The Perception of Material Qualities in Real-World Images

by

Lavanya Sharan

S.M., Massachusetts Institute of Technology (2005)

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

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2009

© Massachusetts Institute of Technology 2009. All rights reserved.
The Perception of Material Qualities in Real-World Images

by

Lavanya Sharan

Submitted to the Department of Electrical Engineering and Computer Science
on August 31st, 2009, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

One can easily tell if a sidewalk is slippery, if food is fresh, if a spoon is made of
plastic or stainless steel, or if a suspicious looking mole warrants a trip to the doctor.
This ability to visually identify and discriminate materials is known as material per-
ception and little is known about it. We have measured human material judgments
on a wide range of complex, real world materials. We have gathered several diverse
image databases and made use of them to conduct psychophysical studies. We asked
observers to classify surfaces and objects as being made of fabric, paper, plastic or
other common material categories. In the first part of this thesis, we present exper-
iments that establish that observers can make these judgments of material category
reliably, quickly and in challenging conditions of rapid presentation. We find that
categorization performance cannot be explained by simple, low-level cues like color
or high spatial frequencies. In the second part of the thesis, we explore judgments
beyond those of common material categories. Observers judged many dimensions of
material appearance such as matte vs. glossy, opaque vs. translucent, rigid vs. non-
rigid, soft vs. rough to touch, and even genuine vs. fake for familiar object categories
like flowers, fruits and dessert. Observers were surprisingly accurate, even in 40 mil-
liisecond presentations. In the final part of this thesis, we compare the performance of
state-of-art computer vision techniques with human performance on our images and
tasks and find current techniques to be severely lacking. Taken together, our findings
indicate that material perception is a distinct mechanism and can be as fast and flex-
ible as object recognition or scene perception. When recognizing materials, low-level
image information is of limited use for both humans and computer vision systems.
We conclude that material recognition is a rich and challenging problem domain and
there is much ground to be covered in both visual perception and computer vision.

Thesis Supervisor: Edward H. Adelson
Title: John and Dorothy Wilson Professor of Vision Science
Acknowledgments

Thank you!

- **Ted Adelson**, for being such a wonderful advisor and mentor, for his time and his support, for his wit and his marvelous intuitions. For teaching me how to ask hard questions, how to give clear presentations, how to write papers, theses and grants, how to take good photographs, how to use Photoshop tools as shortcuts for MATLAB coding, and how to produce bubble-free plastic molds and casts.

- **Ruth Rosenholtz** for being practically a “co-advisor” on this thesis, for teaching me about psychophysics, for advice on writing and presenting, many spontaneous statistics lessons, and numerous enjoyable conversations ranging from women in science to the role of Twinkies in American popular culture.

- **Shin’ya Nishida** and **NTT Communication Science Laboratories** for supporting my graduate studies at MIT. I am grateful to Shin’ya for our collaborations which involved countless Skype meetings, for hosting me at Atsugi and for his invaluable feedback at every stage of my graduate career.

- Thesis committee members, **Bill Freeman** and **Polina Golland**, for advice, support and for letting me do lots of psychophysics while being in the Computer Science graduate program.

- Co-authors, **Ce Liu**, **Isamu Motoyoshi** and **Yuanzhen Li** for all the stimulating discussions.

- **Aseema Mohanty** for her help in scouring Flickr.

- For comments and feedback, **Aude Oliva**, **Michelle Greene**, **Barbara Hidalgo-Sotelo**, members of **CVCL**, **Roland Fleming**, **Molly Potter**, **Nancy Kanwisher**, **Jeremy Wolfe**, **Alvin Raj** and **Kimo Johnson**. A special thank you to Aude Oliva for use of eye tracking equipment.
Members of the T-Ruth group, past and present - Alvin, Kimo, Nat, Krista, Nadja, Ce, Yuanzhen, Lisa, Ronald, Marshall, Josh, Zhenlan - for a fun and productive environment. Alvin and Yuanzhen for sharing offices and for all the silliness.

- **John Canfield**, for being lab administrator *extraordinaire*.

- **Artesa N. Saldivar-Sali** and **Christopher Viscount Throckmorton** for sharing their deep knowledge of material properties.

- All my friends for consistently reminding me of Oscar Wilde’s words, that I am not young enough to know everything. My many housemates at Magazine St and the entire Wine Tasting crew for all the good times.

- My parents, **Basabi** and **Pankaj**, for everything.
Contents

1 Introduction .................................................. 17
  1.1 Previous work in human vision .......................... 18
  1.2 Previous work in computer graphics and computer vision . 21
  1.3 Author’s prior work .................................. 22
  1.4 Outline of thesis .................................. 24

2 Categorizing common materials .......................... 27
  2.1 Stimuli for studying high-level material categorization .... 27
    2.1.1 Determining common material categories ............... 30
    2.1.2 Our database of common material categories .......... 32
  2.2 How good are humans are high-level material categorization? 34
    2.2.1 Related work in object and scene recognition .......... 34
    2.2.2 Experiment 1: Material category judgments during RSVP . 36
    2.2.3 Experiment 2: Material category judgments during rapid presentations ............................. 39
    2.2.4 Experiment 3: Reaction times for material category judgments 43
  2.3 Summary ........................................ 44

3 Role of low-level cues in material categorization ........ 47
  3.1 Experiment 4: Are low-level cues necessary for material categorization? 48
    3.1.1 Methods ...................................... 50
    3.1.2 Results ...................................... 50
    3.1.3 Discussion .................................... 52
3.2 Experiment 5: Are texture and shape cues sufficient for material categorization? .................................................. 53
  3.2.1 Methods ......................................................................... 54
  3.2.2 Results ........................................................................ 56
  3.2.3 Discussion .................................................................... 58
3.3 Summary ........................................................................ 58

4 Role of object knowledge in material judgments ........................................ 61
  4.1 Our image database of ‘real’ and ‘fake’ objects ............................ 62
  4.2 Can observers tell real from fake in brief glances? ....................... 64
    4.2.1 Experiment 6: Descriptions of real-fake objects in brief presentations ........................................................................ 64
    4.2.2 Experiment 7: Real-fake judgments in rapid presentations ..... 68
  4.3 Mechanisms of real-fake judgments ........................................ 73
    4.3.1 Experiment 8: Reaction times for real-fake judgments ........... 73
    4.3.2 Experiment 9: Role of texture and shape cues in real-fake judgments ........................................................................ 74
  4.4 Summary ........................................................................ 81

5 Computational models and material recognition ........................................ 83
  5.1 Previous work .................................................................... 84
  5.2 Testing existing models on our databases ................................... 86
    5.2.1 Features ....................................................................... 87
    5.2.2 Classifiers ..................................................................... 88
    5.2.3 Results ......................................................................... 88
    5.2.4 Comparison with other databases ..................................... 89
  5.3 A second image database for material categories ......................... 93
  5.4 Discussion ......................................................................... 98
  5.5 Summary ......................................................................... 100

6 Conclusions ........................................................................ 107
List of Figures

1-1 Material perception is a constant part of our daily lives. 17
1-2 'Toy' worlds have been used for studying human reflectance perception. 19
1-3 Our previous work on albedo and gloss perception considered somewhat restricted real world stimuli. 23

2-1 Existing image databases for materials. 28
2-2 Stimuli for the material annotation study. 31
2-3 Results of the material annotation study. 32
2-4 Examples from our Material Categories I Database. 33
2-5 Our material categories are more diverse than previous work. 35
2-6 The RSVP procedure. 37
2-7 Accuracy at RSVP material categorization task. 38
2-8 Rapid presentation procedure for material category judgments. 41
2-9 Images used for backward masking in rapid presentation task. 41
2-10 Accuracy at rapid material categorization task. 42
2-11 Results of reaction time study for various judgments. 45
2-12 Distribution of reaction times for various judgments. 46

3-1 Stimuli used in Experiment 4. 49
3-2 Accuracy of material categorization as a function of degradations. 51
3-3 Confusion matrices for material category judgments under various degradations. 52
3-4 Control conditions in Experiment 5. 53
3-5 Accuracy of material categorization is poor when observers are only given low-level information ........................................... 57
3-6 Confusion matrices for material category judgments in severe degradations ................................................................. 57
4-1 Material recognition vs. object recognition .......................... 62
4-2 Humans make complex and often subtle discriminations about materials. 62
4-3 Examples from our Real Fake Database ................................. 63
4-4 Rapid presentation procedure for real-fake judgments ............ 65
4-5 Descriptions made by observers for an image of fake bananas ............................. 66
4-6 Analysis of descriptions of Real Fake Objects ......................... 67
4-7 Authenticity ratings for Real Fake Objects database ............... 69
4-8 Results of rapid real-fake categorization ................................ 71
4-9 Results of reaction time study for various judgments .......... 75
4-10 Distribution of reaction times for various judgments .......... 76
4-11 Stimuli used to test the role of low-level cues in real-fake judgments 77
4-12 Low-level cues are not sufficient for real-fake judgments .... 79
4-13 Low-level cues are not sufficient for real-fake judgments, even when object identity is known ........................................ 80
5-1 Outline shape is not always informative of material identity ...... 84
5-2 A variety of textures maybe observed within a single material category 85
5-3 Some pairs of material categories are easier to distinguish than others 90
5-4 Images from Material Categories II ..................................... 94
5-5 Images from Material Categories II are diverse ..................... 95
5-6 Lighting conditions for Material Categories II Database ........... 95
5-7 Photography setup for Material Categories II ....................... 96
5-8 Our material categories cannot be distinguished by classifier based on low-level image features ..................................... 99
5-9 CURET and MSRC Textile Databases are can distinguished by patch-based features ......................................................... 101
5-10 Our material category databases cannot be distinguished by patch-based features ................................................. 103
5-11 Our real-fake database cannot be distinguished by patch-based features .............................................................. 104
5-12 Reclassifying CURET and MSRC Textile surfaces into meaningful material categories ......................................................... 105
5-13 Reclassifying CURET surfaces in material categories reduces performance ................................................................. 106

6-1 Scenes made from food? ........................................................................................................................................... 109

A-1 Stimuli and results for eye movement study on shape and material judgments .............................................................. 113
A-2 Fixation duration distribution as a function of local image contrast ........................................................................ 114
A-3 Stimuli and performance in the real fake eye movement study ........................................................................... 117
A-4 Analysis of eye movements for real fake task ........................................................................................................ 118
A-5 Stimuli and results for wetness judgment task ........................................................................................................ 120
A-6 Analysis of fixations made during wetness judgments .................................................................................................... 122

B-1 Stimuli for softness judgments ................................................................................................................................ 125
B-2 Results for softness task ............................................................................................................................................ 128

C-1 Stimuli from our Toroidal Objects (“Doughnuts”) dataset .......................................................................................... 131
C-2 Results of matching description to images in “Doughnut” experiment ........................................................................ 133
List of Tables

3.1 The 8 conditions for Experiment 4 ........................................ 48
3.2 The 6 conditions for Experiment 5 ........................................ 54

5.1 Features used by our classifiers. ........................................ 87
5.2 The classifiers we used. .................................................. 88
5.3 Classification performance on pairs of materials ..................... 90
5.4 Features used to compare all databases. .............................. 92
5.5 Databases that were used in the comparison. ......................... 92
5.6 Material classification results on different databases ............... 92
5.7 Material classification results on different databases with maximum
    training set size ....................................................... 93

C.1 Example descriptions of the “doughnut” images ...................... 132
Chapter 1

Introduction

Figure 1-1: We know when (a) a sidewalk is icy, (b) a cheese is moldy, (c) a shirt is clean, (d) a potential mate has healthy looking skin, (e) a knife is rusty, or (f) if a fruit is ripe. These judgments of material appearance are common and important. (Image source: Flickr)

We can easily tell if a sidewalk is slippery, if food is fresh, if a spoon is made of plastic or stainless steel, or if a suspicious looking mole warrants a trip to the doctor. This ability to visually identify and discriminate materials is known as material
perception. Material perception is ubiquitous (see Figure 1-1) and it feels effortless. Presumably, this finely honed perception of materials evolved for practical purposes like choosing what to eat, choosing a mate or choosing a path while walking.

We interact with materials through various sensory modalities - vision, audition and touch. We are sensitive to many aspects of material appearance - optical (e.g. color, glossiness, translucency), mechanical (e.g. slipperiness, softness), chemical (e.g. ripeness of a fruit, rusting metal) and thermal (e.g. boiling water). This rich sense of materials is an important part of our lives, and yet little is known about how material perception is achieved.

In this thesis, we will focus on visual judgments of material properties. Our questions are - what can human observers perceive about materials in images such as those in Figure 1-1? How accurate are they at assessing material properties? What kinds of visual information do observers use to make material judgments? Prior research on this topic has come from several disciplines - visual psychophysics, computer vision and computer graphics. Psychophysical studies aim to uncover the relationship between visual stimulus configurations and perceptual responses to material properties. In computer vision, techniques have been developed to extract the material properties of surfaces from their photographs. In computer graphics, there is great interest in creating simulations of materials like skin, hair, fabrics and so on, that are perceptually convincing. Nearly all prior work in these fields, however, considers a restricted range of surfaces and materials (e.g. synthetic spheres, flat, uniform gray patches; see Figure 1-2). The work in this thesis is the first attempt at studying material judgments for a large selection of complex, real world materials.

1.1 Previous work in human vision

In studies of visual perception, material properties have been examined in the context of surface reflectance. Specific aspects of reflectance such as color, albedo\(^1\) and to a

\(^1\)An ideal diffuse surface reflects light uniformly in all directions regardless of the direction of incident light. This behavior can be described a number - the albedo - which is the fraction of incident light energy reflected by the surface.
lesser extent, gloss have been studied [6, 11, 13, 15, 43, 64, 79]. Figure 1-2 depicts the different types of stimuli that have been used to study reflectance perception. Most studies have assumed ‘toy worlds’ where surfaces are smooth, and have simple shapes and simple reflectance properties (see Figure 1-2a). These works have been important for establishing basic facts about surface perception. For example, it is known that perceived albedo, or lightness, of a flat, diffuse surface depends on a number of factors like the highest luminance in the scene, luminance of the largest area, configuration of the entire scene and so on [43].

Surfaces in the real world, however, often do not obey ‘toy world’ assumptions. Recent studies have attempted to move away from these simple ‘toy worlds’ by using stimuli that incorporate some of the complexity of real world conditions, including synthetic images created by sophisticated graphics software [40, 69, 74, 89, 100, 105]. Nishida and Shinya simulated locally smooth, bumpy surfaces and asked observers to make judgments of diffuse and specular reflectance [74]. They found that image-based information like the shape of the luminance histogram was correlated with observers’ judgments. Fleming et al. synthesized images of smooth, glossy spheres under real world illumination conditions and showed that illumination patterns influence the perception of gloss [40] (see Figure 1-2b). Berzhanskaya et al. manipulated the

Glossy surfaces exhibit specular reflection. Unlike diffuse surfaces, the appearance of specular surfaces is highly dependent on the viewpoint. For example, reflections in a mirror or polished metal or highlights on a shiny surface change with viewing position.
specular highlights on glossy tori and found that the perception of surface gloss tends to drop off with distance from the highlights [7]. Fleming and Bülthoff have used realistic renderings of materials like jade and porcelain to study translucency and found low-level cues to be useful [39]. Xiao and Brainard reported that observers can match color appearance across varying surface gloss conditions for spherical objects in a synthetic scene [114].

A number of studies consider the interaction of material properties with other factors that influence surface appearance. The perception of illumination direction is influenced by reflectance properties for rendered 3-D polyhedra [56]. The perception of reflectance properties of synthetic objects depends on the complexity of surface shapes [107]. The influence of material properties on surface shape is unclear. Fleming et al. suggest that specular reflections for mirror-like surfaces are an important cue for surface shape [41]. Ho et al. find an interaction between perceived surface gloss and perceived bumpiness for synthesized bubble-wrap-like surfaces [49]. Meanwhile, Nefs et al. find that the perceived shape of synthetic, globally convex objects does not depend on whether a surface is matte or glossy [71].

Synthetic stimuli are popular in studies of surface perception because appearance parameters can be varied easily. One concern when using synthetic images is that even with advanced rendering techniques, they do not look natural. Robilotto and Zaidi tried to make their experimental conditions as natural as possible by asking observers to judge the albedo of physical samples of crumpled paper, placed in front of them [89, 90]. Observers were able to do their task, although not perfectly, and seemed to use strategies based on brightness and contrast. In our previous work on albedo and gloss perception, we used photographs of hand-made surfaces with significant mesostructure [69, 100]. For surfaces viewed in isolation, simple image statistics like moments and percentiles of luminance were predictive of our observers’ responses [69, 74, 100].

What do all of these findings tell us about the mechanisms of material perception? One hypothesis, known as ‘inverse optics’, suggests that the visual system estimates the parameters of an internal model of the 3-D layout and illumination of a scene, so
as to be consistent with the 2-D retinal image. This hypothesis has had some success in explaining material judgments [13, 49, 64]. On the other hand, some have argued that in real world scenes, the surface geometry, illumination distributions and material properties are too complex and too uncertain for inverse optics to have much success. The visual system might instead use simple rules like those suggested by Gilchrist et al. for lightness computations [43] or simple image-based information like orientation flows or statistics of luminance [39, 41, 69, 74, 100]. While there is debate over which of these mechanisms is employed by the visual system, it is clear that little is known about material perception beyond the limited laboratory setups such those in Figure 1-2.

