bibliography

Attneave F. (1954). Some informational aspects of visual perception. *Psychol. Rev.* 61:183-93.

Bachmann, T. (1991). Identification of spatially quantized tachistoscopic images of faces: How many pixels does it take to carry identity? *European Journal of Cognitive Psychology*, 3, 85-103.

Barrow, H. G., Tenenbaum, J. M. (1981). Interpreting line-drawings as three-dimensional surfaces. *Artificial Intelligence*, Vol. 17, Nos. 1-3, pp. 75-116.

Barlow H.B. (1961). Possible principles underlying the transformation of sensory messages. In *Sensory Communication*, ed. W.A. Rosenblith, pp. 217-34. Cambridge, MA: MIT Press.

Biederman, I. (1987). Recognition-by-Components: A Theory of Human Image Understanding. *Psychological Review*, 94, 115-147.

Binford, T.O. (1971). Visual perception by computer. Paper presented at the IEEE Systems Science and Cybernetics Conference, Miami, FL.

Brady N., Field D.J. (2000). Local contrast in natural images: normalization and coding efficiency. *Perception* 29:1041-55.

Brainard, D. H. (2004). Color constancy. In L. Chalupa and J. Werner (Eds.), *The visual neurosciences*, pp. 948-961. Cambridge, MA: MIT Press.

Buchsbaum G, Gottschalk A. (1983). Trichromacy, opponent colour coding and optimum colour information transmission in the retina. *Proc. R. Soc. Lond. B Biol. Sci.* 220(1218): 891-13.

- Brooks, M.J., Horn, B.K.P. (1985). Shape and source from Shading. Proceedings of IJCAI-9, Los Angeles, CA, pp. 932-936, August.
- Cavanagh, P. (1991). What's up in top-down processing? Representations of Vision: Trends and Tacit Assumptions in Vision Research. Cambridge, UK: Cambridge University Press, 295-304.
- Christou C., Koenderink J.J., Van Doorn A.J. (1996). Surface gradients, contours and the perception of surface attitudes in images of complex scenes. *Perception*, Vol. 25, 701-713.
- Cole F., Sanik K., DeCarlo D., Finkelstein A., Funkhouser T., Rusinkiewicz S., Singh M. (2009). How Well Do Line Drawings Depict Shape? *ACM Transactions on Graphics (Proc. SIGGRAPH)*. 28(3) August 2009.
- Clatworthy P.L., Chirimuuta M., Lauritzen J.S., Tolhurst D.J. (2003). Coding of the contrasts in natural images by populations of neurons in primary visual cortex (VI). *Vis. Res.* 43:19832001.
- Cutting J., Millard R. (1984). Three gradients and the perception of flat and curved surfaces. *Journal of Experimental Psychology: General*, 113, 198-216.
- DeCarlo D., Finkelstein A., Rusinkiewicz S., Santella A. (2003). Suggestive Contours for Conveying Shape. *SIGGRAPH 2003*. pp. 848-855.
- DeAngelis G.C., Uka T. (2003). Coding of horizontal disparity and velocity by MT neurons in the alert macaque. *J. Neurophysiol.* 89:1094111.
- Derrington AM, Krauskopf J, Lennie P. (1984). Chromatic mechanisms in the lateral geniculate nucleus of macaque. *J. Physiol.* 357: 24165.
- Erens R.G.F., Kappers A.M.L., Koenderink J.J. (1993a). Perception of local shape from shading. *Perception and Psychophysics*, 54 (2), 145-156, 1993.
- Erens R.G.F., Kappers A.M.L., Koenderink J.J. (1993b). Estimating local shape from shading in the presence of global shading. *Perception and Psychophysics*, 54 (3), 334-342.
- Fleming, R. W., Torralba, A., and Adelson, E. H. (2004). Specular reflections and the perception of shape. *Journal of Vision* 4, 9, 798820.
- Frome, A., Singer, Y., Sha, F., Malik, J. (2007). Learning Globally-Consistent Local Distance Functions for Shape-Based Image Retrieval and Classification. *IEEE 11th International Conference on Computer Vision. ICCV 2007*: pp.1-8.

