From Perception to Conception: Learning Multisensory Representations

by

Ilker Yildirim

Submitted in Partial Fulfillment

of the

Requirements for the Degree

Doctor of Philosophy

Supervised by

Robert A. Jacobs
Daniel Gildea

Departments of Brain & Cognitive Sciences
and Computer Science
Arts, Sciences and Engineering
School of Arts and Sciences

University of Rochester
Rochester, New York

2014
Biographical Sketch

The author was born in 1985 in Erzurum, an eastern city of Turkey. He moved to Istanbul in 2003 to study computer science at Bogazici University, Istanbul. There he received Bachelor of Science (2007) and Master of Science (2009) degrees, both in Computer Science. He moved to Rochester in 2009 to study computational principles of human cognition at the doctoral level at the University of Rochester, Rochester, NY. There he received a Master of Arts degree in Brain & Cognitive Sciences (2011), and currently he is studying for a joint PhD degree in Brain & Cognitive Sciences and Computer Science. The main body of his doctoral studies was supervised by Robert Jacobs, alongside his collaborations in language processing projects with faculty members T. Florian Jaeger and Michael K. Tanenhaus. The following list of manuscripts report on his doctoral studies.


Acknowledgments

I would like to thank my advisor, Robbie Jacobs. I cannot be grateful enough for having the opportunity to work with him the last five years. He needed to stay patient for a good part of the first two years into my grad school while subtly brainwashing me to turn into a Cognitive Scientist from my Computer Science background. I believe we both benefited from this journey, which gave the fruit of a “language” oriented view of aspects of perception. Here are a few attributes of Robbie that I should hold on to wherever I go: always an open door, an open mind, letting the students freely pursue their interests, rational encouragement, and realistic feedback.

I would like to thank to my committee chair, Laurel Carney, and members, Dan Gildea, Greg DeAngelis, and Krystel Huxlin. Their feedback and encouragement at our committee meetings guided this thesis. But outside of these meetings, our separate conversations with Dan Gildea, Greg DeAngelis, and Krystel Huxlin on different occasions for different projects have been helpful.

I benefited much from my collaboration with Florian Jaeger, Mike Tanenhaus, and Judith Degen. I enjoyed working in Florian’s lab, which is partly because his lab produces the cutting edge experimental and computational research on human language processing. It was a fruitful experience to extend his framework of adaptation and variability in language processing to semantics, particularly to understanding of quantifier representations, which in fact started with my going to his office and suggesting a language production model of everything. Yes. Judith
has been very helpful in my getting into this project. She did not mind to help in a hands on manner in many aspects of this project (e.g., materials development, interpretations of the results, writing of the manuscript). For Mike, I will just note that it is a recommended experience to work with him.

I benefited much from my collaboration with Goker Erdogan, a fellow Turkish grad student like me in Robbie’s lab, on modeling people’s cross-modal similarity ratings. We shall praise Python over socket connections whenever we get a chance.

I would like to thank many past and present members of Robbie’s lab such as Matt Overlan, Emin Orhan, and Chris Sims. They have been pretty good at tolerating me in whatever I wanted to talk to them about, including cylinders, Fribbles, 3D representations, Bayesian everything. To be fair, I listened a lot about Visual Short Term Memory in return.

I would like to thank Dave Knill, Jurgen Drugowitsch, and Alex Pouget for helpful discussions on Bayesian nonparametrics and multisensory perception.

I would like to thank Steve Piantadosi. Our compositional conversations were perhaps compositionally-intriguing.

I would like to thank Dave Kleinschmidt who arrived later to the program, but taught me much about phonemes, mathematics, and all things American (not in the most canonical sense of it though). Thinking again, I should have initiated more of the food-chemistry discussions with Esteban Buz. I would like to thank all the other past and present grad students and postdocs at the Department of Brain & Cognitive Sciences, from whom I learned much and with whom I shared much.

I would like to thank Kathy Corser, Jennifer Gillis, and Chris Dambra for their approachable and friendly personalities and taking care of secretarial and administrative load. I would also like to thank Chris Freemesser for his invaluable technological support, which often put my mind at ease with the issues I had with my computer. I would also like to thank Ted Twietmeyer, former technician of
the Center for Visual Sciences, for designing and building an audiovisual panel for localized auditory and visual stimulation.