1.2 Previous work in computer graphics and computer vision

In computer graphics, the desire to create convincing simulations of materials like skin, hair, fabrics and so on, has led to several formalizations of the reflectance properties of materials. The most popular one, the bidirectional reflectance distribution function or BRDF\(^3\), characterizes the interaction of light with an opaque surface [72]. For non-opaque surfaces, effects of translucency have been modeled using an extension of the BRDF, the bidirectional surface scattering reflectance distribution function or BSSRDF [52]. The appearance of 3-D textures\(^4\) can be described by the bidirectional texture function or BTF [23].

The BRDF and its variants are of tremendous importance because they allow renderings of materials in synthetic scenes under arbitrary illumination and viewing conditions. As the BRDF is a function of four or more variables, the space of all

---

\(^3\)The BRDF specifies the proportion of light incident on an infinitesimal surface patch for any given direction, given by spherical coordinates \((\theta_1, \phi_1)\), that is reflected in any other direction \((\theta_r, \phi_r)\). The BRDF can also depend on other variables like the wavelength of light or spatial position on the surface.

\(^4\)Texture may result from variations in reflectance properties (wallpaper type) or from variations in fine-scale geometry (3D type) [23, 58]. A surface is a 3-D texture at a scale where the surface roughness can be resolved visually or by a camera. The BTF specifies the two-dimensional image (photograph) of a 3-D texture for all lighting and viewing directions.
realizable BRDFs is vast. Therefore, BRDFs are often approximated by parametric models to allow efficient rendering algorithms [46, 81, 113]. For example, parametric BRDF formulations have been developed for real world diffuse surfaces such as plaster and concrete [58, 78]. Parametrized BRDF models can represent the reflectance properties of several common materials effectively, but they fail to capture a range of real world reflectance phenomena.

An alternative approach has been to use empirically measured BRDFs that make renderings of many complex materials possible. The BRDF of a surface can be measured in the laboratory by a special device, the gonioreflectometer, or estimated directly from images (photographs) of the surface [12, 24, 25, 65, 66, 75, 86, 95, 106, 115]. Image-based techniques for BRDF estimation can be viewed in the ‘inverse optics’ framework. The various techniques differ in the assumptions they make about the illumination, geometry or material properties in the world. Given their assumptions, they fit BRDF parameters to image data. Although BRDF representations are important for rendering purposes, it is reasonable to ask if they are necessary for recognition. Recreating the appearance of a surface might be harder than simply recognizing it, and perhaps simpler representations of surface reflectance would suffice for recognition purposes.

There has been some work on the problem of 3-D texture recognition that has employed filter and patch-based image features to recognize instances of textures like plaster, wool, cork and so on [21, 23, 109, 110]. This work is related to, but does not address the problem of material recognition. A single material category like fabric can contain samples with widely varying 3-D textures and these findings are based on distinguishing texture appearance rather than identifying material classes.

1.3 Author’s prior work

In our earlier work, we have suggested that instead of doing inverse optics, the visual system computes statistics of the 2-D image in a cue-based approach to material perception [69, 99, 100]. It is reasonable that the visual system will use heuristics
based on statistical cues when these cues are informative. There is evidence to suggest
that such statistics may be computed pre-attentively and might mediate perception
of textures and scenes, and might guide visual search [3, 5, 16, 19, 77, 87, 92, 93, 98].
Nishida and Shinya have shown that the shape of the luminance histogram is a cue
to surface gloss and albedo [74]. Dror et al. demonstrated that statistics of real
world illumination can be used to classify images of spheres as shiny, matte, white,
grey, chrome etc. [31]. Pont and Koenderink showed that statistical measurements
on images of rough surfaces can be used to predict the nature of illumination and
surface roughness [82].

Figure 1-3: We have previously shown that for photographs of opaque, glossy surfaces
with surface mesostructure like the ones shown here, a computational model based on
the moment and percentile statistics of the luminance histogram and filter outputs
can predict the perception of albedo and gloss [69, 99, 100].

We hypothesized that there are some relatively simple image measurements that
can provide useful information about surface properties. We considered photographs
of hand-made stucco surfaces and materials like paper, modeling clay, soap, candies,
etc. [69, 99, 100]. These surfaces were opaque, glossy and possessed significant surface
mesostructure (see Figure 1-3). We found that simple statistics of the photographs of
such surfaces, like the moment and percentile statistics of the luminance histogram
or filter outputs, can provide useful information about physical albedo and gloss. We also showed that the same statistics are correlated with human judgments of albedo and gloss. In particular, skewness of the luminance histograms and filter outputs is informative. We demonstrated an aftereffect; adapting to skewed random noise alters the perception of lightness and gloss of a surface viewed subsequently. We proposed a neural computational model for skewness that employs center surround cells and found that center surround cells are better at conveying skewness than Gabor-like cells.

Given the significance of these statistical cues, one might ask what happens when these statistics are modified in an image. Previous work in texture analysis and synthesis shows that imposing statistics on the histograms of an image alters texture perception [48, 83]. Nishida and Shinya demonstrated that applying a look-up table or equivalently changing the shape of the luminance histogram of an image of a surface affects the perceived albedo and gloss [74]. In our work, we developed a modification of the Heeger-Bergen algorithm to manipulate the percept of albedo and gloss [100]. Fleming et al. have shown that similar manipulations affect translucency [39]. Thus, there is a strong connection between these simple histogram statistics and the perception of reflectance at least for the surfaces that we considered.

1.4 Outline of thesis

In this thesis, we depart from what has been the standard approach for studying material perception in a significant way. Most psychophysical studies, some of which were described in the Section 1.1, have considered restricted classes of materials (see Figures 1-2, 1-3). Controlled stimuli are a popular tool to study human vision, but there is also increasing interest in the perception of “natural” images i.e. scenes in the real world that contain man-made and natural objects. Great strides have been made in the fields of object recognition and scene perception by using photographs of the real world. There exist large image collections of object and scene categories that are used in both the psychophysics and computer vision communities [10, 29, 34, 35, 94].
It is unfortunate that there are no such resources to study material perception.

In this work, we introduce new image databases for real world material categories like plastic, metal, glass, wood and so on. We collected these images from online photosharing sources like Flickr.com and by taking photographs of material samples in the laboratory. These databases allow us to study high level judgments of material category, in contrast to past work that has focused entirely on estimating reflectance parameters. In Chapter 2, we present experiments that establish that observers can make these judgments of material category reliably, quickly and in challenging conditions. It is well known that observers can categorize scenes and objects in images that they have never seen before, even when stimuli are presented very briefly [9, 8, 51, 76, 84, 85, 97]. Our results indicate that material categorization can, similarly, be fast and flexible.

Next, we ask how do our observers make these rapid judgments of material category. We examine two extreme possibilities that might explain observers’ performance. One possibility is that observers employ simple, low-level cues like color or high spatial frequencies e.g. wood is brown or fabrics have fine details. The experiments described in Chapter 3 show that categorization performance cannot be explained by such simple cues. Another explanation of our results can be that observers use high-level knowledge about object identity to categorize materials e.g. buckets are made from plastic or chairs from wood. In Chapter 4, we consider special categories of objects - ‘fake’ fruits, flowers and desserts, and demonstrate that shape-based object identity does not account for material categorization performance either.

In Chapter 5, we evaluate state-of-art techniques in computer vision for the purpose of material categorization. There is little prior work in computer vision on the topic of material recognition, except on the specific problem of 3-D texture recognition. We find that current techniques in computer vision are inadequate for categorizing materials on our databases. In Chapter 6, we discuss our findings and conclude that material recognition is a rich and challenging problem domain and there is much ground to be covered in both visual perception and computer vision.
Finally, in the interest of clarity, we now discuss what we mean by the terms \textit{texture}, \textit{shape} and \textit{object} in the context of material recognition. We will use an operational definition of texture, inspired by recent work in texture synthesis techniques [83, 32]. The class of images that can be successfully reproduced by these texture synthesis methods (e.g. a periodic wallpaper pattern) is what we will refer to as a ‘texture’. The distinction between material and texture in real world surfaces is tricky and often we will evoke this operational definition to separate the two. By shape, we will mean the 3-D physical layout of surfaces and by object, we mostly mean shape-based identity. For example, a chair is instantly recognizable by its typical shape and silhouette, as are trees and mugs and cars. In this work, we will assume that a ceramic mug and a plastic mug belong to the same object class, but different material categories.
Chapter 2

Categorizing common materials

When we look at the world around us, we can easily parse it into objects and surfaces, as well as into materials. It is almost trivial to identify if a desk has a wooden surface, a laptop casing is made of painted metal, or a carpet is made from soft, springy fibres. Almost nothing is known about how good we are at making these judgments or how quick. For objects and scenes, it is known that category information (e.g. face or guitar, desert or beach) can be identified in exposures as brief as 50 ms [8, 9, 44, 45, 51, 76, 84, 85]. Can information about material properties be extracted in the same way? Inspired by the work in object and scene recognition, we start by studying judgments of high-level material categories such paper, plastic, fabric, etc. These judgments are common in our visual experience and allow us, as experimenters, to move beyond the constrained stimulus worlds in past work to real world images.

2.1 Stimuli for studying high-level material categorization

Before we can do any psychophysical studies, we need to make decisions about which stimuli to use. There exist far fewer image databases that focus on materials than ones that contain object or scene categories [10, 29, 34, 35, 94]. In the field of computer graphics, descriptions of real world materials exist mostly as BRDF measurements
rather than as photographs [24, 66]. An exception to this trend, is the CURET database that was created by Dana et al. [23] CURET is a database of BTFs i.e. a set of photographs of 61 different material samples illuminated and viewed from many directions (see Figure 2-1a). This database has been popular for 3-D texture modeling and recognition [21, 73, 82, 109, 110].

![Image of material samples](image1.png)

**Figure 2-1**: Image databases that contain material categories: (a) CURET [23] (b) Microsoft Textile Database [96] (c) KTH-TIPS2 database [17] (d) Dror et al. Spheres Set [31].

The CURET database contains a range of real world materials like sponge, cork, aluminium foil, plaster, fabrics and so on. CURET samples vary in reflectance properties, surface structure and spatial homogeneity. In spite of this diverse collection of material samples, CURET is not ideal for studying material recognition due to poor intra-class variation. For example, there are over 200 images of the sponge in
Figure 2-la from different lighting and viewing angles, but there is only one sponge in the database. So, sponge as a material category is poorly represented although the specific sponge has been captured in great detail. This aspect of CURET is not surprising, it was developed for rendering purposes and not for testing recognition.

Caputo et al. constructed an image database of 11 materials, KTH-TIPS2 to overcome some of the limitations of CURET [17]. They used 4 samples per material category (see Figure 2-1c) to increase intra-class variation and photographed the samples in various pose, illumination and scale conditions. The Microsoft Textile database contains photographs of 16 samples of folded textiles (see Figure 2-1b) under diffuse illumination conditions [96]. Dror et al. used photographs of 9 spheres made of different materials such as polished metal, plastic and so on (see Figure 2-1d) [31].

All of these databases contain rather few samples in each material category, all of which look quite similar. This weak intra-class variation is troubling. If we want to study the visual perception of everyday materials, we need databases that contain more than a few surface samples in canonical lighting and viewing positions. It is important to note, however, that the databases in Figure 2-1 were constructed to suit the requirements of computer vision algorithms. These databases were not developed to test or measure human perception.

As there seem to be no appropriate databases for material recognition, one might wonder if object recognition databases can be used instead. There has been great interest in building large databases of objects - CalTech 101 [35], PASCAL [34], LabelMe [94], CBCL Street Scenes [10] and ImageNet [29] are examples. These databases contain thousands of images and on the order of 10 to 200 object categories. Some are available with detailed ground truth annotations and are organized into meaningful conceptual hierarchies [29, 94]. While it is possible to sift through these large image collections and re-organize images into material categories, there are two issues. First, the pixel resolution available for all objects made of, say plastic, varies greatly depending on the context of the object. Since these databases are tailored for object and scene recognition, they do not include many high resolution close-ups. Second, these databases contain great intra-class variation in object and scene
categories but not necessarily material categories. If we want to study the perception of materials systematically, we need new databases that are specifically designed for that purpose.

2.1.1 Determining common material categories

In order to build an image database of material categories, it is reasonable to ask which material categories do humans encounter and recognize in daily life? Little is known about this question. Unlike objects, there is no prior work on distinctions between basic and subordinate level material categories [91]. Rather than arbitrarily picking material categories for our database, we conducted an annotation study to determine the most common materials in everyday experience.

We collected a set of 1000 photographs of daily scenes from the photo sharing website, Flickr.com under the Creative Commons License. To construct a set of images that were representative of the daily experience of an observer, we searched for images that conveyed the following settings - Street, Buildings, Office, Kitchen, Bedroom, Bathroom, People, Shopping, Outdoors and Transportation. In addition to these keywords, we searched for Food, Drinks, Kitchen Equipment, Clothes, Computers, Skin, Store Inside, Trees, Sky and Seat to incorporate close-up images of materials in each of the settings. We collected 50 color images for each of these 20 keywords (see Figure 2-2). The resolution of the images ranged from 300 x 450 to 1280 x 1280 pixels.

Five naive observers participated in this study. Observers were asked to annotate materials in each image for as many images as they could finish in one session. They were given as much time as needed to annotate each image. The order of presentation of images was randomized for each observer. All observers were told to focus on materials that occupied the greatest pixel areas first. They were also told to provide the higher-level category if they were not sure of precise material identity e.g. metal if confused between tin and aluminum.

Observers 1, 2 and 3 annotated non overlapping sets of 300, 299 and 221 images respectively. Responses from these observers were used to create a list of suggestions
for Observers 4 and 5. This was done to facilitate faster annotations since typing is slower than choosing from a pull down menu. Observers 4 and 5 annotated all 1000 images.

Annotations were interpreted by us so that misspellings (e.g. mettal) or scattered instances of subordinate category labels (e.g. aluminum) were counted towards the intended category (e.g. metal). Figure 2-3 plots the frequency of occurrence for the 20 most frequent categories. We see that metal, fabric, wood, glass, plastic and stone are much more frequent than other categories.

We take these results as a first step in the direction of understanding which material categories are common in the real world. There are many choices for experimental designs and assumptions when asking such a question. Our choices were influenced by the desire to create a database of material categories that is representative of an average observer’s experience. It is likely that a different selection of annotation stimuli would yield a different order in Figure 2-3. When analyzing the annotation data, we found that some of our observers named more categories than others (e.g
tin, aluminium, bronze vs. metal). On many images, observers did not use material-specific words, and used object names instead (e.g. wall, food). These observations are important because they tell us something about inter-observer variability on this task and about the deeper question of naming ‘things’ vs. ‘stuff’. We leave the pursuit of these questions for future work. For now, we have a candidate list of material categories that we can begin to study.

![Graph](image)

Figure 2-3: Results of annotation study. Frequency of occurrence per image averaged over all five observers is plotted for the 20 most frequent categories.

### 2.1.2 Our database of common material categories

Based on the results of the annotation study, we constructed a database of 9 material categories - Metal, Fabric, Wood, Glass, Plastic, Stone, Paper, Water and Leather. In choosing this set, we had to balance many constraints such as the availability of images and the diversity of appearance within a category. For example, Greenery, as a category is not as diverse as Plastic or Fabric.

Color photographs of all 9 chosen materials were acquired from the photo sharing
Figure 2-4: Examples from our Material Categories I database. (a) and (b) From top left to bottom right - Fabric, Glass, Leather, Metal, Paper, Plastic, Stone, Water and Wood.

website Flickr.com under the Creative Commons License. There were a total of 100 images in each category. Within each category, 50 images were close-up shots of that particular material and 50 images contained objects made of that material (see Figure 2-4). We made this distinction in order to cover a wide range of appearances for each material. We only chose images that contained a single material in the foreground, so each image was associated with a unique category. All images were cropped down to 512 × 384 pixel resolution.

We took great care in selecting the images. Each image was selected manually from 50-odd candidates to ensure a range of illumination conditions, viewpoints, surface geometries, backgrounds and sub-categories in every material category. This was done to reduce the chances that simple low level information like color could be used to distinguish the categories. Figure 2-5 demonstrates the intra-class variation of our database. It is useful to compare Figure 2-5 to Figure 2-1.

This database will be referred to as Material Categories I in all subsequent sections of this thesis. The nomenclature was chosen to distinguish it from a database to be described in a later chapter, that also contains similar material categories. For all the images in this database, we created binary masks to isolate the pixels that contain the material of interest from the background. These masks allow us flexibility in
experimental design and the ability to use computational models with ease.

2.2 How good are humans at high-level material categorization?

Armed with our image database, we are now in a position to conduct psychophysical studies. We know from our everyday experience, that telling plastic from fabric, is an easy task. Is that true for our challenging and diverse set of images? Is it still true when images are presented rapidly? We start with a discussion of related work on high-level categorization of objects and scenes, and then describe our experiments.

2.2.1 Related work in object and scene recognition

It is known that human observers can quickly extract a lot of information from natural images that they have never seen before. There is a long history of studying such ‘rapid perception’ [9, 8, 51, 84, 85]. Recent work has involved testing the limits of human vision with challenging tasks, such as “is there an animal in this scene?” where the scenes are diverse [4, 27, 28, 33, 37, 104, 108]. Efforts have been made to nail down processing speed by multiple techniques, including reaction times [57] and ERP techniques. Kirchner and Thorpe find reaction times for an animal/no-animal task on the order of 120msec. Grill-Spector and Kanwisher showed that subjects can categorize briefly flashed images into a small set of basic-level object categories almost as fast as they could distinguish an object from noise [45].