- Gerardin, P., de Montalembert, M., and Mamassian, P. (2007). Shape from shading: New perspectives from the Polo Mint stimulus. *Journal of Vision*, 7(11):13, 111.
- Gibson, J. (1950). *The Perception of the Visual World*. Boston: Houghton Mifflin.
- Guyonneau, R., Kirchner, H., and Thorpe, S. J. (2006). Animals roll around the clock: The rotation invariance of ultrarapid visual processing. *Journal of Vision*, 6(10):1008-1017.
- Guzman, A. (1969). Decomposition of a visual scene into three-dimensional bodies. In A. Griselli (Ed.), *Automatic interpretation and classification of images* (pp. 243-276). New York: Academic Press.
- Harmon, L. D. and Julesz, B. (1973). Masking in visual recognition: effects of two-dimensional noise. *Science*, 180:1194-1197.
- Hillis, J. M., Ernst, M. O., Banks, M. S. and Landy, M. S. (2002). Combining sensory information: mandatory fusion within, but not between, senses. *Science*, 298, 1627-1630.
- Horn, B.K.P, Brooks M.J. (1986). *The Variational Approach to Shape from Shading*. *Computer Vision, Graphics and Image Processing*, Vol. 33, No. 2, pp. 174-208.
- Horn, B.K.P. (1970). *Shape from Shading: A Method for Obtaining the Shape of a Smooth Opaque Object from one View*. Technical Report TR-79, MIT Project MAC, Cambridge, MA.
- Horn, B.K.P., Brooks, M.J. (1989). *Shape from Shading*. MIT Press, Cambridge, MA.
- Huffman, D. A. (1971). Impossible objects as nonsense sentences. In M. Meltzer and D. Michie (Eds.), *Machine intelligence* (Vol. 6). Edinburgh, Scotland: Edinburgh University Press.
- Ito, M., Tamura, H., Fujita, I., and Tanaka, K. (1995). Size and position invariance of neuronal responses in monkey inferotemporal cortex. *Journal of Neurophysiology*, 73, 218-226.
- Jaa-Aro, K. and Kjelldahl, L. (1997). Effects of Image Resolution on Depth Perception in Stereo and Non-Stereo Images. *Stereoscopic Displays and Applications VIII*, v3012.
- Johnson J., Olshausen B. (2003). Time courses of neural signatures for object recognition. *Journal of Vision*, 3(7), 499-512.
- Kemelmacher-Shlizerman I., Basri R. and Nadler B. (2008). 3D Shape Reconstruction of Mooney Faces. *IEEE CVPR 2008*.

- Kirchner, H. and Thorpe, S.J. (2006). Ultra-rapid object detection with saccadic eye movements: Visual processing speed revisited. *Vision Research*, 46(11), 1762-1776.
- Kiper DC, Fenstemaker SB, Gegenfurtner KR. (1997). Chromatic properties of neurons in macaque area V2. *Vis. Neurosci.* 14:106172.
- Kleffner, D. A., and Ramachandran, V. S. (1992). On the perception of shape from shading. *Perception and Psychophysics*, 52, 1836.
- Koenderink, J. J., and Van Doorn, A. J. (1982). The shape of smooth objects and the way contours end. *Perception*, 11, 129-137.
- Koenderink, J. J., Van Doorn, A. J., and Kappers, A. M. (1992). Surface perception in pictures. *Perception and Psychophysics*. 52(5), 487-96.
- Koenderink, J.J. (1984). What does the Occluding Contour tell us about Solid Shape. *Perception*, Vol. 13, pp. 321-330.
- Krauskopf J, Williams DR, Heeley DW. (1982). Cardinal directions of color space. *Vision Res.* 22(9):112331
- Lazebnik, S., Schmid, C., and Ponce, J. (2006). Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition: Volume 2* (June 17 - 22, 2006).
- Lennie P, Krauskopf J, Sclar G. (1990). Chromatic mechanisms in striate cortex of macaque. *J. Neurosci.* 10:64969.
- Liu, C. H., Collin, C. A., Burton, A. M., Chaurdhuri, A. (1999). Lighting direction affects recognition of untextured faces in photographic positive and negative *Vision Research* 39, 4003-4009.
- Liu, Y., Bovik, A. C., Cormack, L. K. (2008). Disparity statistics in natural scenes. *Journal of Vision*, 8(11):19, 114.
- Long F., Yang Z., Purves D. (2006). Spectral statistics in natural scenes predict hue, saturation, and brightness. *Proc Natl Acad Sci USA* 103:60136018.
- Lowe, D. (1999). Object recognition from local scale-invariant features. *International Conference on Computer Vision, Corfu, Greece*, pp. 1150-1157.
- Malik J., Rosenholtz R. (1997). Computing Local Surface Orientation and Shape from Texture for Curved Surfaces. *International Journal of Computer Vision*. 23(2), 149-168.