And my family. There is really no way I can thank them. My mother and father have been extremely supportive of all the things I have been doing around in Rochester the last five years. Over our telephone conversations and on my occasional travels to Turkey, they have become pretty knowledgeable of Cognitive Science, and that is just them being supportive again. It has always been motivational whenever they recognized even the slightest of the progress I had. And that is not only during the last five years of my life when we had the Atlantic Ocean in between, but throughout my whole life.

I would like to thank Meltem Yalcin. This last year has been incredible with her in my life. In just one aspect, she brought tranquility to my life. Thank you. I would like to thank my parents-in-law for being supportive of us, and making everything easier for us.
Abstract

The “metamodal” perspective on neural organization states that different neural regions are specialized for performing different functions on sensory data regardless of the modality through which those data were acquired (Pascual-Leone & Hamilton, 2001). If so, this raises an interesting question: Given that we learn about our environments through our senses, how do we acquire abstract concepts that are modality-independent? In other words, how does perception interface with conception?

The “grounded cognition” approach emphasizes that abstract representations must be derived from perceptual signals and representations (Barsalou, 2008). However, it leaves open the question of how these representations are acquired. Moreover, it claims that these representations are never fully modality-independent.

In this thesis, we cast this problem of perception-to-conception interface in the domain of multisensory perception. We study how perception gives rise to abstract concepts by studying the acquisition of multisensory representations. We describe experimental work demonstrating that people exhibit cross-modal transfer of perceptual knowledge, indicating that they have modality-independent, conceptual representations. We also describe our computational work demonstrating that these representations can be acquired from sensory data.
Contributors and Funding Sources

The work described in this thesis are collaborations between me and my advisor Robbie Jacobs. My committee members, Laurel Carney of Biomedical Engineering, Dan Gildea of Computer Science, Greg DeAngelis of Brain & Cognitive Sciences, and Krystel Huxlin of Ophthalmology supervised this work. Chapters 2 and 4 are published in the journals Cognition and Cognitive Science at the present time (see the references in the Biographical Sketch). Chapter 3 is under revision after getting favorable reviews by the journal Psychonomic Bulletin & Review, which we are finishing up our revision of the manuscript. I am the first author in all the studies reported in this thesis. I would like to thank M. Tarr for making the 3-D object files for Fribbles available on his web pages that we used in Chapter 2. I would like to thank J. Movellan for making the Tulips1 data set available on the Web, and to T. Cooke, F. Jäkel, C. Wallraven, and H. Bülthoff for sharing their visual-haptic experimental data with us that we used in Chapter 4.

The research in this thesis is funded by the National Science Foundation (DRL-0817250) and the Air Force Office of Scientific Research (FA9550-12-1-0303).
Table of Contents

Biographical Sketch ii

Acknowledgments iv

Abstract vii

Contributors and Funding Sources viii

List of Tables xi

List of Figures xii

1 Introduction 1

2 Transfer of Object Category Knowledge Across Visual and Haptic Modalities: Experimental and Computational Studies 7

2.1 Introduction 7

2.2 Previous Research on Visual-Haptic Object Perception 9

2.3 Fribbles and the See & Grasp Data Set 11

2.4 Experiment 14

2.5 Preliminary Remarks Regarding the MVH Model 19
List of Tables

2.1 Summary of the variables in the MVH model. . . . . . . . . . . . 25
# List of Figures

2.1 The top row shows computer-generated images of Fribbles which are rendered using the Fribbles’ 3-D object models. The bottom row shows photographs of the physical objects corresponding to these same Fribbles which were fabricated via a 3-D printing process using the same 3-D object models. Pairs of columns illustrate exemplars from different categories (e.g., columns 1-2 illustrate exemplars from Category A). ................................................................. 12

2.2 Learning curves for Groups V-H (mid gray), H-V (dark gray), and Vs-H (light gray) during training. The horizontal axis plots the training block number, and the vertical axis plots the average percent correct. Error bars show the standard errors of the means. ............................................. 17

2.3 (Left) Performances of Groups V-H, H-V, and Vs-H on the final training block. Dots indicate the performances of individual participants. (Right) Performances during testing. ......................................................... 18
2.4 Schematic illustrating steps 3 and 4 in the generative process for the MVH model. (Left) Tree graph characterizing a prototype in terms of its parts (nodes) and spatial configurations among parts (edges). Illustration to the right of the tree shows that the bottom-right edge represents the connection between a child part and its parent part. A subset of the “docking locations” on the parent part are illustrated with black and white dots. The child part is “docked” on a docking location at the top of the parent part. (Right) Haptic and visual features of a prototype are obtained through the use of haptic and visual forward models.