Another set of studies has attempted to answer the question - what information do observers extract in brief stimulus presentations? For natural scenes, the roles of spatial frequencies, color, scene “gist” and global properties like openness or navigability have been examined [44, 76, 97, 103]. When observers were asked to report freely about what they saw, Fei-fei et al. showed that a remarkable amount of detail about objects and scenes is perceived even in a glance [36]. The work in rapid perception has sometimes been controversial. There are debates about the inferences about
Figure 2-5: Here are four examples from the (a) Plastic and (b) Fabric categories. In comparison to other databases (see Figure 2-1), samples from our database exhibit greater within-class diversity.
neural mechanisms [53, 54] and about the connection of these phenomena to issues such as cognitive hierarchies [63, 91]. However, everyone agrees that the findings are important and provocative, and that they force a re-evaluation of many common ideas about visual recognition.

Unfortunately, almost nothing is known about the recognition of real world materials, especially the rapid recognition that would be most useful in daily life. In next section, we describe an experiment that measures rapid recognition of material categories using methods that have been employed to study objects and scenes [84, 85].

2.2.2 Experiment 1: Material category judgments during RSVP

In the rapid serial visual presentation (RSVP) paradigm, observers view a rapidly presented sequence of images, and make judgments about that sequence. There is a rich body of work that shows observers can reliably detect the presence of numbers, words, high-level object and scene categories in RSVP sequences [33, 84, 85]. The RSVP paradigm is useful for testing the limits of human cognitive abilities. In Experiment 1, we used an RSVP design to test material categorization performance on our database. Knowing what observers can or cannot do is important; it can help shape our thinking about how the visual system recognizes materials.

Methods

Four of the categories from our Material Categories I database were chosen as targets (glass, metal, paper and wood) in an RSVP design. The other five categories (fabric, leather, plastic, stone and water) always served as distractors. On each trial, the observer was told the target category (e.g. glass), and then viewed a sequence of 9 images (Figure 2-6). The observer's task was to detect if the target was present in the sequence. The target was present in 50% of the trials and appeared randomly between positions 3 and 7. When the target was present, the sequence contained images from
all 9 material categories. When the target was absent, there were 8 materials in the sequence of 9 images. The target was replaced by a distractor, derived from a pool of 50 images similar to, but not in the database of 900 images. To make sequences as similar as possible, we normalized the mean luminance and Michelson contrast of all 900 images in our database to constant values.

The sequences were presented at either 40 ms per image or 160 ms per image. Each observer performed 400 trials, with 200 trials each for two target materials. For each target, an observer performed 100 trials for every presentation time. These 100 trials, in turn, were evenly split between “Close-up View” and “Regular View” images. Once an image appeared it did not reappear for at least 50 trials. No sequence of 9 images was ever repeated. Trials were blocked by presentation time, target type, and image type (close-up/regular). Presentation order was counterbalanced across observers.

Figure 2-6: The RSVP procedure. Before the sequence, observers were instructed to look for a particular material, e.g. paper. At the end of the sequence, observers responded as to whether they believed the target was present or absent.

Six observers (Group A) searched for glass and paper. Six others (Group B) searched for metal and wood. Before the experiment, observers were shown example images from all 9 material categories. They were also given a brief practice session of 10 trials. Stimuli were displayed centrally against a mid-gray background and
subtended 15 × 12 deg.

Unless mentioned otherwise, all psychophysical experiments presented in this thesis were implemented in MATLAB using Psychophysics Toolbox [14]. Experimental procedures were approved the MIT Committee on the Use of Humans as Experimental Subjects (COUHES).

Results

Figure 2-7 plots average accuracy for each condition. We performed a 2-way ANOVA with factors Time (2 levels = 40 ms, 160 ms), and Zoom (2 levels = Close-up, Regular). All main effects were significant (Time F = 90.63, p ≪ 0.0001, M(40ms) = 0.597, M(160ms) = 0.749; Zoom F = 4.38, p = 0.0392, M(Moderate View) = 0.689, M(Object View) = 0.656). Observer performance was better than chance in all four conditions of Time × Zoom (1-tailed z-test, p < 0.001).

![Figure 2-7: Accuracy at detecting material targets in RSVP sequences. Close-up View images (green) and Regular View images (red). Error bars are 1 s.e.m. Dashed black line shows chance = 50%.](image)

Figure 2-7: Accuracy at detecting material targets in RSVP sequences. Close-up View images (green) and Regular View images (red). Error bars are 1 s.e.m. Dashed black line shows chance = 50%.
Discussion

The results of Experiment 1 show that observers can categorize materials even in challenging conditions (40 ms = 25 images/sec). Note that performance is actually better for Close-up View images, suggesting that observers are not merely doing classic shape-based object recognition. Though we do not here present separate analyses for different material types, trends suggested that some material categories were easier than others and that image type (close-up/regular) interacts with target material category.

The performance of our observers establishes that material categorization is possible even in the rapid presentations. Their performance is certainly impressive given the sheer diversity of the images in our database. Evans and Treisman found performance comparable to our observers at 160 msec, for an animal/no-animal RSVP task when each image was presented for 75 msec [33]. While it is hard to compare numbers across tasks (e.g. fabric vs. non-fabric, animal or no animal) and stimuli (e.g. our images, Corel database), it is fair to conclude that material categorization can be rapid like object and scene categorization. There are a number of known effects in the RSVP paradigm, such as the position of the target image and the interaction between images within a sequence. For example, if images preceding the target image share features that are relevant for categorization, then performance at target detection will be affected. An analysis of these effects on our diverse image database is not fruitful because we are not in a position to even know which features are useful for material identification. Experiment 1 is a broad stroke at attempting to uncover the time course of material judgments.

2.2.3 Experiment 2: Material category judgments during rapid presentations

In Experiment 1, we tested if judgments of material category were possible in a challenging paradigm like RSVP. In this experiment, inspired by studies on object and scene recognition, we explore the time course of material judgments further by
presenting images rapidly in a fixed sequence [44, 45].

Methods

All nine categories from our Material Categories I database were chosen as targets. Half the images in the database served as targets and the rest served as distractors. On each trial, the observer was told the target category (e.g. glass), and then viewed a sequence of 5 images (Figure 2-8) - first, the target (or distractor) followed by four masking stimuli. The observer’s task was to detect if the target was present in the sequence. The target was present in 50% of the trials. Images used for backward masking were derived from our Material Categories I database using the Portilla-Simoncelli texture synthesis algorithm [83] (Figure 2-9). Greene and Oliva have used similar masking stimuli in rapid scene categorization tasks [44]. These masks are matched to the stimulus images in terms of low-level, statistical image measurements and provide more effective masking than other choices such as pink noise. Stimuli were displayed centrally against a mid-gray background and subtended 15 x 12 deg.

Three observers participated. The stimuli were presented at either 40, 80 or 120 ms per image. Each observer performed 900 trials, with 100 trials each for target material. These 100 trials, in turn, were evenly split between “Close-up View” and “Regular View” images. Images were never repeated. The 9 material categories were split into three equal groups - those that were displayed for 40 ms, for 80 ms and for 120 ms. For each target material, distractors were uniformly distributed across the remaining 8 material categories. The split of the database into target images and distractor images and presentation order was balanced across observers. Trials were blocked by presentation time, target material, and image type (close-up/regular).

Results

Figure 2-10 plots observers’ accuracy as a function of stimulus presentation duration. We performed a 2-way ANOVA with factors Time (2 levels = 40 ms, 80 ms, 120 ms), and Zoom (2 levels = Close-up, Regular). The effect of presentation time was significant (Time F = 11.31, p = 0.0001, M(40ms) = 0.833, M(80ms) = 0.881, M(120ms) =
Figure 2-8: Rapid presentation procedure for judgments of material category. Before the sequence, observers were instructed to look for a particular material, e.g., paper. At the end of the sequence, observers responded as to whether they believed the target was present or absent.

Figure 2-9: Images used for backward masking were derived from the Material Categories I database using Portila-Simoncelli texture synthesis method [83].
0.931; Zoom F = 0.05, p = 0.8263, M(Close-up View) = 0.880, M(Regular View) = 0.884). Categorization performance was significantly better than chance (= 50%) and below ceiling (= 100%) for all presentation times (1-tailed z-test, p < 0.001), except at 120 ms, where performance is at ceiling. Planned comparisons with Bonferroni correction for alpha revealed a significant increase in performance from 40 to 80 ms, and 80 to 120 ms (40 vs 80 ms, t = -4.74, p ≪ 0.0001; 80 vs. 120 ms, t = -4.965, p ≪ 0.0001).

![Figure 2-10](image)

Figure 2-10: Accuracy at detecting material targets in rapid presentations averaged across observers and material categories. Error bars are 1 s.e.m. Dashed black line shows chance (= 50%).

**Discussion**

The results of Experiment 2 reinforce what we learnt from Experiment 1, that observers can rapidly categorize materials. Our observers achieve 83% performance in a 2-AFC task at 40 msec presentations. Greene and Oliva find that observers categorize scenes in a 2-AFC design at 75% performance in as little as 30 ms image exposures. Bacon-Mace *et al.* have reported similar numbers. The findings of Experiments 1 and
2.2.4 Experiment 3: Reaction times for material category judgments

In Experiments 1 and 2, we used limited time stimulus presentations to measure material perception abilities. The results of that experiment tell us how long the stimulus needs to be visible in order to categorize materials, but not how long it takes observers to finish processing material properties. Does detecting paper in a stream of other materials require introspection? Or can it be a snap judgment? We can answer that question by measuring observer reaction times (RTs). RTs provide an upper bound on how long the brain processes stimuli to reach a certain level of behavioral performance.

Methods

We used images from the Material Categories I database. Observers were asked whether a certain target (e.g. paper) was present in each trial. Each trial started with a fixation target in the center of the display. Observers pressed a key to display the stimulus. The task was to respond target present or absent as quickly and accurately as possible. Reaction times (RTs) greater than 1 second were discarded. Feedback, in the form of auditory beeps, signaled an incorrect or slow response.

Trials were blocked by target category, with a total of 8 blocks. On 2 of the blocks, observers performed easy target detection tasks (red circle vs. blue circle; bar tilted at 45° or -45°) and RTs on these served as baseline. In both blocks, the 4 x 4 deg stimulus was presented centrally, and the target was present in half the trials. Each block consisted of 50 trials.

In the other 6 blocks, observers looked for a specific material category, the target. For each target, the distractor images were chosen from the other 8 materials. Each observer was assigned three target categories. Each block consisted of 50 trials and the
target was present in half of the trials. All stimuli were 15 x 12 deg. Three observers participated in this experiment. Presentation order and presentation conditions of the blocks was counterbalanced between observers. Examples of targets were shown in the beginning of each block and at the beginning of each trial observers were reminded of the target they were looking for.

Results

Figure 2-11 plots errors at detecting targets versus median RTs for each block. In all cases, observers do well (chance = 50% error for material categorization). RTs and error rates follow the same pattern; there is no evidence of a speed-accuracy tradeoff. A one-way ANOVA with median RT as the dependent variable and task as the factor with 8 levels showed no significant effect of task (F = 2.34, p=.06). Figure 2-12 plots the distribution of reaction times for the baseline tasks (red vs. blue circle, left vs. right oriented bar) and the material category judgments averaged across target material categories.

Discussion

These results establish that not only can observers perceive material properties in fast presentations, they can judge them quickly too. In the case of binary material categorization, the response time is very similar to that for baseline tasks.

2.3 Summary

We have demonstrated evidence that material perception can have the same rapidity that has been documented in object recognition and scene perception. The experiments presented in this chapter concern the rapid perception of material properties, when the stimulus is presented too fast for any eye movements or saccades to happen. Our data make a convincing case for the hypothesis that eye movements are not necessary for many material judgments. It is still reasonable to ask - where do people look when they can make eye movements while judging materials. In Appendix A,
Figure 2-11: Errors (%) made by four observers as a function of their median reaction times (seconds). Chance performance corresponds to 50% error. Red asterisks denote the baseline categorization tasks while the green circles correspond to 6 blocks of the material categorization task for each observer.
we present three additional experiments that examine the role of eye movements. We find no clear evidence for observers' strategies.

In order to understand how our subjects are able to categorize materials quickly and in brief exposures, we examine two explanations. The first, pursued in Chapter 3, concerns the role of low-level cues. If subjects are using something simple like color or simple texture features to recognize material categories, we can test that by manipulating these low-level cues in images. In Chapter 4, we explore a different hypothesis, one based on high-level object knowledge. It is possible that observers use shaped-based object recognition to decide material categories in Experiments 1, 2 and 3. The results in Chapters 3 and 4 indicate neither of these explanations are adequate for explaining our observers' performance.
Chapter 3

Role of low-level cues in material categorization

One reason why our observers do so well in the challenging RSVP task of Experiment 1 might be that they employ low level cues e.g. wood is usually brown or fabric has fine details. If performance is purely due to low-level cues, then altering those cues should cause precipitous performance drops. Even if material categorization performance is not merely due to low-level cues, altering those cues allows us to test the importance of a given cue to a given material judgment; are high spatial frequencies, for instance, important for distinguishing between paper and fabric? Is color important for telling plastic from glass?

Similar strategies have been used in object and scene recognition to get at the underlying visual mechanisms. Schyns and Oliva juxtaposed the spatial frequency components of two different scene categories and found that lower spatial frequencies are accessed first [97]. Oliva and Schyns have studied the role of color in scene categorization [76]. Greene and Oliva demonstrate that global properties of scenes like openness and navigability can mediate rapid scene categorization [44]. For objects, recent work has examined rapid animal detection in situations when stimuli are degraded severely to control the cues available to observers [111, 70]. Elder and Velisavljevic manipulated color, texture, luminance and shape cues independently in photographs and found that shape cues were most useful for rapid animal detection
Table 3.1: The 8 conditions for Experiment 4. Observers responded with a 9-way judgment of material category. Presentation time varied from as long as required (A) to 160 ms (D, F and H). For 160 ms presentations, a pink noise mask followed and was displayed for 1 sec. Images were either color (A,B) or grayscale (C-H), some with additional degradations (E-H).

[111]. Nandakumar and Malik showed images that were degraded by blurring, quantization, spatial or contrast inversion, and showed that categorization performance is robust to these degradations [70].

We employed these ‘lesioning’ strategies in the experiments presented in this chapter. In Experiment 4, we asked if low-level cues like color, high spatial frequencies or luminance contrast are necessary for material categorization by removing these cues in images. In Experiment 5, we asked if simple cues like color and somewhat more complex cues from texture and shape are sufficient for material categorization by constructing images that emphasized only one of these cues.

3.1 Experiment 4: Are low-level cues necessary for material categorization?

In this experiment, we asked observers to categorize materials in conditions where color, high spatial frequencies and the sign of luminance gradients were manipulated.
Figure 3-1: Example of the manipulations in Experiment 4 for an image of plastic toy. (a) Original color version (b) Grayscale version (c) Blurred grayscale version and (d) Inverted grayscale version. Observers viewed only one of these four versions for a given original image.
3.1.1 Methods

This experiment again employed our Material Categories I database. The observer’s task was to identify the material category for each image - fabric, glass, leather, metal, paper, plastic, stone, water or wood. There were 8 viewing conditions (see Table 3.1), differing in presentation time, presence of masking, presence of color, presence of high spatial frequencies and sign of luminance gradients. 18 observers participated. Six of these observers participated in 2 of the 8 conditions while the rest participated in just one condition.

The 900 images of the 9 materials were randomly divided into 3 non-overlapping sets of 300 images each. Each block consisted of categorizing one of these sets, under one condition. Observers participating in more than one block categorized a different set for each block. For any given observer, the presentation time for all blocks was the same.

3.1.2 Results

Figure 3-2 plots the material categorization accuracy, averaged across observers for all eight experimental conditions. In all conditions performance is significantly above chance (11%) and below ceiling (100%) (1-tailed z-tests, p < 0.00001 in all cases). Performance drops significantly when color images are converted to grayscale (Conditions B & C, 2-tailed z-test z = -4.027, p = 0.0001), when the images are blurred (Grayscale to Grayscale with blurring, Conditions C & E, 2-tailed z-test z = -6.359, p < 0.00001), or when contrast is inverted (Grayscale to Inverted Grayscale, Conditions C & G, 2-tailed z-test z = -8.677, p < 0.00001). The accuracy also decreased when the presentation time was reduced from 1 second to 160 ms (Grayscale, Conditions C & D, 2-tailed z-test z = -8.11, p < 0.00001; Grayscale blurred, Conditions E & F, 2-tailed z-test z = -7.555, p < 0.00001; Grayscale inverted, Conditions G & H, 2-tailed z-test z = -7.221, p < 0.00001). However, there was no significant difference in accuracy between 1 sec and unlimited viewing conditions (Conditions A & B, 2-tailed z-test z = -0.815, p = 0.4148).
Figure 3-2: Accuracy of material categorization as a function of experiment conditions listed in Table 3.1 (from left to right A-H). The conditions are grouped by the 4 image degradations (Color, Grayscale, Grayscale Blurred, Grayscale Inverted). For each degradation, the bar on the left corresponds to the longer presentation time. All error bars are 1 s.e.m. Dashed black line indicates chance performance (= 11%).
3.1.3 Discussion

The results of this experiment demonstrate that while low-level cues influence material categorization, they do not completely determine observer performance. Even with severe degradations and rapid viewing conditions, observers are doing much better than chance performance.

