Shape Perception as Bayesian Inference of Modality-Independent Part-Based 3D Object-Centered Shape Representations

by

Goker Erdogan

Submitted in Partial Fulfillment of the

Requirements for the Degree

Doctor of Philosophy

Supervised by Professor Robert A. Jacobs and Professor Jiebo Luo

Department of Brain and Cognitive Sciences

Arts, Sciences and Engineering

School of Arts and Sciences

Department of Computer Science

Arts, Sciences and Engineering

Edmund A. Hajim School of Engineering and Applied Sciences

University of Rochester

Rochester, New York

2017

Table of Contents

Biographical Sketch

The author was born in Izmir, Turkey. He attended Istanbul Technical University, and graduated with a Bachelor of Science degree (2008) in Computer Engineering. He then earned a Master of Science degree (2012) in Computer Engineering from Bogazici University. He began doctoral studies in Brain and Cognitive Sciences and Computer Science at the University of Rochester in 2012. He received a Master of Arts degree in Brain and Cognitive Sciences from the University of Rochester in 2015. He pursued his research in visual and multisensory perception of object shape under the direction of Prof. Robert A. Jacobs (advisor) and Prof. Jiebo Luo (co-advisor).

The following publications were a result of work conducted during doctoral study:

> Erdogan G., Jacobs R. A. (under review) Visual shape perception as Bayesian inference of 3D object-centered shape representations. *Psychological Review.*
>
> Erdogan G., Jacobs R. A. (2016) A 3D shape inference model matches human visual object similarity judgments better than deep convolu-

tional neural networks. In Papafragou, A., Grodner, D., Mirman, D., & Trueswell, J.C. (Eds.) *Proceedings of the 38th Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Erdogan G., Chen, Q., Garcea F. E., Mahon B. Z., Jacobs R. A. (2016) Multisensory part-based representations of objects in human lateral occipital complex. *Journal of Cognitive Neuroscience. 28*(6), 869-881.

Erdogan G., Yildirim I., Jacobs R. A. (2015) From sensory signals to modality-independent conceptual representations: A probabilistic language of thought approach. *PLOS Computational Biology 11*(11): e1004610.

Erdogan G., Yildirim I., Jacobs R. A. (2015). An analysis-by-synthesis approach to multisensory object shape perception. *Multimodal Machine Learning Workshop. NIPS 2015.*

Erdogan G., Yildirim I., Jacobs R. A. (2014). Transfer of object shape knowledge across visual and haptic modalities. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Acknowledgments

First and foremost, I would like to acknowledge my enormous gratitude to my advisor Robbie Jacobs. He transformed me from a mere student to a successful researcher. Robbie somehow always managed to strike the perfect balance in letting me free and directing my efforts towards fruitful directions. As much as he taught me science, he also taught me how to be a good advisor and, most importantly, a good person. Thank you Robbie, for everything.

The Brain and Cognitive Sciences Department at the University of Rochester is a unique place. Not only are all the faculty great scientists, but they are also all good people. I have always been treated as a peer, with the utmost respect, and I am thankful for it. I would like to especially thank two faculty, Brad Mahon and Steve Piantadosi. I worked with Brad and members of his lab, Quanjing Chen and Frank Garcea, on the neuroimaging study reported in Chapter 3 in this thesis, and I learned much from each of them. Steve, who would have been on my committee, if it weren't for an unfortunate scheduling problem, has always provided much helpful feedback on my research, and it has always been a pleasure to talk to him.

I would like to thank the members of my dissertation committee, Professors Krystel Huxlin, Jiebo Luo, Brad Mahon, and Gaurav Sharma for taking the time to be on my committee and their guidance throughout this whole process. I benefited much from my interactions with each of them.

I am thankful for all the great friends I have made in this last five years: past and current members of Robbie's lab, Emin Orhan, Ilker Yildirim, Matt Overlan, and Chris Bates; the class of 2017, Santiago Alonso, Adam Danz, Frank Garcea, and Woon Ju Park; and all former and current graduate students of our wonderful department. You all have provided the much needed, but often neglected, social life to a fellow graduate student. Thank you. I would like to single out one person here, and offer him my deep gratitude. Ilker Yildirim has helped me since the first day of my life as a graduate student; he opened his house to me, he drove me to Wegmans, he cooked delicious Turkish meals for me, and most importantly, he has been a great mentor to me. I am thankful for all of it.

I enjoyed being a graduate student, and that has been possible in large part thanks to the wonderful Kathy Corser and Jennifer Gillis, who took great care of administrative work and were always there when I needed them. I would also like to thank our systems administrator Chris Freemesser for doing such a great job at making sure everything runs smoothly.

Words will never be enough to express my love and gratitude to my family. I consider myself incredibly lucky to have such great parents and a brother. They have always supported and even encouraged my seemingly

endless desire for learning, and have always put me before themselves. Dear mom, dad, and brother, I love you, and I am eternally grateful to you.

As much as I have learned science in this past five years, I have also learned who I am. As much as I have struggled to be a better and better scientist, I have also struggled to become a better and better person, trying to put together the pieces of my life puzzle. Two years ago, I met the last piece of my puzzle. I love you Özlem. You make my life complete.

Abstract

Shape is a fundamental property of physical objects. It provides crucial information for various critical behaviors from object recognition to motor planning. The fundamental question here for cognitive science is to understand object shape perception, i.e., how our brains extract shape information from sensory stimuli and make use of it. In other words, we want to understand the representations and algorithms our brains use to achieve successful shape perception. This thesis reports a computational theory of shape perception that uses modality-independent, part-based, 3D, object-centered shape representations and frames shape perception as Bayesian inference over such representations. In a series of behavioral, neuroimaging and computational studies reported in the following chapters, we test various aspects of this proposed theory and show that it provides a promising approach to understanding shape perception.

Contributors and Funding Sources

This work was supported by a dissertation committee consisting of Professors Robert A. Jacobs (advisor) and Brad Z. Mahon of the Department of Brain and Cognitive Sciences, Professor Jiebo Luo (co-advisor) of the Department of Computer Science, Professor Gaurav Sharma of the Department of Electrical and Computer Engineering, and Professor Krystel Huxlin of the Department of Ophthalmology. Chapters 2 and 3 are published in the journals PLOS Computational Biology and Journal of Cognitive Neuroscience at the present time (see the references in Biographical Sketch). Chapter 4 passed the first round of reviews in the journal Psychological Review, and has been revised and resubmitted for the next round of reviews. The study in Chapter 3 is a collaboration between Professor Robert A. Jacobs' lab and Professor Brad Mahon's lab. Quanjing Chen and Frank E. Garcea collected the neuroimaging data, and Quanjing Chen ran the searchlight analyses. All other analyses and behavioral experiments reported in that chapter were carried out by the author. The stimuli used in studies reported in Chapter 2 and Chapter 3 are based on the set of 3D objects known as Fribbles, and I

would like to thank M. Tarr for making these available on his web pages.

The research in this thesis is supported by research grants from the National Science Foundation (DRL-0817250, BCS-1400784), the Air Force Office of Scientific Research (FA9550-12-1-0303), and the National Institutes of Health (R01 NS089609).

List of Tables

List of Figures

Chapter 1

Introduction

Shape is a fundamental property of physical objects. It provides crucial information for various critical behaviors from object recognition to motor planning. The fundamental question here for cognitive science is to understand shape perception, i.e., how our brains extract shape information from sensory stimuli and make use of it. In other words, we want to understand the representations and algorithms our brains use to achieve successful shape perception. This thesis reports a computational theory of shape perception that uses modality-independent, part-based, 3D, object-centered shape representations, and frames shape perception as Bayesian inference over such representations. In a series of behavioral, neuroimaging and computational studies reported in the following chapters, we test various aspects of this proposed theory and show that it provides a promising approach to understanding shape perception.

Research on shape perception is as old as cognitive science itself (Palmer, 1999). However, we still know very little about it (Peissig & Tarr, 2007;

Gauthier & Tarr, 2016). Researchers have investigated various questions on the nature of shape representations such as whether they are 2D vs. 3D, viewer-centered vs. object-centered, or view-based vs. part-based. Yet, one common feature of all this research is its focus on *visual* shape perception. Even though researchers studied perception of shape through other modalities, there had been little interest on multisensory shape perception until recently (Newell, 2010; Lacey & Sathian, 2011; Yildirim & Jacobs, 2012, 2013). We know both from our daily experience and research on perception that there is significant cross-talk between senses (Amedi, von Kriegstein, van Atteveldt, Beauchamp, & Naumer, 2005; Lacey & Sathian, 2014). Even simple feats like picking up a coffee mug require processing of shape information through both visual and haptic modalities. Empirical support for the multisensory nature of shape perception comes from studies on cross-modal perception and recognition. Subjects can easily recognize an object through a modality different than the one through which the object was initially presented, and cross-modal similarity judgments are strongly correlated with within modality similarity judgments (Cooke, Jakel, Wallraven, & Bulthoff, 2007; Gaissert, Wallraven, & Bulthoff, 2010; Gaissert, Bulthoff, & Wallraven, 2011; Gaissert & Wallraven, 2012; Wallraven, Bulthoff, Waterkamp, van Dam, & Gaissert, 2014). Further support—and perhaps the strongest evidence—for the multisensory nature of shape perception is provided by neuroimaging studies that established lateral occipital complex (LOC) in the brain as a multisensory shape region. Studies have shown that visual and

haptic stimulation lead to similar neural responses in LOC (Amedi, Malach, Hendler, Peled, & Zohary, 2001; Grill-Spector, Kourtzi, & Kanwisher, 2001; Kourtzi & Kanwisher, 2001; Amedi, Jacobson, Hendler, Malach, & Zohary, 2002; Hayworth & Biederman, 2006; Hayworth, Lescroart, & Biederman, 2011).[1]

This multisensory nature of perception requires our brains to abstract away from the sensory-specific inputs to modality-independent representations, and this constitutes the first part of our hypothesis, i.e., *shape representations are modality-independent.* The behavioral experiment reported in Chapter 2 provides support for this hypothesis by showing that object shape is perceived similarly regardless of whether an object is presented visually or haptically. However, a significant question remains. How do our brains extract these modality-independent representations from sensory-specific inputs? The main contribution of the study presented in Chapter 2 is a computational level analysis (in the sense of Marr's levels of analyses (Marr, 1982)) of multisensory perception. We argue that any system capable of learning modality-independent, conceptual representations from modality-specific sensory signals will include three components: a representational language for characterizing modality-independent representations, a set of sensory-specific forward models that map from modality-independent representations to sensory-specific inputs, and an inference algorithm for inverting

[1]However, a recent study (Snow, Goodale, & Culham, 2015) suggests that LOC is not essential for haptic recognition of object shape.

these forward models. We show that a computational model of shape perception built on this framework can explain modality-invariance, and provides an accurate account of people's shape similarity judgments within and across visual and haptic modalities.

The second part of our hypothesis is that *shape representations are part-based*, as opposed to view-based. The part-based hypothesis claims that object parts are represented explicitly. For example, the representation for a human body consists of the representations of its parts (limbs, head, torso etc.) and the spatial relations between these parts. In contrast, the view-based hypothesis argues for a holistic object representation, which simply consists of a collection of stored views (e.g., 2D images) of the object, with no explicit representation of the parts. It might seem obvious that people know about object parts, and that parts play a significant role in cognition (Tversky & Hemenway, 1984). When we see a car, we also see that it has wheels. The view-based hypothesis does not refute the significance of parts but simply claims that the representation for a car and the representation for a wheel are separate; the representation for wheel is not reused or referenced in the representation for car.

The part-based hypothesis has a long history in cognitive science, and various influential models of shape perception are based on the part-based hypothesis (Marr & Nishihara, 1978; Biederman, 1987). Classical evidence for this hypothesis is Biederman's studies that showed primed visual recognition is principally mediated by parts, and recognition suffers dramatically

when part-related information is removed (Biederman & Cooper, 1991; Biederman, 2007). Later studies also found evidence for explicit representation of spatial relations between parts (Hayworth & Biederman, 2006; Hayworth et al., 2011). Overall, the strength of the part-based hypothesis lies in the richness of its representations. The representation for a car and a motorcycle can use the same representation for wheels, and hence part-based representations capture the compositionality of many objects in a natural manner. As pointed out by various researchers, compositionality is crucial for efficient perception and learning (Tenenbaum, Kemp, Griffiths, & Goodman, 2011; Lake, Ullman, Tenenbaum, & Gershman, 2016).

The main motivation behind the view-based hypothesis has been the findings on viewpoint-dependency of object recognition. Various studies have shown that subjects find it harder to recognize an object as it is rotated away from the viewpoint from which it was initially presented (Bulthoff & Edelman, 1992; Tarr, Williams, Hayward, & Gauthier, 1998). This is taken as evidence for a view-based recognition mechanism, in which the incoming image is compared to the stored views of an object to achieve recognition. As the test viewpoint gets farther and farther away from the stored training view, recognition becomes more difficult since the incoming image is less and less similar to the stored training view. One main difficulty with the view-based hypothesis is the purely sensory-specific nature of its representations. This is hard to reconcile with the multisensory nature of shape perception and the claim for modality-independent shape representations.

The part-based hypothesis along with the opposing view-based hypothesis were the subject of fierce debate during the 1980s and 1990s, and the jury still seems to be out on which, if any, of these hypotheses provide a better account of shape perception (Tarr & Bulthoff, 1995; Biederman & Gerhardstein, 1995; Peissig & Tarr, 2007). This is partly because the competing hypotheses were rarely rigorously defined and evaluated (which we attempt to remedy in our study presented in Chapter 4). It is also, to some extent, due to the inherent difficulty of our scientific endeavor. With enough ingenuity, one can almost always come up with an account of some empirical data compatible with either the part-based or view-based hypothesis. In this respect, neuroimaging and especially neural decoding provide unique opportunities. We can look directly into the brain and investigate whether neural shape representations are part-based or view-based. In Chapter 3, we present evidence that human lateral occipital cortex (LOC) carries a part-based shape representation, i.e., *shape representations are part-based*. We show that linear classifiers trained on neural data from LOC on a subset of the objects successfully predict a novel object based on its component part structure. We also show that visual and haptic exploration of objects lead to similar patterns of neural activity in human LOC, which provides further support to the modality-independent nature of shape representations.

The third part of our hypothesis is that *shape representations are 3D and object-centered*, i.e., they encode 3D geometry of objects explicitly in an object-centered coordinate system, as opposed to simply storing 2D views of

objects in a viewer-centered manner. Here again the debate is mainly between part-based and view-based hypotheses. Even though the part-based hypothesis does not necessarily claim 3D shape representations, historically, models based on the part-based hypothesis almost exclusively used 3D representations. Such models have been criticized for failing to capture viewpoint-dependency of object recognition. Several researchers have argued that 3D representations would predict no deterioration in performance for novel viewpoints since the 3D model of the object could be rotated mentally to compensate for viewpoint differences (Bulthoff & Edelman, 1992). As we briefly discussed above, this viewpoint-dependency was in fact the main motivation behind view-based hypothesis, which argued for a recognition mechanism based on stored 2D views. However, view-based representations are rather impoverished and stand in stark contrast to the richness of our perception. In Chapter 4, we argue that the case for 3D, object-centered shape representations is in fact strong, and we provide evidence that shape perception is best understood as statistical inference of 3D shape in an object-centered coordinate system. First, we show that our model accounts for viewpoint-dependency of object recognition, traditionally regarded as evidence against people's use of 3D object-centered shape representations. Second, we report the results of an experiment using a shape similarity task, and present an extensive evaluation of existing models' abilities to account for the experimental data. We find that our shape inference model captures subjects' behaviors better than competing models, including view-based models and

highly successful models of object recognition such as deep convolutional neural networks.

So far, we have described our hypothesis on the *representation* of object shape. However, no model of perception is complete without specifying the *algorithms* that infer these representations from sensory inputs and ultimately use them to produce behavior. This point, unfortunately, has often been overlooked. For example, part-based models generally left the mechanism for extracting and comparing representations unspecified. Such problems made it harder to evaluate existing models and ultimately make progress because the predictions from a model were often unclear. A computational model should be a model of the whole psychological process under investigation, from sensory inputs to behavior. Therefore, we augment our emphasis on representation with an emphasis on the idea that *shape perception is a form of statistical inference*, which constitutes the last part of our hypothesis. This emphasis on statistical inference places our hypothesis in the tradition usually called "perception as inference" or "analysis-by-synthesis" (Yuille & Kersten, 2006). Here perception is characterized as the inference problem of extracting a description of the external world from the sensory inputs. This hypothesis has been fruitfully applied to various aspects of cognition from visual and multisensory perception to high-level cognition (Knill & Richards, 1996; Kersten & Yuille, 2003; Jacobs & Kruschke, 2011). Our work here can be seen as an application of this hypothesis to multisensory shape perception. Our emphasis on statistical inference makes up an important part of our

work, and it is a theme that runs across all chapters in this thesis. For example, in Chapter 2, we argue that Bayesian inference is a crucial component of multisensory perception that enables acquiring modality-independent representations from sensory-specific inputs. And, in Chapter 4, we show that this emphasis on inference enables us to explain how 3D shape representations can give rise to view-dependent recognition.

In Chapter 5, we provide a summary and discussion of the contributions of this thesis and finish by presenting a few promising future directions.

Chapter 2

From Sensory Signals to Modality-Independent Conceptual Representations: A Probabilistic Language of Thought Approach

Introduction

While eating breakfast, you might see your coffee mug, grasp your coffee mug, or both. When viewing your mug, your visual system extracts and represents the shape of your mug. Similarly, when grasping your mug, your haptic system also extracts and represents the shape of your mug. Are the representations acquired when viewing your mug distinct from the representations acquired when grasping your mug? If so, these would be modality-specific representations. Or does there exist a level at which the shape representation of your mug is the same regardless of the sensory modality through which the mug is perceived? If so, this would be a modality-independent representation.

Recent experiments on crossmodal transfer of perceptual knowledge suggest that people have multiple representations of object shape and can share information across these representations. For example, if a person is trained to visually categorize a set of objects, this person will often be able to categorize novel objects from the same categories when objects are grasped but not seen (Wallraven et al., 2014; Yildirim & Jacobs, 2013). Because knowledge acquired during visual training is used during haptic testing, this finding suggests that neither the learning mechanisms used during training nor the representations acquired during training are exclusively visual. To the contrary, the finding indicates the existence of both visual and haptic object representations as well as the ability to share or transfer knowledge across these representations. Successful categorization of objects regardless of whether the objects are seen or grasped illustrates modality invariance, an important type of perceptual constancy.

What type of learning mechanisms and mental representations might underlie modality invariance? One possible answer is that people are able to abstract over their modality-specific representations in order to acquire modality-independent representations. For instance, people might use modality-specific representations of objects as a foundation for inferring modality-independent representations characterizing objects' intrinsic properties. To understand the nature of the latter representations, it is important to recognize the distinction between objects' intrinsic (or "deep") properties and the sensory (or "surface") features that these properties give rise to. The shape

of an object is a modality-independent intrinsic property. Visual and haptic features are modality-specific sensory cues to the object's shape arising when the object is viewed or grasped, respectively.

Once acquired, modality-independent representations may underlie modality invariance. For example, they can mediate crossmodal transfer of knowledge. Consider a person who is first trained to visually categorize a set of objects, and then tested with novel objects (from the same set of categories) when the objects are grasped but not seen. During visual training, the person uses his or her visual representation of each object to infer a modality-independent representation characterizing the object's intrinsic properties, and applies the object's category label to this representation. When subsequently grasping a novel object on a test trial, the person uses the object's haptic representation to infer a modality-independent representation of its intrinsic properties. The novel object is judged to be a member of a category if it has similar intrinsic properties to the training objects belonging to that category.

Because modality-independent representations may underlie modality invariance, they would clearly be useful for the purposes of perception and cognition. Importantly, recent behavioral and neurophysiological data indicate their existence in biological organisms. For instance, behavioral and neural evidence support the idea that object features extracted by vision and by touch are integrated into modality-independent object representations that are accessible to memory and higher-level cognition (Easton, Srini-

vas, & Greene, 1997; Reales & Ballesteros, 1999; Pascual-Leone & Hamilton, 2001; Amedi et al., 2002; James et al., 2002; Norman, Norman, Clayton, Lianekhammy, & Zielke, 2004; Amedi et al., 2005; Taylor, Moss, Stamatakis, & Tyler, 2006; Lacey, Peters, & Sathian, 2007; Ballesteros, Gonzalez, Mayas, Garcia-Rodriguez, & Reales, 2009; Lacey, Pappas, Kreps, Lee, & Sathian, 2009; Lawson, 2009; Tal & Amedi, 2009). Based on brain imaging (fMRI) data, Taylor et al. (2006) argued that posterior superior temporal sulcus (pSTS) extracts pre-semantic, crossmodal perceptual features, whereas perirhinal cortex integrates these features into amodal conceptual representations. Tal and Amedi (2009), based on the results of an fMRI adaptation study, claimed that a neural network (including occipital, parietal, and prefrontal regions) showed crossmodal repetition-suppression effects, indicating that these regions are involved in visual-haptic representation.

Perhaps the most striking data comes from the work of Quiroga and colleagues who analyzed intracranial recordings from human patients suffering from epilepsy (Quiroga, Kraskov, Koch, & Fried, 2009; Quiroga, 2012). Based on these analyses, they hypothesized that the medial temporal lobe contains "concept cells", meaning neurons that are selective for particular persons or objects regardless of how these persons or objects are sensed. For instance, Quiroga et al. (2009) found a neuron that responded selectively when a person viewed images of the television host Oprah Winfrey, viewed her written name, or heard her spoken name. (To a lesser degree, the neuron also responded to the comedian Whoopi Goldberg.) Another neuron

responded selectively when a person saw images of the former Iraqi leader Saddam Hussein, saw his name, or heard his name.

To fully understand modality-independent representations, Cognitive Science and Neuroscience need to develop theories of how these representations are acquired. Such theories would be significant because they would help us understand the relationships between perceptual learning and modality invariance. They would also be significant because they would be early "stepping stones" toward developing an understanding of the larger issue of how sensory knowledge can be abstracted to form conceptual knowledge.

The plan of this paper is as follows. In the Results section, we start by describing a general theoretical framework for how modality-independent representations can be inferred from modality-specific sensory signals. To evaluate the framework, we next describe an instantiation of the framework in the form of a computational model, referred to as the Multisensory-Visual-Haptic (MVH) model, whose goal is to acquire object shape representations from visual and/or haptic signals. Simulation results show that the model learns identical object representations when an object is viewed, grasped, or both. That is, the model's object percepts are modality invariant. We also evaluate the MVH model by comparing its predictions with human experimental data. We report the results of an experiment in which subjects rated the similarity of pairs of objects, and show that the model provides a very successful account of subjects' ratings. In the Discussion section, we highlight the contributions of our theoretical framework in general, and of the MVH

model in particular, emphasizing its combination of symbolic and statistical approaches to cognitive modeling. Due to this combination, the model is consistent with an emerging "probabilistic language of thought" methodology. The Methods section provides modeling and experimental details.

Results

Theoretical framework

According to our framework, any system (biological or artificial) that acquires modality-independent representations from sensory signals will include the following three components: (1) a representational language for characterizing modality-independent representations; (2) sensory-specific forward models for mapping from modality-independent representations to sensory signals; and (3) an inference algorithm for inverting sensory-specific forward models—that is, an algorithm for using sensory signals in order to infer modality-independent representations. These three components are discussed in turn.

(1) Representational language for characterizing modality-independent representations: Although biological representations of modality-specific sensory signals are not fully understood, it is believed that these representations are constrained by the properties of the perceptual environment and the properties of the sensory apparatus. For example, the nature

of biological visual representations depends on the nature of the visual environment and the nature of the eye.

In contrast, constraints on the nature of modality-independent representations are not so easy to identify. One radical view, usually referred to as embodied cognition (Barsalou, 1999), claims that there are no amodal representations; all mental representations consist of sensory representations. However, the majority view in Cognitive Science argues that people possess modality-independent representations (e.g., representations of object shape or representations of abstract concepts such as 'love' or 'justice'), though there is no consensus as to the best way to characterize these representations. Common approaches include both statistical (e.g., distributed representations over latent variables) and symbolic (e.g., grammars, logic) formalisms. These formalisms provide different representational languages for expressing modality-independent thoughts and ideas, each with their own strengths and weaknesses.

(2) Sensory-specific forward models: Modality-independent representations do not make direct contact with sensory signals. To bring them in contact with sensory signals, our framework includes sensory-specific forward models which map from modality-independent representations to sensory features. For example, a vision-specific forward model might map a modality-independent representation of an object's shape to an image of the object when viewed from a particular viewpoint. Similarly, a haptic-specific forward

model might map the same modality-independent representation of an object's shape to a set of haptic features (e.g., hand shape as characterized by the joint angles of a hand) that would be obtained when the object is grasped at a particular orientation. We often find it useful to think of these sensory-specific forward models as implementations of sensory imagery. For instance, if a vision-specific forward model maps an object to its visual features, then that is an implementation of visual imagery.

(3) Inference algorithm for inverting forward models: Sensory-specific forward models map from modality-independent representations to sensory signals. However, perception operates in the opposite direction—it maps from sensory signals to modality-independent representations. Consequently, perception needs to invert the sensory-specific forward models. This inversion is accomplished by a perceptual inference algorithm.

From a larger perspective, our theoretical framework presents a conceptual analysis of the computational problem of multisensory perception. How can we transfer knowledge (category, shape, meaning etc.) from one modality to another? Why are we more accurate when we perceive through more modalities? How can we recognize a novel object crossmodally? Or how can we recognize an object crossmodally from a novel view? We believe our framework is successful in providing a unified account of the answers to these questions and the underlying cognitive processes. Hence, we believe our theoretical framework in itself constitutes a significant contribution to

the understanding of multisensory perception.

Framework applied to visual-haptic object shape perception

To better understand and evaluate our framework, we apply it to the perception of object shape via visual and haptic modalities. This application results in the MVH computational model with the three components outlined above.

We have had to make specific implementation choices to instantiate our theoretical framework as a computational model. To us, these choices are both uninteresting and interesting. On the one hand, the implementation choices that we have made are not essential to the framework. Indeed, other reasonable choices could have been made, thereby leading to alternative framework implementations. On the other hand, we believe that some of our choices are important because they contribute to the study of cognitive modeling. In particular, our computational model combines both symbolic and statistical modeling approaches. Because of this combination, the model can be regarded as falling within a recently emerging "probabilistic language of thought" methodology. This contribution is described in the Discussion section.

One of the implementation choices that we made was a choice as to which stimuli we should focus on. Object shape perception via vision and/or haptics is currently an unsolved problem when considered in its full generality.

Consequently, we focus on a small subset of objects. We designed 16 novel objects, where the set of object parts was based on a previously existing set of objects known as "Fribbles". Fribbles are complex, 3-D objects with multiple parts and spatial relations among parts. They have been used in studies of visual (Hayward & Williams, 2000; Tarr, 2003) and visual-haptic (Yildirim & Jacobs, 2013) object perception. We used part-based objects because many real-world objects (albeit not all) have a part-based structure. In addition, theories of how people visually recognize part-based objects have received much attention and played important roles in the field of Cognitive Science (Marr & Nishihara, 1978; Hoffman & Richards, 1984; Biederman, 1987; Tversky, 1989; Saiki & Hummel, 1998).

Each object that we designed is comprised of five parts (the set of possible parts is shown in Fig. 2.1). One part (labeled P0 in Fig. 2.1), a cylindrical body, is common to all objects. The remaining four parts vary from object to object, though they are always located at the same four locations in an object. A particular object is specified by selecting one of two interchangeable parts at each location (4 locations with 2 possible parts per location yields 16 objects). The complete set of objects is shown in Fig. 2.2.

Shape grammar as a language for characterizing object shape: In the MVH model, object representations have three important properties. The first property is that representations are modality-independent. That is, they are not directly composed from modality-specific features, nor do

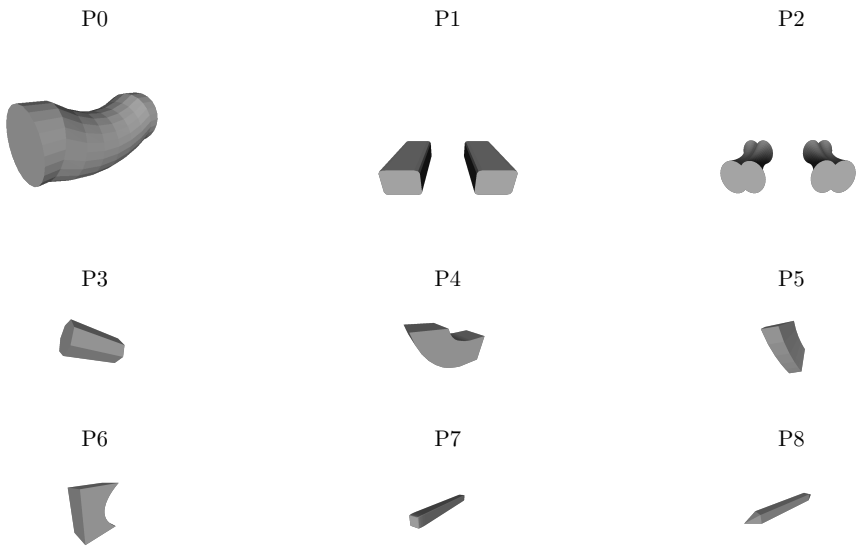


Figure 2.1: Possible object parts. Part P0 is common to all objects. Parts P1-P8 vary from object to object.

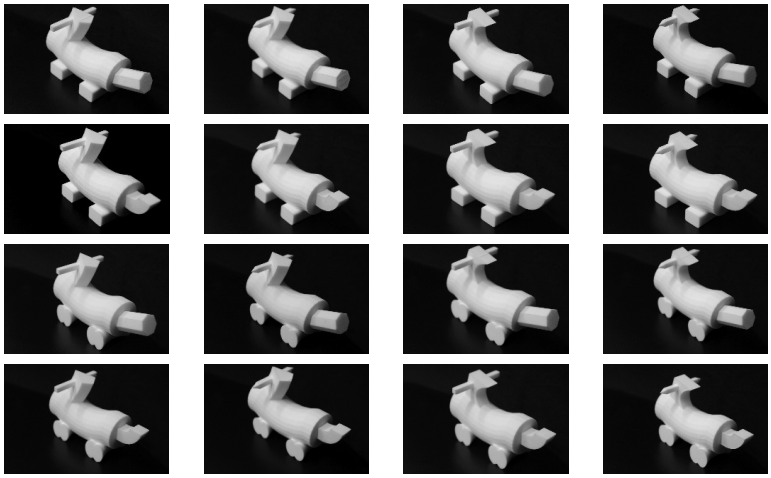


Figure 2.2: Images of objects used in our simulations and experiment.

they directly specify the values of these features.

The second property is that object representations characterize objects in terms of their parts and the spatial relations among these parts. When designing the model, our main focus was not on developing new insights regarding how people represent object shape. Although this is an important research area, many researchers in the Cognitive Science and Artificial Intelligence communities already study this topic (Marr & Nishihara, 1978; Ballard, 1981; Hoffman & Richards, 1984; Biederman, 1987; Kass, Witkin, & Terzopoulos, 1988; Tversky, 1989; Logothetis, Pauls, & Poggio, 1995; Cutzu & Edelman, 1996; Basri, Costa, Geiger, & Jacobs, 1998; Saiki & Hummel, 1998; D. Zhang & Lu, 2004; Feldman & Singh, 2006; Ling & Jacobs, 2007; Op de Beeck, Wagemans, & Vogels, 2008; L. Zhu, Chen, & Yuille, 2009). The scientific literature contains a wide variety of different approaches to object shape representation. To date, there does not appear to be a consensus as to which approach is best.

Instead of researching new ways to represent object shape, our goal is to understand how modality-independent representations can be learned from sensory data. Because the MVH model needs to represent object shape, it necessarily resembles previously existing models that also represent object shape. In particular, like previous models, our model represents objects in terms of their parts and the spatial relations among these parts (Marr & Nishihara, 1978; Hoffman & Richards, 1984; Biederman, 1987; Tversky, 1989; Saiki & Hummel, 1998). In principle, we are not strongly committed to

the hypothesis that people represent objects in a part-based manner. Shape primitives other than parts could have been used in our simulations (as is sometimes done with shape grammars in the Computer Vision and Computer Graphics literatures; e.g., see Felzenszwalb (2013)), albeit at possibly greater computational expense. To us, the use of part-based object representations in our simulations seems reasonable because these representations have played prominent roles and received considerable theoretical and empirical support in the Cognitive Science literature, because the stimuli used in our simulations and experiment were generated in a part-based manner, because the analyses of our experimental data indicate that subjects were sensitive to the part-based structure of the stimuli (see below), and because part-based object representations led to computationally tractable simulations.

The final property is that object representations use a shape grammar to characterize an object's parts and the spatial relations among these parts (Fu, 1986; Bienenstock, Geman, & Potter, 1997; Tu, Chen, Yuille, & Zhu, 2005; Amit & Trouve, 2007; Grenander & Miller, 2007; L. Zhu et al., 2009; Talton et al., 2012; Felzenszwalb, 2013). Grammars are commonly used to characterize human language and language processing (Chomsky, 1965; Pinker, 1994), and are also used in other areas of Cognitive Science (Goodman, Tenenbaum, Feldman, & Griffiths, 2008; Kemp & Tenenbaum, 2008; Piantadosi, Tenenbaum, & Goodman, 2012; T. D. Ullman, Goodman, & Tenenbaum, 2012; Goodman, Tenenbaum, & Gerstenberg, 2015; Yildirim & Jacobs, 2015). In addition, they are used to characterize objects and scenes in fields such as

Computer Vision and Computer Graphics (Fu, 1986; Bienenstock et al., 1997; Tu et al., 2005; Amit & Trouve, 2007; Grenander & Miller, 2007; L. Zhu et al., 2009; Talton et al., 2012; Felzenszwalb, 2013).

The MVH model uses a shape grammar to specify the possible parts and spatial relations among parts. Conventional shape grammars, like other types of symbolic representations, can often be "brittle" when used in noisy environments with significant uncertainty. We ameliorated this problem through the use of a probabilistic approach. The details of the shape grammar are described in the Methods section. For now, note that the grammar is an instance of a probabilistic context-free grammar. Production rules characterize the number of parts and the specific parts comprising an object. These rules are supplemented with information characterizing the spatial relations among parts.

Specifically, an object is generated using a particular sequence of production rules from the grammar. This sequence is known as a derivation which can be illustrated using a parse tree. To fully specify an object, an object's derivation or parse tree is supplemented with information specifying the locations of the object's parts. This specification occurs by adding extra information to a parse tree, converting this tree to a spatial tree representing object parts and their locations in 3-D space (see Methods section).

Vision-specific and haptic-specific forward models: Because object representations are modality independent, they do not make direct contact

with sensory signals. To evaluate and infer these representations, they need to be brought in contact with these signals. For these purposes, the MVH model uses its modality-independent representations to predict or "imagine" sensory features from individual modalities. For example, given a modality-independent representation of a particular object (i.e., a representation of the object's parts and the locations of these parts), the model can predict what the object would look like (perhaps a form of visual imagery) or predict the hand shape that would occur if the object were grasped (perhaps a form of haptic imagery). A mapping from a modality-independent representation to a sensory-specific representation can be carried out by a forward model, a type of predictive model that is often used in the study of perception and action (Jordan & Rumelhart, 1992; Wolpert & Kawato, 1998; Wolpert & Flanagan, 2009). In Cognitive Science, forward models are often mental or internal models. However, forward models exist in the external world too. Our computer simulations made use of two forward models.

The vision-specific forward model was the Visualization Toolkit (VTK; www.vtk.org), an open-source, freely available software system for 3-D computer graphics, image processing, and visualization. We used VTK to visually render objects. Given a modality-independent representation of an object, VTK rendered the object from three orthogonal viewpoints. Images were grayscale, with a size of 200 $\times$ 200 pixels. A visual input to the model was a vector with 120,000 elements (3 images $\times$ 40,000 [200 $\times$ 200] pixels per image).

The haptic-specific forward model was a grasp simulator known as "GraspIt!" (Miller & Allen, 2004). GraspIt! contains a simulator of a human hand. When predicting the haptic features of an object, the input to GraspIt! was the modality-independent representation for the object. Its output was a set of 16 joint angles of the fingers of a simulated human hand obtained when the simulated hand "grasped" the object. Grasps—or closings of the fingers around an object—were performed using GraspIt!'s AutoGrasp function. Fig. 2.3 shows the simulated hand grasping an object at three orientations. In our simulations, each object was grasped 24 times, each time from a different orientation (different orientations were generated by rotating an object 8 times [each time by 45°] around the width, length, and depth axes). The use of multiple grasps can be regarded as an approximation to active haptic exploration. A haptic input to the model was a vector with 384 elements (16 joint angles per grasp × 24 grasps). Our choice of using joint angles as our haptic features follows a common practice in the field of postural hand analysis (Santello, Flanders, & Soechting, 1998; Thakur, Bastian, & Hsiao, 2008).

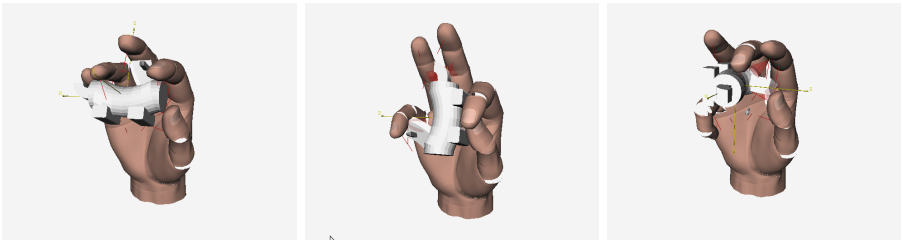


Figure 2.3: GraspIt! simulates a human hand. Here the hand is grasping an object at three different orientations.

Bayes' rule inverts sensory-specific forward models: Importantly, the MVH model learns its object representations. The most influential models of object shape in the Cognitive Science literature, such as those of Biederman (1987) and Marr and Nishihara (1978), used part-based shape representations that were stipulated or hand-crafted by scientific investigators. In contrast, a goal of our model is to learn representations using a probabilistic or Bayesian inference algorithm from visual and/or haptic signals. Using the terminology of Bayesian inference, the model computes a posterior distribution over object representations based on a prior distribution over these representations (indicating which of the representations are more or less likely before observing any sensory data) and a likelihood function (indicating which representations are more or less likely to give rise to observed sensory data).