2.5 Our inference algorithm consists mostly of Metropolis-Hastings (MH) steps. In each iteration $i$, a proposal prototype is drawn from the generative process. Then the visual forward model is used to obtain the visual features of the proposed prototype. These visual features are compared against the visual features of the training exemplars to compute a log-likelihood value. This value is used to evaluate the proposal with respect to the MH acceptance function and the current state of the chain. Details of our inference algorithm can be found in the Appendix.

2.6 Four data items used in our simulations, one from each category of Fribbles. Column 1 of each row shows a Fribble. Columns 2-4 show binary images of the Fribble from top, front, and right viewpoints. Columns 5-6 show the simulated hand grasping the Fribble, once from the Fribble’s front and once from its rear.

2.7 (Top) 3 exemplars from each category. (Bottom) The multisensory prototypes learned by the model, visually rendered at a canonical projection.
2.8 Haptic features for test items and prototypes in a two-dimensional space. .......................... 34

3.1 A. Photos of the audio-visual environment. In the left photo, the speakers, LEDs, and electrical hardware are visible. In the right photo, a scrim conceals the environment, meaning that the speakers, unlit LEDs, and electrical hardware are not visible. B. A schematic of the 7 locations used in our experimental stimuli, and the speaker and LED at each location. .............................. 42

3.2 Learning curves during training for Groups V-A and A-V. .... 46

3.3 (Left) Average performances on the final training block (last 36 trials of training) and on the test block for Groups V-A and A-V (error bars indicate standard errors of the means). (Right) Groups’ final training block and test performances when trials are sorted by the category of the sequence observed on the trial. ............. 47

3.4 Sample programs for characterizing sequence categories. .... 50

3.5 PCFG for multisensory representations of sequence categories. . 53

3.6 First and second rows show posterior probabilities based upon the category-level simulations. The horizontal axis of each graph gives a program identification number, and the vertical axis gives the posterior probability of a program in the posterior sample. The entropy ($H$) of each posterior distribution is shown in each graph. The bottom row shows the program with the highest posterior probability (i.e., the MAP estimate) for each category when samples are combined across all participant-level simulations. The model shows modality invariance as evidenced by the fact that MAP estimates were identical for simulations of visual and auditory training groups. 58
3.7 Modeling results presented in the same format as our experimental results (Figure 3.3). (Left) Average performances of the participant-level simulations on the training and test exemplars for simulations corresponding to Groups V-A and A-V (error bars indicate standard errors of the means). (Right) Participant-level simulations’ training and test performances when trials are sorted by the category of the sequence observed on the trial.

3.8 Results from the model in which multisensory MAP representations are evaluated based on a likelihood score, not a posterior score.

4.1 (Left) A coarse schematic of the multisensory perception model. (Right) A Bayesian network representation of the multisensory perception model.

4.2 The performance of the multisensory perception model on the first synthetic data set. The horizontal axis plots the category from which a test item’s “visual” features were sampled, and the vertical axis plots the probability that the model classified the test item as “sad”. Data points indicated by diamonds correspond to the first test condition (auditory features were sampled from a “happy” category), points indicated by circles correspond to the second test condition (auditory features were sampled from the neutral category), and points indicated by squares correspond to the third condition (auditory features were sampled from a “sad” category).

4.3 The distributions of the numbers of latent features in the vision-only (left), audition-only (middle), and multisensory (right) perception models.
4.4 Categorization performances of the vision-only, audition-only, multisensory, and mixed models on the training set (left) and on the test set (right). The horizontal axis of each graph gives the model, and the vertical axis plots the percent of data items correctly classified (error bars indicate the standard deviations of these percents across iterations of an MCMC chain).