![Confusion matrices](image)

Figure 3-3: Confusion matrices averaged across observers for all 8 experiment conditions listed in Table 3.1. The order of material categories is fabric, glass, leather, metal, paper, plastic, stone, water and wood. The columns in each image represent the averaged responses to all images of the corresponding material. So column 1 contains all responses to images of fabric, column 2 to glass and so on.

It is interesting to ask what errors observers make when they misclassify a material. Figure 3-3 displays the confusion matrices averaged across observers for the eight experimental conditions. On inspecting these matrices visually, one notices that the diagonal is dominant in all conditions; observers, again, are good at this task. Except for certain confusions specific to particular observers (e.g., glass vs. water, Condition A), the confusions become greater with increasing degradations to the image.
3.2 Experiment 5: Are texture and shape cues sufficient for material categorization?

In this experiment, we asked observers to categorize materials in conditions where color, texture and shape information was manipulated in a manner more extreme than Experiment 4. For an example, consider Figure 3-4. Our goal is to test the contribution of each of these cues for material categorization.

Figure 3-4: Example stimuli that were used in the conditions listed in Table 3.2. Shown here is an image of green stone and the derived images for each control condition. The resolution of the images presented to the subjects is much higher than in this figure. The effects of image resolution are particularly noticeable for the Color II condition, which is much noisier when viewed in experimental conditions.
Table 3.2: The 6 conditions for Experiment 5. Observers responded with a 9-way judgment of material category. Presentation time was as long as was required. All images presented were derived from the original images in the database to emphasize color (C1, C2), texture (T1, T2) or shape information (S1, S2). Descriptions of these image manipulations can be found in accompanying text.

### 3.2.1 Methods

This experiment again used our Material Categories I database. The observer’s task was to identify the material category for each image - fabric, glass, leather, metal, paper, plastic, stone, water or wood. There were 6 viewing conditions (see Table 3.1), differing in presence of color, texture and shape information. Five observers participated in all 6 conditions.

The 900 images of the 9 materials were randomly divided into 6 non-overlapping sets of 150 images each. Each block consisted of categorizing one of these sets, under one condition. Each observer completed 6 blocks and categorized a different set of images in each block. Observers were given as much time as needed to categorize images in all blocks. We want to know if observers can do the task at all on our challenging images (see Figure 3-4).

To obtain the images in Figure 3-4, we used the following steps. For Color-emphasis 1 condition, the R, G and B values all pixels in the region containing the material in each image were clustered using the k-means algorithm implementation in MATLAB to identify between 3 to 5 dominant colors. These dominant colors in each image were converted into a visualization inspired by work in information visualization [47]. Each color was conveyed by a circle with radius proportional to
the size of the cluster it represented in RGB space.

In Color-emphasis 2 condition, for each image, we cut up the region containing the material into 16 × 16 image patches. These patches were then scrambled and recombined using techniques from patch-based texture synthesis work in computer vision [32]. The MATLAB implementation for combining patches to minimize differences in pixel values across a seam was made available by Alvin Raj. The order of the image patches in the final image were chosen in locally optimal way, by selecting the patch that fit its neighbors the best. Once this ‘quilted’ version of each image was created, we added pink noise to suppress any texture information that could be gleaned. For the Texture-emphasis 1 and 2 conditions we used the same methods to divide images into patches and recombine them. The difference is that in the Texture-emphasis conditions, no pink noise was added and in the Texture-emphasis 1 condition, patches were arranged in a random order.

Images in the Shape-emphasis 1 condition were simply the binary hand-segmented masks that we created for all images in our database. The mask conveyed the outline or silhouette of the regions containing the material of interest. In the Shape-emphasis 2 condition, we created ‘cartoons’ of all our images using three steps. In the first step, all images were converted to grayscale and filtered using a median filter of support 20 × 20 pixels. This size was sufficient to remove high spatial frequency informations from textures. In the second step, we used the Stamp filter in Adobe Photoshop CS3 to create a black-and-white version of the image that marks strong edges in black and the rest of the image in white. This operation produces a line drawing-like image. In the third step, we combined the median filtered image, the output of the Stamp filter and the binary masks in Shape-emphasis 1 to create an image where the background is blue and the material itself is represented by black solid lines and grayscale shading information.

Our choices for color, texture and shape controls are reasonable ones. Given the diversity of our images it is challenging to come up with a single method for isolating a given cue like color or texture perfectly. Indeed, separating a single image into its shading and reflectance components is a topic of active research [102]. We chose two
conditions to test each cue, one of which was richer than the other. Color-emphasis 2 retains more information about the color distributions in the original image than Color-emphasis 1. Texture-emphasis 2 contains a more globally-ordered texture pattern than Texture-emphasis 1. Shape-emphasis 2 conveys shading and edge distributions unlike Shape-emphasis 1. In making these choices, we were motivated by the desire to understand what is the minimal information that is sufficient for categorization performance. If silhouettes (Shape-emphasis 1) can convey material category information, that could be because of something interesting, like the connection of shape-based object recognition and material perception, or possibly due to something less interesting such our database having confounds that are not representative of the real world. Either way, we stand to gain from the results of Experiment 5.

### 3.2.2 Results

Figure 3-5 plots the material categorization accuracy, averaged across observers for all six experimental conditions. In all conditions performance is significantly above chance (11%) and below ceiling (100%) (1-tailed z-tests, \( p < 0.00001 \) in all cases).

There is no difference in performance in the two Color-Emphasized Conditions (Conditions C1 & C2, 2-tailed z-test \( z = -0.887, p = 0.3750 \)), whereas there is an increase in performance in Texture 2 and Shape 2 conditions over Texture 1 and Shape 1 respectively (Conditions T1 & T2, 2-tailed z-test \( z = -3.641, p < 0.0003 \), Conditions S1 & S2, 2-tailed z-test \( z = -7.155, p < 0.0001 \)). Observers do equally well in the Texture 2 and Shape 2 tasks (Conditions T2 & S2, 2-tailed z-test \( z = -0.604, p = 0.546 \)). Figure 3-6 displays the confusion matrices averaged across observers for all experimental conditions. The diagonal is dominant only in Shape-emphasis 2 condition, where categorization performance was the highest. For the other conditions, observers seem to be guessing and not doing well.
Figure 3-5: Accuracy of material categorization as a function of experimental conditions listed in Table 3.2 (from left to right C1 to S2). All error bars are 1 s.e.m. Dashed black line indicates chance performance (= 11%).

Figure 3-6: Confusion matrices averaged across observers for all 6 experiment conditions listed in Table 3.2. The order of material categories is fabric, glass, leather, metal, paper, plastic, stone, water and wood.
3.2.3 Discussion

The results of Experiment 5 demonstrate that individual cues like color, high spatial frequencies, texture or shape have some utility for material categorization, but that they cannot explain the performance on our images by any means (Color Unlimited condition, Experiment 4, accuracy = 90%). In fact, observers do much worse in Experiment 5 than when viewing negative images in rapid presentations (BW Inverted 160 ms condition, Experiment 3, accuracy = 60%). These results leave open the possibility that observers use multiple cues when judging material categories. Perhaps, any one cue like color or texture is not necessary or sufficient for performance, but a combination of these cues has to exist to enable categorization. It may be the case that the set of cues that are used by observers differ from material to material, or even image to image.

One major challenge in identifying cues in real world images is finding ways to manipulate them in an image without destroying the percept of naturalness. Techniques in computer vision and computer graphics provide some help in this regard, but even the state-of-art methods require heavy user interaction in image editing. An alternative to manipulating images to control cues, is to gather photographs of material samples that vary in specific ways. In our previous work, we created hand-made surfaces where we varied the shapes and material properties like glossiness [69, 99, 100]. There is scope for interesting future work in this direction that can combine photographs of carefully selected real world materials where cues like color and texture can be varied in some parametric manner. For example, one might create surfaces of various shapes in different materials using 3-D printing or molding techniques.

3.3 Summary

The cues that we tested - simple ones like color, high spatial frequencies, contrast sign and somewhat more complex cues like texture and shape - seem to have an influence on material categorization but any one of these cues by itself cannot explain categorization performance. It is impossible to prove that there are no low (or mid)
level cues that allow material categorization, because it is hard to make an exhaustive list of such cues and then manipulate them. The diversity of our image databases reduces the chances that there are simple cues, other than the ones we tested for, that allow subjects to perform the task. In other words, by making the images that belong to each category highly diverse, we make it more likely that any strategy used by the observer is based on complex features.

If observers are not merely using simple, low-level cues, is there another explanation for the rapid categorization performance? We know that object recognition can be accomplished in about 50 ms [45, 84, 104]. Could the rapid material judgments we find really just be rapid object judgments? For example, when looking at an Oreo, do you first recognize it as an Oreo and then infer the materials based on your knowledge of Oreos? We will address this issue in the next chapter.
Chapter 4

Role of object knowledge in material judgments

In real world images, one has the problem of distinguishing between actual material perception (“that is made of flower-petal-like stuff”) from inferences made from object perception (“that’s a flower, so it must be made of flower-petal-like stuff”). This distinction between object recognition and material recognition can sometimes be tricky. If you look at a bagel and a donut, it is obvious which is which (see Figure 4-1). This is object recognition. However, the recognition is not driven by the overall shape of the object, since both objects have the same overall shape. Bagels and donuts differ in the small scale structure of the surface, along with variations in color and texture across the surface. Bagels are made of bagel-stuff, and donuts are made of donut-stuff, and this is how we tell them apart. In this case, object recognition is driven by the material recognition, not the other way around.

What about Oreos? The Oreo cookie has a distinctive shape and coloration, and so it can be recognized by the overall shape and color pattern. However, recognizing the Oreo does not automatically lead to inferences about its materials. Figure 4-2a shows an unusual Oreo cookie. When we look at it, we easily recognize it as an Oreo, and easily recognize it as being made of knit yarn - even though we may never have seen a knit Oreo before, and had no idea that such a thing existed.

In this chapter, we will study judgments like those involved when looking at the
Figure 4-1: The bagel on the left and the doughnut on the right have similar shapes and are easy to distinguish. Is this material recognition or object recognition? *(Image source: Flickr)*

Figure 4-2: The Oreo cookie on the left is made of knit wool whereas the ones on the right are genuine. Both cookies have similar shape and reflectance properties, a fact that may confuse machines but not humans. *(Image source: Flickr)*

Oreos in Figure 4-2. Disentangling object and material information in the real world can be challenging, clothes are usually made from fabric and cars from metal and not the other way around. By conducting studies with real and fake objects, we can decorrelate the perception of objects and object shape from material perception.

### 4.1 Our image database of 'real' and 'fake' objects

There is a vast industry dedicated to creating fake goods like fake flowers that don’t have to be watered, fake leather that is cheaper, faux fur for garments, fake food for restaurant displays and so on. Sometimes, human observers are tricked by these fake materials and objects, but often they can spot the fake items. The study of
fake objects is interesting for our research purposes because it lets us study questions like - what makes a flower look natural? Is it the coloring of the petals or their translucency or the arrangement of the center? Telling real from fake involves subtle material judgments that often require training (e.g. an experienced jeweler can tell a fake gemstone from a real gemstone).

Figure 4-3: Examples from our Real-Fake Objects database: (Top row) A knit wool cupcake, a fabric flower, orange made of clay. (Bottom row) A genuine cupcake, a genuine flower, genuine fruits.

We collected 300 images of real and fake desserts, flowers and fruits, 100 in each object category. These categories are familiar to most observers, both in their genuine and fake forms. Images were high resolution color photographs acquired from the photo sharing website Flickr.com. The images in each category were equally divided between real and fake exemplars. All images were cropped or resized to $1024 \times 768$ pixel resolution while keeping the main object of interest in the center of the image. The fake items were made from materials like plastic, clay, fabric, glass, paper etc.

Figure 4-3 shows examples from this image data set.

In choosing this set of images, we were careful to include both fake items that are easy to distinguish from the genuine items, and fake items that are difficult to distinguish from real. We tried to balance lighting, background, and color cues when
selecting the images. For example, fake flowers could appear next to real leaves and
could be indoors under artificial lights or outdoors in natural light. We also tried to
balance the “content” of each image e.g. for each image of fake oranges, we selected
an image containing real oranges.

4.2 Can observers tell real from fake in brief glances?

In the previous chapter, we learnt that observers can distinguish fabric from plastic
or glass in rapid presentations. We want to understand if the same holds for subtle
distinctions like real or fake. Inspired by a recent study on rapid scene perception
by Fei-fei et al., we asked observers to freely describe the materials and objects in
photographs from our Real-Fake Objects database that were displayed briefly [36].
Knowing which aspects of appearance are perceived in brief presentations is useful,
especially if it helps us understand how observers judge genuine-ness or fake-ness of
objects.

4.2.1 Experiment 6: Descriptions of real-fake objects in brief
presentations

Methods

A total of 72 images were selected from our Real Fake Objects database, 24 from
each object category (Dessert, Flowers and Fruit), 12 real and 12 fake. Images were
displayed for 40 ms, 320 ms, or 2.56 s. Images in each of the six sub-categories (real
dessert, fake dessert, real flowers, fake flowers, real fruit and fake fruit) were divided
equally across the three presentation times. In each trial, a fixation screen appeared
for 1000 ms, followed by the stimulus at one of the 3 presentation times. A colored
pink noise mask was then displayed for 107 ms (see Figure 4-4). Both the stimuli and
the mask subtended 26 x 19.5 degrees of visual angle (deg). Observers then described
what they had seen. Six observers participated. Presentation order and presentation
times were counterbalanced between observers.
To provide some guidance to the observers, they were asked to identify the object category (Dessert, Flowers, or Fruit), color of the material(s), specify (shiny/matte), (translucent/opaque), (soft to touch/feels rough), (rigid/non-rigid), (smooth/bumpy), and whether, for food, they believed it would be warm or cool. They were asked whether the object was real or fake, i.e. whether it was made up of the materials that one might expect for the given object. Observers were told that they could specify that they were unsure of a given attribute, and could provide any other details as they wished. Observers were shown sample descriptions and given a short practice session.

**Results**

Figure 4-5 shows descriptions of all six observers for a photograph of fake bananas. By and large, observers based their descriptions on the list of material attributes provided to them. Some observers provided additional details and even identified the materials of the fake objects.
40 ms condition
S1 Bananas, fruit, yellowish color, somewhat ripened, turning brownish, matte and opaque, feel rough to the touch, somewhat rigid, bumpy cool feel to it, genuine.
S2 Fruit, a banana yellow shiny, opaque, soft to touch, non rigid, smooth, real.

320 ms condition
S3 Bananas, fake bananas, yellow, in between shiny and matte, definitely opaque, not soft, not rigid, bumpy surfaces, room temperature, definitely fake.
S4 Bananas made out of wool yellow with brown tips, wool or maybe felt, matte opaque and soft, non rigid with a bumpy surface, definitely fake.

2.56 s condition
S5 Bananas made out of felt, yellow with litte brown tips matte and opaque, soft and slightly non rigid, smooth surfaces, room temperature and definitely fake.
S6 Bananas definitely fake, made out of fabric, felt maybe, yellow, not particularly shiny definitely opaque, soft to the touch, not bumpy, fairly smooth, definitely non rigid, pretty smooshable, definitely fake.

Figure 4-5: Descriptions by 6 observers of an image of fake felt bananas. Four observers correctly identified the bananas as fake, and three of them identified the material as felt.

As in Fei-fei et al.’s design [36], two independent raters evaluated all descriptions in terms of “overall detail”. Our raters were told about the observers’ task, and provided with all 72 images and the corresponding descriptions but not the presentation time for each image. The raters were asked for each image how the accuracy of a given observer’s description compared to the accuracy of that observer’s descriptions for images in general. The raters gave their answers on a 5-point scale, 1 being “poor”, and 5 being “great”. The results are shown in Figure 4-6. A 2-way ANOVA with factors presentation time and rater found a significant main effect of both time (F = 783.56, p = 0.0013) and rater (F = 411.43, p = 0.0024). Rating of overall detail increases with longer presentation times, i.e. observers provide more accurate details when shown an image for longer (Post hoc tests, Tukey’s HSD, p < 0.05).

A third independent rater was given the same list of material attributes as the observers and asked to code the descriptions according to whether the observer thought each material was (shiny/matte/unsure), (translucent/opaque/unsure), etc. If a particular material attribute was unmentioned by an observer, the result was encoded as “unsure”.

66
We next measured the agreement between observers for all these material attributes in each presentation condition. When agreement is above chance, this implies that the task is meaningful and observers can to some extent perform the task under the given conditions. For color and translucency (see Figure 4-6) agreement between observers is above chance even at 40 ms. Agreement on other judgments increases to above chance with longer presentation times. A 2-way ANOVA with factors presentation time and judgment (color, shininess, etc.) found a significant main effect of both time ($F = 16.56, p < 0.0001$) and judgment ($F = 8.4, p < 0.0001$). Observers agreed more on judgments like color and translucency as compared to smoothness and warm/cool (Post hoc tests, Tukey’s HSD, $p < 0.05$).