- Malik, J. (1987). Interpreting line drawings of curved objects. *International Journal of Computer Vision*, 1(1), 73-103.
- Mamassian P, Kersten D, Knill D C. (1996). Categorical local-shape perception. *Perception*. 25(95107).
- Mamassian, P., and Landy, M. S. (1998). Observer biases in the 3D interpretation of line drawings. *Vision Research*, 38, 2817-2832
- Marr, D. (1977). Analysis of occluding contour. *Proceedings of the Royal Society of London, SeriesB*, 197,441-475.
- Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*, W.H. Freeman, San Francisco.
- Marr, D., and Nishihara, H. K. (1978). Representation and recognition of three dimensional shapes. *Proceedings of the Royal Society of London, SeriesB*, 200,269-294.
- Martin, D., Fowlkes, C., Tal. D. and Malik J. (2001). A Database of Human Segmented Natural Images and its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics. *Proceedings of the 8th International Conference in Computer Vision*. Vol. 2. pp. 416-423.
- Mingolla, E., Todd, J. (1986). Perception of solid shape from shading. *Biological Cybernetics*, Vol. 53, pages 137-151.
- Mooney, C. M. (1957). Age in the development of closure ability in children. *Canadian Journal of Psychology*, 11, 219-226.
- Nandakumar, C., Malik, J. (2009). Understanding rapid category detection via multiply degraded images. *Journal of Vision*, 9(6):19, 1-8.
- Norman, J. F., and Raines, S. R. (2002). The perception and discrimination of local 3-D surface structure from deforming and disparate boundary contours. *Perception and Psychophysics*, 64, 1145-1159.
- Norman, J. F., Bartholomew, A. N., and Burton, C. L. (2008). Aging preserves the ability to perceive 3-D object shape from static but not deforming boundary contours. *Acta Psychologica*, 129, 198-207.
- Norman, J. F., Beers, A. M., Holmin, J. S., and Boswell, A. M. (2010). Effective 3-D shape discrimination survives retinal blur. *Attention, Perception, and Psychophysics*, 72, 1569-1575.