4.5 (Left) Categorization performances of the multisensory and vision-only models on the training (left-side) and test (right-side) data items when visual features were the only inputs to these models. The horizontal axis labels the model, and the vertical axis plots the proportion of data items a model categorized correctly (error bars plot the standard deviations of these proportions across iterations of an MCMC chain). (Right) Categorization performances of the multisensory and audition-only models on the training and test data items when auditory features were the only inputs to these models.

4.6 Graphs in the top row demonstrate that when presented with auditory features of a test item corresponding to one of the digits, the multisensory perception model’s distribution of visual representations was tightly peaked at a representation corresponding to the same digit. Graphs in the bottom row show analogous results for distributions of auditory representations given test items’ visual features.
4.7 Each panel illustrates the 25 objects used in the experiment of Cooke et al. (2007). Objects vary systematically in texture properties along the horizontal axis, and vary in shape properties along the vertical axis. The four panels correspond to the four models that were implemented. Each panel illustrates properties of the latent features acquired by its corresponding model. See text for further explanation. (Figure of Cooke et al., 2007, adapted with permission from Elsevier.)
1 Introduction

The “metamodal” perspective on neural organization states that different neural regions are specialized for performing different functions on sensory data regardless of the modality through which those data were acquired (Pascual-Leone & Hamilton, 2001). If so, this raises an interesting question: Given that we learn about our environments through our senses, how do we acquire abstract concepts that are modality-independent? In other words, how does perception interface with conception?

The “grounded cognition” approach emphasizes that abstract representations must be derived from perceptual signals and representations (Barsalou, 2008). However, it leaves open the question of how these representations are acquired. Moreover, it claims that these representations are never fully modality-independent.

In this thesis, we will cast this problem of perception-to-conception interface in the domain of multisensory perception. We will study how perception gives rise to abstract concepts by studying the acquisition of multisensory representations. We will describe experimental work demonstrating that people exhibit cross-modal transfer of perceptual knowledge, indicating that they have modality-independent, conceptual representations. We will also describe our computational work demonstrating that these representations can be acquired from sensory data.
But before, what is a multisensory representation? Generally speaking, accurately characterizing the nature of mental representations has been elusive (e.g., Pylyshyn, 2003; Kosslyn, Thompson, & Ganis, 2006). In fact, it is unclear as to what term to use exactly when referring to multisensory representations. Different authors used the terms amodal representations, modality-invariant representations, and multisensory representations to refer to subtly different things, leaving the definitions as well as the differences largely vague. Addressing the differences between these terms is outside of the scope of this thesis. Instead, we will work with an operational definition: Multisensory representations are representations that are modality-independent and are accessible to more than one sensory modality. Such representations could underlie some of the fundamental cognitive repertoire such as object perception, event recognition, etc.

Are there multisensory representations in the brain? There is converging neural, behavioral, and computational evidence accumulated mostly in the recent years, which collectively indicates that the answer to this question is a yes. Perhaps most strikingly, Quiroga (2012) argued that the brain contains “concept” cells which are involved in the representation of individual people or objects regardless of the modality used to sense those people or objects. For example, when recording in the human medial temporal lobe, he and his colleagues reported a neuron that selectively responded when a person viewed images of the television host Oprah Winfrey, viewed her written name, or heard her spoken name (Quiroga, Kraskov, Koch, & Fried 2009). These and similar findings indicate that our brains encode abstract representations that are amodal or multisensory in that they are activated by perceptual input spanning multiple modalities in a manner that is modality-invariant.

What might be the nature of these modality-invariant representations? In this

\[1\]We hope that the computational study of multisensory perception such as this thesis will lead to more theoretically grounded definitions of each of these terms.
thesis, we seek an answer to this question under the hypothesis that people extract the intrinsic, modality-independent properties of objects and events, and represent these properties in multisensory representations. To this end, we introduce a computational framework that explores possible forms of multisensory representations by acquiring these conceptual, abstract representations on the basis of sensory data. Our key idea is that modality-independent, conceptual representations can be inferred from sensory data by a process that inverts predictive models, also known as forward models.