**Discussion**

The results of Experiment 6 are promising. They tell us that observers can make rich material judgments in a brief glance, extracting multiple aspects of material properties. These diverse material properties include optical (e.g. shiny), mechanical (e.g. rigid), and thermal (warm/cool). However, one might wonder if we overly constrained our observers by providing them a list of material attributes. Perhaps observers might tell us something different, if they are asked to truly report freely. Would they still use parameters like bumpiness to estimate appearance? An issue that
can affect such an experimental design is that most observers do not have as developed a vocabulary for materials as they do for objects. This was one reason why we gave our observers an attribute list in Experiment 6. A second issue with Experiment 6 was that observers often gave binary responses (e.g. soft/rough). Depending on where each observer sets her/his threshold for a judgment, the agreement we measure will be affected. For the case of softness judgments, we conducted a different experiment to measure the degree of agreement between observers (see Appendix B). We find that observers can rate the softness of surfaces like carpets from pictures alone, in a consistent manner. We explore the methodology of Experiment 6 further in an experiment described in Appendix C. We show that even when observers' descriptions do not contain color or shape information, they are rich enough to allow recognition.

4.2.2 Experiment 7: Real-fake judgments in rapid presentations

Experiment 6 tells us that there is rich, detailed information available in brief glances about color, shininess, translucency, softness etc. What about the complex judgment of genuine or fake? In Experiment 6, we allowed our subjects to express no preference about attributes. In the next experiment, we will ask our observers to be explicit in their responses of ‘genuineness’.

Methods

In our Real Fake Objects database, we were careful to include both fake items that are easy to distinguish from the genuine items, and fake items that are difficult to distinguish from real. To measure this variation in appearance, a set of 4 observers rated the “authenticity” of all images in the Real Fake Objects database. Each image was displayed until observers made a response. They were asked to rate on a scale of 1 (definitely fake) to 5 (definitely genuine) the authenticity of each object. The images from the three object categories (Dessert, Flowers, and Fruit) were presented in random order. Figure 4-7 shows the histogram of responses, averaged across all
four observers. Observers seem to find a spectrum of ‘fakeness’ or ‘genuineness’ in both the real and fake objects. Even when observers are given unlimited time, a significant number (54 out of 300) of images are perceived incorrectly.

Figure 4-7: Averaged histogram of the distribution of authenticity ratings (1 = definitely fake, 5 = definitely genuine). The colors indicate the ground truth for authenticity.

A total of 150 images, that were rated unambiguously by the four observers (Figure 4-7), were selected from our Real Fake Objects database for this experiment: 50 from each object category (Dessert, Flowers, Fruits), divided into half real and half fake. Images in each of the six sub-categories (real dessert, fake dessert, real flowers, fake flowers, real fruit, fake fruit) were divided equally across the three presentation conditions (40 ms, 320 ms and 2.56 s). All stimuli and masks subtended 26 x 19.5 deg.

These 150 images were then presented, one by one, at one of the three presentation times (40 ms, 320 ms and 2.56s) followed by pink noise mask for 1sec. The observer’s task was to identify, for each image, the object category, and whether the object was real or fake. Before the experiment began, observers were familiarized with
examples of real and fake objects in each of the three categories. Observers were asked to base their judgments solely on the material of the main objects. Nine observers participated in this experiment. We counterbalanced presentation order and presentation time between subjects.

Results

Figure 4-8a displays results averaged across all 9 observers. Performance was at ceiling for identifying the object category - dessert, flowers or fruit (1-tailed z-tests, 40 ms $z = -5.72$, $p < 0.0001$; 320 ms $z = -5.27$, $p < 0.0001$; 2.56 s, $z = -5.54$, $p < 0.0001$). Observers perform less well at real fake discrimination, although they are above chance performance in all three presentation conditions (1-tailed z-test, at 40 ms, $z = -9.86$, $p < 0.0001$; 320 ms $z = -19.99$, $p < 0.0001$; 2.56s $z = -24.72$, $p < 0.0001$). A 2-way ANOVA with factors task (Object Category or Real-Fake) and presentation time found significant main effects of task ($F = 20.95$, $p < 0.001$) and presentation time ($F = 5.04$, $p = 0.011$). There was also a significant interaction between task and presentation time ($F = 4.28$, $p = 0.02$). In particular, for real-fake judgments, there is a significant increase in performance between 40 ms ($M=72\%$ correct) and both 320 ms ($M=85\%$), and 2.56s ($M=89\%$) (post-hoc tests, Fisher’s LSD, $p<0.05$), but no significant difference between 320 ms and 2.56 s. This suggests that real-fake judgments improve with longer stimulus exposure, but beyond 320 ms there is no significant improvement.

Discussion

The results of Experiment 7 demonstrate that observers can make real-fake judgments in addition to identifying object categories in brief presentations. Their performance increases with longer presentation times. The errors observers make are not because they could not identify the object category but because some real-fake distinctions were hard.

Is the increase in accuracy from 40 ms to 320 ms due to longer stimulus exposure or the ability to make eye movements? We conducted two control experiments to answer
Figure 4-8: (a) Performance at identifying (left) the object category and (right) if the object was real or fake, in brief presentations. Error bars are 1 s.e.m. Dotted lines indicate chance level. (b) Controls for eye movements (Left) Control Experiment 1. Observers either fixated in the center of the display (Fixation) or were free to move their eyes (Regular). (Right) Control Experiment 2. Observers fixated in the center of either the original image (Fixation Original) or a blurred version (Fixation Blurred), or they fixated in the periphery (Periphery). All error bars 1 s.e.m. and dotted lines indicate chance performance (50%).
this question. In Control Experiment 1, the stimuli and procedure of Experiment 7 were used with the modification that all stimuli were displayed for 600 ms. In half the trials, observers were instructed to maintain fixation at the center of the stimulus for the entire presentation duration (Fixation) and in the other half of the trials, they could move their eyes (Regular). Eye movements were monitored (see Appendix A for methods). A small white fixation target appeared in the Fixation condition. Observers made 2-AFC Real/Fake judgments. Six observers participated. Presentation order and the two viewing conditions were counterbalanced between observers.

We found no significant difference in real-fake performance in the two conditions (Two-tailed z-test, $z = -0.188$, $p = 0.85$; see Figure 4-8b, left panel). Eye movements do not seem to aid performance. Just looking at the central patch for long enough is sufficient. What information do observers gather in longer display times that leads to better performance? Perhaps, it is the fine details on the surfaces that enable observers to make the right decision. These details may not have been available at 40 ms. To investigate this hypothesis further, in Control Experiment 2, we controlled the high spatial frequency content of the stimuli. Again, we used the same basic procedure and stimuli as Experiment 7. In all conditions, observers maintained fixation for the stimulus duration. Presentation times were always 600 ms. Observers fixated in either the center of the original stimulus (Fixation Original), or the center of a blurred stimulus (Fixation Blurred), or in the periphery of the original stimulus (Periphery). We created blurred images by downsampling by a factor of 8, then upsampling to the original size. A fixation target appeared on the stimulus in all conditions. In the Periphery condition, the fixation target was 15 deg from the center of the stimulus. Eight new observers judged real vs. fake on each trial. Presentation order and the three fixation conditions were counterbalanced between observers.

We show results in the right panel of Figure 4-8b. Planned comparisons with Bonferroni correction for alpha revealed a significant effect of removing high frequencies through either blurring the image or viewing it peripherally (Fixation Orig vs Fixation Blur, $t = -5.92$, $p \ll 0.0001$; Fixation Orig vs Periphery, $t = -7.17$, $p \ll$
0.0001). High spatial frequencies facilitate real-fake discriminations. Note, however, that even without high spatial frequencies, performance is still above chance (1-tailed z-test, Fixation Blur z = -13.8, p < 0.0001; Periphery z = -12.5, p < 0.0001).

### 4.3 Mechanisms of real-fake judgments

As with judgments of high-level material categories, we measured reaction times for the real-fake judgments. One might expect that the real-fake distinction is harder than telling say fabric from glass or paper and requires more scrutiny. We used exactly the same design as Experiment 3 to investigate RTs for real-fake judgments.

#### 4.3.1 Experiment 8: Reaction times for real-fake judgments

**Methods**

We used images from the Real Fake Objects database. Observers were asked whether a certain target (e.g. fake fruits or paper) was present in each trial. The experimental methods were identical to Experiment 3 and the same observers participated in both experiments. Trials were blocked by target category, with a total of 3 blocks. In all 3 blocks, observers were asked to detect a fake object. Targets and distracters in each block were derived from the same object category (dessert, flowers or fruit). Each block consisted of 100 trials and target was present in half the trials. Stimuli subtended 15 × 12 deg. Four observers participated in this experiment. Presentation order of the blocks was counterbalanced between observers. Examples of targets were shown in the beginning of each block and at the beginning of each trial observers were reminded of the target they were looking for.

**Results**

Figure 4-9 plots errors at detecting targets versus median RTs for each block. In all cases, observers do well (chance = 50% error for real-fake discriminations). For comparison, the data from the baseline and material categorization conditions in Ex-
periment 3 are reproduced in the plot. RTs and error rates follow the same pattern; there is no evidence of a speed-accuracy tradeoff. Combining the data from Experiment 3 and 8 in a one-way ANOVA with median RT as the dependent variable and task as the factor with 11 levels showed a significant effect of task ($F = 4.84$, $p=3e-4$). Post-hoc tests showed that the baseline tasks (color and orientation detection) were completed faster than real-fake tasks (desserts, flowers and fruits) (Tukey’s HSD, $p < 0.05$). The real-fake discrimination appears to be harder (higher error and longer median RTs) than the other two kinds of tasks. Figure 4-10 plots the distribution of reaction times for the three kinds of judgments.

Discussion

These results demonstrate that for real-fake judgments, unlike judgments of high-level material categories, observers take longer than baseline tasks and are less accurate. This suggests that judgments of ‘fakeness’ are more subtle and challenging. Why do real-fake discriminations take longer even though they are perceived accurately in exposures as brief as for high-level material categories? Perhaps, there is more cognitive processing involved in the decision making for real vs. fake than plastic vs. metal. In Experiment 6, observers often reasoned out loud their strategies and assumptions (e.g. if that banana is fake, it must be made of clay and if it is genuine, it must be stale). Perhaps, these findings make the case for a hierarchy of material categories like the one that exists for objects [91]. Identifying plastic, or metal may be more ‘basic’ than telling different kinds of plastic or metal apart.

4.3.2 Experiment 9: Role of texture and shape cues in real-fake judgments

The results of Experiment 7 and 8 establish that real-fake judgments can be made accurately and reasonably quickly. Experiment 6 gave us some clues about the kinds of information that observers might use, but we do not know which cues are important for real-fake judgments. By design, the high-level object identity information is
Figure 4-9: Errors (%) made by four observers as a function of their median reaction times (sec). Chance performance corresponds to 50% error. Red asterisks denote the baseline categorization tasks, the green circles correspond to the 6 blocks of material categorization tasks (Experiment 3) and blue crosses to the real-fake categorization blocks for each observer.
balanced in our Real-Fake database. But, what about low-level cues like color? Or cues from texture and shape? In Control Experiment 2, we found that removing high spatial frequency information affected real-fake performance. Are observers making real-fake decisions solely on simple, low-level cues? In this experiment, we used the methodology developed in Experiment 5 to test the influence of color, shape and texture cues.

**Methods**

We displayed images in our Real-Fake Objects database in the 6 conditions listed in Table 3.1) and described in Experiment 5. The observer’s task was to identify the object category (Dessert, Fruit or Flowers) and material category (Real or Fake) for each image. Five observers participated in all 6 conditions.

The 300 images in the Real-Fake Objects database were randomly divided into 6 non-overlapping sets of 50 images each. Each block consisted of categorizing one of these sets, under one condition. Each observer completed 6 blocks and categorized a...
different set of images in each block. Observers were given as much time as needed to categorize images in all blocks. Figure 3-4 shows an example of the images used in this experiment.

![Image](image_url)

**Figure 4-11:** Example stimuli derived from the Real Fake database that correspond to the conditions listed in Table 3.2. The original image is that of a knit wool cupcake. The resolution of the images displayed here is inferior to those in experimental conditions.

**Results**

Figure 4-12 plots the real-fake and object categorization accuracy, averaged across observers for all six experimental conditions. For real-fake judgments, performance was at chance in all condition except Texture-emphasis 2 (1-tailed z-test, C1, z = -0.821, p = 0.2059; C2, z = 0.292, p = 0.6150; T1, z = -1.470, p = 0.0708; T2, z = -4.003, p < 0.0001; S1, z = 0.792, p = 0.7857; S2, z = -1.263, p = 0.1034). For object judgments, performance is significantly above chance and below ceiling (1-tailed z-tests, p < 0.0001 in all cases). A 2-way ANOVA with factors task (Object Category
or Real-Fake) and presentation condition found significant main effects of task ($F = 20.82, p \ll 0.001$) and presentation condition ($F = 6.27, p = 0.0001$). There was also a significant interaction between task and presentation condition ($F = 4.04, p = 0.0038$).

In particular, for object judgments, there is a significant increase in performance between Color-emphasis 1 (M=46%) and Texture-emphasis 2 (M=74%) and between Color-Emphasis 1 and Shape-emphasis 2 (M=83%) (post-hoc tests, Fisher’s LSD, p<0.05).

**Discussion**

The results of Experiment 9 tell us that for real-fake judgements cues like color or texture in isolation are not helpful. For object categorization, these cues are informative to varying degrees, with rich shape and texture cues being the most useful.

It is interesting to note that even in the Shape-emphasized 2 condition, observers are unable to identify the object category nearly 20% of the time. This suggests that observers use a combination of texture and shape cues to decide the object category, even when texture information is misleading half the time on our database.

The fact that object identification suffers due to impoverished stimulus conditions, might explain the inability to make real-fake judgments. It is hard to judge if an object is real or fake without knowing what the object is. To test this hypothesis, we conducted Control Experiment 3, with exactly the same design as Experiment 8. The only difference was that observers were told the object category on each trial and only made real vs. fake judgments. Six new observers participated in Control Experiment 3. Figure 4-13 plots the accuracy of real-fake judgments in Control Experiment 3. We find that the results are nearly the same as in Figure 4-12a. This suggests that observers do not merely use individual cues like shape or texture in isolation, even when the object identity is known. Observers’ strategies are more complex, they seem to use multiple cues both for judging materials and objects on our Real-Fake database.
Figure 4-12: Accuracy of (a) real-fake and (b) object categorization as a function of experimental conditions listed in Table 3.2 (from left to right C1 to S2). All error bars are 1 s.e.m. Dashed black line indicates chance performance.
Figure 4-13: Accuracy of real-fake categorization when the object identity is known, as a function of experimental conditions listed in Table 3.2 (from left to right C1 to S2). All error bars are 1 s.e.m. Dashed black line indicates chance performance (= 50%).
4.4 Summary

In this chapter, we presented four experiments that explore the connection between material judgments and object identity. We have shown that a subtle judgment like real vs. fake for familiar objects like dessert or fruit can be made in brief exposures and reasonably quickly. The performance at categorizing objects (Dessert, Fruits or Flowers) is different from that of material categorization (Real vs. Fake) in Experiments 7 and 9. This suggests that material judgments are not merely inferred from object knowledge.
Chapter 5

Computational models and material recognition

In the previous chapters, we learnt a great deal about human abilities at identifying materials and the kinds of information that might be used when making material judgments. We now turn to computational models for material recognition and ask whether machines can mimic human performance. There is a long and rich history in computer vision of studying object recognition [22, 38]. Most of this work has been based on recognizing objects by their outline shapes or shape related features. As a result, most object recognition algorithms would fail to distinguish the two kinds of Oreo cookies in Figure 4-2. Figure 5-1 illustrates two more examples where shape-based object identity is limited in its ability to explain material appearance. Although object identity is correlated with material identity in the real world (e.g. car tires are made of rubber, clothes are made from fabric), in general material recognition is not the same as object recognition.

Material recognition is closely related to texture recognition. Texture has been defined in terms of dimensions like periodicity, orientedness and randomness [61]. Undoubtedly, both the "wallpaper" and 3-D type textures are a significant component of material appearance [82]. The knit texture of the cookie on the left in Figure 4-2 is important for telling the real cookie from the fake one. However, as we learnt in Experiments 5 and 9, material judgments are not solely judgments of texture. If our
Figure 5-1: Outline shape is not always informative of material identity. (Left) We may not know which object is pictured here but we can guess it is made of rusted metal. (Right) The fact that this object is shaped like a frog does not tell us that it is made of glass.

goal is to categorize the images in Figure 5-2 as paper and fabric, traditional texture representations that treat the crumpled brown paper and the patterned origami paper as two widely different classes, will be of limited value.

5.1 Previous work

There is little prior work on material categorization in the computer vision. In Chapter 2, we discussed the paucity of image databases for material recognition. There exist a handful of databases that address material recognition in a manner much more limited than the databases that we have created. We now examine the work that has been based on these few databases. Details about these databases can be found in Section 2.1.

Dror et al. classified 9 spheres of different materials as shiny, matte, white, grey, chrome and so on with a SVM classifier with 94% accuracy. Histograms statistics of the image pixels corresponding to a sphere were used as features [31]. Cula et al. [21] developed bidirectional feature histograms to recognize the materials in the CURET database [23]. Varma and Zisserman [109, 110] developed a classifier based on image patches to classify all CURET materials with impressive accuracy (> 95%). However,
Figure 5-2: A variety of textures maybe observed within a single material category. (Top row) Paper category, a ball of crumpled paper and an origami crane. (Bottom row) Fabric category, crochet scarves and a quilt.
the CURET database contains only one sample per material category and all materials are photographed under the same 200 combinations of lighting and viewing angles (see Figure 2-1a). As a result, there is little intra-class variation and doing well on CURET is not a guarantee of good performance in general.