The model's prior distribution is based on the prior distribution of the Rational Rules model of Goodman et al. (2008). In brief (see the Methods section for full details), the prior distribution is the product of two other distributions, one providing a prior over parse trees and the other providing a prior over spatial models. These priors are Occam's Razors favoring the use of "simple" parse trees and spatial models.

The likelihood function allows the model to use sensory data to evaluate proposed object representations. Object representations which are highly likely to give rise to perceived sensory data are more probable than object representations which are less likely to give rise to these data (ignoring the

prior distribution, for the moment). Sensory-specific forward models play a crucial role in this evaluation. As mentioned above, object representations are modality-independent, and thus do not make direct contact with perceived visual or haptic features. Sensory-specific forward models are needed to relate object representations to their sensory features.

Using Bayes' rule, the MVH model combines the prior distribution and the likelihood function to compute a posterior distribution over object representations. Unfortunately, exact computation of the posterior distribution is intractable. We, therefore, developed a Markov chain Monte Carlo (MCMC) algorithm that discovers good approximations to the posterior. This algorithm is described in the Methods section.

Simulation results: We used the model to infer modality-independent representations of the 16 objects in Fig. 2.2. Object representations were inferred under three stimulus conditions: a vision condition, a haptic condition, and a multisensory (visual and haptic) condition. In all conditions, we inferred the posterior distribution over modality-independent object representations. However, except where explicitly noted, the results reported below are based on maximum a posteriori (MAP) estimates. Because distributions are highly peaked around the MAP estimate, the results are essentially the same when samples from each distribution are used.

The sole free parameter of the model is the variance of the likelihood function. Intuitively, this parameter controls the relative weights of the prior

and likelihood terms. By increasing the variance, thereby increasing the relative weight of the prior, it is possible to constrain the model so that it tends to prefer simple parse trees and spatial models. In contrast, as the variance is decreased, the likelihood becomes more important, thus allowing more complex trees and models to be assigned probability mass. For each stimulus condition, we selected a value for the variance that provides a good balance between prior and likelihood terms. We found that simulation results are robust to the exact choice for the variance value. As long as the variance is small enough to allow object representations which are complex enough, the MVH model produced similar results.

Fig. 2.4 shows the results of a representative simulation in which the model received visual input. This input consisted of three images of an object from orthogonal viewpoints (Fig. 2.4a). The four modality-independent object representations with the highest posterior probabilities are shown in the top row of Fig. 2.4b. The bottom row shows visual renderings of these object representations. The MAP estimate is on the left. Crucially, this estimate represents the object perfectly, successfully inferring both the object parts and their spatial locations. Indeed, we find that the model's MAP estimate always represents an object perfectly for all the objects comprising our stimulus set.

The other estimates in Fig. 2.4b (estimates with smaller probabilities than the MAP estimate) exemplify the robustness of the model. Although imperfect, these estimates are still sensible. When a part is missing from an

object representation, it is often part P8 which is small in size and, thus, has only a small influence on the likelihood function. When a part is mismatched, the model often substitutes part P7 for P8. This is unsurprising given that parts P7 and P8 are visually (and haptically) similar.

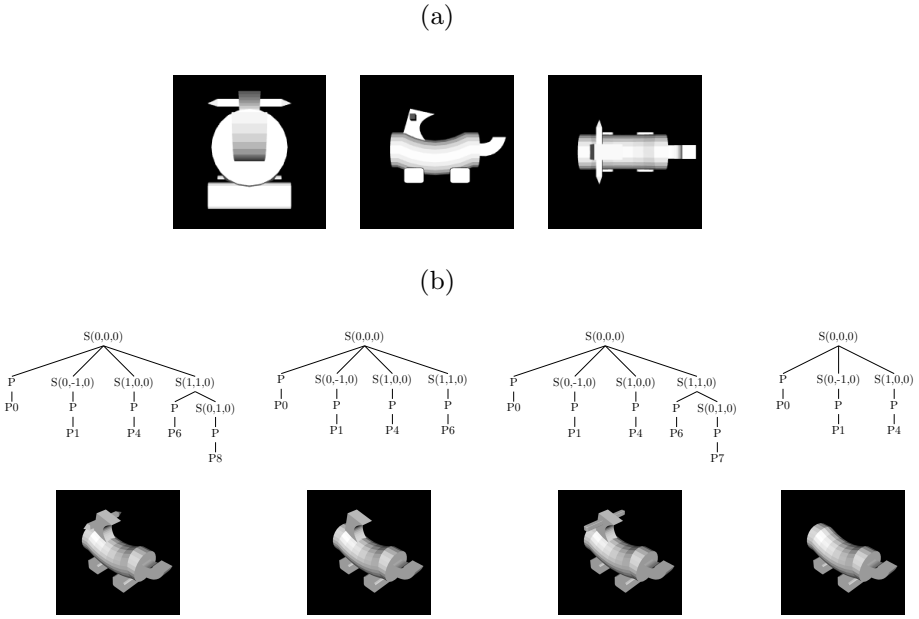


Figure 2.4: Results from a representative simulation of the MVH model. (a) Visual input to the model. (b) Four estimates of modality-independent object representations (parse trees augmented with spatial information) with the highest posterior probabilities (top) and their images (bottom). The MAP estimate is on the left. Each S (spatial) node denotes a position in 3D space relative to its parent S node. P (part) nodes specify the part located at its parent S node position. For example, in all the trees here P0 is located at its ancestor S node's position, which is the origin. The depth of a P node corresponds roughly to its distance from the origin. Please refer to Methods for more details.

Critically, the model shows perfect modality invariance. That is, it performs identically in vision, haptic, and multisensory conditions, meaning the

model produces the same MAP estimate of an object's parts and spatial relations among these parts regardless of whether the object is viewed, grasped, or both. For example, if the model is given the haptic features of the object shown in Fig. 2.4a (instead of images of the object), its MAP estimate is still the parse tree on the left of Fig. 2.4b. This result demonstrates that the object representations acquired by the model are modality independent. For this reason, we do not discuss separately the model's performances in vision, haptic, and multisensory conditions—these performances are identical.

Comparison with human data

Above, the motivations and merits of our computational model were described based primarily on theoretical grounds. Here, we evaluate the MVH model based on its ability to provide an account of human experimental data. The experiment reported here is related to the experiments of Wallraven, Bülthoff, and colleagues who asked subjects to rate the similarity of pairs of objects when objects were viewed, grasped, or both (Cooke, Kannengiesser, Wallraven, & Bulthoff, 2006; Cooke et al., 2007; Gaissert et al., 2010, 2011; Gaissert & Wallraven, 2012). However, our experiment also includes a crossmodal condition in which subjects rated object similarity when one object was viewed and the other object was grasped.

In brief (experimental details are given in the Methods section), the stimuli were the 16 objects described above (Fig. 2.2). On each trial, a subject observed two objects and judged their similarity on a scale of 1 (low similar-

ity) to 7 (high similarity). The experiment included four conditions referred to as the visual, haptic, crossmodal, and multisensory conditions. Different groups of subjects were assigned to different conditions. In the visual condition, subjects viewed images of two objects on each trial. In the haptic condition, subjects grasped physical copies of two objects (fabricated using 3-D printing) on each trial. In the crossmodal condition, subjects viewed an image of one object and grasped a second object on each trial. Finally, in the multisensory condition, subjects viewed and grasped two objects on each trial.

Experimental results: If people's perceptions of object shape are modality invariant, then subjects in all conditions should perform the experimental task in the same manner: On each trial, a subject represents the intrinsic shape properties of the two observed objects in a modality-independent format, and then the two modality-independent object representations are compared to generate a similarity judgment. The goal of the analyses of our experimental data is to evaluate whether subjects in fact based their similarity judgments on modality-independent shape representations. We look at this question by testing various predictions of the modality-invariance hypothesis. First, if people's perceptions of object shape are modality invariant, their similarity judgments should be quite similar regardless of modality. Hence, one would expect to see high correlations between similarity judgments not only within conditions but also across conditions. We test this prediction

with our first analysis below. A much stronger test of modality invariance is possible if we can somehow find the shape representations subjects employed in each condition. We can then simply compare these representations to evaluate modality invariance. In our second set of analyses, we use additive clustering and Multidimensional Scaling (MDS) to infer the perceptual space for each condition and compare them.

First, we looked at the average of subjects' similarity ratings for identical objects; this provides us with a coarse measure of modality invariance as well as a measure of objective accuracy. As expected from modality invariant representations, these ratings were nearly 7 (Visual: 6.89±0.27, Haptic: 6.74±0.47, Crossmodal: 6.71±0.49, Multisensory: 6.82±0.35). To address the question of modality invariance further, we proceeded as follows. First, for each subject in our experiment, we formed a subject-level similarity matrix by averaging the subject's ratings for each pair of objects. Next, we correlated a subject's similarity matrix with the matrices for subjects in the same experimental condition and in other conditions. The average correlations are shown in Table 2.1. These correlations are large, ranging from 0.76 to 0.91 (explaining 58%-83% of the variance in subjects' ratings). To test if these values are significantly greater than zero, we transformed them using the Fisher z-transformation. A t-test using the transformed correlations indicated that all correlations are significantly greater than zero ($p < 0.001$ in all cases). We are primarily concerned with whether subjects from different conditions gave similar similarity ratings, and thus we closely examined

the average correlations when subjects were in different conditions (for example, cells Visual-Haptic or Visual-Crossmodal, but not Visual-Visual or Haptic-Haptic, in the matrix in Table 2.1). Using t-tests, we asked if each of these correlations is "large", which we (arbitrarily, but not unreasonably) defined as meaning that a correlation explains at least 50% of the variance in subjects' ratings. All of these correlations were found to be large by this definition ($p < 0.001$ in all cases). Lastly, for each condition, we also formed a condition-level similarity matrix by averaging the subject-level matrices for the subjects belonging to that condition. As shown in Table 2.2, correlations among these condition-level matrices were extremely high, with the smallest correlation equal to 0.97 (explaining 94% of the variance in subjects' ratings across conditions). Taken as a whole, our correlational analyses strongly suggest that subjects had similar notions of object similarity in all experimental conditions. In other words, subjects' similarity ratings were modality invariant.

We further analyzed the experimental data using a Bayesian nonparametric additive clustering technique due to Navarro and Griffiths (2008). This technique makes use of the Indian Buffet Process (Griffiths & Ghahramani, 2011), a latent feature model recently introduced in the Machine Learning and Statistics literatures. In brief, the technique infers the latent or hidden features of a set of stimuli from their similarities. In our context, the technique assumes that subjects' similarity ratings are generated from hidden or latent binary object representations. Using Bayes' rule, it inverts

	Visual	Haptic	Crossmodal	Multisensory
Visual	0.91 ± 0.0341	0.83 ± 0.0855	0.86 ± 0.0729	0.89 ± 0.0572
Haptic		0.76 ± 0.1032	0.78 ± 0.1052	0.81 ± 0.0962
Crossmodal			0.8 ± 0.0968	0.83 ± 0.0866
Multisensory				0.86 ± 0.0651

Table 2.1: Average correlations within and across conditions among subjects' similarity matrices.

For example, the value in the Visual-Visual cell was calculated by averaging over correlations in subjects' similarity ratings for each pair of subjects in the visual condition (because there were 7 subjects in this condition, there were $42 = 7 \times 6$ such pairs). Similarly, the value in the Visual-Haptic cell was calculated by averaging over correlations for each pair of subjects when one subject was in the visual condition and the other subject was in the haptic condition (because there were 7 subjects in each condition, there were 49 such pairs).

	Visual	Haptic	Crossmodal	Multisensory
Visual	0.99 ± 0.0072	0.95 ± 0.0206	0.96 ± 0.0138	0.97 ± 0.0134
Haptic		0.97 ± 0.0241	0.94 ± 0.023	0.94 ± 0.0232
Crossmodal			0.97 ± 0.0199	0.96 ± 0.017
Multisensory				0.98 ± 0.0136

Table 2.2: Correlations based on condition-level similarity matrices formed by averaging subject-level matrices for the subjects belonging to each condition.

Means and standard deviations are estimated with a bootstrap procedure with 1000 replications.

this generative process so that similarity ratings are used to infer probability distributions over object representations. In other words, the input to the technique is a matrix of similarity ratings. Its output is a probability distribution over object representations where representations that are likely to give rise to the similarity ratings are assigned higher probabilities. The dimensionality of the binary object representations is not fixed. Rather, the technique infers a probability distribution over this dimensionality.

We applied the technique to each of the condition-level similarity matrices. In all conditions, it revealed that the most probable dimensionality was eight (i.e., similarity ratings in all conditions were most likely based on object representations consisting of eight binary features). However, the technique inferred two identical copies of each dimension, a potential problem noted by Navarro and Griffiths (2008). Consequently, the technique actually inferred four-dimensional object representations in all conditions. Interestingly, these object representations can be interpreted as "part based" representations of our experimental stimuli. Recall the structure of the experimental objects. There are four locations on objects at which parts vary. At each location, there are two interchangeable parts, only one of which is present in a given object. As a matter of notation, label the first set of interchangeable parts as {P1, P2}, the second set as {P3, P4}, and so on. An object can, therefore, be represented by four binary numbers. One number indicates which part is present in the set {P1, P2}, another number indicates which part is present in the set {P3, P4}, etcetera. We refer to this as a list-of-parts object representation.

The Bayesian nonparametric additive clustering technique inferred the same list-of-parts object representation as its MAP estimate when applied to every condition-level similarity matrix. This is important because it suggests that the same object representations underlied subjects' similarity ratings in visual, haptic, crossmodal, and multisensory experimental conditions. That is, this analysis of our data suggests that subjects used modality-independent

representations, and thus our data are consistent with the hypothesis that subjects' object perceptions were modality invariant. Importantly, the result did not have to come out this way. If the additive clustering technique inferred different object representations when applied to different condition-level similarity matrices, this outcome would have been inconsistent with the hypothesis of modality invariance.

The fact that the additive clustering technique always inferred *part-based* representations is also noteworthy. In hindsight, however, it might be unsurprising for subjects to have used part-based representations. Recall that our stimuli were generated by combining distinct parts. It seems likely that subjects would be sensitive to the structure of this generative process. Moreover, previous theoretical and empirical studies have indicated that people often use part-based object representations (Marr & Nishihara, 1978; Hoffman & Richards, 1984; Biederman, 1987; Tversky, 1989; Saiki & Hummel, 1998).

Lastly, we analyzed subjects' similarity ratings using non-metric multidimensional scaling (MDS) with the Manhattan distance function. Given a condition-level similarity matrix, MDS assigns locations in an abstract space to objects such that similar objects are nearby and dissimilar objects are far away (Shepard, 1962; Kruskal, 1964; Cox & Cox, 2000). To evaluate the dimensionality of this abstract space, we computed the "stress" value, a goodness-of-fit measure, for several different dimensionalities. In addition, we also calculated the Bayesian Information Criterion (BIC) score for each dimensionality. When using MDS, there are potential pitfalls when averaging

similarity judgments of different subjects. If different subjects use different abstract spaces, then averaging will lose this information. In addition, average similarity ratings can be fit well by MDS regardless of the nature of individual subject's ratings due to the increased symmetry of the average ratings (Ashby, Maddox, & Lee, 1994). Lee and Pope (2003) developed a BIC score that ameliorates these potential pitfalls. This score takes into account both the fit and complexity of an MDS model. The results based on stress values and BIC scores are shown in Figs. 2.5a and 2.5b, respectively. In both cases, values typically reach a minimum (or nearly so) at four dimensions in all experimental conditions. In Fig. 2.6, we plot the MDS space with four dimensions for the crossmodal condition. The results for other conditions are omitted since they are all qualitatively quite similar. In each panel of Fig. 2.6, we plot two of the four dimensions against each other, i.e., project the 4D space down to 2D. What is striking is the clear clustering in all panels. We see four clusters of four objects where each dimension takes one of two possible values. This is precisely the list-of-parts representation found by the Bayesian nonparametric additive clustering technique.

In summary, our correlational analyses of the experimental data reveal that subjects made similar similarity judgments in visual, haptic, crossmodal, and multisensory conditions. This indicates that subjects' judgments were modality invariant. Our analyses using a Bayesian nonparametric additive clustering technique and using multidimensional scaling indicate that subjects formed the same set of modality-independent object representations in

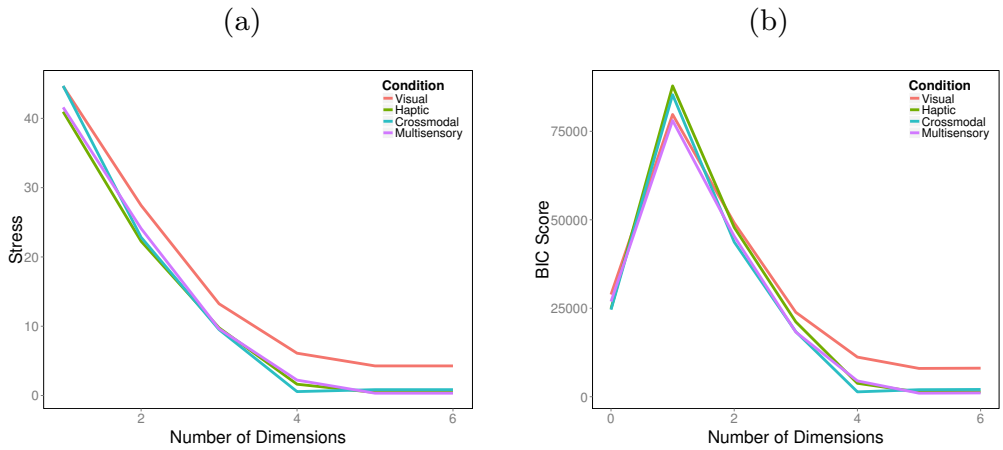


Figure 2.5: Results from MDS analysis. MDS (a) stress values and (b) BIC scores as a function of the number of dimensions.

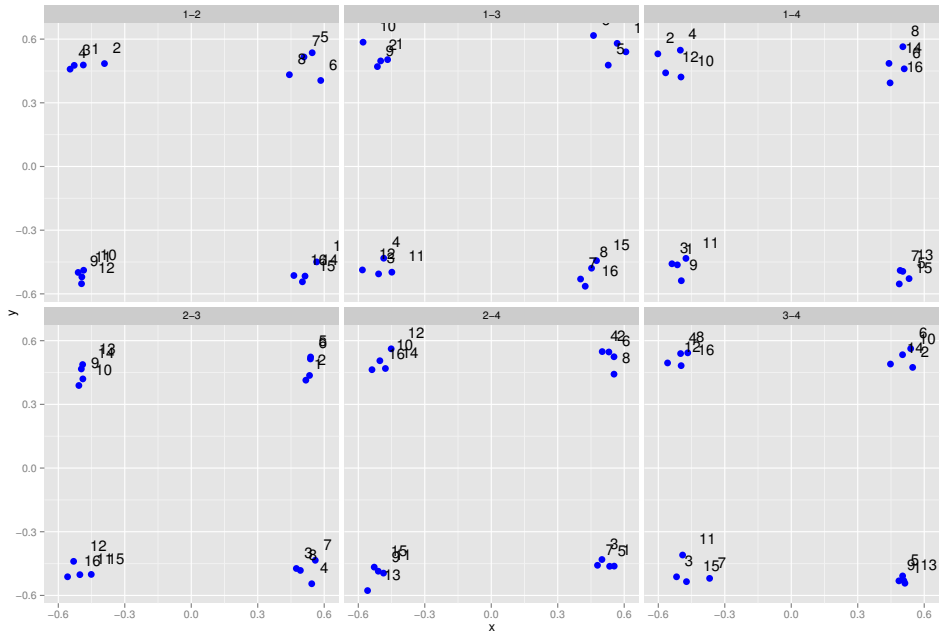


Figure 2.6: 4D MDS space for the crossmodal condition. Each panel plots one of the six possible 2D projections of the 4D MDS space. Each point corresponds to one of the 16 objects.

all conditions.

Simulation results: Here we evaluate whether the MVH model provides a good account of our experimental data. To conduct this evaluation, however, the model must be supplemented with an object similarity metric. Such a metric could potentially take several different forms. For example, object similarity could be computed based on modality-independent features. Alternatively, it could be based on modality-specific features such as visual or haptic features.

Researchers studying how people represent space have made a surprising discovery. Spatial locations can be represented in many different reference frames, such as eye-centered, head-centered, body-centered, or limb-position centered coordinate systems. Counterintuitively, people often transform representations of spatial locations into a common reference frame, namely an eye-centered reference frame, when planning and executing motor movements (Cohen & Andersen, 2000, 2002; Pouget, Ducom, Torri, & Bavelier, 2002; Schlicht & Schrater, 2007).

These studies raise an interesting issue: In what reference frame do people judge object similarity? Do they judge object similarity in a modality-independent feature space? Or do they judge object similarity in a sensory-specific feature space such as a visual or haptic space? Here we address these questions by augmenting the MVH model with different object similarity functions.

The hypothesis that people's percepts are modality invariant predicts that people judge object similarity based on the values of modality-independent features. An alternative possibility is that people acquire modality-independent object representations when objects are viewed and/or grasped, but then re-represent objects in terms of visual features for the purpose of judging object similarity. The mapping from modality-independent to visual features could be achieved by a vision-specific forward model. A second alternative is that people re-represent objects in terms of haptic features (via a haptic-specific forward model) to judge object similarity. Because the MVH model includes modality-independent representations along with vision-specific and haptic-specific forward models, it can be used to evaluate these different possibilities.

In one set of simulations, the model was used to compute object similarity in a modality-independent feature space. On each simulated trial, the model computed modality-independent representations for two objects. Next, the objects' similarity was estimated using a tree-based similarity measure known as "tree edit distance" (K. Zhang & Shasha, 1989). In brief, this measure has a library of three tree-based operators: rename node, remove node, and insert node. Given two modality-independent object representations—that is, two spatial trees or MAP estimates of the shapes of two objects—this similarity measure counts the number of operators in the shortest sequence of operators that converts one representation to the other representation (or vice versa). For similar representations, the representation for object A can be converted to the representation for object B using a short operator sequence, and thus

these representations have a small distance. For dissimilar representations, a longer operator sequence is required to convert one object representation to the other, and thus these representations have a large distance. In our simulations, we first placed the object representations in a canonical form, and then measured pairwise distances between objects using the tree edit distance measure of K. Zhang and Shasha (1989).

In a second set of simulations, the model was used to compute object similarity in a visual feature space. As above, the model was used to acquire modality-independent representations for two objects on each simulated trial. Next, the vision-specific forward model was used to map each object representation to images of the represented object, thereby re-representing each object from a modality-independent reference frame to a visual reference frame. Given three images from orthogonal viewpoints of each object (see Fig. 2.4a), the similarity of the two objects was estimated as the Euclidean distance between the images of the objects based on their pixel values.

In a final set of simulations, the model was used to compute object similarity in a haptic feature space. This set is identical to the set described in the previous paragraph except that the haptic-specific forward model (GraspIt!) was used to map each object representation to sets of a simulated hand's joint angles, thereby re-representing each object from a modality-independent reference frame to a haptic frame. Given sets of joint angles for each object, the similarity of two objects was estimated as the Euclidean distance between the haptic features of the objects based on their associated joint angles.

Which set of simulations produced object similarity ratings matching the ratings provided by our experimental subjects? For ease of explanation, we refer to the model augmented with the modality-independent based, visual-based, and haptic-based similarity functions as the MVH-M, MVH-V, and MVH-H models, respectively. The results for these three models are shown in Figs. 2.7, 2.8, and 2.9. In each figure, the four graphs correspond to the visual, haptic, crossmodal, and multisensory conditions. The horizontal axis of each graph shows subjects' object similarity ratings (averaged across all subjects, and linearly scaled to range from 0 to 1). The vertical axis shows a model's similarity ratings (linearly scaled to range from 0 to 1). Each graph contains 136 points, one point for each possible pair of objects. The correlation (denoted R) between subject and model ratings is reported in the top-left corner of each graph.

A comparison of these figures reveals that the object similarity ratings of the MVH-M model provide an excellent quantitative fit to subjects' ratings. Indeed, the correlation R ranges from 0.975 to 0.987 across the different experimental conditions (explaining 95%-97% of the variance in ratings). In other words, the MVH-M model provides a (nearly) perfect account of our experimental data. The MVH-V model provides a reasonably good fit to subjects' data, though this fit is not as good as the fit provided by the MVH-M model. Based on a two-tailed t-test using the Fisher z-transformation, correlations for the MVH-M model are always greater than the corresponding correlations for the MVH-V model ($p < 0.05$). In addition, correlations for

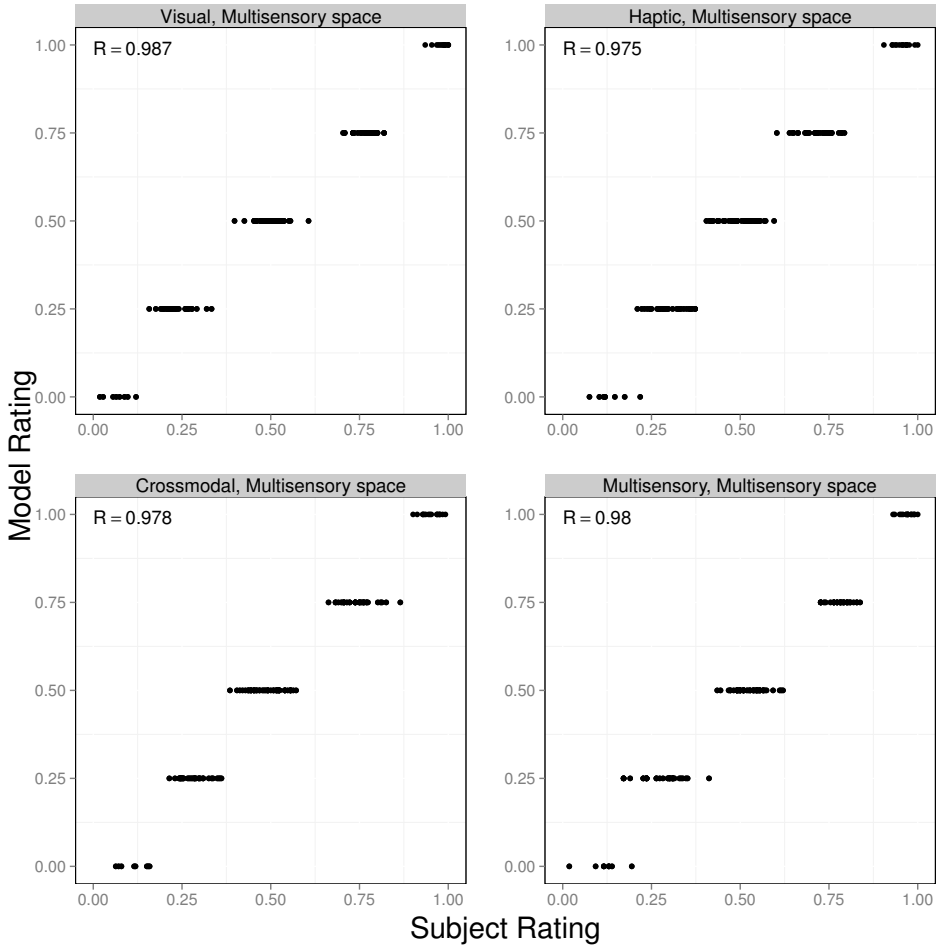


Figure 2.7: Results for the MVH-M model (this model computes object similarity in a modality-independent feature space). The four graphs correspond to the visual (top left), haptic (top right), crossmodal (bottom left), and multisensory (bottom right) experimental conditions. The horizontal axis of each graph shows subjects' object similarity ratings (averaged across all subjects, and linearly scaled to range from 0 to 1). The vertical axis shows the model's similarity ratings (linearly scaled to range from 0 to 1). The correlation (denoted R) between subject and model ratings is reported in the top-left corner of each graph. Note that MVH-M model's similarity ratings take only a finite number of different values since parse trees are discrete structures, and therefore tree-edit distance returns only integer values.

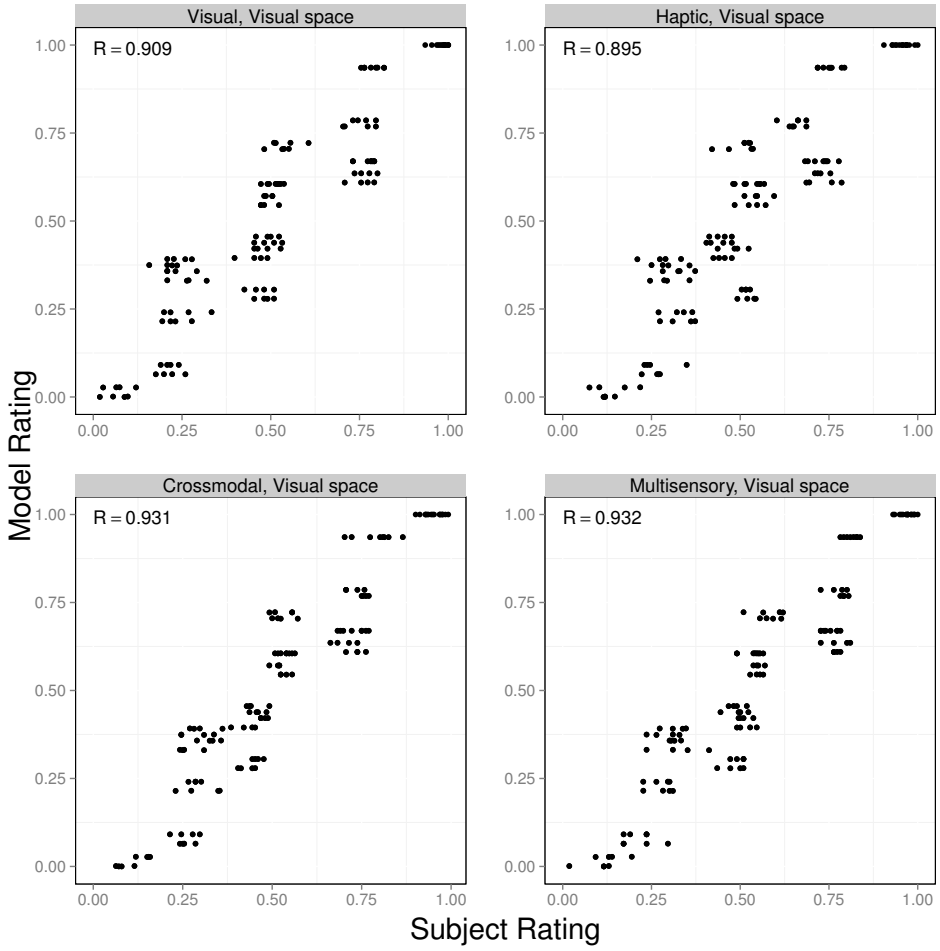


Figure 2.8: Results for the MVH-V model (this model computes object similarity in a visual feature space). The format of this figure is identical to the format of Fig. 2.7.

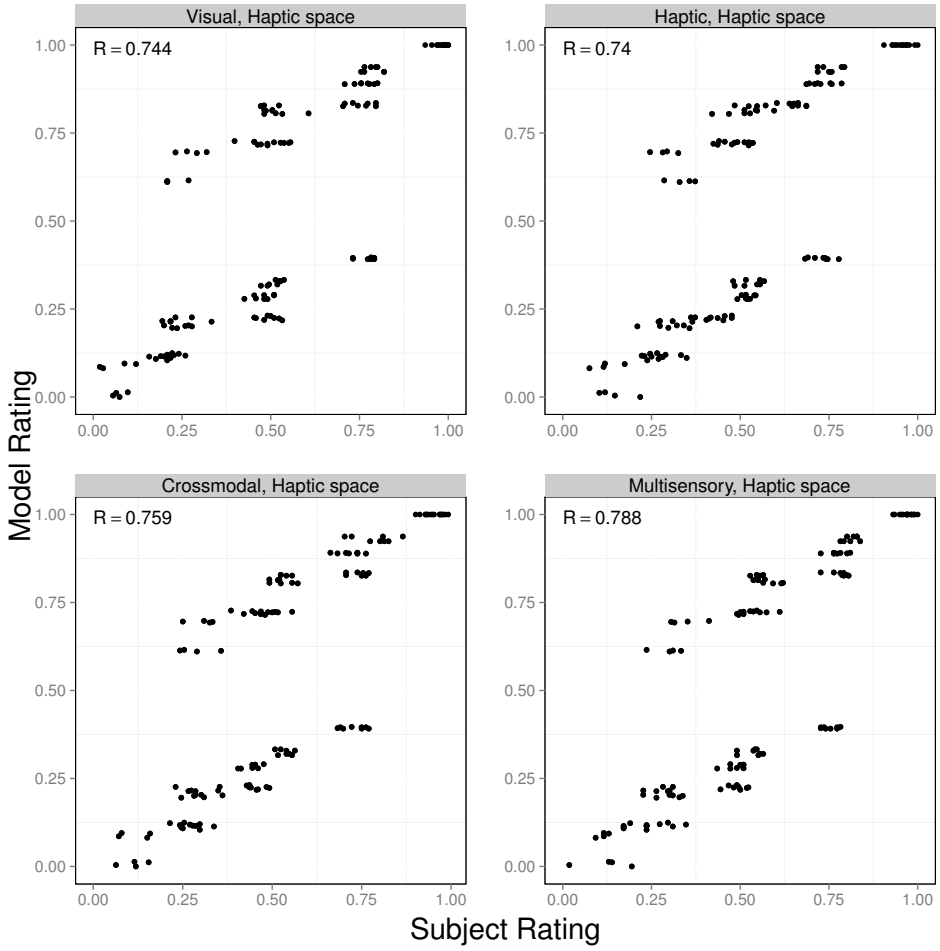


Figure 2.9: Results for the MVH-H model (this model computes object similarity in a haptic feature space). The format of this figure is identical to the format of Fig. 2.7.

the MVH-M model and the MVH-V model are always greater than those of the MVH-H model. That is, the MVH-M model performs best, followed by the MVH-V model, and then the MVH-H model.

In summary, we have compared the performances of three models. All models represent objects in a modality-independent manner. However, the models differ in the space in which they calculate object similarity. One model calculates similarity using modality-independent features (MVH-M), another model maps modality-independent features to visual features and calculates similarity on the basis of these visual features (MVH-V), and a final model maps modality-independent features to haptic features and calculates similarity on the basis of these haptic features (MVH-H). Our results show that the MVH-M model's similarity ratings provide the best quantitative fit to subjects' ratings. Consequently, we hypothesize that subjects computed object similarity in a modality-independent feature space. That is, subjects acquired modality-independent object shape representations based on visual signals, haptic signals, or both, and then compared two objects' shape representations in order to judge their similarity.

Discussion

This paper has studied the problem of learning modality-independent, conceptual representations from modality-specific sensory signals. We hypothesized that any system that can accomplish this feat will include three compo-

nents: a representational language for characterizing modality-independent representations, a set of sensory-specific forward models for mapping from modality-independent representations to sensory signals, and an inference algorithm for inverting forward models (i.e., an algorithm for using sensory signals to infer modality-independent representations).

To evaluate our theoretical framework, we instantiated it in the form of a computational model that learns object shape representations from visual and/or haptic signals. The model uses a probabilistic context-free grammar to characterize modality-independent representations of object shape, uses a computer graphics toolkit (VTK) and a human hand simulator (GraspIt!) to map from object representations to visual and haptic features, respectively, and uses a Bayesian inference algorithm to infer modality-independent object representations from visual and/or haptic signals. Simulation results show that the model infers identical object representations when an object is viewed, grasped, or both. That is, the model's percepts are modality invariant. It is worth pointing out that the particular implementational choices we have made in our model are in some sense arbitrary; any model that instantiates our framework will be able to capture modality invariance. Therefore, from this perspective, our particular model in this work should be taken as one concrete example of how modality independent representations can be acquired and used.

Our work in this paper focused on showing how our framework can capture one aspect of multisensory perception, i.e., modality invariance. We

take this as an encouraging first step in applying our framework to multisensory perception more generally. We believe other aspects of multisensory perception (such as cue combination, crossmodal transfer of knowledge, and crossmodal recognition) can be easily understood and treated in our framework.

The paper also reported the results of an experiment in which different subjects rated the similarity of pairs of objects in different sensory conditions, and showed that the model provides a very good account of subjects' ratings. Our experimental results suggest that people extract modality independent shape representations from sensory input and base their judgments of similarity on such representations. The success of our model in accounting for these results are important from two perspectives. First, from a larger perspective, it is significant as a validation of our theoretical framework. Second, it constitutes an important contribution to cognitive modeling, particularly an emerging probabilistic language-of-thought approach, by showing how symbolic and statistical approaches can be combined in order to understand aspects of human perception.

Related research

Our theoretical framework is closely related to the long standing vision-as-inference (Kersten & Yuille, 2003) approach to visual perception. In this approach, the computational problem of visual perception is formalized as the inversion of a generative process; this generative process specifies how the

causes in the world, e.g., objects, give rise to 2D images on the retina. Then, the purpose of the visual system is to invert this generative model to infer the most likely causes, i.e., the explanation, for the observed sensory data. This approach, also called analysis-by-synthesis, has featured prominently both in cognitive science (Kersten, Mamassian, & Yuille, 2004; Yuille & Kersten, 2006) and computer vision (S.-C. Zhu & Mumford, 2006; Kulkarni, Mansinghka, Kohli, & Tenenbaum, 2014; Kulkarni, Yildirim, Kohli, Freiwald, & Tenenbaum, 2014). Our work here can be seen as the application of this approach to multisensory perception.

Previous research has instantiated our general theoretical framework in other ways. For example, Yildirim and Jacobs (2012) developed a latent variable model of multisensory perception. In this model, modality-independent representations are distributed representations over binary latent variables. Sensory-specific forward models map the modality-independent representations to sensory (e.g., visual, auditory, haptic) features. The acquisition of modality-independent representations takes place when a Bayesian inference algorithm (the Indian Buffet Process (Griffiths & Ghahramani, 2011)) uses the sensory features to infer these representations. Advantages of this model include the fact that the dimensionality of the modality-independent representations adapts based on the complexity of the training data set, the model learns its sensory-specific forward models, and the model shows modality invariance. Disadvantages include the fact that the inferred modality-independent representations (distributed representations over latent vari-

ables) are difficult to interpret, and the fact that the sensory-specific forward models are restricted to being linear. Perhaps its biggest disadvantage is that it requires a well-chosen set of sensory features in order to perform well on large-scale problems. In the absence of good sensory features, it scales poorly, mostly due to its linear sensory-specific forward models and complex inference algorithm.