- Oliva, A., and Torralba, A. (2006). Building the Gist of a Scene: The Role of Global Image Features in Recognition. *Visual Perception, Progress in Brain Research*, vol 155.
- Olshausen, B. A., Anderson, C. H., and Van Essen, D. C. (1995). A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *Journal of Computational Neuroscience*, 2, 4562.
- Olshausen B.A., Field D.J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* 381:6079.
- Panis, S., De Winter, J., Vandekerckhove, J., and Wagemans, J. (2008). Identification of everyday objects on the basis of fragmented outline versions. *Perception*, 37, 271-289.
- Peterson, M. A. (1994). Object recognition processes can and do operate before figureground organization. *Current Directions in Psychological Science*, 3, 105-111.
- Purdy, W. P. (1958). The hypothesis of psychophysical correspondence in space perception. *Dissertation Abstracts International*, 42, 1454 (University Micro- films No. 58-5594).
- Ramachandran V.S. (1988). Perception of shape from shading. *Nature*, 331, 163-166.
- Richards W.A., Koenderink J.J., Hoffman D.D. (1987) Inferring three-dimensional shapes from two-dimensional silhouettes. *J. Opt. Soc. Am. A*4, 1168-1175.
- Rock I., DiVita J. and Barbeito R. (1981). The effect on form perception of change of orientation in the third dimension. *Journal of Experimental Psychology: Human Perception and Performance* 7, pp. 719-732.
- Rosenholtz, R., Malik J. (1997). Surface orientation from Texture: Isotropy or Homogeneity (or Both)?. *Vision Research*. Vol. 37. No. 16. Pp.2283-2293.
- Serre, T., Oliva, A. and Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Science*, 104(15), 6424-6429.
- Ruderman D.L., Cronin T.W., Chiao C.C. (1998). Statistics of cone responses to natural images: implications for visual coding. *J. Opt. Soc. Am. A* 15:2036-45.
- Stevens, K.A. (1981). The information content of texture gradients. *Biological Cybernetics*, 42, 95-105.

- Tadmor Y., Tolhurst D.J. (2000). Calculating the contrasts that retinal ganglion cells and LGN neurons encounter in natural scenes. *Vis. Res.* 40:314557.
- Tarr M., Pinker S. (1989). Mental rotation and orientation-dependence in shape recognition. *Cognitive Psychology*, 2/(2), 233-282.
- Thorpe, S. J. (2002). Ultra-rapid scene categorisation with a wave of spikes. *Proceedings of Biologically Motivated Computer Vision*, pp. 115.
- Thorpe, S. J., Fize, D. and Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, 381(6582), 520-522.
- Todd J., Akerstrom R. (1987). Percpetion of three-dimensional form from patterns of optical texture. *Journal of Experimental Psychology: Human perception and Performance*, 13, 242-255.
- Todd, J. T., and Thaler, L. (2010). The perception of 3D shape from texture based on directional width gradients.
- Todd, J. T., Oomes, A. H. J., Koenderink, J. J., and Kappers, A. M. L. (2004). Perception of doubly curved surfaces from anisotropic textures. *Psychological Science*, 15, 4046.
- Torralba A. (2009). How many pixels make an image? *Visual Neuroscience*, volume 26, issue 1, pp. 123-131.
- Torralba, A. and Sinha. P. (2001). Detecting faces in impoverished images. MIT AI Tech Report, 2001-028.
- Torralba, A., Fergus, R. and Freeman, W. T. (2008). 80 Million Tiny Images: A Large Data Set for Nonparametric Object and Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Vol. 30, No. 11, pp. 1958-1970.
- Van Rullen, R., Thorpe, S. (2001). Is it a bird? Is it a plane? Ultra-rapid visual categorization of natural and artifactual objects. *Perception*. 30(6) 655 - 668.
- Wagemans, J., De Winter, J., Op de Beeck, H., Ploeger, A., Beckers, T., and Vanroose, P. (2008). Identification of everyday objects on the basis of silhouette and outline versions. *Perception*, 37, 207-244.
- Waltz, D. (1975). Understanding line drawings of scenes with shadows. In P. H. Winston (Ed.), *The psychology of computer vision* (pp. 19-92). New York: McGraw-Hill.

Webster, M.A. and Mollon, J.D. (1997). Adaptation and the color statistics of natural images. *Vision Res.* 37, 3283–3298.

Webster, M.A., Webster, S.M., Bharadwaj, S., Verma, R., Jaikumar, J., Madan, G. and Vaithilingham, E. (2002). Variations in normal color vision III. Unique hues in Indian and United States observers. *Journal of the Optical Society of America* 19, pp. 1951-1962.

Witkin, A.P. (1981). Recovering surface shape and orientation from texture. *Journal of Artificial Intelligence*, 17, 17-45.

Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology*, 81, 141-145.

Zaidi Q. (1997). Decorrelation of L- and M-cone signals. *J. Opt. Soc. Am. A* 14(12):3430-31