Varma and Zisserman [110] also tested their classifiers on the Microsoft Textile Database [96] and reported similar performance (> 95%). The Microsoft Textile Database contains photographs of 16 samples of folded textiles (see Figure 2-1b) under various illumination conditions. If each folded textile is considered one class for a classifier, this database again suffers from poor intra-class variation. Caputo et al. showed that a SVM-based classifier can achieve > 80% accuracy on their KTH-TIPS2 database [17].

In our own previous work, we employed moment and percentile statistics of photographs and their filtered versions to estimate the albedo of stucco-like surfaces [69, 100] (see Figure 1-3). A few papers have attempted to search for specific materials in real world photographs like glass or skin [42, 67, 68]. Khan et al. [55] developed a psychophysically-inspired editing method to alter the material properties of objects in photographs.

5.2 Testing existing models on our databases

Although there is no prior work on identifying materials in databases as diverse as ours, a natural place to start is to use computational models that have been used successfully for recognizing textures or object categories. From our human data, we know that texture and shape information play some role in material identification and computational representations for these are well developed. In the next few sections, we describe the models we tested and the results on our database Material Categories 1.
5.2.1 Features

We used three kinds of low-level image features - simple global statistics of images, local jet-like features, and SIFT features [62]. Our simple global statistics comprise the mean, variance, skewness and kurtosis of various subbands of a steerable pyramid decomposition [101]. Such features have been used successfully in texture analysis and synthesis and also recently for material perception [48, 69, 83, 100]. In addition to these global statistics, we used local, pixel-level features that have proven useful in recognizing faces, objects and scenes. We compute jet-like features by considering the coefficients of the Steerable, QMF and Haar pyramids at each pixel location in the original image for 3 scales and 4 orientations per pyramid [1]. In order to do so, we construct oversampled pyramids using the code of Li et al. [59]. Finally, we also use the SIFT descriptor at each pixel location in the image as a feature using the method and code of Liu et al. [62, 60]. Table 5.1 provides more details. SIFT features were computed over grayscale versions of images in the standard way. We did not incorporate color into SIFT features because the two other features already used color and the human classification study showed that color was not crucial.

To ensure that pixel-level features were computed in regions belonging to the material of interest, we used the hand-segmented binary masks for each image in our database. All results for jet and SIFT features used these masks to select the appropriate pixels.

<table>
<thead>
<tr>
<th>Feature Category</th>
<th>Color</th>
<th>Scales</th>
<th>Angles</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Global Statistics</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean, variance, skewness, kurtosis of Steerable pyramid subbands</td>
<td>Yes</td>
<td>2</td>
<td>4</td>
<td>96</td>
</tr>
<tr>
<td><strong>Jet-like Features</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coefficients of Steerable, QMF and Haar pyramids</td>
<td>Yes</td>
<td>3</td>
<td>4</td>
<td>126</td>
</tr>
<tr>
<td><strong>SIFT</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patch size = 8, Grid Spacing = 1</td>
<td>No</td>
<td>4</td>
<td>8</td>
<td>128</td>
</tr>
</tbody>
</table>

Table 5.1: Features used by our classifiers.
We employed three learning techniques - Nearest Neighbor, AdaBoost and Support Vector Machines using the code of Ihler [50], Vezhnevets [112] and Chang and Lin [18] respectively. Table 5.2 lists the combination of features and learning techniques that we used to construct classifiers that we index as A, B, C, D and E. We trained all classifiers on a randomly chosen half of all images in the database and tested on the other half. We repeated these training and test trials 10 times to obtain averages of classification performance. For jet and SIFT features, pixels in each image in the test set were classified and a majority vote was taken to determine the label for the overall image.

For the SVM classifier, we used the Radial Basis Function kernel and selected parameters $C$ and $\gamma$ by performing a grid search using 5-fold cross validation for one choice of the training set. We found that this choice of parameters did not change with different training sets, so we fixed the parameters to the values in Table 5.2.

### 5.2.3 Results

We train our classifiers on two kinds of tasks - binary material classification (e.g. leather vs. water, paper vs. fabric) and the full 9-category classification. Given the challenging nature of our database, it is useful to first consider the case of binary classification. We present results on 12 pairs of materials listed in Table 5.3. These pairs were chosen based on the results of Experiment 4. Six of these pairs were easy for our observers to distinguish and six of them were relatively harder (see Figure

---

Table 5.2: The classifiers we used.

<table>
<thead>
<tr>
<th>Learning Technique</th>
<th>Feature</th>
<th>Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nearest Neighbors</td>
<td>Global</td>
<td>A</td>
</tr>
<tr>
<td>Euclidean distance metric, majority vote of 10 neighbors</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AdaBoost</td>
<td>Jet</td>
<td>B</td>
</tr>
<tr>
<td>Classification trees, 50 iterations</td>
<td>SIFT</td>
<td>D</td>
</tr>
<tr>
<td>C-SVC</td>
<td>Jet</td>
<td>C</td>
</tr>
<tr>
<td>RBF kernel, $C = 0.5$, $\gamma = 0.125$</td>
<td>SIFT</td>
<td>E</td>
</tr>
</tbody>
</table>
Table 5.3 lists the test performance of the 5 classifiers at the binary classification. Data averaged from 10 trials of training and testing are shown. For each pair, the human data in Table 5.3 was obtained from the results of our classification study by considering the $2 \times 2$ submatrix of the original $9 \times 9$ confusion matrix in the Color Unlimited condition of Experiment 4. From Table 5.3, we see that classifier performance is well above chance though significantly lower than human performance.

For the case of full 9-category classification, we present the results of only classifiers A, C and E for brevity. We find that classification performance is well above chance (A = 23.4%, C = 25.6%, E = 28.2%, Chance = 11%) but much below human performance measured in Experiment 4. We attempted to improve the performance of pixel-based classifiers by selecting only those pixels that might be more informative such the ones along strong edges. We chose pixels in the top 20th percentile of an edge energy map obtained by taking the absolute value and blurring a subband in an oversampled Laplacian pyramid [59]. However, we found very minor changes in performance (C = 25.8%, E = 26.9%) with this selection of pixels.

It should be pointed out that 9-way material classification is a much harder problem than binary classification, even for humans. Human observers are at ceiling performance (= 100%) in Table 5.3 for all pairs except Glass vs. Plastic, Glass vs. Water and Stone vs. Wood (1-tailed z-tests, $p < 0.0001$). In comparison, their performance at 9-way classification under the same experimental condition is only 90.6%. Another interesting comparison to be made is with results in Experiment 5 (see Figure 3-5). When humans are provided only low-level image information, which is presumably what our image features are able to extract, the recognition performance is similar.

### 5.2.4 Comparison with other databases

The classifier performance we measured in the previous section is poor, especially when we compare it to the performance that has been achieved on other databases[110]. Is this gap in performance due to the differences in the nature of the databases or in the choice of methods? To answer this question, we used the methods of Varma and
Figure 5-3: (Top row) Leather (Le) and Water (Wa) are easy to tell apart. (Bottom row) Stone (St) and Wood (Wo) are harder to distinguish.

<table>
<thead>
<tr>
<th>Pair</th>
<th>A (%)</th>
<th>B (%)</th>
<th>C (%)</th>
<th>D (%)</th>
<th>E (%)</th>
<th>Human (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Le, Wa</td>
<td>70.8</td>
<td>85.4</td>
<td>77.1</td>
<td><strong>86.0</strong></td>
<td>83.8</td>
<td>99.7</td>
</tr>
<tr>
<td>Pa, Wa</td>
<td>67.2</td>
<td>75.4</td>
<td>69.0</td>
<td><strong>82.3</strong></td>
<td>79.1</td>
<td>100.0</td>
</tr>
<tr>
<td>Gl, Le</td>
<td>61.8</td>
<td><strong>86.8</strong></td>
<td>77.4</td>
<td>81.6</td>
<td>80.9</td>
<td>100.0</td>
</tr>
<tr>
<td>Gl, Me</td>
<td>54.8</td>
<td>73.6</td>
<td>65.7</td>
<td>59.0</td>
<td>53.8</td>
<td>99.5</td>
</tr>
<tr>
<td>Le, Pa</td>
<td>72.9</td>
<td><strong>83.9</strong></td>
<td>78.2</td>
<td>80.0</td>
<td>76.5</td>
<td>99.7</td>
</tr>
<tr>
<td>Fa, Me</td>
<td>59.2</td>
<td>65.7</td>
<td>62.4</td>
<td><strong>70.7</strong></td>
<td>67.6</td>
<td>100.0</td>
</tr>
<tr>
<td>Fa, Le</td>
<td>48.6</td>
<td>71.0</td>
<td><strong>71.3</strong></td>
<td>58.9</td>
<td>61.5</td>
<td>99.2</td>
</tr>
<tr>
<td>Fa, Pa</td>
<td>71.9</td>
<td>73.5</td>
<td>63.5</td>
<td>78.1</td>
<td><strong>79.7</strong></td>
<td>98.7</td>
</tr>
<tr>
<td>Gl, Pl</td>
<td>71.1</td>
<td><strong>78.1</strong></td>
<td>67.6</td>
<td>78.0</td>
<td>77.8</td>
<td>96.1</td>
</tr>
<tr>
<td>Gl, Wa</td>
<td>67.2</td>
<td>77.9</td>
<td>63.6</td>
<td><strong>79.6</strong></td>
<td>73.1</td>
<td>83.0</td>
</tr>
<tr>
<td>Me, Pl</td>
<td>67.9</td>
<td><strong>75.7</strong></td>
<td>73.1</td>
<td>68.0</td>
<td>68.8</td>
<td>97.4</td>
</tr>
<tr>
<td>St, Wo</td>
<td>57.8</td>
<td>70.6</td>
<td>70.2</td>
<td><strong>72.2</strong></td>
<td>68.4</td>
<td>96.1</td>
</tr>
</tbody>
</table>

Table 5.3: Classification performance on pairs of materials. For observers, the top 6 pairs were very easy to distinguish whereas last 6 were relatively harder. Material category names are abbreviated Fa = fabric, Gl = glass, Le = leather, Me = metal, Pa = paper, Pl = plastic, St = Stone, Wa = water and Wo = wood. A, B, C, D and E are the various classifiers as defined in Table 5.2. Entries in bold indicate best performance for each pair. For comparison, human performance in the Color Unlimited condition of Experiment 4 is provided.
Zisserman to benchmark the performance on our database with that on CURET and Microsoft Textile Database. Varma and Zisserman achieve impressive performance on the last two databases, so we implemented their texton-based classifier, henceforth referred to as the VZ classifier.

The VZ classifier works in the following way - for a given set of pixel-based image features and classes, ‘textons’ are obtained by k-means clustering in feature space for each class. Next, pixels in all images in the training set are labeled by identifying the closest ‘texton’. A histogram of texton labelings is created for every image. These histograms then serve as features for a nearest neighbor classifier. This approach is similar to the bag-of-words methods in object recognition.

We used the features shown in Table 5.4, which are similar to ones in Table 5.1. We added a patch based feature that was used by Varma and Zisserman with great success on the CURET database. As before, a nearest neighbor classifier was used for Global Statistics features with the Euclidean distance metric and majority voting by the 10 nearest neighbors. A VZ classifier was constructed for the Steerable, SIFT and Patch features using 5 textons per class. Histograms of texton labelings were matched using a nearest neighbor classifier to the nearest neighbor using the Euclidean distance metric.

We compared many different databases using these methods. We used the same CURET images that were used by Varma and Zisserman and ran our analyses on the Microsoft Textile and our Real-Fake database as well. Table 5.5 lists the databases that were used. Images in all databases were divided into half to separate training and test sets. Table 5.6 shows data averaged from 10 trials of training and testing for various databases. For all databases, only the pixels containing the surface of interest were considered.

The results of our comparison are shown in 5.6. We are able to reproduce Varma and Zisserman’s results on the Microsoft Textile database (row MSRC in the Table). On CURET, our accuracy (= 85%) is lower than what they report. We attribute this gap to the differences in our implementation. We sub-sampled images to achieve speedups which might affect the texton histogram representations. The results on our
### Table 5.4: Features used to compare all databases.

<table>
<thead>
<tr>
<th>Feature Category</th>
<th>Color</th>
<th>Scales</th>
<th>Angles</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Global Statistics</strong></td>
<td>Yes</td>
<td>-</td>
<td>-</td>
<td>21</td>
</tr>
<tr>
<td>Mean, variance, skewness, kurtosis, 10th, 50th &amp; 90th percentile of R, G and B channels</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Steerable</strong></td>
<td>Yes</td>
<td>4</td>
<td>4</td>
<td>48</td>
</tr>
<tr>
<td>Coefficients of Steerable</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>SIFT</strong></td>
<td>No</td>
<td>4</td>
<td>8</td>
<td>128</td>
</tr>
<tr>
<td>Patch size = 8, Grid Spacing = 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Patch</strong></td>
<td>No</td>
<td>-</td>
<td>-</td>
<td>25</td>
</tr>
<tr>
<td>5 x 5 image patch, normalized for Weber contrast</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5.5: Databases that were used in the comparison.

<table>
<thead>
<tr>
<th>Database</th>
<th>Description</th>
<th># Classes</th>
<th># Images per class</th>
</tr>
</thead>
<tbody>
<tr>
<td>CURET</td>
<td>Same cropped images as Varma and Zisserman</td>
<td>61</td>
<td>92</td>
</tr>
<tr>
<td>CURET-M</td>
<td>Same as CURET, but classes are re-defined</td>
<td>3</td>
<td>644</td>
</tr>
<tr>
<td>MSRC</td>
<td>Microsoft Textiles as available</td>
<td>16</td>
<td>20 to 25</td>
</tr>
<tr>
<td>MSRC-M</td>
<td>Same as MSRC, but re-defined classes</td>
<td>2</td>
<td>133 to 140</td>
</tr>
<tr>
<td>MC1</td>
<td>Material Categories I</td>
<td>9</td>
<td>100</td>
</tr>
<tr>
<td>MC2</td>
<td>Material Categories II</td>
<td>8</td>
<td>60</td>
</tr>
<tr>
<td>RF-M</td>
<td>Real-Fake Objects, Material Categorization</td>
<td>2</td>
<td>150</td>
</tr>
<tr>
<td>RF-O</td>
<td>Real-Fake Objects, Object Categorization</td>
<td>3</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 5.6: Comparing various databases at material categorization tasks.

<table>
<thead>
<tr>
<th>Database</th>
<th>Chance (%)</th>
<th>Global Stat + NN (%)</th>
<th>Steerable + VZ (%)</th>
<th>SIFT + VZ (%)</th>
<th>Patch + VZ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CURET</td>
<td>1.64</td>
<td>54.95 ± 0.72</td>
<td>63.46 ± 0.53</td>
<td>66.84 ± 0.78</td>
<td>84.39 ± 0.75</td>
</tr>
<tr>
<td>CURET-M</td>
<td>33.33</td>
<td>57.08 ± 7.85</td>
<td>39.75 ± 4.29</td>
<td>34.57 ± 4.18</td>
<td>46.76 ± 5.59</td>
</tr>
<tr>
<td>MSRC</td>
<td>6.25</td>
<td>72.70 ± 1.86</td>
<td>85.67 ± 1.28</td>
<td>75.17 ± 2.12</td>
<td>99.10 ± 0.54</td>
</tr>
<tr>
<td>MSRC-M</td>
<td>50.00</td>
<td>48.14 ± 12.79</td>
<td>41.54 ± 8.25</td>
<td>51.27 ± 7.10</td>
<td>47.85 ± 10.62</td>
</tr>
<tr>
<td>MC-I</td>
<td>11.1</td>
<td>21.29 ± 1.33</td>
<td>18.96 ± 1.21</td>
<td>22.89 ± 1.54</td>
<td>20.69 ± 1.46</td>
</tr>
<tr>
<td>MC-II</td>
<td>12.5</td>
<td>24.12 ± 1.96</td>
<td>33.04 ± 2.03</td>
<td>33.08 ± 2.23</td>
<td>33.62 ± 2.23</td>
</tr>
<tr>
<td>RealFake-M</td>
<td>50</td>
<td>51.20 ± 1.96</td>
<td>56.60 ± 3.01</td>
<td>50.07 ± 4.92</td>
<td>50.80 ± 4.01</td>
</tr>
<tr>
<td>RealFake-O</td>
<td>33.3</td>
<td>55.53 ± 3.43</td>
<td>38.60 ± 2.66</td>
<td>41.13 ± 3.65</td>
<td>56.40 ± 3.03</td>
</tr>
</tbody>
</table>

Table 5.5: Databases that were used in the comparison.

Table 5.6: Comparing various databases at material categorization tasks.
Table 5.7: Comparing various databases at material categorization tasks with maximal size of training set.

database, Material Categories I (row MC1 in the table) are still the same as before. Even with the more advanced VZ classifier, performance is no better than using simple global statistics with nearest neighbor classifier. Increasing the size of the training set from half the database to its maximum size yields modest increases in recognition accuracy for CURET but not for Material Categories I (Table 5.7, rows CURET and MC1).

Perhaps, the Material Categories I database is too diverse and unconstrained for computational models to have much success. There are many unknown variables in images gathered from the web. The layout of the scene, the illumination, camera curve, digital post-processing and so on are impossible to know for hundreds of images taken by different photographers. Another concern, specific to images on Flickr, is that they may contain aesthetic biases of individual photographers. Certain compositions and colors that are unusual or striking might be favored over familiar views. To address these issues, we created a second database of material categories - Material Categories II - by taking photographs in the laboratory.