As a second example, Yildirim and Jacobs (2013) described a model of visual-haptic object shape perception that is a direct precursor to the MVH model described in this paper. Perhaps its biggest difference with the model presented here is that it represents parts as generalized cylinders, and parts connect to each other using a large number of "docking locations". This strategy for representing object shape provides enormous flexibility, but this flexibility comes at a price. Inference using this model is severely underconstrained. Consequently, the investigators designed a customized (i.e., *ad hoc*) Bayesian inference algorithm. Despite the use of this algorithm, inference is computationally expensive. That is, like the latent variable model described in the previous paragraph, the model of Yildirim and Jacobs (2013) scales poorly.

Probabilistic language-of-thought

We believe that the MVH model described in this paper has significant theoretical and practical advantages over alternatives. These arise primarily due to its use of a highly structured implementation of a representational

language for characterizing modality-independent representations. In particular, the model combines symbolic and statistical approaches to specify a probabilistic context-free object shape grammar. Due to this shape grammar, the model is able to use a principled inference algorithm that has previously been applied to probabilistic grammars in other domains. We find that inference in the model is often computationally tractable. We are reasonably optimistic that the model (or, rather, appropriately extended versions of the model) will scale well to larger-scale problems. Although important challenges obviously remain, our optimism stems from the fact that shape grammars (much more complex than the one reported here) are regularly used in the Computer Vision and Computer Graphics literatures to address large-scale problems. In addition, due to its principled approach, the model should be easy to extend in the future because relationships between the model and other models in the Cognitive Science and Artificial Intelligence literatures using grammars, such as models of language, are transparent. As a consequence, lessons learned from other models will be easy to borrow for the purpose of developing improved versions of the model described here.

In Cognitive Science, there are many frameworks for cognitive modeling. For example, one school of thought favors symbolic approaches, such as approaches based on grammars, production rules, or logic. An advantage of symbolic approaches is their rich representational expressiveness—they can often characterize a wide variety of entities in a compact and efficient manner. A disadvantage of these approaches is that they are often "brittle" when

used in noisy or uncertain environments. An alternative school of thought favors statistical approaches, such as approaches based on neural networks or Bayesian inference. An advantage of statistical approaches is their ability to learn and adapt, and their robustness to noise and uncertainty. Their main disadvantage is that they often require highly structured prior distributions or likelihood functions to work well (Tenenbaum et al., 2011). Advocates of symbolic and statistical schools of thought have often engaged in heated debates (McClelland & Patterson, 2002b, 2002a; Pinker & Ullman, 2002b, 2002a). Unfortunately, these debates have not led to a resolution as to which approach is best.

A recently emerging viewpoint in the Cognitive Science literature is that both symbolic and statistical approaches have important merits, and thus it may be best to pursue a hybrid framework taking advantage of each approach's best aspects (Goodman et al., 2008; Kemp & Tenenbaum, 2008; Piantadosi et al., 2012; T. D. Ullman et al., 2012). This viewpoint is referred to here as a "probabilistic language of thought" approach because it applies probabilistic inference to a representation consisting of symbolic primitives and combinatorial rules (Fodor, 1975). To date, the probabilistic language-of-thought approach has been used almost exclusively in domains that are typically modeled using symbolic methods, such as human language and high-level cognition. A significant contribution of the research presented here is that it develops and applies this approach in the domain of perception, an area whose study is dominated by statistical techniques.

Future research

We foresee at least three areas of future research. First, the framework described here sheds light on modality invariance. Future work will need to study whether this framework also sheds light on other aspects of multisensory perception and cognition. For example, can the framework be used to understand why our percepts based on two modalities are often more accurate than our percepts based on a single modality, why training with two modalities is often superior to training with a single modality (even when testing is conducted in unisensory conditions), or why crossmodal transfer of knowledge is often, but not always, successful? Future work will also need to study the applicability of the framework to other sensory domains, such as visual and auditory or auditory and haptic environments. Future work will also need to consider how our framework can be extended to study the acquisition of other types of conceptual knowledge from sensory signals.

Second, future research will need to study the role of forward models in perception and cognition. For example, we have speculated that sensory-specific forward models may be ways of implementing sensory imagery, and thus our framework predicts a role for imagery in multisensory perception. Behavioral, neurophysiological, and computational studies are needed to better understand and evaluate this hypothesis. From a technological perspective, it is advantageous that we live in a "golden age" of forward models. New and improved forward models are frequently being reported in the scientific literature and made available on the world wide web (e.g., physics

engines providing approximate simulations of physical systems such as rigid body dynamics or fluid dynamics). These forward models will allow cognitive scientists to study human perception, cognition, and action in much more realistic ways than has previously been possible.

Finally, cognitive scientists often make a distinction between rational models and process models (Anderson, 1990). Rational models (or computational theories (Marr, 1982)) are models of optimal or normative behavior, characterizing the problems that need to be solved in order to generate the behavior as well as their optimal solutions. In contrast, process models (or models at the "representation and algorithm" level of analysis (Marr, 1982)) are models of people's behaviors, characterizing the mental representations and operations that people use when generating their behavior. Because the MVH model's inference algorithm is optimal according to Bayesian criteria, and because this algorithm is not psychologically plausible, the model should be regarded as a rational model, not as a process model. Nonetheless, we believe that there are benefits to regarding the MVH model as a rational/process hybrid. Like rational models, the MVH model is based on optimality considerations. However, like process models, it uses psychologically plausible representations and operations (e.g., grammars, forward models).

For readers solely interested in process models, we claim that the MVH model is a good starting point. As pointed out by others (Sanborn, Griffiths, & Navarro, 2010; Griffiths, Vul, & Sanborn, 2012), the MCMC inference algorithm used by the MVH model can be replaced by approximate inference

algorithms (known as particle filter or sequential Monte Carlo algorithms) that are psychologically plausible. Doing so would lead to a so-called "rational process model", a type of model that is psychologically plausible and also possesses many of the advantages of rational models. Future work will need to study the benefits of extending our framework through the use of psychologically plausible and approximately optimal inference algorithms to create rational process models of human perception.

Methods

Ethics statement

The experiments were approved by the Research Subjects Review Board of the University of Rochester. All subjects gave informed consent.

Multisensory-Visual-Haptic (MVH) model

Shape grammar: The production rules of the MVH model's shape grammar are shown in Fig. 2.10. The grammar is an instance of a probabilistic context-free grammar. However, probabilities for each production rule are not shown in Fig. 2.10 because our statistical inference procedure marginalizes over the space of all probability assignments (see below). Production rules characterize the number of parts and the specific parts comprising an object. The rules contain two non-terminal symbols, S and P. Non-terminal

P is always replaced by a terminal representing a specific object part. Non-terminal S is used for representing the number of parts in an object. Production rules are supplemented with additional information characterizing the spatial relations among parts.

$$\begin{aligned} S &\rightarrow S \mid SS \mid SSS \mid SSSS \mid P \mid PS \mid PSS \mid PSSS \\ P &\rightarrow P0 \mid P1 \mid P2 \mid P3 \mid P4 \mid P5 \mid P6 \mid P7 \mid P8 \end{aligned}$$

Figure 2.10: Production rules of the shape grammar in Backus-Naur form. S denotes spatial nodes, and P refer to part nodes. S is also the start symbol of the grammar. P1, P2, etc. are the object parts as seen in Fig. 2.1.

An object is generated using a particular sequence of production rules from the grammar. This sequence is known as a derivation which can be illustrated using a parse tree. To represent the spatial relations among object parts, a parse tree is extended to a spatial tree. Before describing this extension, it will be useful to think about how 3-D space can be given a multi-resolution representation. At the coarsest resolution in this representation, a "voxel" corresponds to the entire space. The center location of this voxel is the origin of the space, denoted (0, 0, 0). At a finer resolution, this voxel is divided into 27 equal sized subvoxels arranged to form a $3 \times 3 \times 3$ grid. Using a Cartesian coordinate system with axes labeled x, y, and z, a coordinate of a subvoxel's location along an axis is either -1, 0, or 1. For example, traversing the z-axis would reveal subvoxels located at (-1, -1, -1), (-1, -1, 0), and (-1, -1, 1). This process can be repeated. For instance, the subvoxel at (-1, -1, -1) can be divided into 27 subsubvoxels. The coordinates

of subsubvoxels would also be either -1, 0, or 1. Note that the location of a subsubvoxel is relative to the location of its parent subvoxel which, in turn, is relative to the location of its parent voxel.

The addition of multi-resolution spatial information to a parse tree converts this tree to a spatial tree. This process is illustrated in Fig. 2.11. Consider the object shown in Fig. 2.11a and the spatial tree for the derivation of this object shown in Fig. 2.11b. The root S node is associated with a voxel centered at the origin (0, 0, 0) of the 3-D space. This node is expanded using the rule $S \rightarrow PSSS$, and locations are assigned to the subvoxels associated with the S nodes [in the figure, these locations are (0, -1, 0), (1, 0, 0), and (-1, 1, 0), respectively]. The P node is replaced with terminal P0 representing the cylindrical body (see Fig. 2.1). This part is placed at the location of its grandparent S node. The two leftmost S nodes in the second level of the tree are eventually replaced with terminals P1 and P3, respectively. These parts are placed at the locations of their grandparent S nodes. The rightmost S node at the second level is expanded using the production $S \rightarrow PS$, and a location is assigned to the S node [(0, 1, 0)]. The P node is replaced with terminal P5. The final S node is eventually replaced with terminal P7.

The multi-resolution representation of 3-D space, and the placement of parts in this space is illustrated in Fig. 2.11c. Two facts about spatial trees are evident from this figure. First, smaller-sized voxels are reached as one moves deeper in a tree, enabling the model to make finer-grained assignments of locations to object parts. Second, with the exception of the root S node,

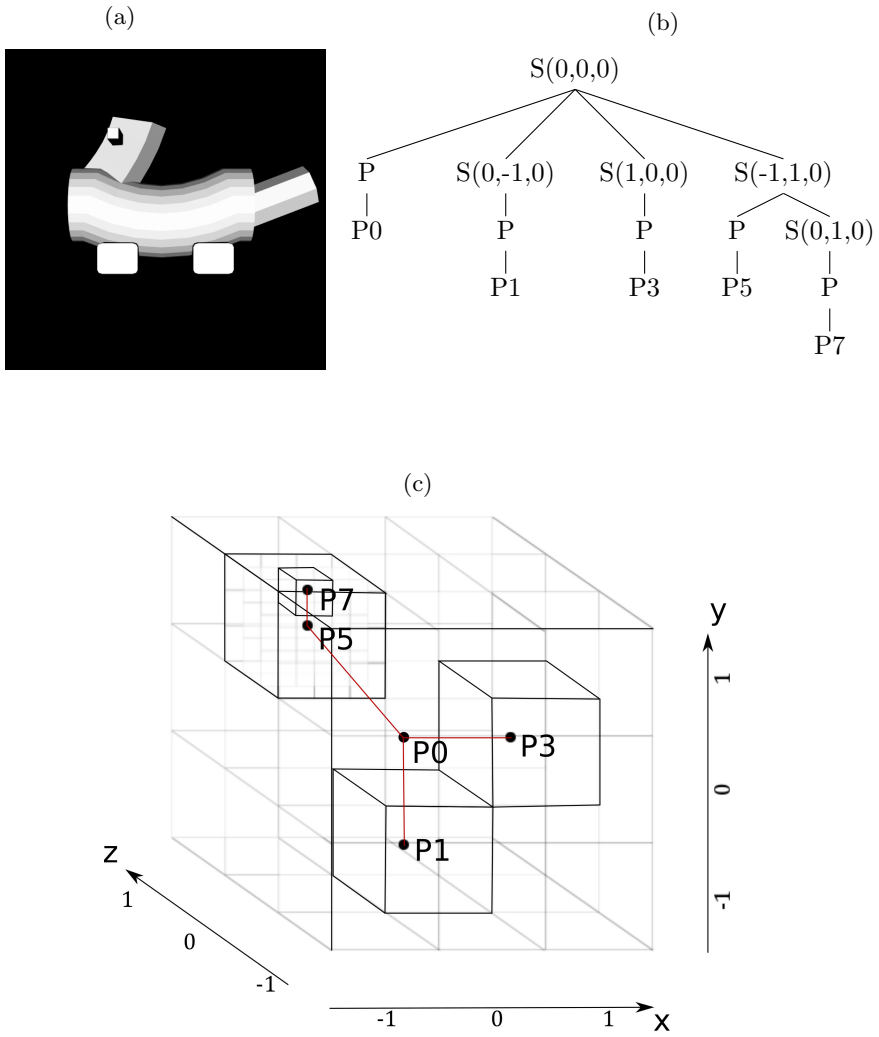


Figure 2.11: Illustration of the multi-resolution representation of 3-D space. (a) Image of an object. (b) Spatial tree representing the parts and spatial relations among parts for the object in (a). (c) Illustration of how the spatial tree uses a multi-resolution representation to represent the locations of object parts.

an S node is never associated with a voxel located at (0, 0, 0) because this would create a situation in which two parts are assigned the same location.

There are several properties of the model's shape grammar and spatial trees that were chosen for convenience: (i) The maximum branching factor of the shape grammar is four; (ii) The creation of spatial trees through the addition of spatial information to parse trees is not strictly necessary. An equivalent representation could be achieved by a more complicated grammar with productions for all possible voxel coordinate assignments to child S nodes; and (iii) Without loss of generality, the set of possible object parts was chosen for convenience. In other situations, other sets could be selected (indeed, one could imagine a system that uses a segmentation algorithm to learn good sets). In addition, the fact that object parts are at fixed scales and orientations is not strictly necessary. More complicated spatial trees could allow for scaling and rotation of parts. Our point here is that the probabilistic shape grammar approach is general and powerful, though the full generality and power of this approach is not needed for our current purposes. Readers interested in how shape grammars can be used to characterize objects and scenes in more realistic settings should consult the Computer Vision and Computer Graphics literatures (Fu, 1986; Bienenstock et al., 1997; Tu et al., 2005; Amit & Trouve, 2007; Grenander & Miller, 2007; L. Zhu et al., 2009; Talton et al., 2012; Felzenszwalb, 2013).

Prior distribution over object representations: An object representation consists of two components, a parse tree, denoted $\mathcal{T}$, and a spatial model, denoted $\mathcal{S}$. The prior probability for an object representation is defined as:

$$P(\mathcal{T}, \mathcal{S}|\mathcal{G}) = P(\mathcal{T}|\mathcal{G})P(\mathcal{S}|\mathcal{T}) \tag{2.1}$$

where $\mathcal{G}$ denotes the shape grammar.

Due to the nature of our grammar, an object has a unique derivation, and thus a unique parse tree. Recall that a derivation is a sequence of productions in the shape grammar that ends when all non-terminals are replaced with terminals. At each step of a derivation, a choice is made among the productions which could be used to expand a non-terminal. Because a probability is assigned to each production choice in a derivation, the probability of the complete derivation is the product of the probabilities for these choices. That is, the probability of a parse tree is:

$$P(\mathcal{T}|\mathcal{G}, \rho) = \prod_{n \in \mathcal{N}_{nt}} P(n \rightarrow ch(n)|\mathcal{G}, \rho) \tag{2.2}$$

where $\mathcal{N}_{nt}$ is the set of non-terminal nodes in the tree, $ch(n)$ is the set of node n's children nodes, and $P(n \rightarrow ch(n)|\mathcal{G}, \rho)$ is the probability for production rule $n \rightarrow ch(n)$. In this equation, ρ denotes the set of probability assignments to production rules. Allowing for uncertainty in these production

probabilities, we integrate over ρ:

$$P(\mathcal{T}|\mathcal{G}) = \int P(\mathcal{T}|\mathcal{G},\rho)P(\rho|\mathcal{G})d\rho. \tag{2.3}$$

Because there is no reason to prefer any specific set of production probabilities, we assume that $P(\rho|\mathcal{G})$ is a uniform distribution. With this assumption, the integral has a Multinomial-Dirichlet form, and thus can be solved analytically:

$$P(\mathcal{T}|\mathcal{G}) = \prod_{s\in\mathcal{G}_{nt}} \frac{\beta(\mathbf{C}(\mathcal{T},s)+\mathbf{1})}{\beta(\mathbf{1})}. \tag{2.4}$$

Here, $\mathcal{G}_{nt}$ is the set of non-terminal symbols in grammar $\mathcal{G}$, $\beta(\cdot)$ is the multinomial beta function, $\mathbf{1}$ is a vector of ones, and $\mathbf{C}(\mathcal{T},s)$ is a vector of counts of the productions for non-terminal s in parse tree $\mathcal{T}$ (the count of a rule increments each time the rule is used).

An advantage of this distribution over parse trees is that it favors "simple" trees, meaning trees corresponding to short derivations. (To see this, note that Equation 2.2 multiplies probabilities [numbers less than one]. The number of terms that are multiplied increases with the length of the derivation.) Consequently, it can be regarded as a type of Occam's Razor.

In addition to the probability of parse tree $\mathcal{T}$, the calculation of the prior probability of an object representation also requires the probability of spatial model $\mathcal{S}$ (Equation 2.1). Recall that model $\mathcal{S}$ contains the voxel coordinates for each S node in a parse tree. Let $\mathcal{V}$ denote the set of possible voxel coordinates, a set with 26 elements (the $3 \times 3 \times 3$ grid yields 27 subvoxels

but the subvoxel centered at (0, 0, 0) is not a valid spatial assignment). Using $\mathcal{N}_S$ to denote the set of S nodes in tree $\mathcal{T}$, and assuming that all voxel coordinates are equally likely, the probability of model $\mathcal{S}$ is:

$$P(\mathcal{S}|\mathcal{T}) = \prod_{n \in \mathcal{N}_S} \frac{1}{|\mathcal{V}|} = \frac{1}{|\mathcal{V}|^{|\mathcal{N}_S|}}. \tag{2.5}$$

As above, this distribution favors spatial models associated with small parse trees, and thus is a type of Occam's Razor.

Likelihood function: Recall that an object representation consists of a parse tree $\mathcal{T}$ and a spatial model $\mathcal{S}$. Let D denote actual sensory data perceived by an observer, either visual features, haptic features, or both. Let $F(\mathcal{T}, \mathcal{S})$ denote predicted sensory features, predicted by the visual-specific forward model (VTK), the haptic-specific forward model (GraspIt!), or both. To define the likelihood function, we assume that perceived sensory data D is equal to predicted sensory features $F(\mathcal{T}, \mathcal{S})$ plus random noise distributed according to a Gaussian distribution:

$$P(D|\mathcal{T}, \mathcal{S}) \propto \exp\left(-\frac{||D - F(\mathcal{T}, \mathcal{S})||_2^2}{\sigma^2}\right) \tag{2.6}$$

where σ^2 is a variance parameter.

MCMC algorithm: Using Bayes' rule, the MVH model combines the prior distribution and the likelihood function to compute a posterior dis-

tribution over object representations:

$$P(\mathcal{T}, \mathcal{S}|D, \mathcal{G}) \propto P(D|\mathcal{T}, \mathcal{S})P(\mathcal{S}|\mathcal{T})P(\mathcal{T}|\mathcal{G}) \tag{2.7}$$

where the three terms on the right-hand side are given by Equations 2.6, 2.5, and 2.4, respectively. Unfortunately, exact computation of the posterior distribution is intractable. We, therefore, developed a Markov chain Monte Carlo (MCMC) algorithm that discovers good approximations to the posterior.

MCMC is a family of methods for sampling from a desired probability distribution by constructing a Markov chain that has the desired distribution as its stationary distribution. A common MCMC method is the Metropolis-Hastings (MH) algorithm (Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953; Hastings, 1970). This algorithm produces a sequence of samples. At each iteration, the algorithm picks a candidate for the next sample value based on the current sample value. With some probability, the candidate is accepted meaning that the candidate value is used in the next iteration or rejected meaning this value is discarded and the current value is reused in the next iteration.

In the context of our simulations, a value is a multisensory object representation—that is, a parse tree $\mathcal{T}$ and a spatial model $\mathcal{S}$. At each iteration, our algorithm proposes a new representation, denoted $(\mathcal{T}', \mathcal{S}')$, based on the current representation $(\mathcal{T}, \mathcal{S})$ with probability given by proposal distribution

$q(\mathcal{T}', \mathcal{S}'|\mathcal{T}, \mathcal{S})$. The new representation is accepted with a probability based on acceptance function $A(\mathcal{T}', \mathcal{S}'; \mathcal{T}, \mathcal{S})$.

We used two different proposal distributions in our simulations, one on even-numbered iterations and the other on odd-numbered iterations (Brooks, 1998; Tierney, 1994). The subtree-regeneration proposal distribution was originally developed by Goodman et al. (2008). When using this proposal distribution, a non-terminal node is randomly selected from parse tree $\mathcal{T}$, all its descendants are removed, and new descendants are generated according to the rules of the shape grammar. Nodes removed from the parse tree are also removed from the spatial model, and random voxel coordinates are sampled for newly added nodes. The new representation is accepted with probability equal to the minimum of 1 and the value of an acceptance function:

$$A(\mathcal{T}', \mathcal{S}'; \mathcal{T}, \mathcal{S}) = \frac{P(D|\mathcal{T}', \mathcal{S}')}{P(D|\mathcal{T}, \mathcal{S})} \frac{P(\mathcal{T}'|\mathcal{G})}{P(\mathcal{T}|\mathcal{G})} \frac{|\mathcal{N}_{nt}|}{|\mathcal{N}'_{nt}|} \frac{P(\mathcal{T}|\mathcal{G}, \rho)}{|P(\mathcal{T}'|\mathcal{G}, \rho)} \tag{2.8}$$

where $\mathcal{N}_{nt}$ and $\mathcal{N}'_{nt}$ are the sets of all non-terminals in tree $\mathcal{T}$ and $\mathcal{T}'$, respectively.

Sole use of the subtree-regeneration proposal did not produce an efficient MCMC algorithm for our problem. This is mainly due to the fact that the algorithm sometimes proposes a new object representation which is very different from the current representation, thereby losing the desirable aspects of the current representation. Consider a scenario in which the current representation is partially correct, such as the parse tree in Fig. 2.12a. Based on

this tree, it is difficult to propose the more correct tree in Fig. 2.12b without losing the desirable aspects of the current tree. To do so, the algorithm would have to choose the root node, thereby deleting nearly all of the current tree, and then generate the proposal tree nearly from scratch.

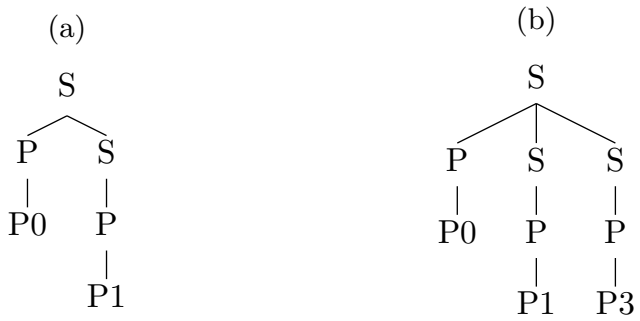


Figure 2.12: Parse trees for illustrating a difficulty with using the subtree-regeneration proposal. (a) Partially correct tree for a hypothetical example. (b) The "true" tree for the example. Note that it is impossible to propose the tree in (b) from the tree in (a) with a subtree-regeneration proposal without deleting and regenerating all the nodes.

This observation led us to design the add/remove-part proposal. This proposal adds or removes object parts to a representation making it possible, for example, to propose the tree in Fig. 2.12b based on the tree in Fig. 2.12a, or vice versa. The proposal starts by randomly choosing whether to add or remove an object part. If adding a part, it draws a random part by choosing a terminal symbol from the grammar. Then it chooses an S node that has less than four children and adds a new S node as a child to the chosen S node. Finally, it expands the child S node to a P node and the P node to the chosen part. If removing a part, an S node that has a P node as its only

child is chosen. This node and its descendants are removed. However, the proposal never chooses the root S node or an S node that is the only child of its parent as these will result in ungrammatical trees. The spatial model is updated accordingly. If a part is added, a random voxel coordinate is sampled for the newly added S node. If a part is removed, the corresponding S node (and its voxel coordinate) is removed. Assuming that representation $(\mathcal{T}', \mathcal{S}')$ is proposed by adding a new part to $(\mathcal{T}, \mathcal{S})$, the new representation is accepted with probability equal to the minimum of 1 and the value of the acceptance function:

$$A(\mathcal{T}', \mathcal{S}'; \mathcal{T}, \mathcal{S}) = \frac{P(D|\mathcal{T}', \mathcal{S}')}{P(D|\mathcal{T}, \mathcal{S})} \frac{P(\mathcal{T}'|\mathcal{G})}{P(\mathcal{T}|\mathcal{G})} \frac{|\mathcal{A}|}{|\mathcal{R}'|} |\mathcal{G}_t| \tag{2.9}$$

where $\mathcal{R}'$ is the set of S nodes in tree $\mathcal{T}'$ that can be removed, $\mathcal{A}$ is the set of S nodes in tree $\mathcal{T}$ to which a new child S node can be added, and $\mathcal{G}_t$ is the set of terminal symbols in the grammar. Similarly, the acceptance function when removing a part is:

$$A(\mathcal{T}', \mathcal{S}'; \mathcal{T}, \mathcal{S}) = \frac{P(D|\mathcal{T}', \mathcal{S}')}{P(D|\mathcal{T}, \mathcal{S})} \frac{P(\mathcal{T}'|\mathcal{G})}{P(\mathcal{T}|\mathcal{G})} \frac{|\mathcal{R}|}{|\mathcal{A}'||\mathcal{G}_t|} \tag{2.10}$$

where $\mathcal{R}$ is the set of S nodes in tree $\mathcal{T}$ that can be removed, $\mathcal{A}'$ is the set of S nodes in tree $\mathcal{T}'$ to which a new child S node can be added.

It is easy to show that our algorithm is a valid Metropolis-Hastings sampler, meaning that it has the posterior distribution over multisensory object representations as its stationary distribution. Derivations for the acceptance

functions for the subtree-regeneration and add/remove-part proposals are straightforward. Readers interested in these topics should contact the first author.

In our simulations, each MCMC chain was run for 10,000 iterations. Samples from the first 6,000 iterations were discarded as "burn-in".

Experimental Details

Stimuli: The experiment used the 16 objects in Fig. 2.2. Visual stimuli consisted of images of objects rendered from a canonical (three-quarter) viewpoint so that an object's parts and spatial relations among parts are clearly visible (Fig. 2.2). Stimuli were presented on a 19-inch CRT computer monitor. Subjects sat approximately 55 cm from the monitor. When displayed on the monitor, visual stimuli spanned about 20 degrees in the horizontal dimension and 15 degrees in the vertical dimension. Visual displays were controlled using the PsychoPy software package (Peirce, 2007).

Subjects received haptic inputs when they touched physical copies of the objects fabricated using a 3-D printing process (Fig. 2.2). Physical objects were approximately 11.5 cm long, 6.0 cm wide, and 7.5 cm high. Subjects were instructed to freely and bimanually explore physical objects.

Procedure: On each experimental trial, a subject observed two objects and judged their similarity on a scale of 1 (low similarity) to 7 (high similarity). Within a block of 136 trials, each object was paired both with itself (16 trials)

and with the other objects (each object could be paired with 15 other objects; ignoring order of object presentation [which was randomized], this results in 120 trials). Pairs were presented in random order. Subjects performed 4 blocks of trials.

The experiment included four conditions referred to as the visual, haptic, crossmodal, and multisensory conditions. Different groups of subjects were assigned to different conditions. We regard the crossmodal condition as the key experimental condition because it is the condition that directly evaluates the modality invariance of subjects' percepts. The visual, haptic, and multisensory conditions are control conditions in the sense that data from these conditions are of interest primarily because they allow us to better understand results from the crossmodal condition.

In the visual condition, subjects saw an image of one object followed by an image of a second object. Images were displayed for 3.5 seconds.

In the haptic condition, physical objects were placed in a compartment under the computer monitor. The end of the compartment closest to a subject was covered with a black curtain. A subject could reach under the curtain to haptically explore an object. However, a subject could not view an object. Messages on the computer monitor and auditory signals indicated to a subject when she or he could pick up and drop objects. On each trial, an experimenter first placed one object in the compartment. The subject then haptically explored this object. The experimenter removed the first object and placed a second object in the compartment. The subject explored this

second object. Each object was available for haptic exploration for 7 seconds. As is common in the scientific literature on visual-haptic perception, the haptic input in the haptic experimental condition was available for longer than the visual input in the visual condition (Freides, 1974; Newell & Ernst, 2001; Lacey et al., 2007; Gaissert et al., 2011).

In the crossmodal condition, objects in a pair were presented in different sensory modalities. For one subgroup of three subjects, the first object was presented visually and the second object was presented haptically. For another subgroup of four subjects, this order was reversed. We checked for a difference in ratings between the two subgroups. A two-tailed Welch's t-test (used when two samples have possibly unequal variances) did not find a significant effect of the order of the modalities in which objects were presented ($t = 0.087$, $p = 0.935$). We, therefore, grouped the data from these subgroups.

In the multisensory condition, both objects were presented both visually and haptically. During the 7 seconds in which an object could be touched, the visual image of the object was displayed for the final 3.5 seconds.

Visual and crossmodal conditions were run over two one-hour sessions on two different days, each session comprising two blocks of trials. For haptic and multisensory conditions, an individual block required about an hour to complete. These conditions were run over four one-hour sessions. Although subjects performed four blocks of trials, we discarded data from the first block because subjects were unfamiliar with the objects and with the exper-

imental task during this block. Results reported above are based on data from blocks 2-4.

Subjects: Subjects were 30 students at the University of Rochester who reported normal or corrected-to-normal visual and haptic perception. Subjects were paid $10 per hour. Of the 30 subjects, 2 subjects provided similarity ratings that were highly inconsistent across blocks (one subject in the visual condition and the other in the multisensory condition). A Grubbs test (Grubbs, 1950) using each subject's correlations among ratings in different blocks revealed that these two subjects' ratings are statistical outliers (Subject 1: $g = 2.185$, $p < 0.05$; Subject 2: $g = 2.256$, $p < 0.05$). These ratings were discarded from further analyses. The remaining 28 subjects were divided among the four experimental conditions, seven subjects per condition.

MVH-V and MVH-H models applied to the experimental data: The MVH-V and MVH-H models are equivalent to alternative models. For instance, consider a model that computes object similarity based solely on the pixel values of images of those objects. In fact, this is equivalent to MVH-V. This equivalency arises from the fact the the MVH model's MAP estimates of object shape are always correct (given an object, this estimate is the correct representation of the object in terms of the shape grammar). When MVH-V obtains images of two objects (by rendering the object representations using the vision-specific forward model), these images are also always correct (they are identical to the true images of the objects). Consequently, MVH-V

performs no differently than a model that rates object similarity based on the pixel values of images of objects. Given this fact, why is MVH-V needed? It is because people do not always have images of two objects (consider a case where one object is viewed and the other object is grasped). Analogous remarks apply to MVH-H.

Chapter 3

Multisensory Part-based Representations of Objects in Human Lateral Occipital Cortex

Introduction

While eating breakfast, the object shape you perceive when viewing your coffee mug is the same as the shape you perceive when grasping your mug. This phenomenon illustrates modality invariance, an important type of perceptual constancy. Modality invariance suggests that people have representations of objects that are multisensory (i.e., with a significant degree of modality independence).

From behavioral studies, we know that participants trained in the visual modality to recognize novel objects show partial or near-complete transfer to the haptic modality, and vice versa (Lawson, 2009; Lacey et al., 2007; Norman et al., 2004), and that object similarity is judged in similar ways across modalities (Gaissert & Wallraven, 2012; Gaissert et al., 2011, 2010;

Cooke et al., 2007, 2006). Those findings suggest that participants base their similarity judgments on a multisensory representation. Where is the neural substrate for these representations and how are the representations structured?

Prior brain imaging work suggests that human lateral occipital cortex (LOC) is one seat of multisensory representations of object shape, at least across the visual and haptic modalities. Previous research shows that LOC represents visual information about object shape (Grill-Spector et al., 2001; Kourtzi & Kanwisher, 2001) and responds to haptic exploration of objects in sighted and congenitally blind individuals (Naumer et al., 2010; Amedi et al., 2002; James et al., 2002; Amedi et al., 2001). Furthermore, neural shape similarity matrices from blind participants are correlated with neural shape similarity matrices from sighted individuals (Peelen, He, Han, Caramazza, & Bi, 2014), suggesting that LOC is biased to represent object shape even if the principal modality of input is not vision.

To date, researchers have relied mainly on two measures—amount of neural "activation" (e.g., BOLD contrast) and correlations between neural similarity matrices—to argue for the multisensory nature of representations in LOC. Most studies compared the amount of BOLD contrast in LOC in response to visually and haptically presented stimuli. For example, James et al. (2002) showed that both visual and haptic exploration of objects led to neural activity in LOC. Similarly, Amedi et al. (2001, 2002) argued for multisensory shape representations in LOC on the basis of increased neural

activity in response to objects compared with textures for visual and haptic stimuli. In a more recent study, Naumer et al. (2010) showed that the amount of neural activation when stimuli are presented through both visual and haptic modalities is higher than the amount of neural activation when stimuli are presented through a single modality.

Importantly, comparing the amount of activation in response to visual and haptic presentation of objects is an indirect test of multimodality of neural representations. It is quite possible that LOC carries distinct modality-specific representations for both visual and haptic object shape. A stricter test is possible by measuring the similarity in patterns of neural activity. Recently, Peelen et al. (2014) calculated neural similarity matrices for a set of objects presented visually and verbally to blind and sighted individuals. By measuring the correlations between these neural similarity matrices, Peelen et al. (2014) argued that LOC carries a cross-modal shape representation. With respect to our current goals, there are two limitations associated with this study. First, Peelen et al. (2014) did not measure neural activity in response to haptic stimuli. Second, the correlation between two neural similarity matrices is a measure of second-order relations between two representations. It is possible for visual and haptic neural similarity matrices to be highly correlated even though the visual and haptic representations themselves are not. Here, we present a stricter test of the multisensory nature of object representations in LOC by correlating activations from different modalities directly to form cross-modal neural similarity matrices. Our analyses show

that cross-modal correlation of an object with itself is larger than the cross-modal correlations among different objects and that objects can be decoded cross-modally from neural activations in LOC.

The second question we focus on is concerned with the structure of multisensory shape representations in LOC. Two competing theories emerge from previous research on object shape representations. First, view-based theories argue that the representation for an object is a collection of 2D images of the object from different views (Peissig & Tarr, 2007). View dependency of object recognition is usually advanced as the main evidence for the view-based hypothesis. For example, a previous study (Bulthoff & Edelman, 1992) showed that the recognition performance for previously seen views of an object is better than the performance for views of the same object not previously seen. However, the view-based hypothesis is difficult to reconcile with the hypothesis that LOC encodes multisensory object representations, because the view-based hypothesis presumes a strictly visual nature of object representations.

Alternatives to the view-based hypothesis are part-based or structural description theories (e.g., Peissig & Tarr, 2007; Riddoch & Humphreys, 1987). These theories assume that objects are represented as collections of parts and the spatial relations among these parts. There is behavioral and neural evidence for both aspects of the part-based theory: representation of parts and spatial relations among those parts. An influential study by Biederman (1987) showed that priming is principally mediated by parts, and recognition

suffers dramatically when part-related information is removed. Later studies also investigated whether spatial relations are explicitly represented. For example, Hayworth et al. (2011) found that it was impossible for participants to ignore relations between objects in a scene even when that information was irrelevant. Importantly for our current study, previous work has found evidence that LOC encodes object parts and spatial relations explicitly. Using fMRI adaptation, Hayworth and Biederman (2006) found that, when part-related information was removed from an image, there was a release from adaptation in LOC, suggesting that different parts involve different LOC representations. A separate study (Hayworth et al., 2011) showed that a comparable amount of release from adaptation in LOC is observed when the spatial relation between two objects is changed as when one of the objects is replaced with a new object. This suggests that spatial relations are encoded explicitly by this region. More recently, Guggenmos et al. (2015) tested whether LOC encodes objects in a part-based or holistic manner by measuring decoding accuracy for split and intact objects. They showed that a classifier trained on neural activations for intact objects can successfully discriminate between activations for split objects (e.g., a camera with its lens and body separate) and vice versa. These studies suggest that LOC represents objects in a part-based format. Here, we provide further evidence for this hypothesis by showing that a novel object can be decoded from the neural activations in LOC based on part-based representations.

Methods

Participants

Twelve (six in Experiment 1 and six in Experiment 2) University of Rochester students (mean age = 21.5 years, SD = 1.57 years, five men) participated in the study in exchange for payment. All participants were right-handed (assessed with the Edinburgh Handedness Questionnaire), had normal or corrected-to normal vision, and had no history of neurological disorders. All participants gave written informed consent in accordance with the University of Rochester research subjects review board.

General Procedure

Stimulus presentation was controlled with "A Simple Framework" (Schwarzbach, 2011) written in MATLAB Psychtoolbox (Brainard, 1997; Pelli, 1997) or E-Prime Professional Software 2.0 (Psychology Software Tools, Inc., Sharpsburg, PA). For all fMRI experiments with visual presentation of stimuli, participants viewed stimuli binocularly through a mirror attached to the head coil adjusted to allow foveal viewing of a back-projected monitor (temporal resolution = 120 Hz). Each participant completed four 1-hr sessions: one session for retinotopic mapping and somatosensory and motor cortex mapping (data not analyzed herein), one session for an object-responsive cortex localizer, and two sessions for the experiment proper (visual and haptic ex-

ploration of objects).

Object-responsive Cortex Localizer (LOC Localizer)

The session began with (i) one 6-min run of resting state fMRI, (ii) eight 3-min runs of the object-responsive cortex localizer experiment, and (iii) one 6-min run of resting state fMRI. The resting state fMRI data are not analyzed herein.

To localize object-responsive areas in the brain, participants viewed scrambled and intact images of tools, animals, famous faces, and famous places (see Q. Chen, Garcea, & Mahon, 2016; Fintzi & Mahon, 2013). For each of four categories (tools, animals, faces, and places) 12 items were selected (e.g., hammer, Bill Clinton, etc.), and for each item, eight exemplars (gray-scale photographs) were selected (e.g., eight different hammers, eight different pictures of Bill Clinton, etc.). This resulted in a total of 96 images per category and 384 total images. Phase-scrambled versions of the stimuli were created to serve as a baseline condition. Participants viewed the images in a miniblock design. Within each 6-sec miniblock, 12 stimuli from the same category were presented, each for 500 msec (0 msec ISI), and 6-sec fixation periods were presented between miniblocks. Within each run, eight miniblocks of intact images and four miniblocks of phase-scrambled versions of the stimuli were presented with the constraint that a category of objects did not repeat during two successive miniblock presentations. All participants completed eight runs of the object-responsive cortex localizer experiment (91 volumes per

run).

Experimental Materials

The stimuli used in Experiment 1 were taken from the set of objects known as Fribbles (Tarr, 2003). We picked 12 Fribbles (four objects from three categories) for Experiment 1. For the stimuli used in Experiment 2, we created a new set of objects by taking parts from Fribbles and combining them in the following way. Each object is made up of five components where the body (one component) is common to all objects. The remaining four components are located at four fixed locations on the body. For each location, there are two possible parts or values that the component can take (i.e., 2×4 , hence 16 objects). Figures 3.1 and 3.2A show the entire set of objects used in Experiments 1 and 2, respectively. Figure 3.7 shows how we constructed the set of objects for Experiment 2 from the parts and how these were combined to create an example object. For the haptic stimuli, we used 3D-printed plastic models of the objects. The physical objects were approximately 11.5 cm long, 6.0 cm wide, and 7.5 cm high.

To summarize, the stimuli used in Experiment 1 were drawn from three "categories" of objects (four items per category) but the part structure was not explicitly (i.e., factorially) manipulated across the stimulus set. In contrast, in Experiment 2, the materials were created by creating all possible combinations of part values (two values) at each of four possible locations, leading to a factorial stimulus space defined by part structure.

Visual and Haptic Exploration of Novel Objects (Two Sessions)

Each participant completed two 1-hr sessions of the experiment proper. Each session was composed of four runs, two runs dedicated to visual exploration of objects and two runs dedicated to haptic exploration of objects. In the first experiment, the participants observed each novel object stimulus in the visual and haptic conditions; that is, all 12 objects were presented in each run. In the second experiment, the novel object stimuli were divided (arbitrarily) into two sets, A and B. Within a given scanning session, a participant was presented (for instance) Set A for haptic exploration and Set B for visual exploration; that is, in each run, participants saw eight objects. In their second session for the experiment proper, that same participant was presented Set B for haptic exploration and Set A for visual exploration. This ensured that participants only viewed or only haptically explored a given object in a given scanning session. The order of a given item set (Set A first, Set B first) by modality (visual, haptic) was also counterbalanced across participants. For both Experiments 1 and 2, visual and haptic exploration was blocked by run, organized in an ABBA/BAAB fashion, and counterbalanced evenly across participants.

While laying supine in the scanner, participants were visually presented with the objects or were required to keep their eyes closed while haptically exploring the objects. In the haptic condition, the objects were handed to the participant by the experimenter. For runs in which items were visually pre-

sented, participants were instructed to deploy their attention to the features of the object.

In the visual condition in Experiment 1, the objects were presented in the center of the screen for the participants to fixate upon. Miniblocks were 4-sec long and were interspersed by 8-sec fixation periods. Each object was presented in four miniblocks per run, with the constraint that the same object did not repeat on two successive miniblocks. This meant that there were a total of 48 (12×4) object presentations in each run. In Experiment 2, the objects were presented centrally and rotated 40 degrees per second along the vertical axis (i.e., the objects revolved in the depth plane). Miniblocks in the visual condition were 9-sec long and were interspersed by 9-sec fixation periods. Each object was presented in four miniblocks per run, in a similar manner to Experiment 1. Therefore, there were in total 32 (8×4) object presentations in each run. In the haptic condition, participants were instructed to form a mental image of the plastic object while haptically exploring the object with their hands. In Experiment 1, miniblocks were 12-sec long and were interspersed by 9-sec periods in which their hands were unoccupied. Each plastic object was presented in four miniblocks per run, with the constraint that the same item did not repeat across two successive miniblock presentations. Miniblocks in Experiment 2 were 16-sec long and were interspersed by 16-sec periods in which their hands were unoccupied. Each plastic object was presented in four miniblocks per run, in a similar manner to Experiment 1.

In our experiments, participants performed no explicit task other than visually or haptically exploring the presented objects. We believe such a design enables us to investigate visual-haptic processing without any potential task-related effects. Previous research shows that, even in the absence of any explicit task, visual and haptic processing converges in LOC (Naumer et al., 2010). Although our participants did not perform an explicit task, we asked them to mentally picture the object they were exploring in the haptic condition. This might raise suspicions about whether the activation in LOC was due to mental imagery rather than haptic processing. However, previous research suggests that LOC is minimally activated by mental imagery (James et al., 2002; Amedi et al., 2001).

Before the experiment began, participants were introduced to comparable plastic objects outside the scanner. These objects were not used in the experiment proper and were dissimilar to the experimental stimuli. Visual analogs of the objects were also presented to the participants to inform them of the format of the visual experiment and to practice the implicit task that they were required to carry out while in the scanner.

MR Acquisition and Analysis

MRI Parameters

Whole-brain BOLD imaging was conducted on a 3-T Siemens (Amsterdam, The Netherlands) MAGNETOM Trio scanner with a 32-channel head coil

located at the Rochester Center for Brain Imaging. High-resolution structural T1 contrast images were acquired using a magnetization prepared rapid gradient-echo pulse sequence at the start of each participant's first scanning session (repetition time = 2530, echo time = 3.44 msec, flip angle = 7°, field of view = 256 mm, matrix = 256 × 256, 1 × 1 × 1 mm sagittal left-to-right slices). An EPI pulse sequence was used for T2* contrast (repetition time = 2000 msec, echo time = 30 msec, flip angle = 90°, field of view = 256 × 256 mm, matrix = 64 × 64, 30 sagittal left-to-right slices, voxel size = 4 × 4 × 4 mm). The first six volumes of each run were discarded to allow for signal equilibration (four at acquisition and two at analysis).

fMRI Data Analysis

fMRI data were analyzed with the BrainVoyager software package (Version 2.8) and in-house scripts drawing on the BVQX toolbox written in MATLAB (`wiki2.brainvoyager.com/bvqxtools`). Preprocessing of the functional data included, in the following order, slice scan time correction (sinc interpolation), motion correction with respect to the first volume of the first functional run, and linear trend removal in the temporal domain (cutoff: two cycles within the run). Functional data were registered (after contrast inversion of the first volume) to high-resolution deskulled anatomy on a participant-by-participant basis in native space. For each participant, echo-planar and anatomical volumes were transformed into standardized space (Talairach & Tournoux, 1988). Functional data for the localizer experiment

(object-responsive cortex localizer) were smoothed at 6 mm FWHM (1.5 mm voxels) and interpolated to 3 mm^3 voxels; functional data for the experiment proper (visual and haptic exploration of objects) were interpolated to 3 mm^3 but were not spatially smoothed.

For all experiments, the general linear model was used to fit beta estimates to the experimental events of interest. Experimental events were convolved with a standard 2-gamma hemodynamic response function. The first derivatives of 3D motion correction from each run were added to all models as regressors of no interest to attract variance attributable to head movement. Thus, all multi-voxel pattern analyses were performed over beta estimates.

In all multivoxel analyses, we normalized individual voxel activations within a run to remove baseline differences across runs. In other words, for each voxel, we subtracted the mean activation for that voxel over all objects in the run and divided it by the standard deviation of that voxel's activation across objects. Additionally, for linear correlation multivoxel analyses, activations for all eight repeats of a single item (in a given modality, i.e., visual/haptic) were averaged to obtain a single activation vector for each item. In our correlation analyses, we transformed correlation values using Fisher's z transformation and ran all statistical tests on those transformed values. When calculating correlations between correlation matrices, we used only the upper triangles of matrices. All statistical tests were two-tailed. For training the support vector machine (SVM) for decoding, we used the

library libsvm (`www.csie.ntu.edu.tw/~cjlin/libsvm/`). We used linear kernels with cost parameter set to 1.

Whole-brain pattern analyses were performed using a searchlight approach (Kriegeskorte et al., 2006). Whole-brain searchlight maps were computed with a mask fit to the deskulled Talairach anatomy of individual participants. The "searchlight" passes over each voxel (in each participant) and extracts the beta estimates (for 16 items) for the cube of voxels (n = 125) that surround the voxel. The analysis was carried out based on the pattern of responses across the 125 voxels, and the results were assigned to the center voxel of that cube. All whole-brain analyses were thresholded at $p < .005$ (corrected), cluster threshold for nine contiguous voxels. If no regions were observed at that threshold, a more lenient threshold was used ($p < .05$, uncorrected, nine voxels).

Definition of ROIs (LOC)

Left and right LOC were identified at the group level using the object-responsive localizer experiment with the contrast of [intact images] > [scrambled images]. The result used cluster size corrected alpha levels by thresholding individual voxels at $p < .05$ (uncorrected) and applying a subsequent cluster size threshold generated with a Monte Carlo style permutation test (1000 iterations) on cluster size to determine the appropriate alpha level that maintains Type I error at 1% (using AlphaSim as implemented in Brain Voyager). The Talairach coordinates were as follows: left LOC: $x = 40, y = 71, z = 9$;

right LOC: $x = 38, y = 65, z = 12$. We note as well that none of the results in this study change qualitatively if LOC is defined individually for each participant, rather than at the group level.

Results

Our study consisted of two experiments. In both experiments, participants either viewed or haptically explored a set of objects during fMRI. The stimuli for Experiment 1 consisted of 12 objects (four objects from three categories; see Figure 3.1) picked from the set of objects known as Fribbles (Tarr, 2003). For Experiment 2, we created a novel set of objects based on Fribbles. Each object in this set was composed of one component that was common to all objects and four components that varied across objects. The variable components were located at four fixed locations (Figure 3.2A), and there were two possible parts (or values) that each component could take (i.e., $2^4 = 16$ objects in total).

Cross-modal Decoding of Novel Objects in LOC

If object representations in LOC are multisensory across haptic and visual modalities, it should be possible to decode object identity using cross-modal representational similarity analyses. To that end, we correlated the voxel patterns in LOC elicited when a participant was viewing objects with the voxel patterns elicited when the same participant haptically explored the

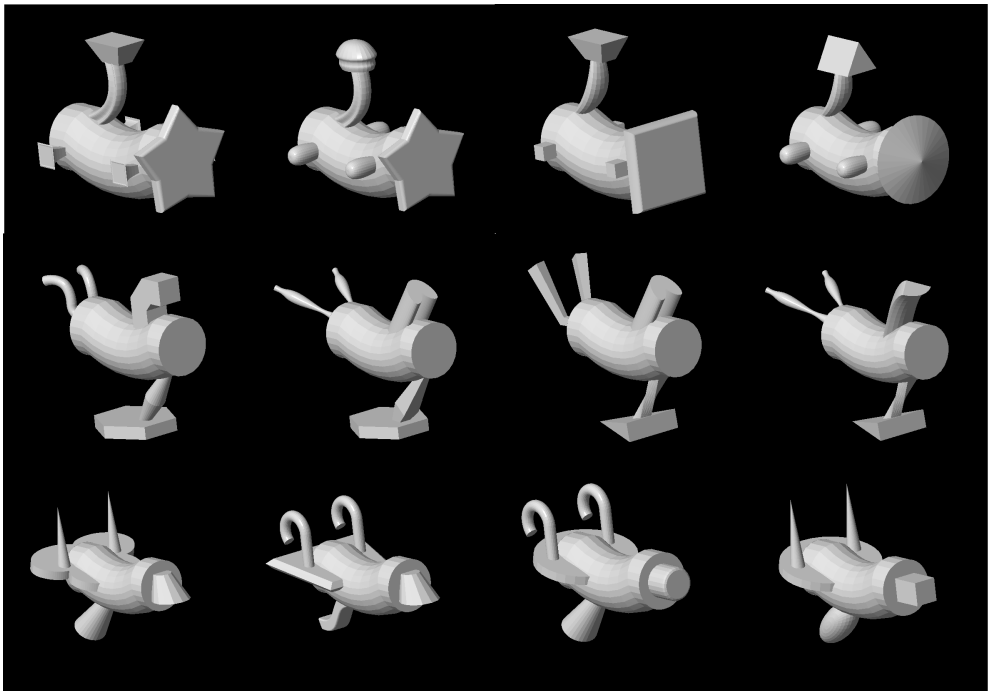


Figure 3.1: Experimental stimuli used in Experiment 1. The stimuli are taken from the set of novel objects known as Fribbles (Tarr, 2003).

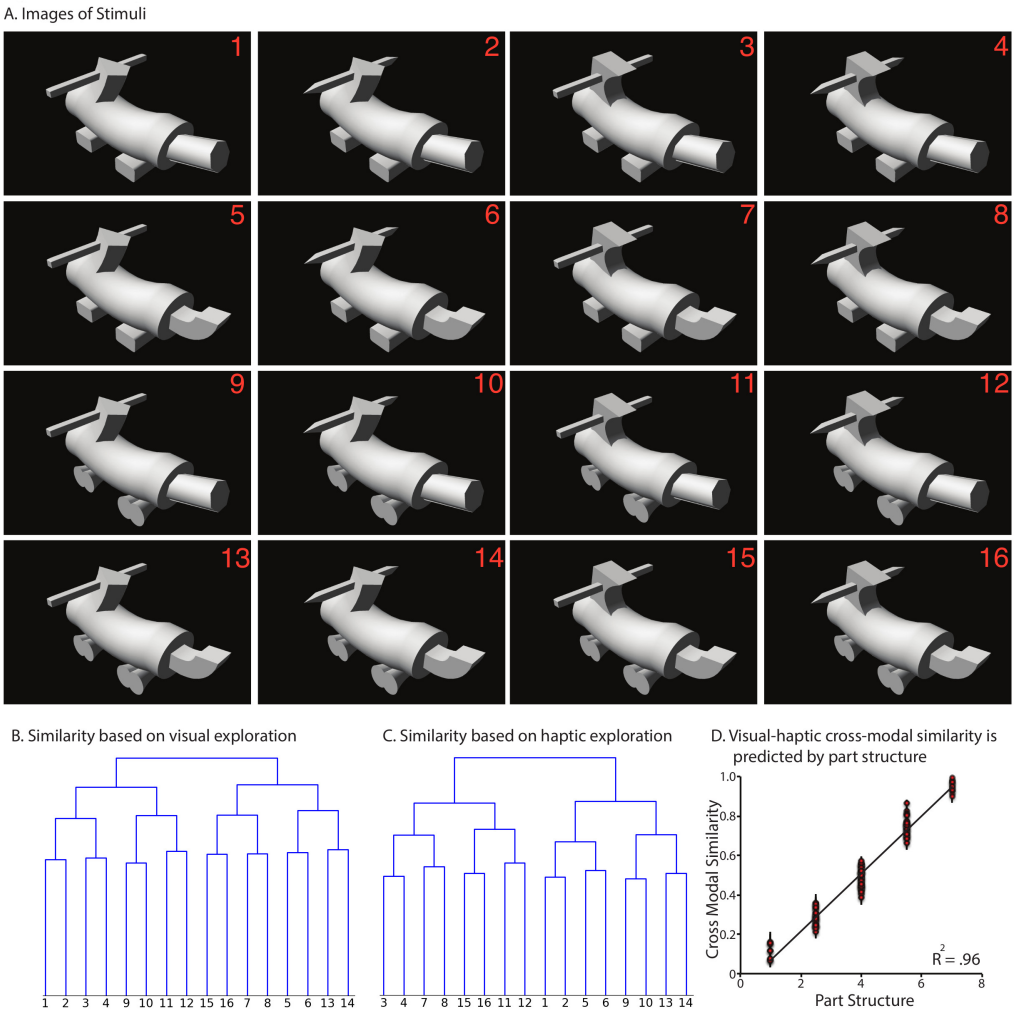


Figure 3.2: (A) Experimental stimuli used in Experiment 2. The stimuli are based on Fribbles (Tarr, 2003). (B) Results of agglomerative clustering applied to behavioral similarity data from the visual condition. (C) Results of agglomerative clustering applied to haptic behavioral similarity data. (D) Scatter plot of cross-modal behavioral similarity judgments versus similarities calculated from part structure. Similarities based on part structure are calculated by counting the number of shared parts between pairs of objects.

objects. The resulting representational similarity analysis quantifies the similarity of voxel patterns across modalities, comparing every object to every other object as well as to itself. Previous studies have calculated neural similarity matrices separately for each modality and then subsequently correlated those matrices (e.g., Peelen et al., 2014). Such an approach amounts to showing that neural correlations among objects in one modality correlate with the neural correlations among objects in another modality. The goal of the current analysis is to run a stricter test of the hypothesis that LOC encodes objects in a multisensory manner by correlating voxel patterns from different modalities directly to form a cross-modal neural similarity matrix.

Two predictions are made by the hypothesis that object representations in LOC are multisensory. First, cross-modal correlations between the visual and haptic voxel patterns for the same object will be higher than cross-modal correlations among the voxel patterns for different objects (i.e., the diagonal values will be greater than the nondiagonal values in the cross-modal representational similarity matrix). The results of this analysis for each participant in Experiments 1 and 2 can be seen in Figure 3.3. For every participant in right LOC and for 10 of 12 participants in left LOC, cross-modal correlations were in fact higher for identical objects than they were for different objects (see Figure 3.4 for average cross-modal correlation matrices). Because an initial ANOVA analysis found no effect of Experiment (L-LOC, $F = 0.17, p = .69$; R-LOC, $F = 0.48, p = .50$), we combined the results from both experiments. Diagonal versus non-diagonal differences reached

statistical significance in both L-LOC and R-LOC (L-LOC, difference = 0.06; $t = 3.86, p < .004$; R-LOC, difference = 0.05; $t = 5.08, p < .001$), indicating that LOC contains multisensory representations of objects. A second and stricter prediction is that it should be possible to decode object identity using the representational similarity matrix by testing whether each object is more correlated with itself (across modalities) than it is with each of the other objects in the set (also across modalities). We calculated the decoding accuracies for each participant and compared these to the chance decoding accuracy (1/12 for Experiment 1 and 1/16 for Experiment 2). Again, because an initial ANOVA analysis found no effect of Experiment (L-LOC, $F = 0.67, p = .43$; R-LOC, $F = 0.82, p = .39$), we combined the results from both experiments. Our results showed that it is possible to decode object identity cross-modally in both L-LOC and R-LOC (L-LOC, difference from chance accuracy = 0.09, $t = 2.48, p < .04$; R-LOC, difference = 0.10, $t = 3.48, p < .006$). These data indicate that LOC contains multisensory representations of objects.

We then tested whether multisensory coding of novel objects was specific to LOC or was a property observed throughout the brain. To that end, a whole brain searchlight analysis was conducted in which each voxel was coded according to whether it (and its immediate neighbors) showed higher pattern similarity for an object correlated with itself (across modalities) than with other objects (also cross modality). Converging with the ROI analyses, the results (Figure 3.5) identified the right LOC in both experiments (see

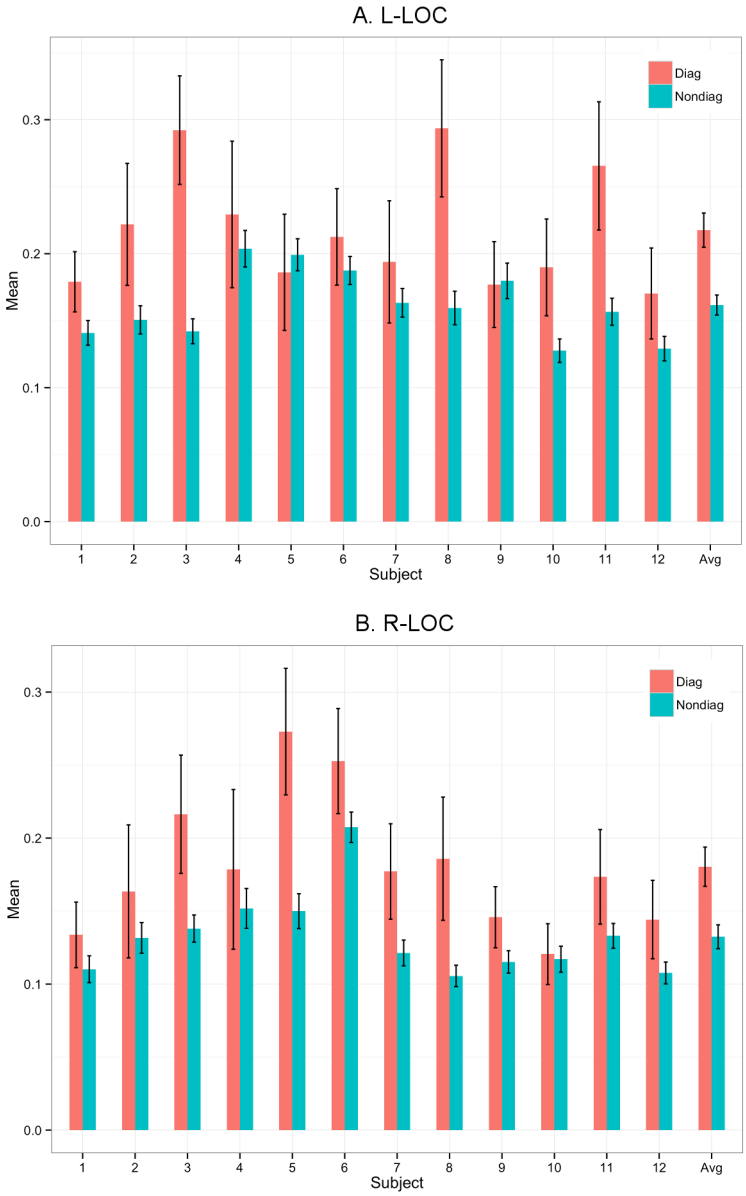


Figure 3.3: Comparison between diagonals and nondiagonals of cross-modal similarity matrices for both experiments. Participants 1-6 are in Experiment 1, and participants 7-12 are in Experiment 2. Avg = average of all 12 participants. (A) Results for left LOC. (B) Results for right LOC.

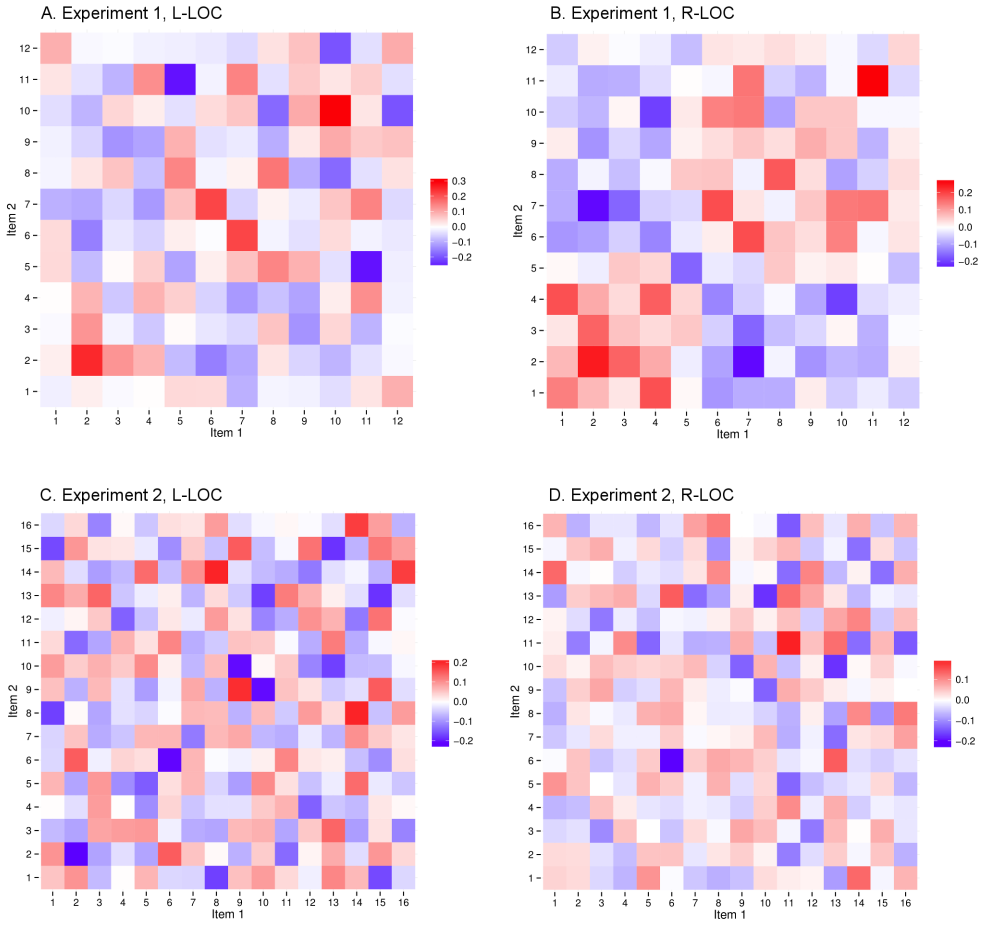


Figure 3.4: Cross-modal similarity matrices for both experiments. (A, B) Cross-modal similarity matrices calculated from left (A) and right (B) LOC activations from Experiment 1. (C, D) Cross-modal similarity matrices calculated from left (C) and right (D) LOC activations from Experiment 2.

Table 3.1 for coordinates). The left posterior temporal-occipital cortex was also identified in the searchlight analyses from both experiments.

A Common Similarity Space of Novel Objects as Derived from Neural and Behavioral Metrics

The stimuli used in Experiment 2 were designed to have a clear part-based structure for the purpose of testing the part-based hypothesis through representational similarity and neural decoding analyses. In a prior study (Erdogan, Yildirim, & Jacobs, 2015), we collected behavioral similarity judgments for these stimuli while participants viewed or haptically explored the objects. Similarity ratings consisted of Likert similarity ratings (range 1:7) for each pair of objects. We evaluated how well participants' judgments of the similarity among the objects were explained by the part-based structure of the objects. As shown in Figure3.2D, the agreement was extremely good ($R^2 = .96$). This indicates that participants perceive the similarity among these object stimuli in terms of their part structure. Therefore, a significant agreement between the neural similarity matrices and behavioral similarity judgments will lend support to both the hypothesis that LOC representations are multisensory and to the hypothesis that they are part based. We tested this prediction by calculating correlations between behavioral similarity judgments and measures of object similarity derived from neural data. A visual similarity matrix was formed by correlating voxel patterns when participants viewed the objects during fMRI, and a haptic similarity ma-

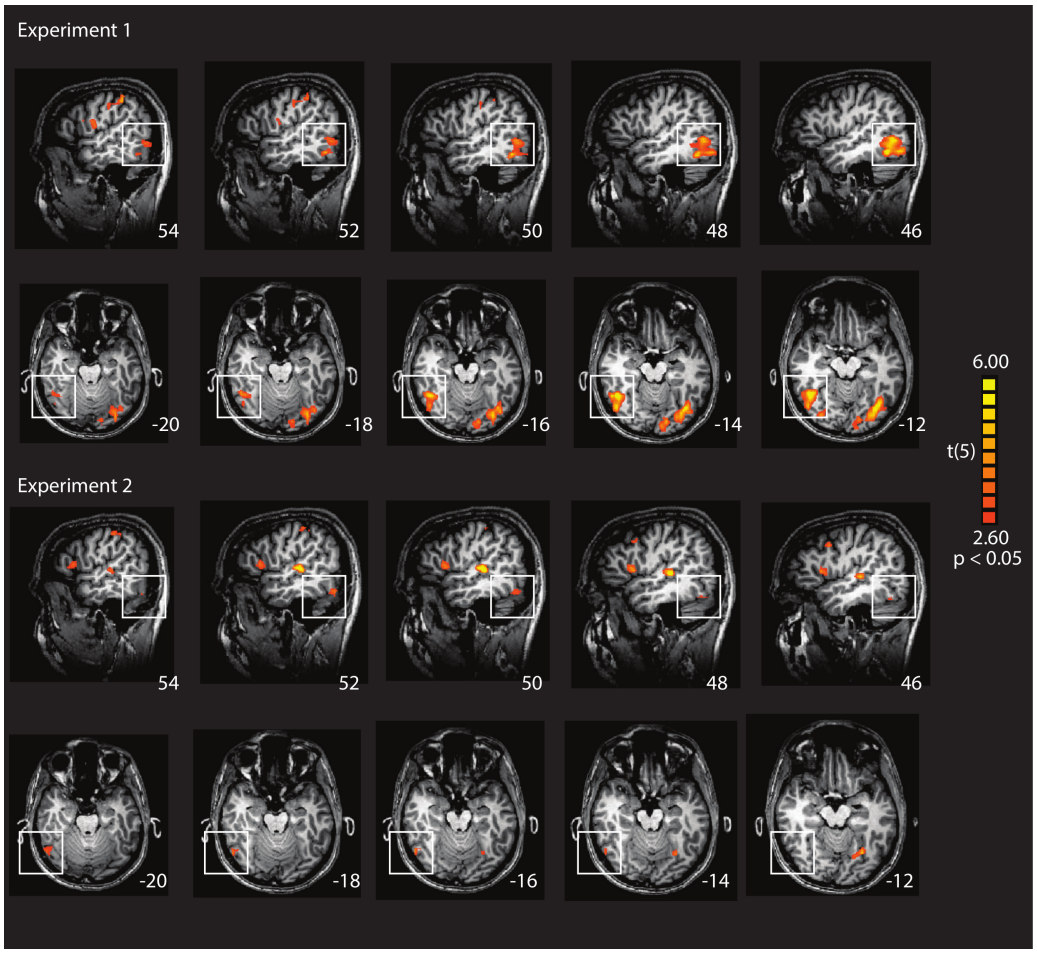


Figure 3.5: Whole searchlight analysis of brain regions in which the diagonal of the cross-modal neural similarity matrix is greater than the off-diagonal values. The cross-modal similarity matrix was created by correlating the voxel patterns elicited when visually exploring objects with the voxel patterns elicited when haptically exploring objects. If the diagonal of the matrix is greater than the off-diagonal values, that means that the pattern of voxel activations elicited by an object (across modalities) is more similar than the patterns elicited by two different objects.

Region	Talairach Coordinates x	y	z	Cluster Size (mm^2)	t	p
Exp1: Diagonal of the Cross-modal Neural Similarity Matrix > Off-diagonal Values ($p <$ .05, Cluster > 9 Voxels)						
Precentral gyrus LH	-51	-13	34	6790	8.55	$< .001$
Middle occipital gyrus LH	-24	-88	19	25007	12.28	$< .001$
Lateral occipital cortex LH	-39	-67	-14		12.28	$< .001$
Precentral gyrus RH	57	-1	19	1469	7.06	$< .001$
Postcentral gyrus RH	63	-25	38	3175	7.51	$< .001$
Lateral occipital cortex RH	39	-55	-5	23568	13.83	$< .001$
Exp2: Diagonal of the Cross-modal Neural Similarity Matrix > Off-diagonal Values ($p <$ .05, Cluster > 9 Voxels)						
Inferior frontal gyrus LH	-39	20	10	1100	5.47	$< .01$
Precentral gyrus LH	-30	-16	52	1514	7.30	$< .001$
Superior parietal lobule LH	-21	-58	58	4417	12.27	$< .001$
Inferior frontal gyrus RH	39	17	16	1450	6.19	$< .002$
Superior temporal gyrus RH	51	-25	7	877	10.04	$< .001$
Lateral occipital cortex RH	50	-62	-18	257	3.88	$< .01$
Exp2: Correlation between Neural and Behavioral Similarity for Visual Exploration of Objects ($p < .05$, Cluster > 9 Voxels)						
Parietal lobe LH	-18	-58	46	539	8.67	$< .001$
Lateral occipital cortex RH	42	-70	1	742	9.84	$< .001$
Lingual gyrus RH	0	-73	-11	2607	6.02	$< .002$
Exp2: Correlation between Neural and Behavioral Similarity for Haptic Exploration of Objects ($p < .05$, Cluster > 9 Voxels)						
Lateral occipital cortex LH	-39	-67	-14	1230	9.99	$< .001$
Precentral gyrus RH	42	-13	34	1577	10.78	$< .001$
Postcentral gyrus RH	51	20	34	2323	19.19	$< .001$
Parietal lobe RH	9	-37	61	3143	12.86	$< .001$
Superior temporal gyrus RH	42	-49	19	2110	13.79	$< .001$
Lateral occipital cortex RH	33	-73	-8	2190	11.84	$< .001$

Table 3.1: Talairach Coordinates, Cluster Sizes, Significance Levels, and Anatomical Regions for the Searchlight Results (LH=left hemisphere, RH=right hemisphere)

trix was formed when participants haptically explored the objects during fMRI. As predicted by the hypothesis that LOC encodes multisensory, part-based representations of objects, the neural similarity matrices obtained from R-LOC for both modalities were correlated with the behavioral similarity matrices (neural similarity measures based on visual exploration: L-LOC: $r = .02, t = 1.10, p = .33$, R-LOC: $r = .08, t = 4.21, p < .009$; Haptic condition, L-LOC: $r = .08, t = 1.52, p = .187$, R-LOC: $r = .14, t = 3.28, p < .03$).

To evaluate the degree to which the observed relationship between behavioral and neural similarity measures was specific to LOC, we again carried out a whole-brain searchlight analysis that maps how similar the neural similarity matrices were to the behavioral similarity matrices. The most stringent test of whether LOC encodes multisensory representations of novel objects is to test whether LOC is identified by two independent searchlight analyses: The first analysis relates neural and behavioral similarity data for visual exploration of objects, and the second analysis relates neural and behavioral similarity data for haptic exploration of objects. Thus, the key test is whether these two independent searchlight analyses overlap in LOC. The results indicate overlap in right LOC (see Table 3.1 for Talairach coordinates). As can be seen in Figure 3.6, there is good overlap (35 voxels, 958 mm^3 , across the maps in Figure 3.6A, B, and C) between the independent functional definition of right LOC (objects $>$ scrambled images) and right LOC as identified by the two independent multivoxel pattern searchlight analyses. Interestingly, the whole-brain searchlight analysis over haptic data also

identified several other regions in the temporal and frontal lobes involved in sensory processing (see Table 3.1 for coordinates).

Object Category Representations in LOC

Stimuli in Experiment 1 formed three families or categories of objects (Figure 3.1). This raises the possibility of evaluating whether LOC object representations encode category structure. Using analyses of the LOC cross-modal similarity matrix, we found that neural activations were more similar when considering two objects belonging to the same category than when considering two objects belonging to different categories. Using decoding analyses, we found that we can decode the category to which an object belongs at above-chance levels. However, because we are uncertain about the proper interpretation of these results, we do not study LOC object category representations here. One possibility is that LOC encodes the category structure of objects. Another possibility is that LOC encodes object shape and that the results regarding category structure are due to the fact that objects belonging to the same category have similar shapes in our experiment and objects belonging to different categories have dissimilar shapes. Because we cannot distinguish these two possibilities based on the stimuli used here and because there is substantial evidence indicating that LOC represents object shape, a stronger test of the nature of LOC object representations is provided by fine-grained analysis of the part structure within the materials from Experiment 2.

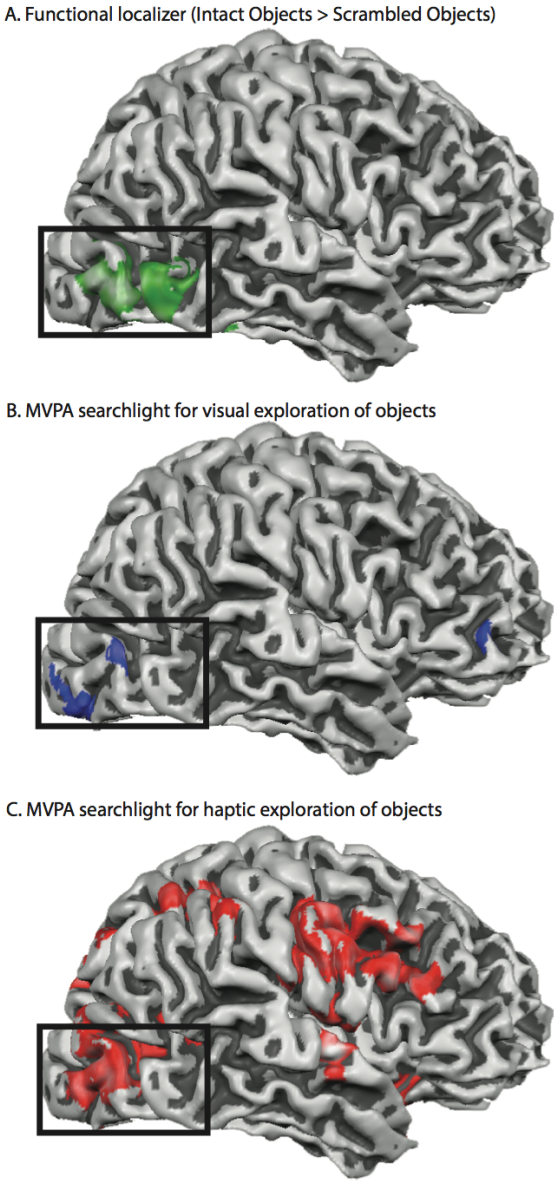


Figure 3.6: Overlap in right LOC for the (A) functional localizer (i.e., objects $>$ scrambled objects), (B) a whole brain searchlight analysis of the correlation between neural similarity matrices and behavioral similarity for visual exploration of objects, and (C) a whole brain searchlight analysis of the correlation between neural similarity matrices and behavioral similarity for haptic exploration of objects.

Part-based Object Representations in LOC

Finally, we sought to directly test the hypothesis that LOC encodes objects in a part-based manner. If the shape representations in LOC are encoding object parts, we should be able to decode the parts that make up an object from neural activations. We focused these analyses only on our second experiment because the stimuli in our first experiment are not suited to testing the part-based hypothesis. Although all objects used in Experiment 1 have a clear part-based structure, each part is at most shared by two objects, which drastically limits the amount of data available for decoding part identities. However, the stimuli in our second experiment were designed specifically to test the part-based hypothesis, with each part being shared by 8 of 16 objects in the stimulus set. The objects in our second experiment can be represented as four binary digits with each digit coding which one of the two possible parts for each of the part locations is present (see Figure 3.7 for a schematic of this analysis approach). In our decoding analyses, we thus sought to predict the four-digit binary representation of each object using neural activity patterns. We trained four separate linear SVMs, one for each location. Each SVM model was trained to predict which of the two possible part values for that location was present in an object. Each of the four classifiers was trained on 15 of the 16 objects, and the classifiers were tested by having them jointly predict the four-digit binary representation for the 16th object. If all four of the predictions (one for each location) were correct, we counted that as a successful decoding of the object (see Fig-

ure 3.7B). Thus, chance for this classification test was $0.5^4 = 0.0625$. This analysis approach was performed using 16-fold leave-one-out cross-validation, each time leaving one object out (for test) and training the classifiers on the remaining 15 objects. We then averaged the classification accuracies over folds to obtain an estimate of the classification accuracy across all objects for each participant. Statistical analysis was then performed over subject means. The results of this analysis indicated that it was possible to decode novel objects in LOC, both for fMRI data obtained during visual and during haptic exploration of the objects (visual condition, L-LOC: classification accuracy = 0.198, $t = 3.61, p < .016$, R-LOC: classification accuracy = 0.250, $t = 5.81, p < .003$; haptic condition, L-LOC: classification accuracy = 0.167, $t = 2.50, p = .055$, R-LOC: classification accuracy = 0.302, $t = 5.86, p < .003$).