5.3 A second image database for material categories

When acquiring photographs for a database, there are many choices one might make. Most databases in Figure 2-1 focus on the changes in appearance with illumination and viewpoint variation, rather than variation in the material samples themselves. We want to concentrate on representing a wide range of materials samples and categories.
Figure 5-4: Material Categories II database. (from top row to bottom) Fabric, Leather, Metal, Plastic, Stone and Wood. Two objects per category are shown in two poses in each row. All images were taken under a diffuse halogen lamp (Light 3, see text).
Figure 5-5: Objects in the Ceramic category in Material Categories II database. All images were taken under a fluorescent light (Light 1, see text).

Figure 5-6: Three lighting conditions were used in the Material Categories II database: (a) Light 1, an overhead fluorescent light (b) Light 2, a halogen spotlight (c) Light 3, a diffuse halogen lamp
Figure 5-7: Measurements are indicated in inches for the photography setup (a) Stand on which objects were photographed (b) Position of camera and Light 3. Light 3 setup includes a white foamboard placed on one side of the stand. (c) Position of Light 1. (d) Position of Light 2.
So for simplicity, we decided to fix viewpoint of the camera and vary illumination across three representative conditions (see Figure 5-6). There are eight material categories in this database - Ceramic, Fabric, Leather, Metal, Paper, Plastic, Stone and Wood - and ten everyday objects in each category (see Figure 5-4). Seven of these categories are represented in Material Categories I as well. Each object was photographed in two different poses under each of three lighting conditions. Objects were chosen to be between a few inches to 1 ft in at least one dimension. This was necessary because the viewpoint, hence the scale, was fixed. Objects in our database are items that may be found in a home or an office. We asked our co-workers to bring in items that matched any of our categories. Items were required to be made of only one of the materials (e.g. fabric bags with leather straps were not appropriate). Figure 5-5 shows examples from the Ceramic category.

Images were acquired in a RAW 12-bit format by a Canon EOS 10D camera. The RAW images were linearized using dcrw software [20]. Some of our surfaces have deep shadows and strong specular highlights. In order to capture these surfaces with a limited dynamic range camera, we used the technique of multiple exposure imaging. Multiple exposures were combined into a single high dynamic range image using HDRShop [26]. The final HDR image was hand-segmented to create binary masks.

We used three lighting conditions - Lights 1, 2 and 3 (see Figure 5-6). Light 1 was an overhead fluorescent light source (Kino Flo Diva Lite 200). Light 2 was a halogen spotlight (LTM Pepper 300 W Quartz-Fresnel Light). Light 3 was a diffuse soft light source (Lowel Rifa Lite 66, 750 W Tungsten Halogen Lamp). The position of the lights and the camera viewpoint were held fixed. We used a fixed focal length lens (50 mm) and chose an aperture setting of \( f/22 \) so all objects in our database could be in focus. White balance was set for each lighting condition in reference to a standard white paper. Figure 5-7 details the photography setup.

We ran the analyses described in the previous section on this newer - Material Categories II database (see Table 5.6, row MC2). We found somewhat improved performance than for Material Categories I, but this increase is rather modest (see
Figure 5-8). It should clarified that when dividing our Material Categories II database into training and test sets, all 6 views (3 lights × 2 poses) of any given object occurred in only of the two sets. Unlike CURET and MSRC, we do not define each object as a category by itself.

5.4 Discussion

The numbers in Table 5.6 tell us two things - 1) the same classifiers yield high recognition accuracy on CURET and Microsoft Textile databases but not on our databases and 2) performance is low even on the ‘simpler’ Material Categories II database. Are these differences due to poor choice and use of classification techniques or do they reflect something more fundamental about our databases vs. other databases. We can try to answer this question by examining the how well our features spaces distinguish categories in various databases. Of course, the results in Table 5.6 speak to this question and provide numbers for how discernable the different databases are. But to gain intuitions, in Figures 5-9, 5-10 and 5-11, we plot the first two principal components of the various feature spaces for our databases and for CURET and Microsoft Textiles. These visualizations are useful because they give us a sense of how discriminable material categories are in different databases. It should be pointed that these visualization only show two dimensions of the multi-dimensional features spaces we have used. For example, the patch-based texton histogram representation for CURET in Figure 5-9f contains 305 dimensions and 61 categories. While viewing these plots it is important to remember that if categories do not seem separable in two dimensions, it is not necessary that they will not be separable in a higher dimension. Moreover, principal components are linear projections and our classifiers were based on non-linear methods. However, if classes do look separable in these plots, that shows ease of discrimination for our methods. We find that the Microsoft Textiles database and CURET can be discriminated by patch-based classifiers, mirroring the results in Table 5.6. On our material category databases, the plots show no clustering patterns. In Figure 5-11 we plot the Real-Fake Objects database viewed as material
Figure 5-8: The categorization performance on our databases is poor (a) Material Categories I and (b) Material Categories II. Dotted line indicates chance = 11% for (a) and 12.5% for (b). All error bars are 1 standard error of mean. Refer to Table 5.6 and the text for details about the classifiers.
category and as an object category database (Table 5.6, rows RF-M and RF-O). The performance at real-fake categorization is dismal although there is some information that can be gleaned about object categories with our classifiers.

One reason why there are better results on CURET and Microsoft Textiles can be that in these databases, the training and test sets contain sufficient examples of the same surface. For example, of the 200 similar images of one sponge, 100 are used for training and the rest are used for test. On our databases, objects and surfaces are never repeated in the training and test sets. It is reasonable to ask what would happen if we used the CURET or Microsoft Textile categories in the way we define categories on our databases. Figure 5-12 shows a partitioning of these databases into high-level material categories. In making these partitions we had limited choices and the ones shown in Figure 5-12 were the most reasonable ones we could make. By re-defining the categories in this database, we can run the analyses as before (Table 5.6, see rows CURET-M and MSRC-M). We did not repeat any surfaces in the training and test sets. We find that these new classes that are more representative of high-level material categories, are no longer discriminable by the classifiers (see Figure 5-13). These experiments suggest that performance depends greatly on how images and categories are chosen. Our databases vary hugely along the appearance dimension and generalizing across them is extremely hard for machines.

5.5 Summary

We have examined the merit of low-level image information for discriminating materials categories. Both from our human data (Experiments 4 and 5) and our experiments with standard computer vision algorithms, we conclude that low-level features like color and high spatial frequencies are not adequate for material recognition. The same features and classifiers that achieve good categorization accuracy on texture databases like CURET do not succeed on our challenging material databases. These results open the door for mid-level image representations that capture some aspects of material appearance. In the human vision literature, specular highlights, inter-
Figure 5-9: The first two principal components for the different features described in Table 5.4 and accompanying text - (a) & (b) Global image statistics, (c) & (d) 5 x 5 image patches and (e) & (f) Texton histograms for 5 x 5 image patches - are shown for the MSRC Textiles (left column) and CURET (right column) databases. Texton histogram representations are particularly effective for separating categories on these databases.
reflections and context are known to play a role in the perception of materials. Image features that are correlated with these physical variables are likely to be useful for material recognition algorithms. Material recognition is an important problem in computer vision and compared to the well-studied topics of texture, object and scene recognition, we have only just scratched the surface.
Figure 5-10: The first two principal components for the different features described in Table 5.4 and accompanying text - (a) & (b) Global image statistics, (c) & (d) 5 × 5 image patches and (e) & (f) Texton histograms for 5 × 5 image patches - are shown for our Material Categories I (left column) and Material Categories II (right column) databases. None of these representations can separate our material categories.
Figure 5-11: The first two principal components for the different features described in Table 5.4 and accompanying text - (a) & (b) Global image statistics, (c) & (d) 5 x 5 image patches and (e) & (f) Texton histograms for 5 x 5 image patches - are shown for our Real-Fake database. In the left column, material categories are shown while on the right object categories are considered. None of these representations can separate our material categories. There is some modest success at distinguishing the object categories.
Figure 5-12: (a) CURET-M, we reclassify CURET surfaces into three groups - Fabric, Stone and Food items (top, middle and bottom rows). We used 7 CURET surfaces in each group. (b) MSRC-M, we divided textile into natural fibres (top row) and synthetic fibres (bottom).
Figure 5-13: The first two principal components for $5 \times 5$ image patch features are shown for the original MSRC Textile database (left column) and our interpretation of their textile categories, MSRC-M (right column) databases. Refer Table XX for details about databases and features. When we group surfaces according to broad level material categories, recognition performance drops and the same features are no longer as useful.
Chapter 6

Conclusions

In this thesis, we have explored many aspects of material recognition on real world images. Before we started, little was known about judgments of materials in the real world and there were no databases to even study material judgments. We have gathered diverse sets of photographs to create three new databases of material categories. In deciding how to start and which questions to ask, we were greatly influenced by the results and methods in the more developed fields of object recognition and scene perception. We used limited time presentations and found that judgments of high-level material categories, like plastic or fabric, can be made quickly and in challenging conditions. Even subtle judgments like real vs. fake flowers or fruits, can be made in brief exposures, suggesting that material information, like that for objects and scenes, can be extracted rapidly. We examined the role of cues like color, high spatial frequencies, texture and shape in material categorization. We found that these mostly low-level cues, in isolation, cannot explain performance. On the other hand, material judgments are not simply judgments of object identity. These findings leave open the possibility that observers use multiple low-level cues in conjunction with high-level object knowledge to recognize materials. For example, for real vs. fake flowers, the appearance of fakeness may be conditioned on the low-level cues like color as well as on the fact that the object is supposed to be a flower. We have also examined the utility of low-level image-based information for computational models of material recognition. Our findings indicate that feature spaces that have been used for texture
recognition are not adequate for our material databases.

The work in this thesis is an attempt to frame questions about material perception as much as it is about trying to answer them. Our results point us in exciting directions for future work. Now that we know material categorization can be rapid, how does it interact with judgments about objects and scenes (see Figure 6-1)? In terms of cues, how do different cues like color, shape, texture or even object identity combine to enable material recognition? Building material samples that vary in shape, wallpaper texture and material properties in a systematic way will be useful for studying the relationship between these variables. Another approach that can be fruitful is to study image patches from our databases. Knowing when the patches are big enough to enable shape, texture and material recognition can lead to hypotheses about diagnostic image features for material perception. Based on our results, it is likely that for the diverse images in our databases, the answers will not be simple. Perhaps, the strategy for now is to pick a simpler ‘stimulus world’. Following the work in scene recognition [77], it might be useful to identify prototypical images of materials e.g. wooden furniture, plastic keyboards, ceramic mugs and so on. It is likely that one might discover strong correlations between shape, texture and material categories in such a database. Such correlations certainly exist in the real world and it is plausible that observers make use of them. When choosing prototypes for material categories, there is always the danger of picking a stimulus world that is too far removed from the real one. Nevertheless, simpler stimulus worlds might give us intuitions to tackle the complexity in our current databases.

What about computational models - it is clear that current methods and techniques do not work. There are two directions, one is to build exhaustive databases for materials as has been accomplished for objects and scenes. This approach allows computational models to have access to some of the visual training that observers in our experiments walk in with. The other approach is to design more sophisticated features tailored for material recognition. The binary material categorization results in Chapter 5 point to the possibility that low-level features are used as a first cut to eliminate hypotheses (e.g. that is not wood, stone or leather, but it could be glass,
Hierarchical models of material recognition, like the ones that exist for objects [98], might be a solution.

It is our hope that the work in this thesis is an invitation for other researchers to pursue problems in material recognition. It is unfortunate that this topic has received so little attention because the real world is made up not only of things in various contexts, but also as in Figure 6-1, of stuff.

Figure 6-1: An image inspired by Joan Steiner’s famous Look-Alikes book. This scene is made entirely from food items, the boulders in the back are bread. Images like these play with our expectations of the real world. Questions for future work can be - is scene information prioritized over material information? Is it due to attention? *(Image Source: Google Images)*
Appendix A

Eye movements and material judgments

There are virtually no studies of eye movements during material perception tasks. Previous work on eye movements for shape perception has employed stimuli and tasks that come closest to what we wish to pursue in the context of material perception [30, 88]. The questions we pose are - what regions of an image are fixated the most when we make a material judgment say an albedo or gloss judgment? Our work on reflectance perception has shown that luminance contrast and skewness are predictive of albedo and gloss [69, 100]. Do observers fixate on regions of high contrast or skewness when making albedo and gloss judgments? Are the same regions fixated when making a different judgment, say a shape judgment? What about more complex judgments like real-or-fake or wet-or-dry? The experiments presented here explore the role of eye movements for different kinds of material judgments.

In analyzing the data for eye movement studies, there are several options. We may get hints of what features people are using for a material task simply by looking at where they fixate. We can also attempt to correlate their fixations with features that might be cues to material perception.
A.1 Methods

A video based eye tracker (ISCAN RK-464) was used to record eye movements. The eye tracker works as follows; a beam of infrared light is projected on the observer’s right eye. A camera records the reflections off the observer’s eye. Tracking software is then used to analyze the video input from the camera. The center of the pupil as well as the corneal reflection caused by the IR beam are located and tracked in the video. The relative position of the pupil center with respect to the corneal reflection changes as the eye moves around so it can be used to calculate the direction of the observer’s gaze.

Observers are seated 75 cm from the display monitor and 65 cm eye tracking camera. A chin rest is provided to minimize accidental head movements that can affect eye tracker performance. Stimuli were presented on a LCD monitor with 1280 X 1024 pixel resolution and refresh rate of 70 Hz. The eye tracker records eye movements at 240 Hz sampling frequency. At the beginning of each experimental session, all observers were calibrated using a routine that required them to gaze at five fixed locations on the screen for 0.5 seconds each. If the eye movements gathered in these 0.5 second intervals land within a 1 degree of visual angle at all five locations, the observer is deemed calibrated and allowed to proceed to the experiment.

After the experiment, eye movement analyses were conducted. The eye tracker records the eye position every 4.2 msec. This raw eye position data is first smoothened by a moving window of 33 msec (8 data points). Next, an acceleration criterion is used to divide the eye movements into saccades and fixations [2]. Saccades are very fast eye movements and fixations are the eye movements that occur in between saccades. Only the fixations above the threshold of 50 ms were considered.
Figure A-1: Eye movement study on shape and material judgments (a) Two sample stimuli (b) The fixations for four observers are shown superimposed on the original stimulus. Intensity of each fixation corresponds to the total time spent at that location. Observers in top row performed a material judgment task and those in the bottom row a shape task. Discs shown have radius = 1 degree of visual angle.
A.2 Eye movements for unfamiliar 3-D shapes during material and shape judgments

To examine potential differences between eye movements during a shape judgment task and during a material judgment task, we conducted eye movement studies with images of unfamiliar, three-dimensional objects. We constructed shapes by adding randomized spherical harmonics and rendered these shapes using PBRT under different illumination and viewing conditions [80]. The reflectance properties, albedo and gloss, of these shapes were varied, as were the spherical harmonic coefficients in order to generate different shapes. Figure A-1a shows some example shapes.

![Figure A-1a: Example shapes.](image)

Figure A-2: Histograms of local contrast of the fixated regions. The baseline is estimated by simulating an observer who fixates at random locations on the object. If the subject were fixating on regions of higher contrast, we would see a difference in the shapes of the histograms between subjects and the baseline.

![Figure A-2: Histograms of local contrast.](image)

We set up our experiment as follows. Observers were shown 96 images of our synthesized shapes for 5 seconds each, during which time eye movements were recorded. At the end of 5 seconds, the stimulus disappeared and observers were asked to make either an albedo judgment (light or dark) about the shape they just saw or a shape judgment (smooth vs. spiky). Two observers participated in the albedo judgment.
task and two other observers in the shape judgment task. Each shape was viewed from two viewpoints (front and back). There were 48 distinct shapes in all. Half the shapes were viewed under diffuse lighting conditions and the other half under directional lighting. Half the shapes were light (high albedo) and half dark (low albedo). Finally, half the shapes were glossy and half matte. The combinations of lighting, albedo and gloss were chosen to ensure a balanced design. The resolution of all images was 1024 × 768 pixels were displayed centrally on the LCD that had 1280 × 1024 pixel resolution.

Based on previous work in shape perception, one might expect that certain image regions like occluding boundaries, high contrast areas and corners, are more useful than others for shape judgments. For material judgments, we know from our previous work that regions of high skewness and contrast which often correspond to specular highlights and prominent edges in an image are informative. Therefore, it is plausible that observers look in different places during shape and material perception tasks. Figure A-1b shows the fixations for a particular shape by all four subjects. Fixations are non-random, as verified by simulating a subject making random fixations, and correlated between subjects. This result is typical for our stimuli and tasks. Observers seem to be looking at the same places during different tasks, at least on our shapes. For our simple shapes, it is likely that the regions that convey shape information like occluding boundaries and corners are also indicative of material properties like albedo.

Next, we tried to predict the fixation locations for all subjects using simple low-level image features - such as the local mean luminance or local energy. The goal is to discover correlations, if any, between image features and what observers look at. Figure A-2 shows histograms of local contrast of the fixated regions for three observers and the baseline. The baseline is estimated by simulating a subject who fixates randomly all over the shape and its boundaries, but not on the background. Results of Figure A-2 do not show any significant correlation between local contrast and subject fixations. Similar analyses were run with local mean, local energy and local skewness. It appears that there are certain regions in our shapes that elicit
eyemovement during shape and material perception tasks. However, they cannot be predicted by simple, low-level image features.