Discussion

We have shown that it is possible to decode object identity from a cross-modal similarity matrix created by correlating LOC voxel patterns during visual and haptic exploration of the same set of objects. This suggests that there is a unique neural code generated during perceptual exploration of each of the novel objects that is similar regardless of whether the sensory modality is vision or touch. We also found that linear classifiers successfully predict a novel object based on its part structure. Thus, the fundamental units

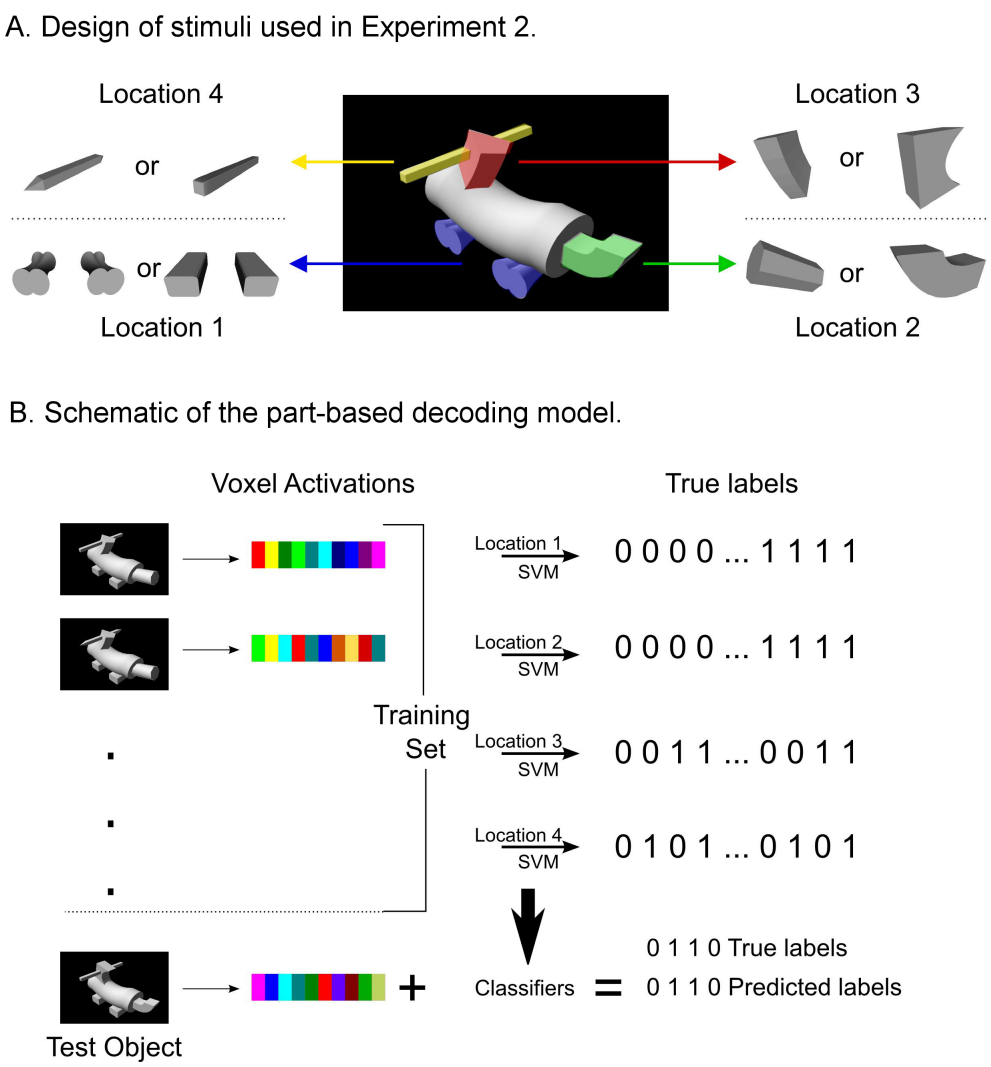


Figure 3.7: (A) Design of stimuli. Each object is composed of four components at four fixed locations. (Parts are colored for illustration purposes. All images were grayscale in the experiment.) (B) Schematic of the decoding model. Neural activations for 15 of the objects are used as the training set to train four linear SVMs to predict parts at each location. Then, the trained classifiers are used to predict the parts of the left-out test object, and these predictions are compared with the true parts of the object.

of object representation in LOC are expressed in terms of an object's composite parts. These findings provide further evidence for part-based visual representations of objects in LOC and multisensory representations of whole objects, at least across the haptic and visual modalities (Peelen et al., 2014; Naumer et al., 2010; Amedi et al., 2001, 2002; James et al., 2002). Crucially, our cross-modal decoding analyses relied on a direct comparison between activations from different modalities, representing a more direct test of the multi-sensory nature of object representations in LOC than was present in prior studies. Additionally, we believe our part-based decoding of novel objects presents a significant step towards understanding the nature of object representations in LOC. The only previous study that used a similar decoding analysis (Guggenmos et al., 2015) employed simpler stimuli (two-part objects) and presented objects only visually. Our study used a richer set of stimuli and showed that decoding of a novel object is possible from both visual and haptic activation in LOC. We believe that the findings we have reported strongly suggest that object representations in LOC are multisensory and part based.

Our results show an interesting hemispheric asymmetry; in most of our analyses, the findings are stronger in R-LOC. We do not have a clear understanding of why this is the case. A recent study suggests that haptic processing is stronger in LOC for the nondominant hand (Yalachkov, Kaiser, Doehrmann, & Naumer, 2015). However, it is important to note that participants in our experiment used both of their hands to explore objects. Addi-

tionally, these hemispheric differences are seen in the visual condition as well, making an explanation based on haptic processing unlikely. Future research should investigate whether this hemispheric asymmetry is a consistent characteristic of object shape processing or merely an artifact of our particular sample.

Although we have referred to the object representations in LOC as multisensory, it is worth pointing out that our study focused on visual and haptic processing, simply because shape information is conveyed mainly through these two modalities. For example, as previous research (Naumer et al., 2010; Amedi et al., 2002) shows, LOC does not respond to auditory stimulation. Similarly, our study says little about the representation of objects that lack a clear part-based structure, for example, bell peppers, or that are processed holistically, for example, faces. The question of how an object without a clear part-based structure is represented lies at a finer level than that on which our study focused; we did not investigate how an individual part might be neurally represented but whether parts are explicitly represented in the first place. Future research should focus on this more difficult question of how individual parts are represented.

In this study, we focused mainly on LOC and the nature of object representations in this region. However, looking at Table 3.1, we see that our searchlight results identified other regions, for instance, the precentral gyrus and the left posterior temporal-occipital cortex. Although none of those regions show the consistent activity that LOC shows across various analyses,

it is possible that multi-sensory object representations reside in a larger network of brain regions and likely that multisensory object representations in LOC are embedded in a broader network of regions that support multisensory processing. This is an empirical question that needs to be addressed by future research.

A key claim of the part-based hypothesis is that objects are represented as a combination of shape primitives from a finite set. Although our data cannot speak to the inventory of shape-based primitives that the brain may encode, further research using the methods we have developed may be able to describe that inventory. A second key aspect of part-based theories of object representation is that spatial relations among parts are directly represented. The findings we have reported motivate a new approach to test whether the spatial arrangement among an object's parts are encoded in the same region (LOC) that encodes the part information. Alternatively, information about the spatial arrangement of parts may be stored elsewhere in the brain.

Our findings also bear on the principal alternative theoretical model to part-based object representations: image- or view-based models. View-based theories argue that the representation of an object is a concatenation of 2D images of the object from different views (for discussion, see Peissig & Tarr, 2007). View dependency in object recognition is advanced as the main evidence for the view-based hypothesis. However, view-based models have difficulty accounting for our finding that there is a high degree of similarity in the voxel patterns elicited by haptic and visual exploration of objects and

that the shared variance in voxel pattern maps onto the part structure of the stimuli.

In this study, we have presented evidence that LOC carries multisensory and part-based representations of objects. In addition to the empirical evidence presented here and in earlier studies, we believe this hypothesis is also appealing from a theoretical perspective as it elegantly captures how information can be transferred across modalities, how inputs from multiple modalities can be combined, and more generally, how we cope with a world that is in its essence multisensory.

Chapter 4

Visual Shape Perception as Bayesian Inference of 3D Object-Centered Shape Representations

Introduction

Consider the objects in Figure 4.1. Even though you have not previously encountered these objects, you can readily perceive that the object in Figure 4.1c is more similar to the object in Figure 4.1a than the object in Figure 4.1b. However, the ease with which people make this judgment belies the complexity of the mental operations involved in this task. People's visual systems need to extract a representation of these objects from 2D images, and compare these representations to make a similarity judgment. This task illustrates the essence of the computational problem of object shape perception.

How people perceive object shape is one of the most fundamental questions about human visual perception. However, as evidenced by decades of

research, this simple question is surprisingly difficult to answer. Researchers have proposed numerous hypotheses about shape perception, and much research has focused on proving or disproving particular hypotheses. These efforts have led the field toward theoretical dichotomies such as whether people's shape representations are "view-based" or "structural", or whether these representations code two-dimensional or three-dimensional information. To date, investigations into such dichotomies have rarely produced clear outcomes. For example, after a long line of research on whether people's shape representations are view-based or structural, Peissig and Tarr (2007) summarized the state of the debate as follows: "In the end, it is unclear whether the large body of work focused on view-based models is compatible with, incompatible with, or just orthogonal to structural models of object representation". Which approach, if either, properly characterizes human shape perception is still a matter of fierce debate.

Here, we argue that existing models of shape perception are inadequate in important respects, and we propose a new model based on the hypothesis that shape perception of unfamiliar objects can be best understood as Bayesian inference of 3D shape in an object-centered coordinate system. This hypothesis includes four important components: (i) Our hypothesis is a hypothesis about shape representations of **unfamiliar** objects. Shape representations of familiar objects might be best understood in other ways. Coverage of this topic is deferred until the "Discussion" section. (ii) Shape perception for unfamiliar objects is a form of statistical inference which can

be characterized as Bayesian inference. This implies that people's shape representations are probabilistic, and thus contain information about certainty or confidence. For example, the shape properties of one portion of an object (e.g., the portion of an object facing a viewer) might be represented with high certainty, whereas the shape properties of another portion of the same object (e.g., a portion seen in peripheral vision, or a portion that is partially or fully occluded) might be represented with low certainty. It also implies that shape representations are influenced by a person's prior beliefs about shape properties. (iii) Shape representations code information about an object's three-dimensional structure, not the two-dimensional structure of its retinal image. (iv) Shape representations code shape properties in an object-centered coordinate system, not a viewer-centered coordinate system.

Although each of these components has been studied previously in the scientific literature, their combination has not. Indeed, as demonstrated below, their combination gives rise to interesting and unexpected results. For example, we have found that probabilistic object-centered representations can underlie viewpoint-dependency, suggesting that the distinction between view-based and view-independent representations is less useful than commonly believed when applied to the study of viewpoint invariance.

This article provides support for our hypothesis along two lines. First, we show that the use of 3D object-centered shape representations does **not** imply viewpoint-invariant object recognition. As demonstrated below, a person may, for example, attempt to infer a 3D object-centered shape representa-

tion from a 2D image in which one portion of a viewed object is clearly visible whereas another portion is not. If shape representations are treated in a probabilistic manner, the person's shape representation will have high certainty about shape properties in the former portion and low certainty about shape properties in the latter portion, thereby leading to viewpoint-dependent object recognition. We find that a computational model based on our hypothesis successfully accounts for the finding that people's object recognition performances can be viewpoint-dependent. Consequently, viewpoint-dependency should not be regarded as evidence for a view-based account of object recognition, as is typically done in the scientific literature.

Second, we report the results of an experiment using a shape similarity task, and evaluate a broad array of existing models of shape perception for their abilities to account for the experimental data. This evaluation provides compelling empirical support for our 3D object-centered shape inference model. Because the model captures subjects' judgments better than its competitors, our results support the hypothesis that people's object shape representations for unfamiliar objects are probabilistic, 3D, and object-centered. We conclude that our hypothesis is unique in its explanatory power and scope, and provides a promising approach for future investigations of object shape perception.

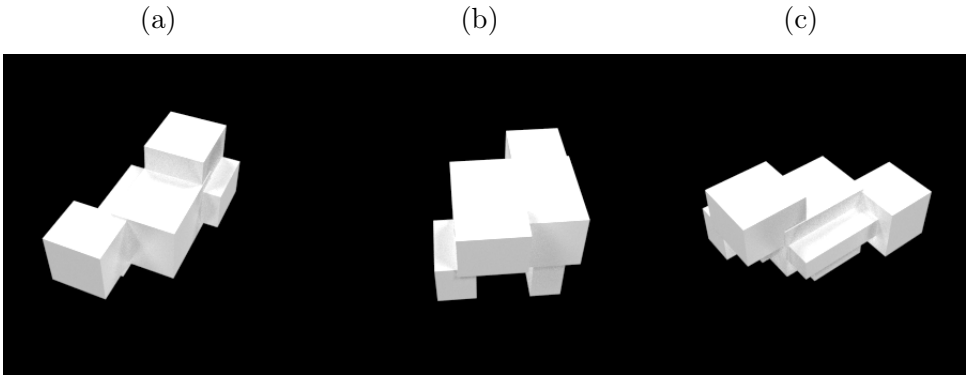


Figure 4.1: Is the shape of the middle or rightmost object more similar to the shape of the leftmost object?

Theoretical Background

It is frustratingly difficult to present a clear and well-organized analysis of hypotheses on shape perception. This is mostly because research on shape perception has revolved around dichotomies that are rarely rigorously defined, such as whether shape representations code 2D or 3D information, whether these representations are view-based or view-independent, or whether these representations are holistic or structural. These poorly defined dichotomies make the boundaries between different hypotheses hard to discern. In this section, we follow the analysis provided by Palmer (1999) and discuss three classes of shape perception hypotheses: feature-based, view-based, and structural description hypotheses. We present a critical review of each class, highlighting a class's strengths and weaknesses. For each class, we first present its main claims and then discuss computational models based on that class.

Feature-based hypotheses

Feature-based hypotheses claim that object shape is represented by a list of feature values extracted from 2D input images. These values are calculated by feature extractors through multiple layers of processing in the visual system. To compare the shapes of objects, one needs to specify a procedure for evaluating the similarity between two feature-based representations. In concrete models using a feature-based approach, feature values are usually real-valued and dissimilarity is quantified as Euclidean distance between representations. Feature-based hypotheses take their inspiration directly from what we know about biological visual systems, and this class of hypotheses represents the dominant perspective in the field of neuroscience. Building on the early work of Hubel and Wiesel (1962), neuroscientists have investigated visual perception by seeking to understand the neural feature detectors implemented by our visual systems. To date, this project faces major challenges in understanding cortical regions beyond primary visual cortex (Kourtzi & Connor, 2011).

To be meaningful, a feature-based hypothesis needs to specify the particular features that the hypothesis claims to be involved in shape perception. One popular proposal claims that what characterizes these features is that they are invariant to shape-preserving transformations such as translation and rotation (Palmer, 1999). Previous research has shown that some neurons in inferotemporal cortex (IT) are significantly position and scale invariant (Riesenhuber & Poggio, 2002). However, recent research suggests

that the extent of the invariance exhibited by these neurons is significantly less than previously believed (Lehky & Tanaka, 2016). Moreover, the naive invariance hypothesis cannot be the whole story because features that are fully invariant to shape-preserving transformations are inadequate for visual object recognition. For example, features that are fully position-invariant cannot distinguish between two objects that consist of the same features but in different spatial arrangements.

Feature-based models

In the field of computational neuroscience, an influential example of a feature-based model is Riesenhuber and Poggio (1999)'s HMAX (hierarchical MAX) model. HMAX extends Hubel and Wiesel (1962)'s ideas about simple and complex cells to higher level visual areas by proposing a sequence of template matching and pooling operations that build position and scale invariant features. HMAX consists of alternating layers of what are called S and C layers. Units in an S layer implement template matching. These templates can be simple Gabor filters (as in early layers) or more complex features (as in later layers) that are either specified by hand or learned. C layers play a key role in building invariant features since these pool over multiple units in the previous S layer and apply "max-pooling" (i.e., select the maximum input activation). By pooling over units tuned to different positions and scales, HMAX builds position and scale invariant features. Riesenhuber and Poggio (1999) showed that HMAX captures tuning and invariance properties of IT

neurons, and later work provided further evidence that HMAX is a good model of higher level processing in biological visual systems (Cadieu et al., 2007; Riesenhuber & Poggio, 2000, 2002; Serre, Oliva, & Poggio, 2007; Serre, Wolf, Bileschi, Riesenhuber, & Poggio, 2007).

Feature-based hypotheses are also popular in the study of computer vision. Recently, multi-layer artificial neural networks known as convolutional neural networks (CNNs) have achieved state-of-the-art object categorization performances (Krizhevsky, Sutskever, & Hinton, 2012; LeCun, Bengio, & Hinton, 2015). These models are similar to HMAX in the sense that they implement a sequence of feature extraction and pooling operations. However, these models are much deeper (containing tens to hundreds of layers), and features are learned from large amounts of labeled image data to maximize performance. Given their successes in computer vision and their similarity to hierarchical processing in biological visual systems, recent work in cognitive science and neuroscience has started to investigate the extent to which these models provide insights into biological vision (Kriegeskorte, 2015). Khaligh-Razavi and Kriegeskorte (2014) compared a large set of models from computer vision and computational neuroscience (including HMAX) on how well they account for human fMRI and monkey neural data from cortical area IT. Results showed that AlexNet (Krizhevsky et al., 2012), a popular CNN trained on 1.2 million images, captured the most variance in IT activities. In a related study, Cadieu et al. (2014) showed that CNNs rival the representational performance of IT, matching the object categorization performance

of neural responses from IT.

Feature-based hypotheses are appealing in multiple respects. From a neuroscience perspective, they build object representations hierarchically through multiple layers of processing, and thus resemble biological visual systems. They have been found to provide useful models of neural processing at all levels of the visual cortical hierarchy. From an engineering perspective, CNN implementations of feature-based hypotheses provide state-of-the-art performances, sometimes achieving object recognition and categorization performances comparable to those of people. Additionally, these implementations are appealing because they do not require preprocessing of the input image, and they can work directly on natural images.

The main weakness of feature-based hypotheses is that they are too unconstrained. Many feature-based models, such as CNNs, use adaptive features that are learned from data to maximize performance on a specified task. The shape perception procedure acquired by a feature-based model is determined by its training, including its training data and adaptation procedure (e.g., loss function and optimization procedure). Therefore, a feature-based model needs to specify not only its structural architecture (e.g., how many layers of units, how are units in one layer connected to units in the next layer, etc.), but also its training procedure in detail. Even when these details are specified, there is reason to doubt whether current feature-based models provide good scientific models of biological shape perception. These models usually have large numbers of parameters (e.g., 60 million in Krizhevsky et

al., 2012) that adapt with nonlinear dynamics, meaning that the models are complex. To date, it is nearly impossible to know how and why these models achieve what they achieve. Understanding why feature-based models work so well is the focus of much current research (Anselmi, Rosasco, Tan, & Poggio, 2015; Mehta & Schwab, 2014; Patel, Nguyen, & Baraniuk, 2015; Yuille & Mottaghi, 2016).

View-based hypotheses

View-based hypotheses claim that people's shape representation for an object consists of a collection of memorized "views" of the object from different viewpoints. Recognition is achieved by comparing the observed view of an object to these stored views. View-based hypotheses focus on this comparison procedure rather than on how each view of an object is mentally represented. Indeed, view-based hypotheses are agnostic with respect to how views are represented (referred to as the "view encoding scheme"; see Tarr & Bulthoff, 1995, and Edelman, 1997). Different instantiations of view-based hypotheses have proposed different view comparison procedures (see below).

View-based hypotheses are motivated primarily by experimental findings demonstrating that visual object recognition performance can depend on the viewpoint from which an object is observed (Edelman, Bulthoff, & Weinshall, 1989; Edelman & Bulthoff, 1992; Rock & DiVita, 1987; Tarr & Bulthoff, 1995; Tarr et al., 1998). These studies have shown that it becomes harder to recognize an object as it is rotated away from its training view. View-dependent

recognition has been presented as evidence for view-based hypotheses, and proponents of view-based hypotheses have argued that their findings provide strong evidence against approaches that use 3D, object-centered shape representations. However, as we discuss below in more detail, this view has been challenged by various researchers, and we demonstrate below with our simulation study that 3D, object-centered shape representations can in fact give rise to viewpoint dependency.

View-based models

Although view-based hypotheses do not make representational commitments, most view-based models have assumed that views are stored as lists of 2D features. These models have focused on how a test image can be compared with the stored 2D views in order to recognize objects. The "alignment-based" approach (S. Ullman, 1989) claims that the similarity between two view-based representations is calculated by first aligning the views and then comparing them. The alignment step aims to achieve robustness to shape-preserving transformations (e.g., scaling, translation, rotation), thereby enabling recognition despite such variation. S. Ullman (1989) has presented simple examples of how the alignment-based approach can be used to recognize objects but this model has not been evaluated for its ability to account for people's recognition performances.

Another approach is recognition by linear combination of views (S. Ullman & Basri, 1991). S. Ullman and Basri (1991) showed that under ortho-

graphic projection, views of an object span a linear subspace. Therefore, one can evaluate whether a test view depicts an object simply by checking if the test view can be represented as a linear combination of stored views of the object. Since this process requires multiple views of an object, this model cannot explain object recognition when relatively few views of an object reside in memory. For example, this model cannot recognize objects that are seen from a single view.

Another influential view-based model is that of Poggio and Edelman (1990). The model is an artificial neural network that is trained to map the input image of an object to an image depicting what the object would look like from a canonical viewpoint. The network is a "radial basis function" network in which the basis functions are centered around the stored views. The model has been used to replicate the experimental findings in Bulthoff and Edelman (1992) demonstrating that people's object recognition performances can be viewpoint-dependent. Despite its strengths, the model can be regarded as unsatisfactory in multiple respects. First, one needs hundreds of views of an object to train the network (Longuet-Higgins, 1990). Even if this might be possible for objects we encounter daily, it does not explain how people recognize objects that are seen only a few times or perhaps only once. Second, the model requires a separate network to be trained for each object. Even if this is plausible, training separate networks for each object ignores generalization across objects.

All view-based models suffer from a common problem—they all assume

that the same set of features can be extracted from all views. This requires determining the same set of features in all views, and also the correspondences between features across different views. S. Ullman (1989) argued that our visual systems can achieve this feature extraction easily. However, Poggio and Edelman (1990) admitted that this is a non-trivial task. It might be easy to extract and match features in the case of simple images, but it is unclear whether feature extraction and matching can be so easily achieved in natural settings.

Structural description hypotheses

Structural description hypotheses claim that object shape can be analyzed using a finite set of simple shape primitives. The structural description of an object consists of a list of the primitives making up that object and the spatial relations among them. A structural description model needs to specify three components: the structural description format (i.e., the set of primitives and possible spatial relations between primitives); the shape extraction procedure (i.e., how structural descriptions are extracted from 2D images); and the shape comparison procedure (i.e., how similarity between structural descriptions is measured). In principle, the structural description of an object can characterize either 2D or 3D information in either viewer-centered or object-centered coordinate systems. However, structural description hypotheses have almost always used 3D, object-centered shape representations. Structural description hypotheses, along with the opposing

view-based hypotheses, were the subject of fierce debate during the 1980s and 1990s (Biederman & Gerhardstein, 1993, 1995; Tarr & Bulthoff, 1995). The main point of contention was the viewpoint dependence of object recognition. Structural description hypotheses were interpreted as implying that recognition would be viewpoint invariant since a full 3D, object-centered shape representation is used in the recognition process. However, as we have remarked above and will discuss in detail below, this conclusion is mistaken. 3D, object-centered representations can, in fact, account for viewpoint-dependency.

Structural description models

Structural description models have a long history starting with the early works of Binford (1971) and Marr and Nishihara (1978). Arguably the most famous and detailed proposal is Biederman's recognition-by-components (RBC) theory (Biederman, 1987, 2007). RBC claims that objects are represented as collections of 3D volumetric primitives called geons and the spatial relations among them. Crucially, structural descriptions in RBC represent shape only qualitatively. Geons do not encode metric properties such as the exact values of a part's width, height, depth, or aspect ratio. Similarly, relations between geons are encoded in coarse terms such as above, below, left-of, and right-of. Biederman (1987) presented a detailed account of the structural description format and a sketch of how these representations might be extracted from 2D images on the basis of "non-accidental" features. Similarity between two structural representations was assumed to depend on the degree of match be-

tween representations, but the similarity measure was not specified in detail.

RBC has been at the center of the debate between structural description and view-based hypotheses. It has been criticized because it fails to explain viewpoint-dependency. RBC predicts view-invariant recognition in Bulthoff and Edelman (1992)'s study because all stimuli used in the experiment have the same structural description. In response to this criticism, Biederman and Gerhardstein (1993) argued that RBC did not apply to the set of objects used in these experiments because RBC was intended as a model of "entry-level" categorization in which different objects have different structural descriptions and where all geons are visible in all images. Thus, in Bulthoff and Edelman (1992)'s experiment, subjects must be relying on a different shape perception mechanism.

The argument provided by Biederman and Gerhardstein (1993) is an instance of a two-process account of shape perception (Foster & Gilson, 2002; Marsolek, 1999; Palmeri & Gauthier, 2004). According to such an account, shape perception consists of two distinct processes. One is responsible for what is usually called "metric" recognition (mainly concerned with within-category discrimination, such as discrimination of objects that differ in metric properties such as length, size, and aspect ratio). The second process is responsible for discriminating between objects that are qualitatively different (e.g., across category discrimination). Biederman and Gerhardstein (1993) argued that RBC concerns this non-metric, qualitative recognition process. For this process, one should expect view-invariant recognition given that all

geons of an object are visible in an image. However, although acknowledging that such a two-process system is possible, Tarr and Bulthoff (1995) argued that Biederman's theory failed to explain what it purported to explain. There are examples of objects (e.g., cow and horse) that have the same geon structural description but nonetheless belong to different categories. Additionally, Biederman's two-process account, although plausible, is far from elegant. It is unclear why there should be two processes in the first place, apart from the fact that RBC fails to adequately account for the data from some experiments. Obviously, a far more satisfactory theory would capture both metric and non-metric recognition, and explain under which circumstances viewpoint-dependency is or is not obtained.

Overall, the strength of structural description hypotheses lies in the richness of their representations. Experimental data indicates that people seem to think of many natural objects as composed of parts (Tversky & Hemenway, 1984), some of which may be considered objects in their own right. For example, people think of bodies as consisting of parts such as limbs, torso, and head. Structural descriptions capture the compositionality of many objects in a natural manner. Compositionality is also crucial for efficiency since object representations can refer to other object representations, and object parts can be shared across objects. Additionally, structural descriptions make information about shape explicit. For example, a structural description model can discriminate objects and also explain why they are different. However, the power of structural description hypotheses can also be considered their

weakness. The shape extraction problem is very difficult when the goal is to extract rich shape representations from realistic 2D images, and this might explain why there have been so few implementations of structural description hypotheses (Hummel & Biederman, 1992). Perhaps more importantly, it is unclear whether such powerful representations are needed for shape perception. One might argue that structural description hypotheses make the shape perception problem more difficult than is necessary in many circumstances, and people could do well enough at object recognition with simpler representations.

This section has presented a critical analysis of existing hypotheses on shape perception. We believe that the above exposition shows that existing hypotheses are inadequate in important respects. This conclusion will be reinforced in Section 4.5 where we present an empirical evaluation of a broad array of models using data from an experiment on people's judgments of shape similarity. In the next section, we outline our own hypothesis claiming that shape perception for unfamiliar objects should be characterized as Bayesian inference of 3D object-centered shape representations.

Shape Perception as Bayesian Inference of 3D Object-Centered Shape Representations

Many researchers have argued that a fruitful approach to understanding biological visual perception is provided by the vision-as-inference hypothesis

(Von Helmholtz, 1867). This hypothesis characterizes the task facing our visual systems as the inference problem of extracting a description of (the task-relevant portions of) the external world from the visual stimulations on our retina. Using tools from the calculus of probability, modern research has implemented and transformed this idea into the "visual perception as Bayesian inference" hypothesis (Jacobs & Kruschke, 2011; Kersten & Yuille, 2003; Kersten et al., 2004; Knill & Richards, 1996; Yuille & Kersten, 2006). According to this hypothesis, perception is understood as the inversion of a generative model of how events in the visual environment give rise to retinal stimulations. Visual-perception-as-Bayesian-inference has been fruitfully applied to various aspects of visual perception, and past studies have shown that many perceptual phenomena can be understood from a probabilistic perspective as Bayesian inference under different probability models (Kersten & Yuille, 2003; Kersten et al., 2004; Knill & Richards, 1996). We believe that the visual-perception-as-Bayesian-inference hypothesis provides a promising approach to shape perception as well. We argue that shape perception can be best understood as the inference problem of extracting a description of object shape from 2D retinal stimulations.

The combination of this hypothesis with computational modeling provides natural cures for many of the problems we identified in our discussion of existing hypotheses in the previous section. We have seen that many models often leave important details unspecified. For example, RBC does not present an account of how two structural descriptions are compared, or view-based

models do not specify how views are encoded. Building computational models forces researchers to specify their theories clearly and rigorously, and the visual-perception-as-Bayesian-inference hypothesis makes it especially easy to do so. All that is required is to specify the generative model of how causes (e.g., objects) in the world give rise to visual stimulations (i.e., images) on the retina. Once a generative model is specified, the calculus of probability provides equations for inferring the values of task-relevant variables. For instance, one can categorize or identify objects, one can judge the similarity between two shapes, and one can study the conditions under which recognition should be viewpoint-dependent versus viewpoint-invariant.

Here, we argue that shape representations for unfamiliar objects can be characterized as coding 3D shape properties in an object-centered coordinate system. An unusual feature of our approach is that these are probabilistic representations, inferred using a statistical—specifically Bayesian—inference mechanism. As a result, shape properties are random variables, meaning that their values have distributions. The variances of these distributions carry information about the certainty of knowledge regarding these properties. For instance, a shape property for the portion of an object that is clearly visible may be inferred to have a distribution with a small variance, indicating relative certainty of knowledge about this property. At the same time, a property for a portion that is less visible (e.g., it may be visible in peripheral vision, or it may be partially or fully occluded) may be inferred to have a distribution with a large variance, suggesting a lack of certainty of knowledge about this

property. As discussed below, this aspect of our theory allows us to account for viewpoint-dependent object recognition (despite our theory's use of an object-centered coordinate system). In addition, our Bayesian approach implies that an observer's prior beliefs about shape properties influence his or her inferences about these properties.

To our knowledge, there are few previous articles in the psychology literature with an approach to shape perception that is closely similar to our own. In fact, the only one that we are aware of is the work of Feldman, Singh, and colleagues (Feldman & Singh, 2006; Feldman et al., 2013). These authors also treat shape perception as a form of Bayesian inference. In their model, observers infer 2D skeletal shape representations from 2D silhouettes of objects. These representations are based on medial-axis representations first introduced by Blum and Nagel (1978). Feldman et al. (2013) showed that their model is able to capture coarse shape similarity, and can also account for how some objects are decomposed into parts. While we have great admiration for this work (indeed, it has inspired our own efforts), it also has important shortcomings. To date, this model has not been tested as a general theory of object shape perception. Although Feldman et al. (2013) argued that their model can (eventually) be extended to handle 3D shape, their model is currently limited to inferring 2D shape representations. Section 4.5 presents an evaluation of their shape skeleton model on a shape similarity task.

Viewpoint-Dependency with Probabilistic 3D Object-Centered Representations

In this section, we show that a 3D object-centered shape inference model can account for the viewpoint-dependency of visual object recognition. We first discuss why 3D object-centered shape representations do not necessarily imply viewpoint-invariant recognition. Then we replicate an influential experimental finding regarding viewpoint-dependency with our shape inference model, and show that viewpoint-dependency of visual object recognition does not rule out probabilistic 3D object-centered shape representations.

Experiments showing that people's object recognition can be viewpoint dependent are often presented as evidence against shape perception models that use 3D object-centered representations. The reasoning underlying this claim is as follows. Because the 3D object-centered model of an object can be mentally rotated, recognition performance will not depend on viewpoint as long as a test object's true 3D shape representation can be extracted from the test viewpoint (Bulthoff & Edelman, 1992). In other words, differences between the viewpoint of an object at the time of study and the viewpoint of an object at the time of test can always be compensated for via mental rotation.

To us, this claim is poorly conceived. The claim assumes that the same 3D shape representation is extracted regardless of viewpoint. This is not necessarily the case and, in fact, is not perceptually (or computationally)

plausible. Different views of an object are not equally informative about the object's shape. Some properties of an object's shape may be easy to infer (i.e., can be inferred with low variance or high confidence) from a particular viewpoint, but difficult to infer (i.e., are inferred with high variance) from other viewpoints. Importantly, shape properties for one portion of an object might be easy to infer from an image of the object at a particular viewpoint, whereas the properties for another portion of the object are difficult to infer from the same image. A good illustration of this point is the canonical view effect. Previous research shows that even if all views of an object are presented an equal number of times during training, recognition performance depends significantly on viewpoint (Edelman & Bulthoff, 1992; Bulthoff, Edelman, & Tarr, 1995). These findings suggest that not all views of an object are equally informative. Therefore, one should generally expect that an observer will infer different 3D shape representations from different views of the same object. If so, one should expect object recognition to be viewpoint dependent. Furthermore, as long as the 3D shape inference procedure extracts more similar representations for closer views, one should expect object recognition to fall off gradually with viewpoint. That is, object recognition should be best when study and test viewpoints are most similar, should be moderate when these viewpoints are moderately similar, and should be worst when these viewpoints are least similar.

To illustrate these points, consider the three views of a paperclip object in Figure 4.2. To us, it seems intuitive that an observer's 3D shape

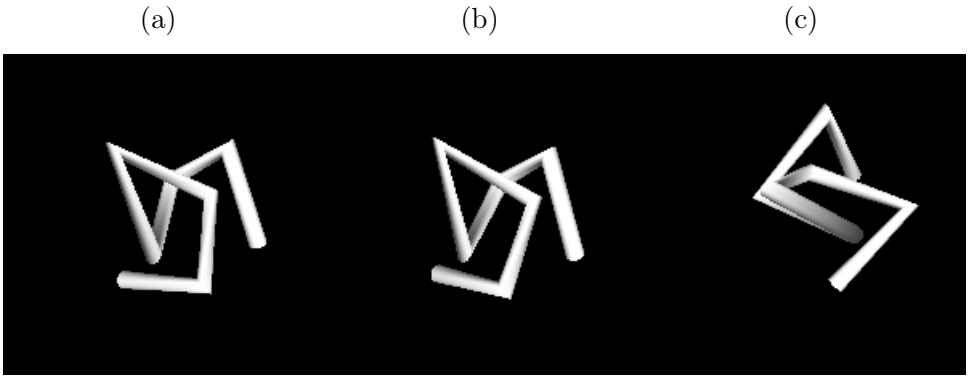


Figure 4.2: Three views of a paperclip object. Viewpoint differences between (a)-(b), (a)-(c), (b)-(c) are 10°, 70°, 80° respectively.

representations for the first and second views will be more similar than the representations for the first and third views, and hence recognition will be viewpoint-dependent. We show below that is, in fact, the case for a shape inference model that infers 3D object-centered shape representations. Therefore, the use of probabilistic 3D object-centered shape representations does not imply viewpoint-invariant object recognition.

To our knowledge, similar points have been made by a few researchers in the past. Z. Liu, Kersten, and Knill (1999) and Tjan and Legge (1998) argued that not only the internal representation of shape but also the information available in the stimuli mattered for viewpoint-dependency of recognition. They presented ideal observer analyses and experimental findings that suggest, depending on the complexity of a stimulus set, one would expect object recognition to be more or less viewpoint-dependent. Even though such findings can explain why recognition performance for stimuli like paperclips are

much worse than say objects made up of Biederman's geons, they do not speak to the issue of why recognition performance for a given object should get worse as the difference in viewpoint between the training and test views increases. Similarly, in a study investigating whether object representations are viewpoint-dependent, Z. Liu (1996) argued that a viewpoint-independent representation can also give rise to viewpoint-dependent performance. This point was repeated in a more recent article (Ghose & Liu, 2013). Unfortunately, neither of these articles provided an account of how this might happen. Bar (2001) also argued that viewpoint-dependency is not necessarily an indication of view-based representations. Bar (2001) presented an argument based on neural priming to show how object-centered representations can lead to viewpoint-dependent recognition. Although neural priming might be a plausible explanation for viewpoint-dependency, here we argue for an inference-based account where viewpoint-dependency follows from probabilistic inference of shape.

We show how our shape inference model accounts for viewpoint-dependency by replicating the main experimental findings from an influential study by Bulthoff and Edelman (1992). During training, subjects viewed two animations of a paperclip object. In one animation, the viewpoint of the object oscillated between -15° and 15° around the vertical axis. In the other animation, the viewpoint oscillated between -60° and -90°. During the test phase, subjects were presented with static test images in three conditions, and judged whether each test image depicted the same object as

observed during training. In the *interpolation* condition, test viewpoints spanned the range between the two training viewpoints in 15° increments (i.e., $0°, -15°, \ldots, -90°$ around the vertical axis) . In the *extrapolation* condition, test viewpoints spanned the range outside the training viewpoints in 15° increments (i.e., $0°, 15°, \ldots, 90°$ around the vertical axis). Finally, in the *orthogonal* condition, test viewpoints differed from training viewpoints because they were rotations around the horizontal axis ($0°, 15°, \ldots, 90°$ around the horizontal axis). Bulthoff and Edelman (1992) argued that a view-based model predicts slower and less accurate recognition as the object is rotated away from its training views, but a recognition scheme using 3D object-centered models would predict no effect of viewpoint as long as subjects were able to extract the true 3D model from training images. They used paperclip objects comprised of multiple tubular segments to make sure that the true 3D model can, in principle, be extracted from any viewpoint (similar to the objects shown in Figures 4.2 and 4.3).

Computational model

For our simulations, we generated ten paperclip objects similar to the stimuli used by Bulthoff and Edelman (1992). Each object consisted of seven segments, and each segment's length was sampled from a normal distribution around a mean segment length. We started by placing one segment at the origin. Two new segments pointing in randomly selected directions were joined to this center segment, one on each side. These directions were se-

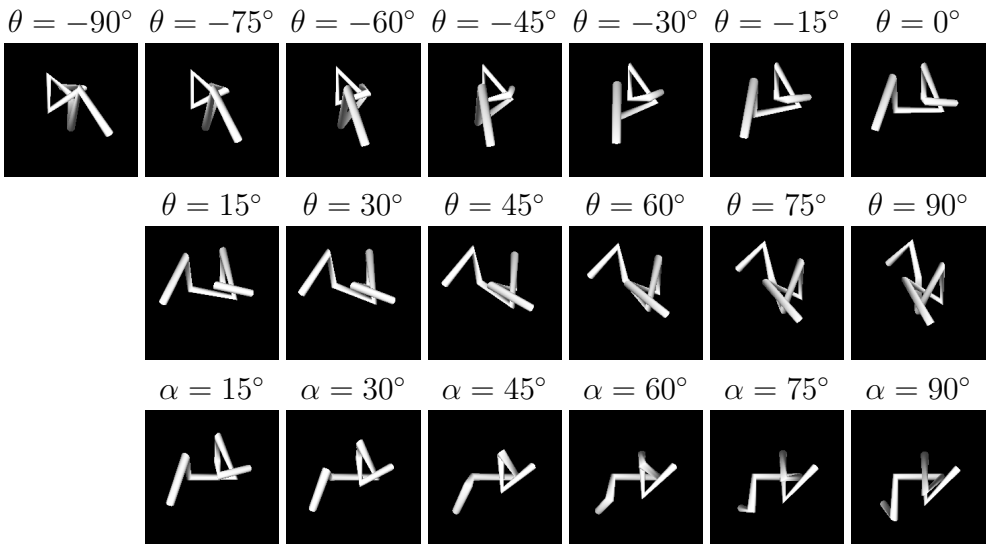


Figure 4.3: All views of an object used in our viewpoint-dependency simulations. θ refers to the angle around the vertical axis, and α refers to the angle around the horizontal axis.

lected such that the angles between segments were neither too small nor too large. We continued in this fashion by adding two segments to each end of the object until an object had seven segments. An object depicted from all simulated viewpoints is shown in Figure 4.3.[1]

Given the image of an object, our computational model infers the object's 3D structure in an object-centered coordinate system. In the model, an object is represented as a list of segment endpoint positions. For example, a 5-segment object shape S is represented as a list of six endpoint positions, $S = \{\vec{p}_1, \vec{p}_2, \ldots, \vec{p}_6\}$, with $|S|$ denoting the number of endpoints. (Although objects in our simulations always contained 7 segments, this information was

[1]The full set of stimuli can be seen online at `http://gokererdogan.github.io/ShapePerceptionAsBayesianInference/`.

not provided to the model. Instead, the model infers a posterior distribution over object shapes, meaning that shapes with, for example, 6, 7, or 8 segments might all be assigned non-zero probabilities.)

Prior distribution: In general, the model assumes that the number of segments comprising an object is sampled from a uniform distribution over integers in the interval $[2, 12]$, and that the coordinates of endpoint positions (i.e., the components of each vector $\vec{p}_i$) are sampled from a uniform distribution over $[-0.5, 0.5]$. However, without loss of generality, the model assigns the middle segment of an object to lie along the horizontal axis and to be centered at the origin. This enables the model to represent an object in a viewpoint-independent manner—that is, in an object-centered coordinate frame—and to easily "mentally" rotate the object to a canonical view if necessary. These assumptions define a prior probability distribution over possible object shapes:

$$P(S) \propto \frac{1}{|S| - 1}. \tag{4.1}$$

Likelihood function: To produce an image of shape S, we need to specify the viewpoint from which it is viewed. We denote viewpoint with $\vec{\phi} = (r, \theta, \alpha)$ using polar coordinates, and assume that the distance to the origin r is fixed. The prior probability distribution over viewpoint is assumed to be independent of shape S, and uniform over the sphere with radius r. The visual "forward model" $\mathcal{F} : (S, \vec{\phi}) \rightarrow I$ renders images by mapping a shape S and viewpoint $\vec{\phi}$ to image I. We implemented the forward model using

the Visualization Toolkit (VTK; `http://www.vtk.org`), a software package for 3D computer graphics, image processing, and visualization. Assuming an observed image is corrupted by Gaussian pixel noise with variance σ^2, the likelihood of shape S and viewpoint $\vec{\phi}$ is:

$$P(I|(S,\vec{\phi})) \propto \exp\left(-\frac{||\mathcal{F}(S,\vec{\phi}) - I||_F^2}{\sigma^2}\right) \tag{4.2}$$

where $||\cdot||_F$ denotes the Frobenius norm.

Posterior distribution: Combining the prior distribution and likelihood function via Bayes' rule, the posterior distribution of S and $\vec{\phi}$ is:

$$P((S,\vec{\phi})|I) \propto P(S)P(\vec{\phi})P(I|(S,\vec{\phi})) \tag{4.3}$$

where $P(S)$ and $P(I|(S,\vec{\phi}))$ are given by Equations 4.1 and 4.2, respectively, and $P(\vec{\phi})$ is uniform. Samples from this distribution were obtained using Markov chain Monte Carlo techniques (see Appendix 4.A.1 for details of the sampling procedure).[2] Figure 4.4 provides examples of samples for three objects.

Modeling results

We evaluated the model as if it was a subject in Bulthoff and Edelman (1992)'s experiment. During the training stage of the experiment, the model

[2]Implementation of our 3D shape inference model is available online at `https://github.com/gokererdogan/Infer3DShape/releases/tag/ro3Dpaper`.

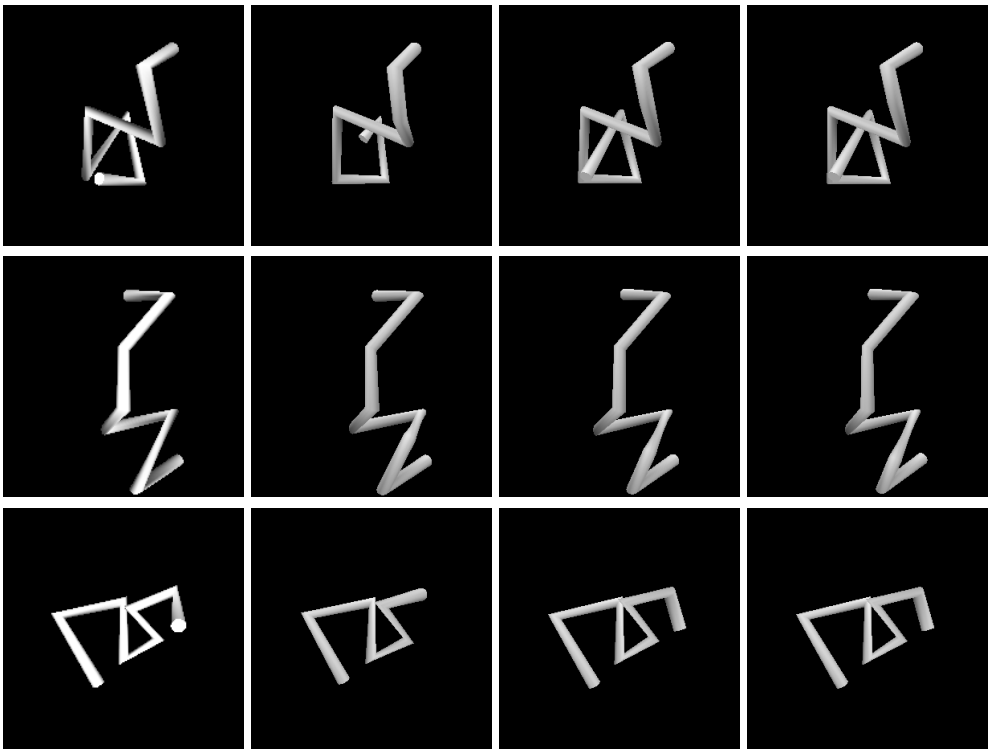


Figure 4.4: Examples of samples from the inferred posterior distribution $P(S|I_{\text{train}})$ for three objects. Each row depicts one object and three samples. The leftmost column shows the object from viewpoint $\theta = 0°$. Here, I_{train} consists of six views of an object from $\theta \in \{-90°, -75°, -60°, -15°, 0°, 15°\}$.

inferred the posterior distribution $P((S,\vec{\phi})|I_{\text{train}})$ over 3D shapes from the set of training images. The training images I_{train} consisted of six images at views $\theta \in \{-90°, -75°, -60°, -15°, 0°, 15°\}$. On a test trial, the model was presented with test image I_{test}, and it judged whether the image depicted the same object as observed during training.

We implemented this decision process as a comparison between two probabilities: (i) the probability that the test image depicted the same object as depicted in the training images, versus (ii) the probability that the test image depicted any other object. These probabilities were formalized as $P(I_{\text{test}}|I = I_{\text{train}})$ and $P(I_{\text{test}}|I \neq I_{\text{train}})$, respectively. We estimated $P(I_{\text{test}}|I = I_{\text{train}})$ as follows:

$$\begin{aligned} P(I_{\text{test}}|I_{\text{train}}) &= \int P(I_{\text{test}}|S)\ P(S|I_{\text{train}})\ dS \\ &\approx \frac{1}{N}\sum_{i=1}^{N} P(I_{\text{test}}|S_i) \end{aligned} \tag{4.4}$$

where S_i is a sample from the posterior $P(S|I = I_{\text{train}})$.[3] Because an object can be depicted from any viewpoint on a test trial, viewpoint needs to be taken into account when calculating $P(I_{\text{test}}|S)$. In our simulations, we found the viewpoint that best aligned the object with the observed image (i.e., we used $P(I_{\text{test}}|S) = \max_{\vec{\phi}} P(I_{\text{test}}|S,\vec{\phi})$). To find the best viewpoint, we carried

[3]We sampled from the posterior $P((S,\vec{\phi})|I = I_{\text{train}})$ but we ignored viewpoint $\vec{\phi}$ and treated S as a sample from $P(S|I = I_{\text{train}})$. This is equivalent to approximating $P(S|I)$ with $P((S,\vec{\phi}_{\text{MAP}})|I)$. Since $P((S,\vec{\phi})|I)$ is highly peaked around the MAP sample, this is a very good approximation. Our results do not change if we integrate out $\vec{\phi}$ to get $P(S|I) = \int p(S,\vec{\phi}|I)d\vec{\phi}$ instead of using the approximation.

out a search over the whole viewing sphere (θ and α were each discretized into 5° bins).

We calculated $P(I_{\text{test}}|I \neq I_{\text{train}})$ in a similar manner:

$$P(I_{\text{test}}|I \neq I_{\text{train}}) = \int P(I_{\text{test}}|S)\ P(S|I \neq I_{\text{train}})\ dS. \tag{4.5}$$

To approximate this integral, samples from the posterior $P(S|I \neq I_{\text{train}})$ are needed. Because it is unlikely that any shape except the true shape was depicted in the training images, $P(S|I \neq I_{\text{train}})$ is close to the prior $P(S)$. Using this approximation, $P(I_{\text{test}}|I \neq I_{\text{train}})$ can be estimated as follows:

$$\begin{aligned} P(I_{\text{test}}|I \neq I_{\text{train}}) &\approx \int P(I_{\text{test}}|S)\ P(S)\ dS \\ &\approx \frac{1}{M}\sum_{i=1}^{M} P(I_{\text{test}}|S_i) \end{aligned} \tag{4.6}$$

where S_i is a sample from prior $P(S)$.[4]

Bulthoff and Edelman (1992) reported error rates in their experiment. As shown in Figure 4.5a, subjects' performances were excellent in the *interpolation* condition, but these rates were significantly higher in the *extrapolation* and *orthogonal* conditions. Importantly, performances in the *interpolation* condition were relatively unaffected by viewpoint. However, error rates rose

[4]In our simulations, we used an additional approximation based on the fact that for a random shape S_i, $P(I_{\text{test}}|S_i)$ is nearly proportional to $\exp\left(-||I_{\text{test}}||_F^2/2\sigma^2\right)$ since there will be little overlap between the image of a random shape and a test image (i.e., $P(I_{\text{test}}|S_i)$ is nearly independent of S_i). Simulations confirm that this approximation is in general quite good—the results are virtually the same as when we approximate $P(I_{\text{test}}|I \neq I_{\text{train}})$ with samples from $P(S)$.

with the difference in viewpoint between training and test in the other conditions. Performance was worst in the *orthogonal* condition. At first, this might seem to be due to the fact that subjects observed two sets of views varying along the horizontal axis during training, hence receiving more information about side views of objects. However, Bulthoff and Edelman (1992) ran a variant of their experiment where the training views varied along the vertical axis, and subjects still performed worse for test views varying along this axis. This finding suggests that people find it harder to generalize to top/bottom views than to side views. To account for this finding, Bulthoff and Edelman (1992) restricted their model's generalization capability along the vertical axis to be significantly less than what it is along the horizontal axis.

To compare our model's performances with those of the subjects in Bulthoff and Edelman (1992)'s experiment, we need to calculate an error measure for our model. Because an observer is expected to make more errors as the observer becomes less confident about whether a test image depicts the training object, we used the posterior ratio $\frac{P(I_{\text{test}}|I\neq I_{\text{train}})}{P(I_{\text{test}}|I=I_{\text{train}})}$ as an error measure. For each test image in the three experimental conditions, this error measure was calculated. The results are summarized in Figure 4.5b. Overall, our model provides a good qualitative account of the experimental data. Its performance is best in the *interpolation* condition and markedly worse in the *extrapolation* and *orthogonal* conditions.[5] Moreover, its performance in the

[5]Given that we used a uniform prior over viewpoint, the lack of difference in per-

interpolation condition was relatively unaffected by viewpoint. However, its error measure rose with the difference in viewpoint between training and test in the other conditions.

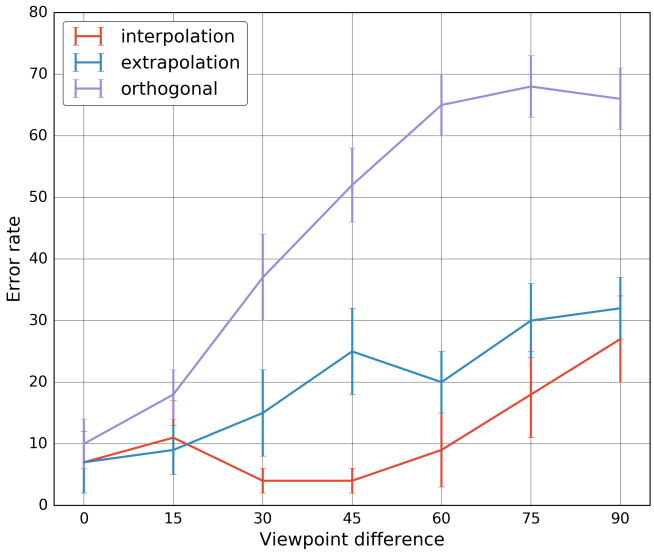
Overall, these results show that viewpoint-dependency does not imply that an observer is using 2D or viewpoint-dependent object representations. Our model, using probabilistic 3D object-centered representations, accounts for viewpoint-dependency of visual object recognition. Contrary to received wisdom in the field, viewpoint-dependency does not provide compelling evidence about whether object shape representations are 2D versus 3D, nor does it provide evidence about whether these representations are view-dependent or view-independent.

Behavioral Experiment and Model Comparisons

The previous section reported results indicating that it is erroneous to claim that viewpoint-dependent visual object recognition suggests the use of view-based shape representations. Indeed, either view-based or probabilistic 3D object-centered representations can underlie viewpoint-dependency, particularly when such a representation is inferred from an image. The goal of this section is to report results strengthening our hypothesis that people's

formances between the *extrapolation* and *orthogonal* conditions is unsurprising. We could have captured this difference by assuming a non-uniform prior over viewpoint (like Bulthoff and Edelman (1992) do in their view-approximation model). However, we chose not to do so because our primary aim here is not to capture this difference but to account for view-dependency (i.e., increases in error rate with increases in viewpoint difference between training and test views).

(a)



(b)

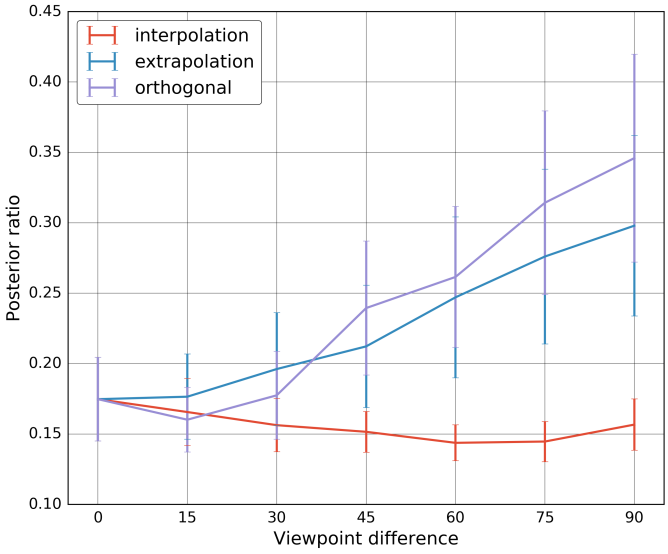


Figure 4.5: (a) Experimental results from Bulthoff and Edelman (1992). (b) Simulation results from our model.

shape representations of unfamiliar objects are probabilistic, 3D, and object-centered. We present a behavioral experiment, along with an extensive evaluation of a diverse array of computational models based on how well the models account for the experimental data. We show that our probabilistic, 3D, object-centered inference model captures subjects' performances better than all other models.

Behavioral experiment: Stimuli and procedure

Experimental stimuli were objects built from rectangular blocks. They were generated as follows. Each object started with a single fixed-size block centered at the origin. Then, one or more faces of this root block were randomly selected, and one or more new blocks with randomly sampled sizes were connected to the selected faces. This procedure was applied recursively—after child blocks were connected to a parent block, each child became a parent and had one or more child blocks connected to it. In practice, a parent block was restricted to have at most three child blocks. We also restricted the depth of each object to three (i.e., an object consisted of its root block, the root block's child blocks, and the root block's grandchild blocks). A sample object and its corresponding shape tree representation are shown in Figure 4.6.

We generated 10 target objects in this manner. Comparison objects with shapes similar, but not identical, to target objects were also created. They were generated by applying the following four manipulations to each target object. Each manipulation was applied at levels two and three in the shape

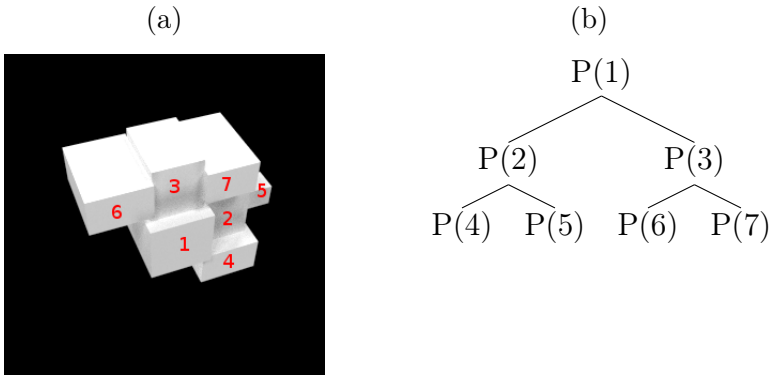


Figure 4.6: (a) Example of an object. The numbers on parts refer to the part numbers in its shape tree. (b) The shape tree representing the object in (a).

trees, resulting in 8 comparison objects generated from each target object. When using the *change part size* manipulation, one object part was randomly selected, and its size was set to a random value. This operation might change the positions of the selected part's descendants. When using the *change connecting face of part* manipulation, we again picked one part randomly, picked a new connecting face for it from the unoccupied faces of its parent part, and moved the part to this new location. Again, this manipulation moves all descendants of the selected part. The *add part* manipulation added one part randomly to the desired level in the tree. For example, to add a new part to level 3, we picked one of the parts at level 2 randomly, picked one of its unoccupied faces randomly, and connected a new part with a random size to the chosen face. When using the *remove part* manipulation, we picked one part randomly and removed it and all of its descendants. Figure 4.7

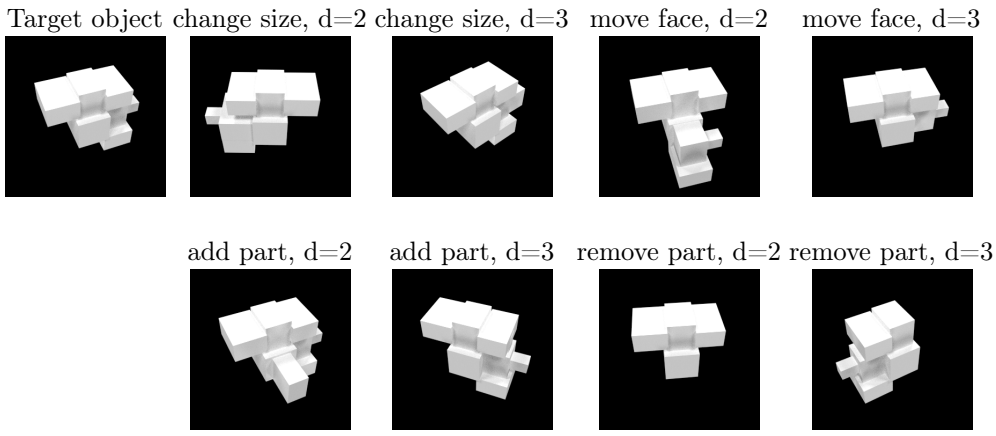


Figure 4.7: Target object (upper left) and its 8 comparison objects. The comparison objects were created using the four manipulations applied at levels two and three of the target object's shape tree. For example, "add part, d=2" refers to the object created by adding a new part to depth 2 in the shape tree.

illustrates a target object and examples of its 8 comparison objects.[6]

The experiment used a shape similarity judgment task. On each trial, a subject viewed images of three objects, one target and two comparisons. Subjects judged which comparison was most similar in shape to the target. Images were rendered from a random viewpoint on the 45° parallel ($\alpha = 45°$) along the viewing sphere using Blender (`http://www.blender.org`), a 3D graphics and animation software package. Each subject performed 100 trials, 16 of which were catch trials where one of the comparison objects was identical to the target. Forty-one subjects participated in the experiment, but data from five subjects were discarded because they failed to achieve 85%

[6]The full set of stimuli can be seen online at `http://gokererdogan.github.io/ShapePerceptionAsBayesianInference/`.

accuracy on catch trials. Subjects participated in this web-based experiment via Amazon Mechanical Turk.

Competing computational models

A diverse array of models of shape perception was simulated, and each model was evaluated based on how well it accounts for subjects' responses in our experiment. We include the following models in our evaluation.

Pixel-based model: The pixel-based model compares two objects by calculating the Euclidean distance between the pixel values in their images. This model can be regarded as an implementation of a view-based hypothesis that stores images as views. Subjects in our experiment saw each object only from a single viewpoint, and thus the shape representation for an object in the pixel-based model consists of a single image. Because there is only one image stored for each object, the pixel-based model is also an implementation of a particular version of Poggio and Edelman (1990)'s view-approximation model that works directly on raw images.

Alignment-based models: Another set of models that use 2D representations is motivated by the recognition-by-alignment approach (S. Ullman, 1989). Here, images of objects are aligned before they are compared. This alignment process requires a set of image features to be labeled in the images. The best alignment is calculated on the basis of these features. The dissimilarity between two images is taken to be the Euclidean distance in pixel space between the images after alignment. In our simulations, we used the corners

of the root block as features for alignment since these corners are present in every image. One can imagine allowing various types of transformations in the alignment process. Here we tried two transformations: one allowing only scaling and translation, and another allowing any affine transformation.

We also tried a third method that does not do any alignment. Instead, this no-alignment model simply calculates the Euclidean distances between feature lists (i.e., the coordinates of corners of root blocks). The model is an implementation of a view-based hypothesis that uses a simple feature-based representation for views. For this reason, the model is also referred to as a naive feature-based model. Since Poggio and Edelman (1990)'s original view approximation model worked on similar feature-based representations, the no-alignment model also provides a test of the view approximation model.

HMAX: An influential example of a feature-based model is HMAX (Riesenhuber & Poggio, 1999).[7] The model is a type of artificial neural network consisting of four layers of units: S1, C1, S2, C2. We used the outputs from the C1 and C2 layers (as is generally done in previous work evaluating HMAX models). The particular implementations we used applied feature extraction at the C1 layer at eight different spatial scales. We treated each scale as a separate model, and also combined all eight scales into a single C1 layer representation. Our HMAX implementation also used eight different patch (i.e., feature) sizes at the C2 layer. Again, we treated activations for

[7]We used the implementation provided by the authors at `http://maxlab.neuro.georgetown.edu/docs/hmax/hmaxMatlab.tar`

each of these patch sizes as a separate model, but also combined all of these models to form a single C2 layer representation. Therefore, in total, there are 18 versions of HMAX (8 scales for C1, all scales combined, 8 patch sizes for C2, and all patch sizes combined). We used the feature dictionary provided with the HMAX implementation. These features were extracted from random natural images, and are intended as a universal set of features. To get feature-based representations for each object in our experiment, we fed each image of an object to an HMAX model and calculated the responses of the C1 and C2 layers. These responses constitute objects' shape representations. We used Euclidean distance to compute dissimilarities between two such shape representations.

Convolutional neural networks: We evaluated two convolutional neural networks (CNNs) that are regarded as state-of-the-art computer vision systems: AlexNet (Krizhevsky et al., 2012) and GoogLeNet (Szegedy et al., 2014).[8] AlexNet is an eight-layer (five convolutional, three fully connected layers) CNN trained on 1.2 million images in the ImageNet dataset. AlexNet achieved the best performance on the 2012 ImageNet Large Scale Visual Recognition Challenge, and was in large part responsible for the recent surge of interest in deep neural networks. We treated each of its 14 layers (making the three max-pooling and two normalization layers explicit) as a separate model. Using the standard terminology in the deep neural network literature, these layers are: conv1, pool1, norm1, conv2, pool2, norm2, conv3,

[8]We use the pre-trained networks provided by the Caffe framework (Jia et al., 2014).

conv4, conv5, pool5, fc6, fc7, fc8, and prob. The set of unit activations in the last layer, prob, is a 1000-dimensional vector encoding the probability of belonging to each of 1000 object categories in ImageNet. The second CNN that we tested was GoogLeNet by Szegedy et al. (2014). This model set the state-of-the-art performance on the 2014 ImageNet Large Scale Visual Recognition Challenge. GoogLeNet has 22 layers (with an additional five pooling layers). Our simulations used 16 layers: pool1, conv2, inception3a-b, pool3, inception4a-e, pool5, inception5a-b, pool5, loss3 and prob. To make predictions from AlexNet and GoogLeNet, we input each image to the networks and performed a feedforward pass to calculate each layer's responses. The dissimilarity between two objects is computed as the Euclidean distance between vectors of these responses.

Structural distance-based model: We implemented a structural distance-based model that calculates object similarity using the structural descriptions of objects. Unfortunately, there are no concrete proposals in the literature for how this should be done. Because the objects in our experiment can be represented as shape trees (see Figure 4.6), one plausible way is to use the distance between these trees as a measure of dissimilarity. We used one such measure referred to as tree-edit distance (K. Zhang & Shasha, 1989). Using this measure, the distance between two shape trees is the total cost of operations needed to turn one tree into the other.[9] Tree-edit distance allows

[9]Tree-edit distance considers two nodes to be equal if their labels are the same. In the case of our shape trees, this means that two P nodes need to have the same connection face to be considered equal.

add-node, remove-node and change-node operations, and we assumed that each operation has equal cost.

Shape skeleton model: As discussed above, Feldman et al. (2013) proposed to represent the 2D shape of a 2D object as a shape skeleton. This skeleton is inferred from an image silhouette using Bayesian inference. To calculate similarities between shapes, we first extracted the boundaries of objects in images to create 2D silhouettes. Then we used Feldman et al. (2013)'s model[10] to find the maximum-a-posteriori (MAP) shape skeleton for each silhouette. The similarity between two shapes can be formalized as the probability of observing the image for one shape given the image for the other shape. For example, the similarity between the target I_t and a comparison I_c can be evaluated by calculating either $P(I_t|I_c)$ or $P(I_c|I_t)$. $P(I_t|I_c)$ (and similarly $P(I_c|I_t)$) can be approximated on the basis of an estimated MAP shape skeleton for each shape as follows:

$$P(I_t|I_c) \approx P(I_t|Sk_{\text{MAP}})P(Sk_{\text{MAP}}|I_c) \tag{4.7}$$

where Sk_{MAP} denotes the MAP skeleton for I_c. We tried three similarity measures based on these probabilities: $P(I_t|I_c)$, $P(I_c|I_t)$, and their average $\frac{1}{2}[P(I_t|I_c) + P(I_c|I_t)]$.

3D shape inference model: Lastly, we describe our proposed model that treats shape perception as Bayesian inference of 3D shape in an object-

[10]We used the implementation provided by the authors at `http://ruccs.rutgers.edu/images/ShapeToolbox1.0.zip`

centered coordinate system. To specify our model, we need to describe the representation for object shape as well as the generative process or forward model mapping these representations to images. We assume that shape representations consist of the positions and sizes of a collection of rectangular blocks. Each object S is represented by a tuple (T, M) where T is a string from a probabilistic shape grammar with production rules: $P \rightarrow P \mid PP \mid PPP \mid \epsilon$. In these rules, P is a non-terminal symbol and ϵ is a terminal null symbol. In a string T generated by this grammar, each P symbol corresponds to an object part (i.e., a rectangular block). Hence, the string T characterizes the parent-child relations between parts in an object. The grammar follows closely our stimulus generation procedure, with each part being constrained to have at most three children. The sizes and positions of each part are specified in spatial model M. The spatial model associates a size $s \in \mathbf{R}^3$ and a connecting face of a block $f_i \in \{1, 2, 3, 4, 5, 6\}$ with each P node in T (see Figure 4.8 for an example object and its associated (T, M) shape representation).

The prior probability of shape S is:

$$P(S) = P(T)P(M|T). \tag{4.8}$$

The probability of producing T from the shape grammar, $P(T)$, is calculated as follows:

$$P(T) = \prod_{n \in \mathcal{P}} P(n \rightarrow ch(n)) \tag{4.9}$$

where $\mathcal{P}$ is the set of P nodes in tree T, $ch(n)$ are the children of node n, and $p(n \rightarrow ch(n))$ is the probability of the production rule $n \rightarrow ch(n)$. We assume production probabilities to be uniform (i.e., each of the four production rules has a probability of 0.25) which simplifies $P(T)$ to:

$$P(T) = \frac{1}{4^{|\mathcal{P}|}}. \tag{4.10}$$

The probability for spatial model M, $P(M|T)$, consists of the probabilities of picking part sizes and connecting faces. Because we assumed part sizes to be uniform over the interval [0, 1], we only need to focus on the probabilities for connecting faces. For a part with k available faces and c children, there are $\binom{k}{c}$ possible combinations of face assignments to its children. Since we have six empty faces for the root P node and five empty faces for the remaining P nodes (because one face is occupied by the parent), the probability of spatial model M is

$$P(M|T) = \frac{1}{\binom{6}{|\mathcal{O}_{\text{root}}|} \prod_{n \in \{\mathcal{P} \setminus \text{root}\}} \binom{5}{(|\mathcal{O}_n|-1)}} \tag{4.11}$$

where $\mathcal{O}_i$ refers to the set of occupied faces of node i.

Given a shape S and a viewpoint $\vec{\phi}$, forward model $\mathcal{F} : (S, \vec{\phi}) \rightarrow I$ maps 3D shape representations to 2D images. As above, we used the Visualization Toolkit software package to implement the forward model. Assuming Gaussian noise on images, the likelihood function $\mathcal{L}(H, \theta; I)$ is:

$$\mathcal{L}(S, \vec{\phi}; I) = P(I|S, \vec{\phi}) \propto \exp\left(\frac{1}{\sigma^2}||I - \mathcal{F}(S, \vec{\phi})||_F^2\right) \tag{4.12}$$

where σ^2 denotes the variance of the noise on I and $||\cdot||_F$ is the Frobenius norm.

The posterior distribution over shapes given an image can be calculated via Bayes' rule:

$$P(S, \vec{\phi}|I) \propto P(I|S, \vec{\phi})P(S)P(\vec{\phi}). \tag{4.13}$$

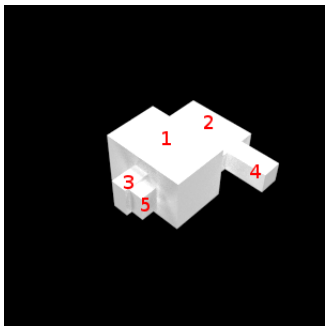
We assumed that $P(\vec{\phi})$ is a uniform distribution, and that viewpoint $\vec{\phi}$ is independent of shape S. We sampled from this posterior using MCMC techniques (see Appendix 4.A.2 for details). Figure 4.9 shows samples from the posterior over shapes for various objects in our experiment.

To calculate the similarity between target and comparison objects, we evaluated how likely it is to observe the image for one object given the image of the other object. Denoting the images for target and comparison by I_{t} and I_{c}, respectively, we calculated three similarity measures: $P(I_{\mathrm{t}}|I_{\mathrm{c}})$, $P(I_{\mathrm{c}}|I_{\mathrm{t}})$, and their average. We calculated $P(I_{\mathrm{c}}|I_{\mathrm{t}})$ as follows (and similarly for $P(I_{\mathrm{t}}|I_{\mathrm{c}})$):

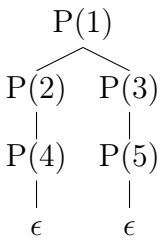
$$P(I_{\mathrm{c}}|I_{\mathrm{t}}) = \int P(I_{\mathrm{c}}|S, \vec{\phi})P(S|I_{\mathrm{t}})P(\vec{\phi})dSd\vec{\phi}. \tag{4.14}$$

In a similar vein to Equation 4.4, the value of this integral was approximated using samples from $P(S|I_{\mathrm{t}})$.

(a)



(b)



(c)

Part Id	Size	Conn. face
P1	[0.75, 0.50, 0.50]	n/a
P2	[0.64, 0.18, 0.45]	3
P3	[0.14, 0.28, 0.23]	2
P4	[0.37, 0.19, 0.18]	1
P5	[0.13, 0.24, 0.15]	1

Figure 4.8: (a) An example object. The numbers on parts refer to the part numbers in its parse tree. (b) Parse tree T associated with the object in (a). (c) Spatial model M associated with the object in (a). "Conn. face" is shorthand for "connection face" (i.e., the parent's face to which a part is connected).

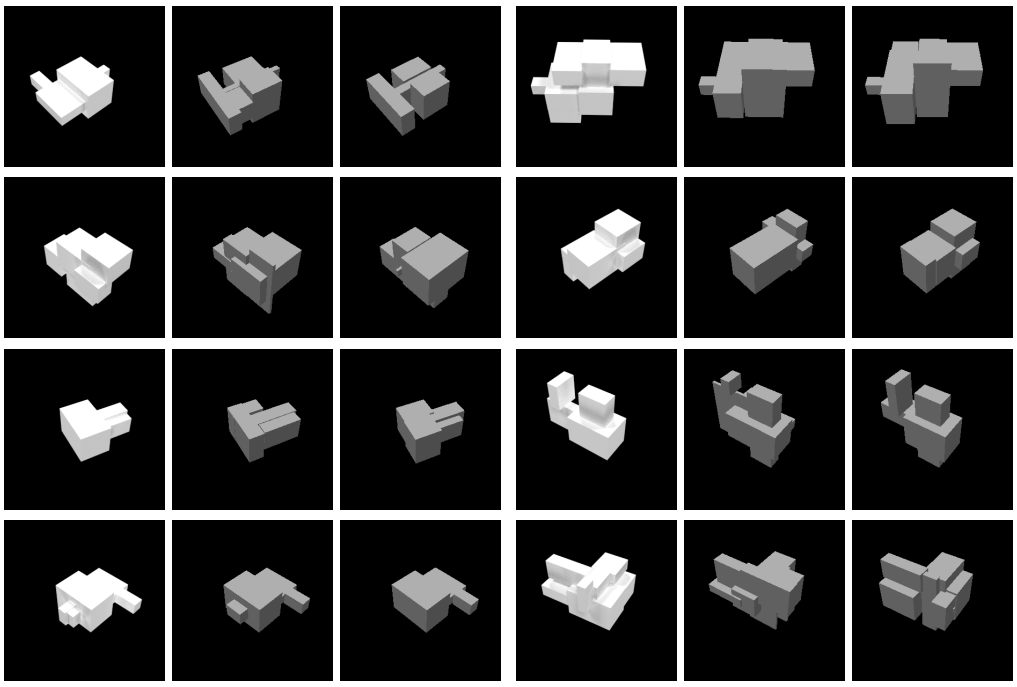


Figure 4.9: Samples from the posterior over shapes for various objects in our experiment. Each row contains two sets of one object followed by two samples.

Simulation results

For each computational model described above, we calculated its predictions as follows. For each simulated trial, we computed the similarities between a target object and each comparison object, and used the most similar comparison as a model's prediction. We evaluated the performance of each model by calculating the percentage of trials in which a model and our experimental subjects made the same judgment (i.e., they picked the same comparison object as most similar to the target). The results are shown in Figure 4.10.

Clearly, our proposed computational model significantly outperformed all other models (particularly the version whose similarity measure averaged $p(I_t|I_c)$ and $p(I_t|I_c)$; binomial test, $p < 0.005$ for all comparisons). The pixel-based (i.e., view-based) model performed at 58%. Even though this is significantly better than chance, it still lags far behind our model's performance of 72%. Similarly, the best alignment-based model only reached an accuracy of 59%.[11] The structural distance-based model lagged even the pixel-based model at 54% accuracy, which is not significantly better than chance. Similarly, the best version of the shape skeleton model performed worse than the pixel-based model with 56% accuracy (with similarity measure based on $P(I_c|I_t)$). However, this performance is significantly better

[11] Interestingly, allowing only translation and scaling transformations led to better performance than allowing any affine transformation. This might seem implausible because translation and scaling transformations are special cases of affine transformations. However, the alignment-based method simply finds the transformation that aligns two images as well as possible. This is not necessarily the alignment that makes the Euclidean distances between images reflect subjects' judgments.

than chance ($p = 0.035$). The best version of HMAX also performed worse than the pixel-based model and naive feature-based model (i.e., no alignment model) with an accuracy of 57% (with layer C1, s=5). Convolutional neural networks (CNNs) performed slightly better than pixel-based and alignment-based models. The best version of AlexNet reached an accuracy of 62% using its output layer prob, and the best version of GoogLeNet achieved 64% using layer inception5a. However, neither of these accuracies are significantly better than the pixel-based model's performance (binomial test, $p > 0.05$).