### A.3 Eye movements during real-fake judgments

Telling real from fake involves subtle material judgments that often require training (e.g. an experienced jeweler can tell a fake gemstone from a real one). At the level of eye movements, we want to determine, if there are certain image regions that are more informative than others for real-vs.-fake tasks. These regions might depend on the material e.g. one might look in the body of an object for a fake leather bag or at the edges in a fake gemstone.

To pursue these questions, we collected images of real and fake plants, as well as real and fake leather bags from the photo sharing website, Flickr.com. Figure A-3a shows examples of our stimuli. There were 48 images of plants (half fake, half real) and 18 images of leather bags (half fake, half real). The images were scaled to have height of 768 pixels, and were displayed centrally on the LCD display of resolution 1280 × 1024 pixels. The images were cropped appropriately to remove contextual cues and the background was made as uniform as possible. We displayed these images, in a random order, for 3 seconds each, during which time the eye movements were recorded. At the end of 3 seconds, the stimulus disappeared and observers were asked to make a real or fake decision.

Observers were instructed that lighting was not an informative cue, especially for the real and fake plants, since in our data set real and fake plants could occur indoors or outdoors. Two observers participated in this experiment. All observers reported that leather bags were harder to judge than plants, perhaps because of their limited experience with leather goods. An additional observer who participated viewed only the plant images.

Given the observers’ reports, we first examined their performance at telling real from fake. Figure A-3b shows accuracy of real-fake judgments for plants and leather. For plants, the performance ranges from 65-75% (well above chance) whereas for
Figure A-3: (a) Examples of real and fake objects (Top row) Fake items (Bottom row) Genuine items (b) Performance at real-fake discrimination. (Left panel) Plants (Right Panel) Leather. The different colors identify real and fake items that were identified correctly.
Figure A-4: (Top row) Fixations of one observer on a fake plant on left and a real plant on right. Discs have radius = 1 degree of visual angle. The observer identified both of these plants as real. Bottom row shows histogram analyses for the images in the first row. For the fake plant image, one can see that the observer fixated on high contrast (local standard deviation) regions more than baseline. In the real plant image, the observer fixated on higher mean luminance regions more than baseline. In general, observers tended to fixate on higher luminance and higher contrast regions more for all images, real or fake, plant or leather.
leather, the performance (approx. 55%) is close to chance. Therefore, it is likely that for plants, observer eye movements are less random and more informative. Figure A-4 shows fixations of one observer on a fake plant that was mistaken for a real plant. We found that all our observers seem to fixate on edges and specular highlights of leaves. In order to test if observers do look at certain regions more than an observer who fixates randomly, we conducted histogram analyses as before. Figure A-4 shows results for one observer. Regions of higher mean luminance and higher contrast are fixated more than baseline. This is consistent with the hypothesis that observers look at highlights and edges. Even on leather bags, observers spent more time on similar regions. It is likely that observers know ‘where to look’ in order to spot a fake plant or leather bag. They fixate on certain regions that they expect to be useful, and most of the time, they seem to make the right decisions based on the information they gather in those locations.

A.4 Eye movements during wetness judgments

Judging if the floor we are about to step on is slippery or not, is a part of daily life. It is a decision we make often, such as when we are walking on icy pavements, stepping out of the shower, walking on newly mopped hallways in office buildings and so on. Identifying what cues make this decision easy or hard is important, especially in the context of the elderly. In this experiment, we tested the extent to which observers can identify a surface as wet or dry from a single image.

We acquired 11 samples of bathroom tiles of varying patterns, materials and textures. We photographed them under three different illumination conditions and two viewpoints both in the dry and wet conditions. Figure A-5a shows examples from our image data. The images were acquired by a Canon EOS-10D camera in the RAW format. The RAW images were processed using Dcraw software to linearize the pixel values and make them proportional to intensity [20]. These linearized images were then cropped to a pixel resolution of 512 × 512 pixel resolution. In total, we had 11 samples × 3 lighting conditions × 2 viewpoints × 2 wetness condition = 132 images.
Figure A-5: (a) Examples of tiles in our dataset. The two tiles on the left are dry and the two on the right are wet. All of these images were acquired from the same viewpoint, although under different lighting conditions. (b) Percentage correct responses at the dry-wet classification task. The lighter color corresponds to the images of dry surfaces and the darker color to images of wet surfaces. As viewing time and stimulus resolution increase from subjects AS to BS, performance improves.
The experimental setup was as before - observers viewed these images for certain time interval (0.5 seconds or 2 seconds). During this interval, eye movements were recorded. After the interval elapsed, the stimulus disappeared and observers were required to make a wet or dry judgment about the image just viewed. The images were displayed in a random order, either at the original resolution or at twice the resolution (1024 × 1024 pixels). Stimuli were displayed centrally on an LCD monitor of 1280 × 1024 pixel resolution. Three observers participated in this experiment. The first observer, AS, viewed the stimuli for 0.5 seconds at 512 × 512 pixel resolution. The second observer, BHS, viewed images for 0.5 seconds also, but at 1024 × 1024 pixel resolution. The third observer, BS, was given 2 seconds to look at the stimuli at 1024 × 1024 pixel resolution. In a sense, the third observer was given the ‘maximum’ information to perform the wet-dry task.

Figure A-5b shows the performance of our observers. All observers can do the task reasonably well; the performance gets better as the viewing time and image resolution increase. Observers also tend to err on the safe side, so more dry images are marked wet than the other way around. In terms of where observers look, a qualitative examination of their fixations reveals that - white lines, dark lines and certain albedo changes - seem to be informative. If a surface has splashes of water on it, then it tends to have strong specular highlights which manifest as white spots, curves or lines in the image. The edges of a water droplet on a tile tend to be dark, so looking for dark lines makes sense. Lastly, a water droplet can reduce the scattering of the tile by index matching, giving an apparent reduction in albedo. Figure A-6a illustrates these effects. It is plausible that observers use some of these cues when making wetness judgments.

In addition to the qualitative observations we made in Figure A-6a, we conducted histogram analyses of the fixations and low-level image features like local mean luminance, local contrast and local skewness. We find that there is a correlation between the standard deviation of a patch and the time spent fixating on it. Plots in Figure A-6b demonstrate that our observers spent more time looking at patches with high standard deviation (contrast) than baseline (i.e. a simulated observer who looks
Figure A-6: (a) Fixations made by observer BS on dry black tile and a wet white tile. The observer thought both tiles were wet. On examining the fixations more closely, in the timelines on the right, we find that the observer seemed to be looking for 'white streaks' as cues to water spills. The black tile happens to have a pattern that makes it look wet at first glance. (b) Histograms of the standard deviation of fixated patches on two wet tiles look significantly different from baseline.
randomly all over the image).

A more detailed analysis, or different statistics, may yield additional insight, but these preliminary results suggest a relationship between fixation and image statistics that can be explored more fully. Furthermore, they suggest that one can learn interesting features for a material perception task by analyzing fixations during that task.
Appendix B

Judgments of softness

Figure B-1: (a) A subset of the original rug images used in Condition 1 (b) Degradations applied to the first two images in (a). The first and third images were obtained by blurring and the second and fourth images by contrast equalization.

We collected images of rugs (see Figure B-1 for examples) and asked observers to make judgments about the material the rugs were made of. Observers were shown color images of the rugs on an LCD monitor and were asked to sort the images in
the order of decreasing softness. We used three experimental conditions. In the first condition, non-degraded images of rugs, like those in Figure B-1a, were ranked relative to each other. We ask whether visual judgments of material softness are meaningful, i.e. to what degree do observers agree in their judgments, and to what degree are the rugs perceived as having different softness? In the second condition, two types of images degradations, blurring and contrast equalization, were applied to the non-degraded images. Here we ask whether our (modest) degradations impair judgments of softness. In the final condition, the original rug images as well their degraded versions were judged relative to each other. This allows us to judge whether particular degradations lead to bias in softness judgments, e.g. whether blur leads to rugs being perceived as softer.

B.1 Methods and Stimuli

Condition 1 used 25 original images of rugs downloaded from the website of an online store. For Condition 2, the 25 original images were subjected to two types of degradation - blurring and contrast equalization. Figure B-1b demonstrates these degradations. The blurring was accomplished by convolving each image with a Gaussian low pass filter (= 1.5 pixels). Contrast was manipulated by linear scaling; the Michelson contrast of each original image was set to 0.5. For Condition 3, 10 images from the original twenty-five were chosen. Blurred and contrast equalized versions of these ten images were created. Blur in this condition was greater than in Condition 2 (= 3 pixels). For the contrast-equalized images, the Michelson contrast of each original image was set, as before, to 0.5. The color channels of the degraded images were swapped randomly to avoid repetition of colors and reduce (but most likely not eliminate) the recognition of a particular degraded image as merely being a degraded version of another image in the set. We wanted to minimize effects of observers ranking image n and its degraded versions as having similar softness simply because they recognized the images as having come from the same rug. Images were displayed at 320 × 320 pixel resolution on a 24-inch wide screen LCD monitor. Observers received
the following instructions

_Rank these images of rugs from the softest rug to the roughest rug. Imagine touching these rugs. How soft would they feel under your fingers? Disregard the color of the rugs._

A given observer participated in only one of the three experimental conditions. In each condition, observers viewed all the images at the same time. The initial spatial arrangement was randomized for each observer. Observers were allowed as much time as required to sort all images (typically 10-15 minutes). Condition 1 had 12 observers. Condition 2 had 8 observers viewing the blurred rugs, and 10 viewing the contrast equalized rugs. Condition 3 had 10 observers. Observers had normal or corrected-to-normal vision, and ranged in age from 18 to 55.

B.2 Predictions

We do not have ground truth for the softness of our rugs, but we can ask how much observers agree with each other in their softness judgments. If observers agree with each other, this indicates that the softness judgment is a meaningful judgment; there is a significant difference in the perceived softness of different rugs. To measure agreement between observers, we used Kendall’s coefficient of concordance, W, which ranges from 0 (no agreement) to 1 (unanimous). Furthermore, we plot the median rank for each rug vs. the image index, with the image index sorted in order of increasing rank, as shown in Figure B-2. As shown in Figure B-2(a), for high agreement between observers, this plot should approach a 45 degree line. For low agreement between observers, i.e. essentially meaningless random ranking of the images, the plot should approach a horizontal line. If image degradation makes it more difficult for observers to judge softness, this should lead to a reduction in W. Furthermore, some degradations might lead to a bias in softness judgments. Condition 3 allows us to test this hypothesis, by allowing observers to rank original images against degraded images.
B.3 Results and analysis

Figure B-2: The median rank assigned to each image is plotted versus image index. The image index was sorted in the order of increasing median ranks to obtain monotonic curves. (a) The curves for unanimous agreement (best) and no agreement (worst) are shown (b) Condition 1 results (c) Condition 2 results (c) Condition 3 results. The right pointing triangles refer to the original images, the circles to their blurred versions and the stars to the contrast equalized version. It is clear that blurred images get a lower median rank than the originals.

Observers significantly agreed on their rankings of the original images in Condition 1(Kendall's $W = 0.42, = 121.6, p<0.001$). Thus, it is meaningful for observers to visually judge softness of these rugs. In Figure B-2(b) one can see the plot of the median rank for each image. Observer agreement on ranking the softness of the blurred images in Condition 2 was also significant, though modestly reduced (Kendall's $W = 0.34, = 64.9, p<0.001$). The same was true for the contrast-equalized images ($W = 0.35, = 85.9, p<0.001$). Figure B-2(c) plots the sorted median ranks for both groups. We find a relatively small change in agreement from the rankings of the original
image, but the degradations were also quite modest, as shown in Figure B-1.

In Conditions 1 and 2, observers saw only original images, or only one type of degradation. Thus we can compare their effectiveness at judging rug softness, but not whether the degradation introduced any bias in their judgments. Condition 3 allows us to test for these biases. For this condition, observers again had significant agreement in their softness judgments (Kendall’s $W = 0.34$, $p<0.001$), indicating that observers had no difficulty ranking original images against degraded. A Friedman test shows that there is a significant effect of degradation on rug ranks ($p<0.001$). Post-hoc testing (Fishers LSD) finds that blurred images (mean rank $M = 11.86$) are significantly softer than original ($M = 19.82$, $p<0.05$). The difference between original and blurred images can be seen in Figure B-2 (d), which shows the sorted median ranks, with each image labeled according to its type: original, blurred, or contrast equalized.
Appendix C

Judgments beyond shape and color

We have rich impressions of the materials around us. We can assess the finish of leather, the crispness of onion rings, and fluffiness of a shag carpet. In this experiment, we probe the richness of material perception in a new task, using an experimental design with free-form reports.

C.1 Methods

We collected a database of photographs of 60 objects, all of which were toroidal in shape. These objects differed in the materials they were made of and their surface texture. We chose objects of a specific outline shape in order to minimize variation along dimensions other than the material attributes we want to study.

![Toroidal Objects dataset](image)

(a) (b) (c)

Figure C-1: Toroidal Objects dataset (Left panel) Ring made of knitted wool (Middle panel) Doughnut with chocolate glaze (Right panel) Bagel.

Color photographs of these toroidal objects were acquired from the photo sharing
Table C.1: Example descriptions of the images in Figure C-1. For these examples, the images were presented for the amount of time specified in the second column.

<table>
<thead>
<tr>
<th>Image in Fig. C-1</th>
<th>Display time</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Left panel</td>
<td>320 ms</td>
<td>Wool knitted ring, rough but regularly rough because of the knitting stitches and looks soft.</td>
</tr>
<tr>
<td>Middle panel</td>
<td>40 ms</td>
<td>Irregular, soft and matte, malleable.</td>
</tr>
<tr>
<td>Right panel</td>
<td>320 ms</td>
<td>Looks like a bagel, sliced in half with some sort of sprinkling on top, maybe sprinkled with grainy flour, soft and warm, opaque and matte.</td>
</tr>
</tbody>
</table>

website Flickr. These objects were then cropped from their original backgrounds and placed on a uniform white background (see Figure C-1). The final images had resolution $640 \times 480$ pixels ($19 \times 14$ degrees of visual angle) and all objects were resized to have 450 pixels width. These images were displayed briefly to an observer (AS) who was particularly articulate in describing the appearance of materials. Half of the objects (30) were presented for 40 msec, and the other half for 320 msec. All presentations were followed by a colored pink noise mask for 1 second. Observer AS was given the following instructions

"Pretend you are talking to your friend over the phone. Describe each object in a way that your friend can pick it out from a set of objects of similar shape. Focus on the appearance of each object, excluding color".

To provide some guidance, AS was given a list of attributes - shiny/matte, translucent/opaque, soft to touch/feels rough, and rigid/non-rigid - that could be employed in the descriptions. Color was excluded for the purposes of this experiment since the database is small and color by itself could be a good identifier.

Descriptions made by AS were then edited to remove information about color and outline shape, which was roughly the same for all objects in the database. Table C.1 shows these edited descriptions for the objects in Figure C-1.

The set of 60 descriptions by AS was randomly divided into 2 sets - A and B. Half the images in each set had been viewed for 40 msec and half for 320 msec. These descriptions were then given to 4 observers (2 observers were given set A, the rest set
B) along with all the images in the Toroidal Objects dataset. These observers were asked to match the 30 descriptions to 60 images. They were told that

"The descriptions were made by someone who viewed these images very briefly. So, the descriptions may not be as detailed or as accurate as one might expect. In particular, the descriptions are missing information about color and outline shape".

C.2 Results

Figure C-2 plots the average accuracy for all observers at matching the descriptions made by AS at 40 ms and 320 ms presentations. The matching accuracy is significantly above chance in both conditions (1-tailed z-test, 40 ms condition, $z = -3.76$, $p = 0.0001$; 320 ms $z = -7.75$, $p < 0.0001$). There is a significant increase in performance between 40 ms ($M = 21.67\%$) and 320 ms ($M = 51.67\%$) (2-tailed z-test, $z = -3.588$, $p = 0.0003$).

![Figure C-2: Fraction of descriptions matched correctly for the two conditions - 40 ms viewing and 320 ms viewing. Dotted line indicates chance. Error bars are 1 s.e.m.](image)
C.3 Discussion

These results demonstrate that the richness and accuracy of descriptions made by AS increase with presentation time. The descriptions made when images were viewed for 320 msec are more informative about the identity of the original stimuli than the ones made in the 40 msec condition. However, even in the 40 ms presentations, descriptions can be matched to the images well above chance performance. Therefore, there is useful information about the material appearance of surfaces (other than color) that can be perceived and conveyed in short presentations.


135


[47] M. Hearst. Tilebars: Visualization of term distribution information in full text
information access. In Proceedings of the ACM SIGCHI Conference on Human

22nd Annual Conference on Computer Graphics and Interactive Techniques,
pages 229–238.

[49] Y. X. Ho, M. S. Landy, and L. T. Maloney. Conjoint measurement of gloss and

uci.edu/~ihler/code/kde.html.

Journal of Experimental Psychology: Human Perception and Performance,

for subsurface light transport. In 28th Annual Conference on Computer Graph-
ics and Interactive Techniques.


[54] J. S. Johnson and B. A. Olshausen. The earliest eeg signatures of object recog-
nition in a cued-target task are postsensory. Journal of Vision, 5(4):299–312,
2005.

material editing. In ACM SIGGRAPH, pages 654–663.

[56] B. G. Khang, J. J. Koenderink, and A. M. L. Kappers. Perception of illumina-
tion direction in images of 3-d convex objects: Influence of surface materials