We also looked at the performance of each model on trials with high between-subject agreement. Even though average agreement between subjects was high (75%), it might be unfair to expect models to predict subjects' judgments on trials when subjects did not clearly prefer either comparison object significantly more than the other. The following analysis focuses on "high confidence" trials where at least 80% of subjects picked the same comparison object. Model accuracies on these high-confidence trials are shown in Figure 4.11. Our model significantly outperformed all other models with an accuracy of 87% ($p < 0.001$ for all comparisons). Pixel-based and alignment-based models achieved accuracies of 62% and 64%, respectively. Both of these values are significantly better than chance ($p = 0.01$ for pixel-based; $p = 0.002$ for alignment-based). Similarly, the structural distance-based model and shape skeleton model achieved an accuracy equal to that of the pixel-based model at 62%. The best version of HMAX performed at 57% which is not significantly different from the performance of either the pixel-

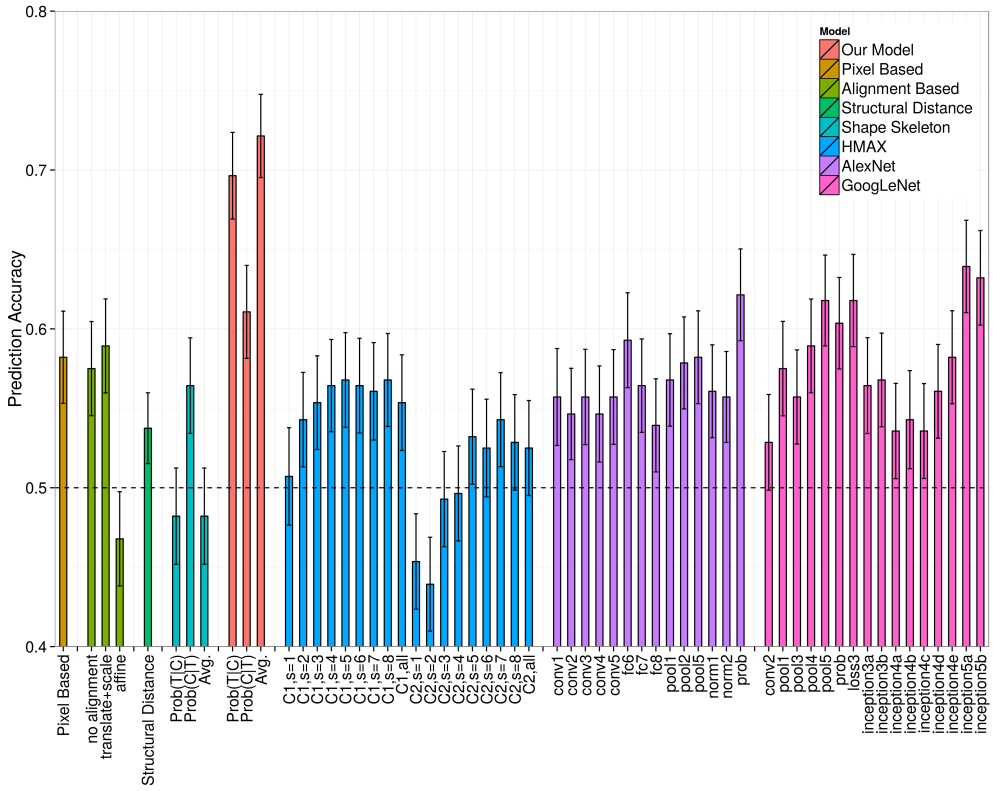


Figure 4.10: Predictions accuracies for each model on all trials. Error bars show SEMs estimated by a bootstrap procedure with 1000 replications. Note that the y-axis starts from 0.4.

based or alignment-based model. The best version of AlexNet reached an accuracy of 73% (with layer prob) which is significantly better than both pixel-based and alignment-based models ($p = 0.005$ for comparison with pixel-based; $p = 0.017$ for comparison with alignment-based). However, the best version of GoogLeNet reached an accuracy of 68% (with layer inception5b) which is not significantly better than the performance of either pixel-based or alignment-based models ($p = 0.11$ for comparison with pixel-based; $p = 0.24$ for comparison with alignment-based).

In the evaluations presented so far, object similarity was computed using the Euclidean similarity metric for several models. What would happen, however, if these models used a more powerful metric such as the Mahalanobis similarity metric? Would their performances significantly improve? The Euclidean metric is a special case of the Mahalanobis metric. Let $\vec{r}_M(I_i)$ denote a vector coding model M's shape representation based on image I_i. The Mahalanobis metric for the similarity of shape representations based on images I_i and I_j is:

$$[\vec{r}_M(I_i) - \vec{r}_M(I_j)]^T \, \Sigma^{-1} \, [\vec{r}_M(I_i) - \vec{r}_M(I_j)] \tag{4.15}$$

where Σ is a covariance matrix. The Euclidean metric is obtained by setting Σ to the identity matrix.

In the next analysis, we re-evaluated those models that previously used a Euclidean metric by allowing the models to use a Mahalanobis metric. For

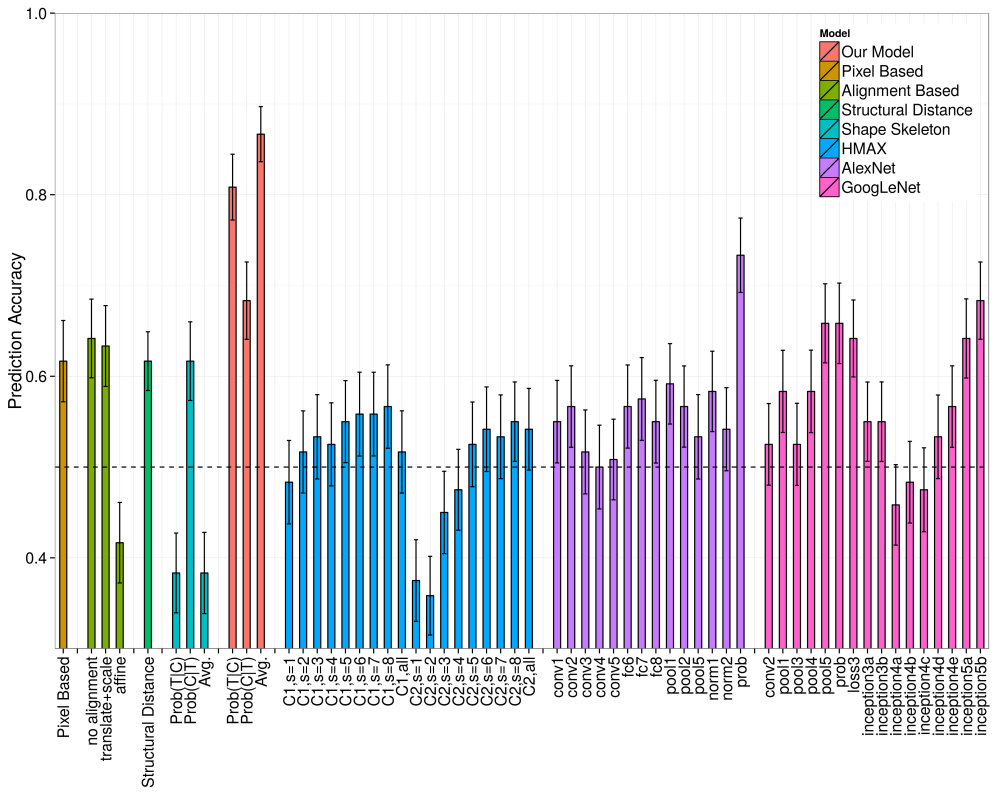


Figure 4.11: Predictions accuracies for each model on only high confidence trials. Error bars show SEMs estimated by a bootstrap procedure with 1000 replications. Note that the y-axis starts from 0.3.

each model, the covariance matrix Σ was obtained as follows. Subjects' judgments in our experiment can be thought of as relative similarity constraints. For example, if subjects picked object O_j to be more similar to O_i than O_k is, this can be characterized by a constraint of the form $s(O_i, O_j) > s(O_i, O_k)$, where s measures similarity between two objects. Using these constraints, it is possible to learn a Mahalanobis metric (i.e., learn a covariance matrix Σ) that satisfies as many of these constraints as possible. This problem is known as "metric learning" in the literature on Machine Learning (Kulis, 2013) where it is treated as an optimization problem that can be solved by iterative methods (see Appendix 4.A.3 for details on how we solved this problem).

To evaluate each model, 70% of subjects' similarity judgments selected at random were placed in a training set and the remaining judgments formed a test set. Using the training set, a model learned a Mahalanobis metric, and then this model was evaluated using the test set. This procedure was repeated 50 times to get a performance estimate for each model. We tried both diagonal and low-rank Σ matrices with varying rank values and report the best results. Table 4.1 shows the performances on all trials for the pixel-based model, the no-alignment (naive feature-based) model, HMAX, AlexNet, and GoogLeNet. (Recall that metric learning cannot be applied to the shape skeleton models, to the structural distance-based models, and to our proposed model because these models do not use vectors to represent shapes.)

Model	Metric type	Accuracy	Accuracy w/o metric learning
Pixel-based	low rank, r=10	0.566	0.582
Naive feature-based	diagonal	0.568	0.575
HMAX (C2, s=3)	diagonal	0.595	0.568
AlexNet (prob)	low rank, r=20	0.660	0.621
GoogLeNet (inception5b)	low rank, r= 20	0.633	0.639

Table 4.1: Best metric learning prediction accuracies on all trials.

Model	Metric type	Accuracy	Accuracy w/o metric learning
Pixel-based	low rank, r=10	0.698	0.616
Naive feature-based	diagonal	0.648	0.642
HMAX (C1, s=6)	diagonal	0.714	0.567
AlexNet (prob)	low rank, r=5	0.752	0.733
GoogLeNet (inception5b)	diagonal	0.715	0.683

Table 4.2: Best metric learning prediction accuracies on high-confidence trials.

Metric learning seems to help only AlexNet and, to a lesser extent, HMAX. However, neither of these increases in performance are statistically significant ($p = 0.18$ and $p = 0.37$ respectively). Importantly, our proposed computational model still outperforms all other models significantly ($p = 0.03$ for comparison with AlexNet). If we focus on only high confidence trials (see Table 4.2), metric learning improves the performances of all models, albeit not significantly for any model except HMAX ($p > 0.05$ for all other comparisons). Again, our 3D shape inference model is still significantly better than all other models ($p = 0.003$ for comparison with AlexNet). These results show that—even if we fit the similarity metric used by competing models to subject data—our shape inference model still provides a better account of subjects' judgments.

We believe that our results are significant in multiple respects. First,

our results suggest that people's shape representations for unfamiliar objects code 3D, rather than 2D, shape properties. Models that use 2D representations (i.e., pixel-based, alignment-based, and shape skeleton models)[12] were far inferior to our 3D shape inference model. Even if we allowed these models to fit their similarity metrics to subjects' data, our model still significantly outperformed them. These results strongly suggest that people do not represent shape for unfamiliar stimuli using 2D representations.

Second, our results raise doubts as to the promise of feature-based models. Even though these models tended to perform better than other models, they were still significantly behind our 3D shape inference model. This result is especially interesting for CNNs, which have attracted interest in the cognitive science and neuroscience communities as good models of biological visual systems. Their poor performance at accounting for our experimental data suggests that these models might be representing visual objects in a manner that is different from how people represent visual objects. Further evidence for this claim is provided by studies showing that CNNs are easily fooled by images that seem indistinguishable or unrecognizable to the human eye (Nguyen, Yosinski, & Clune, 2014; Szegedy, Zaremba, & Sutskever, 2013).

[12]It is worth emphasizing here that we base our claim on the 3D nature of shape representations, *not* on a comparison between our model and deep neural networks because it is unclear whether deep neural network models of shape perception use 2D or 3D representations. We touched upon this difficulty of knowing how and why these models achieve what they achieve in our review of feature-based models in Section 2. In fact, one line of research (Patel et al., 2015) suggests that deep neural networks are implementing an approximate version of probabilistic inference in a hierarchical probabilistic rendering model, similar to our proposed approach.

We should stress that we are not arguing that feature-based models (e.g., convolutional neural networks) are fundamentally inadequate for modeling biological visual systems. Since these models are universal approximators (i.e., can implement any input-output mapping), given the right model architecture, training data and optimization procedure, they can be trained to capture any empirical data. For example, one might fine tune AlexNet and GoogLeNet on the images of objects used in our experiment. However, there are two problems with such an approach. First, subjects in our experiments have never seen objects like the ones in our experiment either. Second, in cognitive science and neuroscience literatures, these models are presented as good models of our visual systems *without* any further training (Khaligh-Razavi & Kriegeskorte, 2014; Cadieu et al., 2014). Our results here suggest that this is not the case.

Third, the structural-description based model's poor performance suggests that it is not adequate to represent objects as lists of parts and the coarse spatial relations among parts. Subjects' similarity judgments in our experiment seem to be based on finer-scale information than encoded in these structural descriptions, including the probabilistic information inferred by our proposed model.

Finally, our results have implications for the view-based hypothesis. Here we tested several view-based models. Alignment-based models tested S. Ullman (1989)'s approach, and our pixel-based model and no-alignment models tested two versions of Poggio and Edelman (1990)'s influential view approx-

imation model. Our results show that none of the view-based models can account for subjects' judgments, and strongly suggest that view-based models do not provide good models of human shape perception.

Discussion

In summary, we have pursued an approach to investigating shape perception based on the "visual perception as Bayesian inference" framework. We hypothesized that shape perception of unfamiliar objects is well characterized as statistical inference of 3D shape in an object-centered coordinate system. The article provided evidence for this hypothesis along two lines. It first showed that a shape inference model that uses probabilistic, 3D, object-centered shape representations can account for view-dependency. This is a surprising result because previous researchers have interpreted view-dependency as incompatible with 3D, object-centered representations. Based on this result, we argued that view-dependency is not diagnostic of whether shape representations are 2D versus 3D, nor is it diagnostic of whether these representations are view-based versus view-independent. In addition, the article reported the results of a behavioral experiment using a shape similarity task, and compared the predictions of a diverse array of computational models to the experimental data. We found that our proposed shape inference model captures subjects' behaviors better than competing models. In conjunction, our experimental and computational results illustrate the promise of our ap-

proach and suggest that people's shape representations of unfamiliar objects are probabilistic, 3D, and object-centered.

Research on the visual perception of object shape has a long history. However, in terms of understanding the representations and algorithms involved in shape perception, it often seems as if we have made little progress (Peissig & Tarr, 2007; Gauthier & Tarr, 2016). We believe this is largely due to a lack of rigorous and quantitative approaches addressing the whole shape perception process from images to behavior. For example, view-based hypotheses rarely made commitments on the representation of individual views, or structural description hypotheses never completely specified how structural descriptions can be extracted from 2D images or how such descriptions can be compared. Hence, it became difficult to test these hypotheses, since without a clear specification of the whole perception process, their predictions were subject to interpretation. We believe progress is possible only if we build rigorous computational models, and our study is significant because it presents one such rigorous model of shape perception. As argued by Gauthier and Tarr (2016), we need to move away from unproductive dichotomies such as view-dependent versus view-invariant representations towards understanding the nature of the representation and algorithms involved in shape perception, which ultimately will explain when and why view-invariant or view-dependent performance is obtained. Our rigorous and quantitative approach here enables us to do exactly that.

We believe our work here is also significant because it presents a con-

ceptual framework for understanding shape perception in its totality, rather than one aspect of it such as view-dependency or behavior on some single task such as object recognition. For example, view-based models focused almost exclusively on view-dependency of object recognition. Similarly, popular feature-based models are all models of object recognition. However, there is much more to shape perception than view-dependency or mapping images of objects to labels. We believe our approach is significant because it addresses shape perception in its totality, not just one aspect of it. By treating shape perception as inference of 3D, object-centered representations, we can explain not only view-dependency but also capture perceived similarities between unfamiliar objects. This is possible because our framework presents a generative model of shape perception, capturing how causes in the world give rise to retinal stimulations. Such models are often contrasted with discriminative approaches (such as popular feature-based models like AlexNet and GoogLeNet) that are built for individual tasks (such as object recognition) and cannot be easily adapted to new tasks (Lake et al., 2016).

Our work directly or indirectly addresses or raises a large number of questions about the representation of object shape. Here we address several of these questions.

Previous research in the psychology literature has focused on how people might represent object shape, but has largely ignored the question of how people might acquire these representations. Why does the hypothesis proposed here emphasize that shape perception is a form of statistical inference? We

believe that focusing on visual representations without also focusing on the acquisition of these representations is misguided. For example, it led researchers to develop theories of shape perception based on complete and accurate 3D, object-centered shape representations despite the fact that the acquisition of such representations is perceptually (and computationally) implausible, especially from a small number of viewpoints. If one augments an emphasis on representation with an emphasis on inference, one quickly realizes that people's shape representations will rarely be complete and accurate. For example, when a person views an object from a single viewpoint, the person is likely to infer a relatively accurate representation of some portions of the object but an inaccurate representation of other portions (e.g., portions seen in the periphery, or portions that are partially or fully occluded). We claim that this shape-inference problem underlies view-dependency.

The proposed computational model uses a specific approach, namely one based on probabilistic shape grammars. Why adopt this approach? Our proposed model uses a probabilistic shape grammar for several reasons. First, a shape grammar characterizes knowledge of possible object parts and of how parts might be combined to form objects. Part-based shape representations have previously received considerable theoretical and empirical support in the psychology literature (Biederman, 1987; Hoffman & Richards, 1984; Marr & Nishihara, 1978; Saiki & Hummel, 1998; Yildirim & Jacobs, 2013). Second, we represent shape in a probabilistic manner because probabilistic approaches are robust in noisy and uncertain environments, and because

probabilistic inference algorithms often show excellent performances (as evidenced by the tremendous progress in the fields of Machine Learning and Statistics over the past few decades). Third, we are reasonably optimistic that the proposed model (or, rather, appropriately extended versions of the model) will scale well to larger-scale settings. Although important challenges obviously remain (too many to be mentioned here), our optimism stems from the fact that probabilistic shape grammars (much more complex than the one reported here) are regularly used in the Computer Vision and Computer Graphics literatures to address large-scale problems (Amit & Trouve, 2007; Bienenstock et al., 1997; Fu, 1986; Grenander & Miller, 2007; Talton et al., 2012; Tu et al., 2005; L. Zhu, Chen, & Yuille, 2007; L. Zhu et al., 2009).

The proposed computational model seems restricted to part-based objects. Is this a significant shortcoming? Can this model be scaled up to handle natural objects? Our main focus in this study was to argue for probabilistic, 3D, and object-centered shape representations. We have chosen the particular part-based shape representations used in this work because these are both powerful enough to capture 3D geometry of the stimuli we used and simple enough to make inference computationally feasible. Our mental shape representations are no doubt much richer than the representations we used here. A comprehensive understanding of object shape perception will require future work on shape representations that are rich enough to represent natural objects.

It is notoriously hard to predict the future[13] but we are hopeful that our approach can be scaled up to deal with the full complexity of natural objects. 3D volumetric representations similar to ours are being scaled to larger and larger settings by computer vision researchers (Rezende et al., 2016; Qi et al., 2016; Wu, Zhang, Xue, Freeman, & Tenenbaum, 2016). Moreover, recent research in Machine Learning and Statistics is leading to exciting advances in efficient inference in generative models. For example, fast, discriminative models can be trained to speed up inference dramatically in generative models (Kingma & Welling, 2014; Kulkarni, Yildirim, et al., 2014; Yildirim, Kulkarni, Freiwald, & Tenenbaum, 2015).

The proposed computational model makes use of a powerful "forward model" that maps shape representations and viewpoints to visual images. Is this realistic? We believe that it is. Our results show that people discount viewpoint to a large extent when judging similarities which suggests such a forward model is implemented by our visual systems. In other settings, this mapping is referred to as visual imagery. Visual imagery is a type of mental simulation which researchers are increasingly hypothesizing as playing an essential role in human perception and cognition (Battaglia, Hamrick, & Tenenbaum, 2013).

The hypothesis proposed here is restricted to unfamiliar objects. Why?

[13]Minsky and Selfridge (1961) famously predicted that hill-climbing approaches will never scale beyond the simple neural networks of the time. The current ubiquitous use of the backpropagation algorithm for training deep neural networks illustrates how wrong well-intentioned predictions can be.

There are at least two reasons for this choice. First, our focus on unfamiliar objects provides a setting where potential confounding factors are controlled. Given past experience with familiar objects and their possible semantic significance, it is difficult (perhaps impossible) to dissociate the representation of shape from other possible relevant factors such as object category, object function, and developmental and evolutionary significance. Indeed, previous research clearly shows that conceptual knowledge affects visual perception (Dixon, Bub, & Arguin, 1997; Gauthier, James, Curby, & Tarr, 2003; Goldstone, Lippa, & Shiffrin, 2001; Wiseman, MacLeod, & Lootsteen, 1985). Second, and perhaps more important, we believe that it is unrealistic to expect that people's visual systems use a single shape representation for all objects. For example, given the significance of some familiar objects—such as faces—and the difficulty of the associated visual recognition problem, it seems likely that people have specialized mechanisms and representations for these highly significant and familiar objects.

The hypothesis proposed here does not take into account an observer's task or goal. Is this a significant shortcoming? Yes and no. Consistent with the "active vision" approach to the study of perception (Findlay & Gilchrist, 2003; Hayhoe & Ballard, 2005), we believe that visual perception is often task-based. At the same time, we also believe that people use multiple representations of object shape, including representations that are not strongly dependent on task. Among other sources, evidence for this claim comes from our own recent brain-imaging research showing that cortical re-

gion LOC forms similar (and part-based) object shape representations when people visually or haptically perceive an object's shape in the absence of a task (Erdogan, Chen, Garcea, Mahon, & Jacobs, 2016).

Object shape can be perceived visually but it can also be perceived haptically. What is the relationship between visually-based and haptically-based shape representations? We believe that behavioral and computational studies (Erdogan et al., 2015; Yildirim & Jacobs, 2013) as well as brain imaging studies (Erdogan et al., 2016) suggest that people acquire and use modality-independent object shape representations. These representations underlie behavioral phenomenon, such as cross-modal transfer of shape knowledge (Lacey & Sathian, 2011; Newell, 2010; Wallraven et al., 2014), and seem to reside in neural region LOC as well as other regions (Amedi et al., 2001; Erdogan et al., 2016; Grill-Spector et al., 2001; James et al., 2002). Our own previous work has shown that a computational model related to the one proposed here can infer shape representations from visual information, from haptic information, or both, and can account for an array of experimental data on cross-modal transfer of shape knowledge (Erdogan et al., 2015; Yildirim & Jacobs, 2013).

Is the proposed computational model psychologically plausible? Is it neurally plausible? Cognitive scientists often make a distinction between rational models and process models. Rational models are models of optimal or normative behavior, characterizing the problems that need to be solved in order to generate the behavior as well as their optimal solutions. In con-

trast, process models are models of people's behaviors, characterizing the mental representations and operations that people use when generating their behavior. Because our model's inference algorithm is optimal according to Bayesian criteria, and because this algorithm is not psychologically plausible, the model should be regarded as a rational model, not as a process model. Nonetheless, we believe that there are benefits to regarding the model as a rational/process hybrid. Like rational models, our model is based on optimality considerations. However, like process models, it uses psychologically plausible representations and operations (e.g., grammars, forward models).

For readers solely interested in process models, we claim that our model is a good starting point. As pointed out by others (Griffiths et al., 2012; Sanborn et al., 2010), the MCMC inference algorithm used by our model can be replaced by approximate inference algorithms (known as particle filter or sequential Monte Carlo algorithms) that are psychologically plausible. Doing so would lead to a so-called "rational process model", a type of model that is psychologically plausible and also possesses many of the advantages of rational models.

In regard to neural plausibility, an important trend in computational neuroscience is to interpret neural activity in terms of probabilistic representations and operations (Pouget, Beck, Ma, & Latham, 2013). We, therefore, regard our model as at least potentially neurally plausible.

What are some important areas for future studies? We have emphasized the need to augment an emphasis on visual representation with an emphasis

on the idea that shape perception is a form of statistical inference. This perspective leads to at least two areas for future research. First, any statistical inference mechanism needs to contain inductive biases in order to be effective. Future research needs to study the biases that play a role when people infer shape. These biases might take the form of "generic view" assumptions (Freeman, 1996) or "simplicity" assumptions (Feldman, 2000; Feldman et al., 2013; Griffiths, Chater, Kemp, Perfors, & Tenenbaum, 2010). Second, the fact that our shape representations are the product of an inference process means that these representations may be inaccurate or incomplete (highlighting an advantage of probabilistic representations which directly code uncertainty). Here, we showed that an important consequence of this fact is that our percepts are view-dependent. Future research will need to study other perceptual consequences of our visual inference mechanisms.

Appendix to Chapter 4

Details of MCMC algorithm for viewpoint-dependency simulations

In this appendix, we present the details of our Markov chain Monte Carlo (MCMC) procedure for inferring posterior probability distributions over the shapes of paperclip stimuli used in our viewpoint dependency simulations.[14]

[14]Implementations of the inference procedure for both paperclip and block stimuli are available online at `https://github.com/gokererdogan/Infer3DShape/releases/tag/ro3Dpaper`

To sample from the posterior distribution $P(S, \vec{\phi}|I)$ over shape representations given a 2D image, we use MCMC techniques (J. S. Liu, 2004). These techniques produce samples from a desired probability distribution by constructing a Markov chain whose stationary distribution is the distribution of interest. In our inference procedure, we use the Metropolis-Hastings (MH) algorithm, a popular algorithm for constructing such Markov chains (Metropolis et al., 1953; Hastings, 1970).

An MH algorithm proposes a new hypothesis H' based on the current hypothesis H at each iteration, and accepts or rejects the proposed hypothesis with some probability. This accept/reject probability, called the acceptance ratio, is designed in such a way as to ensure that the stationary distribution of the Markov chain is the distribution of interest. Denote the probability of proposing hypothesis H' given the current hypothesis H with $q(H'|H)$ and the distribution of interest with $\pi(H)$. The MH acceptance ratio is:

$$a(H \to H') = \min\left(1, \frac{\pi(H')q(H|H')}{\pi(H)q(H'|H)}\right). \tag{4.16}$$

In our case, the target distribution $\pi(H)$ is the posterior $P(S, \vec{\phi}|I)$, and we need to design a proposal function q to move efficiently in the space of hypotheses. We use a mixture proposal (Tierney, 1994; Brooks, 1998) that consists of multiple proposals where one proposal is picked randomly at each iteration. Below we discuss each proposal function and its associated acceptance ratio.

Add/remove endpoint proposal: Given a shape $S = \{\vec{p}_1, \vec{p}_2, \ldots, \vec{p}_{|S|}\}$ consisting of $|S|$ endpoints, the add/remove endpoint proposal adds or removes a single endpoint. We allow only the "free" endpoints (i.e., $\vec{p}_1$ or $\vec{p}_{|S|}$) to be removed, and a new endpoint can only attach to one of these free endpoints. We calculate the probability for this proposal by considering the probability of each step in the procedure for adding/removing an endpoint. The add/remove endpoint proposal first randomly picks whether an add or remove endpoint manipulation should be carried out. We set each manipulation to be equally likely (i.e., $P(\text{add}|H) = P(\text{remove}|H) = 0.5$).[15] For a remove endpoint manipulation, the next step is to pick the endpoint to remove. Since there are two free endpoints, one of these is picked at random. For an add endpoint manipulation, again we first need to pick the free endpoint. In addition, we need to pick the position (x', y', z') of the new endpoint. A random vector on the unit sphere is picked randomly and added to the picked free endpoint to determine the position of the new endpoint.

[15]For some shapes, it might not be possible to add or remove a endpoint. For example, we never allow shapes with no endpoints. Therefore, we cannot apply a remove endpoint manipulation to a shape with only a single segment. In such cases, add and remove manipulation probabilities need to be modified accordingly. Similar modifications may be required for other steps in the add/remove endpoint proposal as well. See the implementation of our model for details.

The proposal probabilities for add and remove endpoint manipulations are:

$$q_{\text{add}}(H'|H) =$$
$$P(\text{add}|H)\ P(\text{pick endpoint}|H, \text{add})\ P(x', y', z'|H, \text{add}, \text{pick endpoint}) \tag{4.17}$$

$$q_{\text{remove}}(H'|H) = P(\text{remove}|H)\ P(\text{pick endpoint}|H, \text{remove}). \tag{4.18}$$

However, we cannot simply plug these into the MH acceptance ratio formula because the add/remove endpoint proposal manipulations move between spaces with different numbers of dimensions—shapes with different numbers of endpoints live in spaces with different number of dimensions. Therefore, we use a variant of the MH algorithm called "reversible jump MCMC" that can move between such spaces (Green, 1995). To see how it is applied for our add/remove endpoint proposal, assume that we have a shape $S = \{\vec{p}_1, \vec{p}_2, \ldots, \vec{p}_{|S|}\}$ that consists of $|S|$ endpoints, and we add a new endpoint to get the proposed hypothesis $S' = \{\vec{p'}_1, \vec{p'}_2, \ldots, \vec{p'}_{|S|}, \vec{p'}_{|S|+1}\}$. Reversible jump MCMC assumes that we have sampled random variable $\vec{u}$ to make the number of dimensions equal in both hypotheses. In our case, we sampled $\vec{u} = (x', y', z') \in \mathbf{R}^3$ and added it to shape S (i.e., $S = \{\vec{p}_1, \vec{p}_2, \ldots, \vec{p}_{|S|}, \vec{u}\}$). We define a function $h : \vec{p}_1, \vec{p}_2, \ldots, \vec{u} \rightarrow \vec{p'}_1, \vec{p'}_2, \ldots, \vec{p'}_{|S|+1}$ that maps shape

S to shape S'. Then, the reversible jump acceptance ratio is:

$$a(S \to S') = \min\left(1, \frac{\pi(S')q(S|S')}{\pi(S)q(S'|S)} \left|\det\left(\frac{\partial(\vec{p'}_1, \vec{p'}_2, \ldots, \vec{p'}_{|S|+1})}{\partial(\vec{p}_1, \vec{p}_2, \ldots, \vec{u})}\right)\right|\right) \quad (4.19)$$

where the rightmost term in this equation is the absolute value of the determinant of the Jacobian of the mapping h. Since in our case h is the identity function, its Jacobian is 1. Therefore, the acceptance ratio for the add endpoint manipulation is:

$$a(H = (S, \vec{\phi}) \to H' = (S', \vec{\phi})) = \min\left(1, \frac{\pi(H')q_{\text{remove}}(H|H')}{\pi(H)q_{\text{add}}(H'|H)}\right) \quad (4.20)$$

where q_{add} and q_{remove} are given by Equations 4.17 and 4.18, respectively. The acceptance ratio for the remove endpoint manipulation from H' to H is the inverse of the above expression.

Move endpoint proposal: This proposal picks one endpoint randomly and moves it a random amount $\vec{m} \in \mathbf{R}^3$ sampled from a normal distribution $\mathcal{N}(0, \sigma^2\mathbf{I})$. Hence, the proposal probability $q(H'|H)$ is:

$$q(H'|H) \propto \exp\left(-\frac{\sum_{i=1}^{3} m_i^2}{2\sigma^2}\right). \quad (4.21)$$

Since this proposal is symmetric (i.e., $q(H'|H) = q(H|H')$), the MH acceptance ratio is:

$$a(H \to H') = \min\left(1, \frac{\pi(H')}{\pi(H)}\right). \quad (4.22)$$

Rotate viewpoint proposal: This proposal changes the viewpoint $\vec{\phi} = (r, \theta, \alpha)$ from which a shape is viewed. We sample two random angles from a von Mises distribution with mean zero and variance κ, and add these to the polar coordinates θ and α. Since this proposal is symmetric, the acceptance ratio is again given by Equation 4.22.

Details of MCMC algorithm for shape similarity task

In this appendix, we present the details of our MCMC inference procedure for block stimuli used in our behavioral experiment. See Appendix 4.A.1 for a short discussion of the Metropolis-Hastings algorithm. Similar to our inference procedure for paperclip stimuli, we use a mixture proposal that consists of multiple proposals, one of which is picked randomly at each iteration. Below we provide the details for each proposal procedure.

Add/remove part proposal: Let $S = (T, M)$ denote a shape where T refers to the parse tree associated with the shape, and M is the spatial model that consists of one size vector $\vec{s}_i \in \mathbf{R^3}$ and connecting face $f_i \in \{1, 2, 3, 4, 5, 6\}$ for each P node in parse tree T. The add/remove part proposal first randomly picks whether an add or remove manipulation will be carried out. We assume each manipulation is equally likely. For a remove part manipulation, a P node is picked randomly from the set $\mathcal{R}$ of P nodes with no children, and this part is removed. For an add part manipulation, a P node is picked randomly from the set $\mathcal{A}$ of P nodes that have fewer

than three child P nodes. Then, a new child P node is added to the picked P node. This requires randomly sampling a size $\vec{s}$ for the new part and a connecting face f from the unoccupied connecting faces of its parent. The proposal probabilities for add and remove manipulations are:

$$q_{\text{add}}(H'|H) = P(\text{add}|H)\ P(\text{pick part}|H, \text{add})\ P(\vec{s})\ P(f|H, \text{add}, \text{pick part}) \tag{4.23}$$
$$= \frac{1}{2}\frac{1}{|\mathcal{A}|}\frac{1}{(6 - |O_P|)}$$

$$q_{\text{remove}}(H'|H) = P(\text{remove}|H)\ P(\text{pick part}|H, \text{remove}) \tag{4.24}$$
$$= \frac{1}{2}\frac{1}{|\mathcal{R}|}$$

where we assume $P(\vec{s})$ is uniform and use O_P to denote the set of occupied faces of the picked parent P part for add part manipulation.[16] Similar to the add/remove endpoint proposal for paperclip stimuli discussed above, we cannot simply plug these proposal probabilities into the MH acceptance ratio because hypotheses H and H' reside in spaces with different numbers of dimensions. Therefore, we use the reversible jump MCMC algorithm. A derivation similar to the one discussed for the add/remove endpoint proposal

[16] In some cases, it might not be possible to add or remove parts for a shape S. The proposal probabilities need to be modified accordingly in such cases.

shows that the acceptance ratio for the add part manipulation is:

$$a(H = (T, M, \vec{\phi}) \rightarrow H' = (T', M', \vec{\phi})) = \min\left(1, \frac{\pi(H')\ q_{\text{remove}}(H|H')}{\pi(H)\ q_{\text{add}}(H'|H)}\right) \tag{4.25}$$

where q_{add} and q_{remove} are given by Equations 4.23 and 4.24, respectively. The acceptance ratio for the remove part manipulation from H' to H is the inverse of the above expression.

Change part size proposal: This proposal picks one P node randomly from shape $S = (T, M)$ and resamples its size $\vec{s}$ from a uniform distribution over $[0, 1] \times [0, 1] \times [0, 1]$. Since this proposal is symmetric, the MH acceptance ratio is given by Equation 4.22.

Change connecting face of part proposal: This proposal picks one P node randomly from the set of P nodes whose parent P node has at least one empty face. A new connecting face is picked randomly from the set of empty faces of its parent, and the P node is connected to this new face. Again, because this proposal is symmetric, the MH acceptance ratio is given by Equation 4.22.

Rotate viewpoint proposal: This proposal changes the viewpoint $\vec{\phi} = (r, \theta, \alpha)$ from which a shape is viewed. In contrast to the proposal we used for paperclip stimuli, here we allow rotations only around the vertical direction. We sample a random angle from a von Mises distribution with mean zero

and variance κ and add this to the polar coordinate θ. Since this proposal is symmetric, the acceptance ratio is given by Equation 4.22.

Details of metric learning

In our evaluation of shape perception models, we use metric learning to fit the representations learned by models to behavioral data. Metric learning (Kulis, 2013) aims to learn a linear transformation of input data such that the distances between data points in the transformed space capture similarity/dissimilarity relations as well as possible. More formally, denote the representation for stimuli i with $\vec{r}_i$, and the distance between stimuli i and j with $d(\vec{r}_i, \vec{r}_j)$. Assume that we are given a set of relative similarity constraints of the form $d(\vec{r}_i, \vec{r}_j) < d(\vec{r}_i, \vec{r}_k)$. Our aim is to learn a linear mapping A such that the distances $d_A(\vec{r}_i, \vec{r}_j), d_A(\vec{r}_i, \vec{r}_k)$, etc., in this new space will satisfy as many of these relative similarity constraints as possible. Here $d_A(r_i, r_j)$ is the Mahalonobis distance between $\vec{r}_i$ and $\vec{r}_j$ which is given by $(\vec{r}_i - \vec{r}_j)^T A(\vec{r}_i - \vec{r}_j)$. Because there might not be a linear mapping A satisfying all constraints, we introduce slack variables ξ_{ijk} to express the metric learning problem as the following optimization problem (Schultz & Joachims, 2003):

$$\begin{aligned} \min_{A,\{\xi_{ijk}\}} \quad & \frac{1}{2}||A||_F^2 + C \sum_{ijk \in R} \xi_{ijk} \\ \text{s.t.} \quad & d_A(\vec{r}_i, \vec{r}_k) - d_A(\vec{r}_i, \vec{r}_j) \leq 1 - \xi_{ijk} \\ & \xi_{ijk} \geq 0 \end{aligned} \tag{4.26}$$

where $||\cdot||_F$ denotes the Frobenius norm and C is a cost parameter controlling how much we care about violations of the relative similarity constraints. We consider two variants of this problem. In the first variant, we constrain A to be a diagonal matrix. In that case, this problem becomes equivalent to the one treated in Schultz and Joachims (2003). We find the optimal diagonal A by solving the dual of the optimization problem using the L-BFGS-B algorithm (Byrd, Lu, Nocedal, & Zhu, 1995) provided in the "scipy" open-source package of scientific tools (Jones, Oliphant, Peterson, & others, 2001). The second variant constrains A to be a low rank matrix. This can be achieved by writing A as $G^T G$ where G has fewer rows than columns. To solve this problem, we rewrite it in the following unconstrained form:

$$\min_{G} \quad \frac{1}{2}||G^T G||_F^2 + C \sum_{ijk \in R} \max(0, d_A(\vec{r}_i, \vec{r}_j) - d_A(\vec{r}_i, \vec{r}_k) + 1) \tag{4.27}$$

and again use L-BFGS-B to find the optimal A matrix. Implementations of these metric learning methods are provided online at `https://github.com/gokererdogan/gmllib`.

Chapter 5

Discussion

In this thesis, we put forward a computational theory of shape perception and tested it in a series of behavioral, neuroimaging, and computational studies. We argued that shape perception is best understood as *statistical inference over modality-independent, part-based, 3D, and object-centered representations*. In Chapter 2, we have shown that a computational shape perception model based on our hypothesis accounts very well for within- and cross-modal shape similarity judgments, and that our framework can explain how modality-independent representations are acquired from sensory-specific inputs. In Chapter 3, we turned to neuroimaging and have shown that visual and haptic stimulation lead to similar neural activity in lateral occipital complex (LOC). In addition, we have demonstrated that the part-based structure of an object can be decoded from visual and haptic neural activations in LOC. These findings provide evidence for multisensory and part-based shape representations in the brain. Chapter 4 has focused on visual shape perception and shown that a computational shape perception model based on statisti-

cal inference over 3D, object-centered shape representations accounts better for shape similarity judgments than other competing models. We have also shown that such a model can explain view-dependency of object recognition, a finding that is usually taken as evidence against 3D, object-centered shape representations. We believe, taken altogether, these studies present a strong case for our hypothesis of shape perception as statistical inference over modality-independent, part-based, 3D, and object-centered representations.

As we have remarked in previous chapters, the pieces of this hypothesis appeared in earlier research on shape perception. Our shape representations are clearly inspired by volumetric and hierarchical representations of Marr and Nishihara (1978); Biederman (1987). Our emphasis on modality-independence is shared by earlier work (Yildirim & Jacobs, 2012, 2013). And "perception as Bayesian inference" is a well-established approach in cognitive science and especially in research on visual perception (Knill & Richards, 1996; Kersten & Yuille, 2003). However, what makes our work novel is the combination of all these pieces and our emphasis on the algorithms as well as the representations that take role in shape perception.[1] For example, in Chapter 4, we have seen that this emphasis on the whole shape perception process allowed us to explain view-dependency of object recognition with 3D, object-centered representations.

[1] A hypothesis on representation by itself is incomplete without an accompanying hypothesis on algorithm, i.e., on how representations are used. This point was forcefully made by Anderson (1978) in the context of analog vs. symbolic representations debate in mental imagery.

An important feature of our hypothesis here is that it uses rich, symbolic shape representations that encode 3D geometry and are modality-independent and compositional. Such representations can be contrasted with the 2D and flat representations of view-based or some feature-based models. Exactly how rich our shape representations should be is a long-standing question that featured prominently in various debates in cognitive science (McClelland et al., 2010; Griffiths et al., 2010). We believe that the efficiency and flexibility of perception call for rich, structured representations. For example, our shape similarity experiment in Chapter 4 suggests that people extract a rich representation of the 3D structure of an object even from a single image. Similarly, previous research has shown that people can learn new object concepts from just a few examples (Tenenbaum et al., 2011). Such feats are possible only if we employ rich representations that impose strong inductive biases on our generalizations. Perhaps an equally forceful argument for rich representations can be made on the basis of the richness of the external world and the diversity of the tasks we face. Research in shape perception generally focused on one particular task, object recognition. However, our interactions with the world require more information than only the identities of objects. Object shape is crucial not only for recognition but also for scene understanding, motor planning and many other tasks. This perspective on perception necessitates rich representations that capture all we know about the external world.

Computational models developed in earlier chapters use various represen-

tational schemes from ones based on a fixed set of parts and a discretized 3D space (in Chapter 2) to ones that model objects out of blocks connected to each other at fixed docking locations (in Chapter 4). No doubt, such representational schemes are not adequate for fully capturing the rich structure of objects we encounter in the real world. This has been and still is the fundamental question in shape perception. We need representational schemes that are powerful enough to capture the richness of natural objects and yet at the same time allow efficient inference and learning. Even though we might seem quite far from this goal, we are reasonably optimistic that, with the framework we presented in this thesis, we can make progress on this fundamental question. Future work should challenge our computational model with more and more empirical data (e.g., similarity judgments on much larger sets of highly diverse and naturalistic objects and under more naturalistic settings) and refine our hypothesis on shape perception accordingly.

One reason for our optimism is the dramatic advances made in AI and particularly in computer vision in the past decade. Advances in generative modeling and variational techniques are making it possible to scale models like ours to larger and larger settings (Kingma & Welling, 2014; Goodfellow et al., 2014; Angelino, Johnson, & Adams, 2016; Mnih & Rezende, 2016; Rezende et al., 2016). And models that are trained on large amounts of data lead to more and more improvements in accuracy on many tasks, from object recognition to segmentation (Bengio, 2009; LeCun et al., 2015). We believe there is an opportunity for cognitive science here. Most of our computational

models in cognitive science can account for only a very small set of empirical data, perhaps from one or two experiments. However, large datasets used in computer vision provide a far larger set of empirical data we can use to test and improve our models. If we can fit computational models that use rich representations like ours to large amounts of empirical data, we can gain further insights into the nature of our shape representations. This would require either writing shape grammars for large numbers of different objects or learning shape grammars via induction directly from data, both of which are still significant challenges. Recent models that learn deep generative models with interpretable representations represent promising early steps in this direction (X. Chen et al., 2016; Siddharth et al., 2016; Kusner, Paige, & Hernandez-Lobato, 2017). If the rapid progress made in computer vision in the past decade is any indication, grammar induction on large datasets will become a possibility soon enough, and we will get one step closer to understanding object shape perception.

References

Amedi, A., Jacobson, G., Hendler, T., Malach, R., & Zohary, E. (2002). Convergence of visual and tactile shape processing in the human lateral occipital complex. *Cerebral Cortex*, *12*(11), 1202–12.

Amedi, A., Malach, R., Hendler, T., Peled, S., & Zohary, E. (2001). Visuo-haptic object-related activation in the ventral visual pathway. *Nature Neuroscience*, *4*(3), 324–330.

Amedi, A., von Kriegstein, K., van Atteveldt, N. M., Beauchamp, M. S., & Naumer, M. J. (2005). Functional imaging of human crossmodal identification and object recognition. *Experimental Brain Research*, *166*(3-4), 559–571.

Amit, Y., & Trouve, A. (2007). POP: Patchwork of parts models for object recognition. *International Journal of Computer Vision*, *75*(2), 267–282.

Anderson, J. R. (1978). Arguments concerning representations for mental imagery. *Psychological Review*, *85*(4), 249–277.

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, N.J:

Psychology Press.

Angelino, E., Johnson, M. J., & Adams, R. P. (2016, November). Patterns of scalable bayesian inference. *Foundations and Trends in Machine Learning*, *9*(2-3), 119–247.

Anselmi, F., Rosasco, L., Tan, C., & Poggio, T. (2015). Deep convolutional networks are hierarchical kernel machines. *arXiv:1508.01084*.

Ashby, F. G., Maddox, W. T., & Lee, W. W. (1994, May). On the dangers of averaging across subjects when using multidimensional scaling or the similarity-choice model. *Psychological Science*, *5*(3), 144–151.

Ballard, D. H. (1981, January). Generalizing the Hough transform to detect arbitrary shapes. *Pattern Recognition*, *13*(2), 111–122.

Ballesteros, S., Gonzalez, M., Mayas, J., Garcia-Rodriguez, B., & Reales, J. M. (2009). Cross-modal repetition priming in young and old adults. *European Journal of Cognitive Psychology*, *21*(2-3), 366–387.

Bar, M. (2001). Viewpoint dependency in visual object recognition does not necessarily imply viewer-centered representation. *Journal of Cognitive Neuroscience*, *13*(6), 793–799.

Barsalou, L. W. (1999). Perceptual symbol systems. *The Behavioral and brain sciences*, *22*(4), 577–660.

Basri, R., Costa, L., Geiger, D., & Jacobs, D. (1998, August). Determining the similarity of deformable shapes. *Vision Research*, *38*(15-16), 2365–2385.

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as

an engine of physical scene understanding. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(45), 18327–32.

Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, *2*(1), 1–127.

Biederman, I. (1987). Recognition-by-Components: A theory of human image understanding. *Psychological Review*, *94*(2), 115–147.

Biederman, I. (2007). Recent psychophysical and neural research in shape recognition. In N. Osaka, I. Rentschler, & I. Biederman (Eds.), *Object Recognition, Attention, and Action.*

Biederman, I., & Cooper, E. (1991). Priming contour-deleted images: evidence for intermediate representations in visual object recognition. *Cognitive Psychology*, *23*, 393–419.

Biederman, I., & Gerhardstein, P. (1993). Recognizing depth-rotated objects: evidence and conditions for three-dimensional viewpoint in variance. *Journal of Experimental Psychology: Human Perception and Performance*, *19*(6), 1162–1182.

Biederman, I., & Gerhardstein, P. C. (1995). Viewpoint-dependent mechanisms in visual object recognition: Reply to Tarr and Bulthoff (1995). *Journal of Experimental Psychology: Human Perception and Performance*, *21*(6), 1506–1514.

Bienenstock, E., Geman, S., & Potter, D. (1997). Compositionality, mdl priors, and object recognition. In *Advances in Neural Information Pro-*

cessing Systems (pp. 838–844). MIT Press.

Binford, T. O. (1971). Visual perception by computer. In *Proceedings of IEEE Conference on Systems and Control*. Miami, USA.

Blum, H., & Nagel, R. N. (1978). Shape description using weighted symmetric axis features. *Pattern Recognition*, *10*(3), 167–180.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*(4), 433–436.

Brooks, S. (1998). Markov chain Monte Carlo method and its application. *Journal of the Royal Statistical Society: Series D*, *47*(1), 69–100.

Bulthoff, H. H., & Edelman, S. (1992). Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proceedings of the National Academy of Sciences of the United States of America*, *89*(1), 60–4.

Bulthoff, H. H., Edelman, S. Y., & Tarr, M. J. (1995). How are three-dimensional objects represented in the brain? *Cerebral Cortex*, *5*(3), 247–260.

Byrd, R., Lu, P., Nocedal, J., & Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, *16*(5), 1190–1208.

Cadieu, C. F., Hong, H., Yamins, D. L. K., Pinto, N., Ardila, D., Solomon, E. A., ... DiCarlo, J. J. (2014). Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLoS Comput Biol*, *10*(12), e1003963.

Cadieu, C. F., Kouh, M., Pasupathy, A., Connor, C. E., Riesenhuber, M., & Poggio, T. (2007). A model of V4 shape selectivity and invariance. *Journal of Neurophysiology*, *98*(3), 1733–1750.

Chen, Q., Garcea, F. E., & Mahon, B. Z. (2016). The representation of object-directed action and function knowledge in the human brain. *Cerebral Cortex*, *26*(4), 1609–1618.

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., & Abbeel, P. (2016). InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. *arXiv:1606.03657 [cs, stat]*.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Cohen, Y. E., & Andersen, R. A. (2000). Reaches to sounds encoded in an eye-centered reference frame. *Neuron*, *27*(3), 647–652.

Cohen, Y. E., & Andersen, R. A. (2002). A common reference frame for movement plans in the posterior parietal cortex. *Nature Reviews Neuroscience*, *3*(7), 553–562.

Cooke, T., Jakel, F., Wallraven, C., & Bulthoff, H. H. (2007). Multimodal similarity and categorization of novel, three-dimensional objects. *Neuropsychologia*, *45*(3), 484–95.

Cooke, T., Kannengiesser, S., Wallraven, C., & Bulthoff, H. H. (2006). Object feature validation using visual and haptic similarity ratings. *ACM Transactions on Applied Perception*, *3*(3), 239–261.

Cox, T. F., & Cox, A. A. (2000). *Multidimensional scaling*. Boca Raton,

FL: Chapman & Hall/CRC.

Cutzu, F., & Edelman, S. (1996). Faithful representation of similarities among three-dimensional shapes in human vision. *Proceedings of the National Academy of Sciences of the United States of America*, *93*(21), 12046–50.

Dixon, M., Bub, D. N., & Arguin, M. (1997). The interaction of object form and object meaning in the identification performance of a patient with category-specific visual agnosia. *Cognitive Neuropsychology*, *14*(8), 1085–1130.

Easton, R. D., Srinivas, K., & Greene, A. J. (1997). Do vision and haptics share common representations? Implicit and explicit memory within and between modalities. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *23*(1), 153–163.

Edelman, S. (1997). Computational theories of object recognition. *Trends in Cognitive Sciences*, *1*(8), 296–304.

Edelman, S., & Bulthoff, H. H. (1992). Orientation dependence in the recognition of familiar and novel views of three-dimensional objects. *Vision Research*, *32*(12), 2385–2400.

Edelman, S., Bulthoff, H. H., & Weinshall, D. (1989). *Stimulus familiarity determines recognition strategy for novel 3d objects* (Tech. Rep. No. AI Memo No 1138). Boston, MA: MIT.

Erdogan, G., Chen, Q., Garcea, F. E., Mahon, B. Z., & Jacobs, R. A. (2016). Multisensory part-based representations of objects in human lateral

occipital cortex. *Journal of Cognitive Neuroscience*, *28*(6), 869–881.

Erdogan, G., Yildirim, I., & Jacobs, R. A. (2015). From sensory signals to modality-independent conceptual representations: A probabilistic language of thought approach. *PLoS Comput Biol*, *11*(11), e1004610.

Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. *Nature*, *407*(6804), 630–3.

Feldman, J., & Singh, M. (2006). Bayesian estimation of the shape skeleton. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(47), 18014–9.

Feldman, J., Singh, M., Briscoe, E., Froyen, V., Kim, S., & Wilder, J. (2013). An integrated Bayesian approach to shape representation and perceptual organization. In S. J. Dickinson & Z. Pizlo (Eds.), *Shape Perception in Human and Computer Vision* (pp. 55–70). London: Springer London.

Felzenszwalb, P. F. (2013). A stochastic grammar for natural shapes. In S. J. Dickinson & Z. Pizlo (Eds.), *Shape Perception in Human and Computer Vision* (pp. 299–310). Springer London.

Findlay, J. M., & Gilchrist, I. D. (2003). *Active vision: The psychology of looking and seeing*. New York, NY: Oxford University Press.

Fintzi, A. R., & Mahon, B. Z. (2013). A bimodal tuning curve for spatial frequency across left and right human orbital frontal cortex during object recognition. *Cerebral Cortex*, *24*(5), 1311–1318.

Fodor, J. (1975). *The language of thought*. Cambridge, MA: Harvard Uni-

versity Press.

Foster, D. H., & Gilson, S. J. (2002). Recognizing novel three-dimensional objects by summing signals from parts and views. *Proceedings of the Royal Society B: Biological Sciences*, *269*(1503), 1939–1947.

Freeman, W. T. (1996). The generic viewpoint assumption in a Bayesian framework. In D. C. Knill & W. Richards (Eds.), *Perception As Bayesian Inference* (pp. 365–389). New York, NY, USA: Cambridge University Press.

Freides, D. (1974). Human information processing and sensory modality: cross-modal functions, information complexity, memory, and deficit. *Psychological Bulletin*, *81*(5), 284–310.

Fu, K. S. (1986). A step towards unification of syntactic and statistical pattern recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *8*(3), 398–404.

Gaissert, N., Bulthoff, H. H., & Wallraven, C. (2011). Similarity and categorization: from vision to touch. *Acta Psychologica*, *138*(1), 219–30.

Gaissert, N., & Wallraven, C. (2012). Categorizing natural objects: a comparison of the visual and the haptic modalities. *Experimental Brain Research*, *216*(1), 123–34.

Gaissert, N., Wallraven, C., & Bulthoff, H. (2010). Visual and haptic perceptual spaces show high similarity in humans. *Journal of Vision*, *10*(2), 1–20.

Gauthier, I., James, T. W., Curby, K. M., & Tarr, M. J. (2003). The

influence of conceptual knowledge on visual discrimination. *Cognitive Neuropsychology*, *20*(3-6), 507–523.

Gauthier, I., & Tarr, M. J. (2016). Visual object recognition: Do we (finally) know more now than we did? *Annual Review of Vision Science*, *2*(1), 377–396.

Ghose, T., & Liu, Z. (2013). Generalization between canonical and non-canonical views in object recognition. *Journal of Vision*, *13*(1), 1–1.

Goldstone, R. L., Lippa, Y., & Shiffrin, R. M. (2001). Altering object representations through category learning. *Cognition*, *78*(1), 27–43.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2014). Generative adversarial networks. *arXiv:1406.2661 [cs, stat]*.

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, *32*(1), 108–54.

Goodman, N. D., Tenenbaum, J. B., & Gerstenberg, T. (2015). Concepts in a probabilistic language of thought. In E. Margolis & S. Laurence (Eds.), *The Conceptual Mind: New Directions in the Study of Concepts* (pp. 623–655). Cambridge, MA: MIT Press.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, *82*(4), 711–732.

Grenander, U., & Miller, M. (2007). *Pattern theory: From representation to inference*. New York, NY: Oxford University Press.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences*, *14*(8), 357–64.

Griffiths, T. L., & Ghahramani, Z. (2011). The Indian buffet process: An introduction and review. *Journal of Machine Learning Research*, *12*, 1185–1224.

Griffiths, T. L., Vul, E., & Sanborn, A. N. (2012). Bridging levels of analysis for probabilistic models of cognition. *Current Directions in Psychological Science*, *21*(4), 263–268.

Grill-Spector, K., Kourtzi, Z., & Kanwisher, N. (2001). The lateral occipital complex and its role in object recognition. *Vision Research*, *41*(10-11), 1409–1422.

Grubbs, F. (1950). Sample criteria for testing outlying observations. *The Annals of Mathematical Statistics*, *21*(1), 1–164.

Guggenmos, M., Thoma, V., Cichy, R. M., Haynes, J. D., Sterzer, P., & Richardson-Klavehn, A. (2015). Non-holistic coding of objects in lateral occipital complex with and without attention. *Neuroimage*, *107*, 356–363.

Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*(1), 97–109.

Hayhoe, M., & Ballard, D. (2005). Eye movements in natural behavior. *Trends in Cognitive Sciences*, *9*(4), 188–94.

Hayward, W. G., & Williams, P. (2000). Viewpoint dependence and object

discriminability. *Psychological Science*, *11*(1), 7–12.

Hayworth, K. J., & Biederman, I. (2006). Neural evidence for intermediate representations in object recognition. *Vision Research*, *46*(23), 4024–31.

Hayworth, K. J., Lescroart, M. D., & Biederman, I. (2011). Neural encoding of relative position. *Journal of Experimental Psychology. Human Perception and Performance*, *37*(4), 1032–50.

Hoffman, D., & Richards, W. (1984). Parts of recognition. *Cognition*, *18*, 65–96.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, *160*(1), 106–154.2.

Hummel, J. E., & Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review*, *99*(3), 480.

Jacobs, R. A., & Kruschke, J. K. (2011). Bayesian learning theory applied to human cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(1), 8–21.

James, T. W., Humphrey, G. K., Gati, J. S., Servos, P., Menon, R. S., & Goodale, M. A. (2002). Haptic study of three-dimensional objects activates extrastriate visual areas. *Neuropsychologia*, *40*(10), 1706–1714.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., ... Darrell, T. (2014). Caffe: Convolutional architecture for fast feature

embedding. *arXiv:1408.5093*.

Jones, E., Oliphant, T., Peterson, P., & others. (2001). *Scipy: Open source scientific tools for Python.*

Jordan, M. I., & Rumelhart, D. E. (1992). Forward models: Supervised learning with a distal teacher. *Cognitive Science*, *16*(3), 307–354.

Kass, M., Witkin, A., & Terzopoulos, D. (1988). Snakes: Active contour models. *International Journal of Computer Vision*, *1*(4), 321–331.

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(31), 10687–92.

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annual Review of Psychology*, *55*, 271–304.

Kersten, D., & Yuille, A. (2003). Bayesian models of object perception. *Current Opinion in Neurobiology*, *13*(2), 150–158.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Comput Biol*, *10*(11), e1003915.

Kingma, D. P., & Welling, M. (2014). Auto-encoding variational Bayes. *arXiv:1312.6114 [cs, stat]*.

Knill, D. C., & Richards, W. (Eds.). (1996). *Perception as Bayesian inference.* New York, NY, USA: Cambridge University Press.

Kourtzi, Z., & Connor, C. E. (2011). Neural representations for object perception: Structure, category, and adaptive coding. *Annual Review*

of Neuroscience, *34*(1), 45–67.

Kourtzi, Z., & Kanwisher, N. (2001). Representation of perceived object shape by the human lateral occipital cortex. *Science*, *293*(5534), 1506–1509.

Kriegeskorte, N. (2015). Deep neural networks: A new framework for modelling biological vision and brain information processing. *Annual Reviews of Vision Science*, *1*, 417–446.

Kriegeskorte, N., Kriegeskorte, N., Goebel, R., Goebel, R., Bandettini, P., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(10), 3863–3868.

Krizhevsky, A., Sutskever, I., & Hinton, G. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25* (pp. 1097–1105).

Kruskal, J. (1964). Multidimensional scaling by optimizing goodness of fit to a nonparametric hypothesis. *Psychometrika*, *29*(1), 1–27.

Kulis, B. (2013). Metric learning: A survey. *Foundations and Trends in Machine Learning*, *5*(4), 287–364.

Kulkarni, T., Mansinghka, V., Kohli, P., & Tenenbaum, J. (2014). Inverse graphics with probabilistic CAD models. *arXiv:1407.1339*, 1–10.

Kulkarni, T., Yildirim, I., Kohli, P., Freiwald, W., & Tenenbaum, J. (2014). Deep generative vision as approximate Bayesian computation. In *NIPS 2014 ABC Workshop.*

Kusner, M. J., Paige, B., & Hernandez-Lobato, J. M. (2017). Grammar variational autoencoder. *arXiv:1703.01925 [stat]*.

Lacey, S., Pappas, M., Kreps, A., Lee, K., & Sathian, K. (2009). Perceptual learning of view-independence in visuo-haptic object representations. *Experimental Brain Research*, *198*(2-3), 329–37.

Lacey, S., Peters, A., & Sathian, K. (2007). Cross-modal object recognition is viewpoint-independent. *PLoS One*, *2*(9), e890–e890.

Lacey, S., & Sathian, K. (2011). Multisensory object representation: Insights from studies of vision and touch. In *Progress in Brain Research* (1st ed., Vol. 191, pp. 165–76). Elsevier B.V.

Lacey, S., & Sathian, K. (2014). Visuo-haptic multisensory object recognition, categorization, and representation. *Frontiers in Psychology*, *5*, 730–730.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2016). Building machines that learn and think like people. *arXiv:1604.00289 [cs, stat]*.

Lawson, R. (2009). A comparison of the effects of depth rotation on visual and haptic three-dimensional object recognition. *Journal of Experimental Psychology. Human Perception and Performance*, *35*(4), 911–930.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Lee, M. D., & Pope, K. J. (2003). Avoiding the dangers of averaging across subjects when using multidimensional scaling. *Journal of Mathematical*

Psychology, *47*(1), 32–46.

Lehky, S., & Tanaka, K. (2016). Neural representation for object recognition in inferetemporal cortex. *Current Opinion in Neurobiology*, *37*, 23–35.

Ling, H., & Jacobs, D. W. (2007). Shape classification using the inner-distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(2), 286–299.

Liu, J. S. (2004). *Monte Carlo strategies in scientific computing*. New York, NY: Springer New York.

Liu, Z. (1996). Viewpoint dependency in object representation and recognition. *Spatial Vision*, *9*(4), 491–521.

Liu, Z., Kersten, D., & Knill, D. C. (1999). Dissociating stimulus information from internal representation—a case study in object recognition. *Vision Research*, *39*(3), 603–612.

Logothetis, N. K., Pauls, J., & Poggio, T. (1995). Shape representation in the inferior temporal cortex of monkeys. *Current Biology*, *5*(5), 552–563.

Longuet-Higgins, H. C. (1990). Recognizing three dimensions. *Nature*, *343*(6255), 214–215.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. Cambridge, Massachusetts: MIT Press.

Marr, D., & Nishihara, H. K. (1978). Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London. Series B*, *200*(1140), 269–94.

Marsolek, C. J. (1999). Dissociable neural subsystems underlie abstract and specific object recognition. *Psychological Science*, *10*(2), 111–118.

McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S., & Smith, L. B. (2010). Letting structure emerge: connectionist and dynamical systems approaches to cognition. *Trends in Cognitive Sciences*, *14*(8), 348–56.

McClelland, J. L., & Patterson, K. (2002a). Rules or connections in past-tense inflections: what does the evidence rule out? *Trends in Cognitive Sciences*, *6*(11), 465–472.

McClelland, J. L., & Patterson, K. (2002b). 'Words or Rules' cannot exploit the regularity in exceptions. *Trends in Cognitive Sciences*, *6*(11), 464–465.

Mehta, P., & Schwab, D. J. (2014). An exact mapping between the variational renormalization group and deep learning. *arXiv:1410.3831*.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, *21*(6), 1087.

Miller, B. Y. A. T., & Allen, P. K. (2004). Graspit! a versatile simulator for robotic grasping. *IEEE Robotics Automation Magazine*, *11*(December), 110–122.

Minsky, M. L., & Selfridge, O. G. (1961). Learning in random nets. In *Fourth London Symposium on Information Theory.* London: Butterworth Ltd.

Mnih, A., & Rezende, D. J. (2016). Variational inference for Monte Carlo objectives. *arXiv:1602.06725 [cs, stat]*.

Naumer, M. J., Ratz, L., Yalachkov, Y., Polony, A., Doehrmann, O., Van De Ven, V., ... Hein, G. (2010). Visuohaptic convergence in a cortico-cerebellar network. *European Journal of Neuroscience*, *31*(10), 1730–1736.

Navarro, D. J., & Griffiths, T. L. (2008, November). Latent features in similarity judgments: A nonparametric bayesian approach. *Neural computation*, *20*(11), 2597–628.

Newell, F. N. (2010). Visuo-haptic perception of objects and scenes. In J. Kaiser & M. J. Naumer (Eds.), *Multisensory Object Perception in the Primate Brain* (pp. 251–271). New York, NY: Springer New York.

Newell, F. N., & Ernst, M. O. (2001). Viewpoint dependence in visual and haptic object recognition. *Psychological Science*, *12*(1), 37–42.

Nguyen, A., Yosinski, J., & Clune, J. (2014). Deep neural networks are easily fooled: high confidence predictions for unrecognizable images. *arXiv:1412.1897*.

Norman, J. F., Norman, H. F., Clayton, A. M., Lianekhammy, J., & Zielke, G. (2004). The visual and haptic perception of natural object shape. *Perception & Psychophysics*, *66*(2), 342–51.

Op de Beeck, H. P., Wagemans, J., & Vogels, R. (2008). The representation of perceived shape similarity and its role for category learning in monkeys: a modeling study. *Vision Research*, *48*(4), 598–610.

Palmer, S. E. (1999). *Vision science: Photons to phenomenology*. Cambridge, MA: The MIT Press.

Palmeri, T. J., & Gauthier, I. (2004). Visual object understanding. *Nature Reviews Neuroscience*, *5*(4), 291–303.

Pascual-Leone, A., & Hamilton, R. (2001). The metamodal organization of the brain. *Progress in Brain Research*, *134*, 427–445.

Patel, A. B., Nguyen, T., & Baraniuk, R. G. (2015). A probabilistic theory of deep learning. *arXiv:1504.00641*.

Peelen, M. V., He, C., Han, Z., Caramazza, A., & Bi, Y. (2014). Nonvisual and visual object shape representations in occipitotemporal cortex: evidence from congenitally blind and sighted adults. *The Journal of Neuroscience*, *34*(1), 163–70.

Peirce, J. W. (2007). PsychoPy—Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1-2), 8–13.

Peissig, J. J., & Tarr, M. J. (2007). Visual object recognition: Do we know more now than we did 20 years ago? *Annual Review of Psychology*, *58*, 75–96.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*(4), 437–442.

Piantadosi, S. T., Tenenbaum, J. B., & Goodman, N. D. (2012). Bootstrapping in a language of thought: A formal model of numerical concept learning. *Cognition*, *123*(2), 199–217.

Pinker, S. (1994). *The language instinct*. New York, NY: Harper Perennial

Modern Classics.

Pinker, S., & Ullman, M. (2002a). Combination and structure, not gradedness, is the issue. *Trends in Cognitive Sciences*, *6*(11), 472–474.

Pinker, S., & Ullman, M. T. (2002b). The past and future of the past tense. *Trends in Cognitive Sciences*, *6*(11), 456–463.

Poggio, T., & Edelman, S. (1990). A network that learns to recognize three-dimensional objects. *Nature*, *343*(6255), 263–6.

Pouget, A., Beck, J. M., Ma, W. J., & Latham, P. E. (2013). Probabilistic brains: Knowns and unknowns. *Nature Neuroscience*, *16*, 1170–8.

Pouget, A., Ducom, J. C., Torri, J., & Bavelier, D. (2002). Multisensory spatial representations in eye-centered coordinates for reaching. *Cognition*, *83*(1), B1–B11.

Qi, C. R., Su, H., Niessner, M., Dai, A., Yan, M., & Guibas, L. J. (2016). Volumetric and multi-view CNNs for object classification on 3d data. *arXiv:1604.03265 [cs]*.

Quiroga, R. Q. (2012). Concept cells: The building blocks of declarative memory functions. *Nature Reviews Neuroscience*, *13*(8), 587–97.

Quiroga, R. Q., Kraskov, A., Koch, C., & Fried, I. (2009, August). Explicit encoding of multimodal percepts by single neurons in the human brain. *Current Biology*, *19*(15), 1308–13.

Reales, J. M., & Ballesteros, S. (1999). Implicit and explicit memory for visual and haptic objects: Cross-modal priming depends on structural descriptions. *Journal of Experimental Psychology: Learning, Memory,*

and Cognition, *25*(3), 644–663.

Rezende, D. J., Eslami, S. M. A., Mohamed, S., Battaglia, P., Jaderberg, M., & Heess, N. (2016). Unsupervised learning of 3d structure from images. *arXiv:1607.00662 [cs, stat]*.

Riddoch, M. J., & Humphreys, G. W. (1987). A case of integrative visual agnosia. *Brain: A Journal of Neurology*, *110*(6), 1431–1462.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*(11), 1019–25.

Riesenhuber, M., & Poggio, T. (2000). Models of object recognition. *Nature Neuroscience*, *3*, 1199–204.

Riesenhuber, M., & Poggio, T. (2002). Neural mechanisms of object recognition. *Current Opinion in Neurobiology*, *12*(2), 162–8.

Rock, I., & DiVita, J. (1987). A case of viewer-centered object perception. *Cognitive Psychology*, *19*(2), 280–293.

Saiki, J., & Hummel, J. E. (1998). Connectedness and the Integration of Parts with Relations in Shape Perception. *Journal of Experimental Psychology: Human Perception and Performance*, *24*(1), 227–51.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*, *117*(4), 1144–67.

Santello, M., Flanders, M., & Soechting, J. F. (1998). Postural hand synergies for tool use. *Journal of Neuroscience*, *18*(23), 10105–10115.

Schlicht, E. J., & Schrater, P. R. (2007). Impact of coordinate transformation

uncertainty on human sensorimotor control. *Journal of Neurophysiology*, *97*(6), 4203–4214.

Schultz, M., & Joachims, T. (2003). Learning a distance metric from relative comparisons. In *Advances in Neural Information Processing Systems.*

Schwarzbach, J. (2011). A simple framework (ASF) for behavioral and neuroimaging experiments based on the psychophysics toolbox for MATLAB. *Behavior Research Methods*, *43*(4), 1194–1201.

Serre, T., Oliva, A., & Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(15), 6424–9.

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., & Poggio, T. (2007). Robust object recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(3), 411–26.

Shepard, R. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika*, *27*(2), 125–140.

Siddharth, N., Paige, B., Desmaison, A., Van de Meent, J.-W., Wood, F., Goodman, N. D., ... Torr, P. H. S. (2016). Inducing interpretable representations with variational autoencoders. *arXiv:1611.07492 [cs, stat]*.

Snow, J. C., Goodale, M. A., & Culham, J. C. (2015). Preserved haptic shape processing after bilateral LOC lesions. *Journal of Neuroscience*, *35*(40), 13745–13760.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... Rabinovich, A. (2014). Going deeper with convolutions. *arXiv:1409.4842*.

Szegedy, C., Zaremba, W., & Sutskever, I. (2013). Intriguing properties of neural networks. *arXiv:1312.61*.

Tal, N., & Amedi, A. (2009). Multisensory visual-tactile object related network in humans: insights gained using a novel crossmodal adaptation approach. *Experimental Brain Research*, *198*(2-3), 165–182.

Talairach, J., & Tournoux, P. (1988). *Co-planar stereotaxic atlas of the human brain.* New York, NY: Thieme.

Talton, J., Yang, L., Kumar, R., Lim, M., Goodman, N., & Mech, R. (2012). Learning design patterns with bayesian grammar induction. In *Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology - UIST '12* (pp. 63–74). ACM Press.

Tarr, M. J. (2003). Visual object recognition: Can a single mechanism suffice? In M. Peterson & G. Rhodes (Eds.), *Perception of Faces, Objects, and Scenes: Analytic and Holistic Processes* (pp. 177–211). New York, NY: Oxford University Press.

Tarr, M. J., & Bulthoff, H. H. (1995). Is human object recognition better described by geon structural descriptions or by multiple views? Comment on Biederman and Gerhardstein (1993). *Journal of Experimental Psychology: Human Perception and Performance*, *21*(6), 1494–1505.

Tarr, M. J., Williams, P., Hayward, W. G., & Gauthier, I. (1998). Three-dimensional object recognition is viewpoint dependent. *Nature Neuro-*

science, *1*(4), 275–277.

Taylor, K. I., Moss, H. E., Stamatakis, E. A., & Tyler, L. K. (2006). Binding crossmodal object features in perirhinal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(21), 8239–8244.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279–85.

Thakur, P. H., Bastian, A. J., & Hsiao, S. S. (2008). Multidigit movement synergies of the human hand in an unconstrained haptic exploration task. *Journal of Neuroscience*, *28*(6), 1271–1281.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *The Annals of Statistics*, *22*(4), 1701–1728.

Tjan, B. S., & Legge, G. E. (1998). The viewpoint complexity of an object-recognition task. *Vision Research*, *38*(15-16), 2335–2350.

Tu, Z., Chen, X., Yuille, A. L., & Zhu, S.-c. (2005). Image parsing: Unifying segmentation, detection, and recognition. *International Journal of Computer Vision*, *63*(2), 113–140.

Tversky, B. (1989). Parts, partonomies, and taxonomies. *Developmental Psychology*, *25*(6), 983–995.

Tversky, B., & Hemenway, K. (1984). Objects, parts, and categories. *Journal of Experimental Psychology: General*, *113*(2).

Ullman, S. (1989). Aligning pictorial descriptions: An approach to object

recognition. *Cognition*, *32*(3), 193–254.

Ullman, S., & Basri, R. (1991). Recognition by linear combinations of models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*(10), 992–1006.

Ullman, T. D., Goodman, N. D., & Tenenbaum, J. B. (2012). Theory learning as stochastic search in the language of thought. *Cognitive Development*, *27*(4), 455–480.

Von Helmholtz, H. (1867). *Handbuch der physiologischen optik.* Leipzig, Germany: Leopold Voss.

Wallraven, C., Bulthoff, H. H., Waterkamp, S., van Dam, L., & Gaissert, N. (2014). The eyes grasp, the hands see: Metric category knowledge transfers between vision and touch. *Psychonomic Bulletin & Review*, *21*(4), 976–85.

Wiseman, S., MacLeod, C. M., & Lootsteen, P. J. (1985). Picture recognition improves with subsequent verbal information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*(3), 588–595.

Wolpert, D. M., & Flanagan, J. R. (2009). Forward models. In T. Bayne, A. Cleermans, & P. Wilken (Eds.), *The Oxford Companion to Consciousness* (pp. 295–296). New York: Oxford University Press.

Wolpert, D. M., & Kawato, M. (1998). Multiple paired forward and inverse models for motor control. *Neural Networks*, *11*(7-8), 1317–1329.

Wu, J., Zhang, C., Xue, T., Freeman, W. T., & Tenenbaum, J. B. (2016). Learning a probabilistic latent space of object shapes via 3d generative-

adversarial modeling. *arXiv:1610.07584 [cs]*.

Yalachkov, Y., Kaiser, J., Doehrmann, O., & Naumer, M. J. (2015). Enhanced visuo-haptic integration for the non-dominant hand. *Brain Research*, *1614*, 75–85.

Yildirim, I., & Jacobs, R. A. (2012). A rational analysis of the acquisition of multisensory representations. *Cognitive Science*, *36*(2), 305–32.

Yildirim, I., & Jacobs, R. A. (2013). Transfer of object category knowledge across visual and haptic modalities: Experimental and computational studies. *Cognition*, *126*, 135–148.

Yildirim, I., & Jacobs, R. A. (2015). Learning multisensory representations for auditory-visual transfer of sequence category knowledge: A probabilistic language of thought approach. *Psychonomic Bulletin and Review*, *22*(3), 673–686.

Yildirim, I., Kulkarni, T. D., Freiwald, W. A., & Tenenbaum, J. B. (2015). Efficient and robust analysis-by-synthesis in vision: A computational framework, behavioral tests, and modeling neuronal representations. In *Proceedings of the Thirty-Seventh Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: Analysis by synthesis? *Trends in Cognitive Sciences*, *10*(7), 301–8.

Yuille, A., & Mottaghi, R. (2016). Complexity of representation and inference in compositional models with part sharing. *Journal of Machine Learning Research*, *17*(11), 1–28.

Zhang, D., & Lu, G. (2004). Review of shape representation and description techniques. *Pattern Recognition*, *37*(1), 1–19.

Zhang, K., & Shasha, D. (1989). Simple fast algorithms for the editing distance between trees and related problems. *SIAM Journal on Computing*, *18*(6), 1245–1262.

Zhu, L., Chen, Y., & Yuille, A. (2007). Unsupervised learning of a probabilistic grammar for object detection and parsing. In *Advances in Neural Information Processing Systems 19* (pp. 1617–1624).

Zhu, L., Chen, Y., & Yuille, A. (2009). Unsupervised learning of probabilistic grammar-Markov models for object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(1), 114–128.

Zhu, S.-C., & Mumford, D. (2006). A stochastic grammar of images. *Foundations and Trends in Computer Graphics and Vision*, *2*(4), 259–362.