Chapter 2 studies people’s abilities to transfer object category knowledge across visual and haptic domains. If a person learns to categorize objects based on inputs from one sensory modality, can the person categorize these same objects when the objects are perceived through another modality? Can the person categorize novel objects from the same categories when these objects are, again, perceived through another modality? This chapter makes three contributions. First, by fabricating Fribbles (3-D, multi-part objects with a categorical structure), we developed visual-haptic stimuli that are highly complex and realistic, and thus more ecologically valid than objects that are typically used in haptic or visual-haptic experiments. Based on these stimuli, we developed the See & Grasp data set, a data set containing both visual and haptic features of the Fribbles, and are making this data set freely available on the world wide web. Second, complementary to previous research such as studies asking if people transfer knowledge of object identity across visual and haptic domains, we conducted an experiment evaluating whether people transfer object category knowledge across these domains. Our data clearly indicate that we do. Third, we developed a computational model that learns multisensory representations of prototypical 3-D shape. Similar to previous work, the model uses shape primitives to represent parts, and spatial relations among primitives to represent multi-part objects. However, it is distinct in its use of a Bayesian inference algorithm allowing it to acquire multi-
sensory representations, and sensory-specific forward models allowing it to predict visual or haptic features from multisensory representations. The model provides an excellent qualitative account of our experimental data, thereby illustrating the potential importance of multisensory representations and sensory-specific forward models to multisensory perception.

Chapter 3 takes on the learning and transfer of sequence category knowledge between vision and audition. If a person is trained to recognize or categorize objects or events using one sensory modality, the person can often recognize or categorize those same (or similar) objects and events via a novel modality. This phenomenon is an instance of cross-modal transfer of knowledge. We conducted an experiment evaluating whether people transfer sequence category knowledge across auditory and visual domains. Our experimental data clearly indicate that we do. We also developed a computational model quantitatively accounting for our experimental results. Our model formalizes multisensory representations as symbolic “computer programs” and uses Bayesian inference to learn these representations. The model demonstrates how the acquisition and use of amodal, multisensory representations can underlie cross-modal transfer of knowledge. The model also accounts for subjects’ experimental performances.

Chapter 4 takes a step back and evaluates the acquisition of multisensory representations more broadly. How do people learn multisensory, or amodal, representations, and what consequences do these representations have for perceptual performance? We address this question by taking a statistical approach to the problem of multisensory representations in the general case. This analysis extends the Indian Buffet Process (IBP), a Bayesian nonparametric approach, such that the model acquires latent multisensory features that optimally explain the unisensory features arising in individual sensory modalities. Moreover, differently from our previous models, this model also learns the sensory-specific forward models, albeit in a tractably simpler linear form. The model qualitatively accounts for
several important aspects of multisensory perception: (i) it integrates information from multiple sensory sources in such a way that leads to superior performances in, for example, categorization tasks; (ii) its performances suggest that multisensory training leads to better learning than unisensory training, even when testing is conducted in unisensory conditions; (iii) its multisensory representations are modality invariant; and (iv) it predicts “missing” sensory representations in modalities when the input to those modalities is absent. Analysis in this chapter indicates that all of these aspects emerge as part of the optimal solution to the problem of learning to represent complex multisensory environments.

Chapter 5 provides a summary of this thesis while discussing its implications for the field. It ends with pointing to new exciting research avenues linking perception and conception in general.

There are three common themes encompassing this thesis. First, the computational and the experimental studies described in the Chapters 2 through 4 support the Multisensory Hypothesis, which states that people automatically extract and represent objects’ and events’ intrinsic properties, and use these properties to process and understand the same (and similar) objects and events when they are perceived through novel sensory modalities. Indeed, our computational framework—consisting of multisensory representations plus sensory-specific forward models—shows how such representations could be acquired and be useful for perception.

Second, in addition to contributing to our understanding of multisensory perception and the acquisition of abstract concepts, our modeling work (Chapters 2 and 3) also contributes to an emerging perspective—referred to as the probabilistic “language of thought” (pLOT) perspective (Goodman, Tenenbaum, Feldman, & Griffiths, 2008; Piantadosi, Tenenbaum, & Goodman, 2012)—attempting to combine the advantages of probabilistic and symbolic modeling approaches. It shows how such inherently amodal representations can be acquired on the ba-
sis of sensory data, thereby extending the pLOT framework beyond higher-level cognition to the realm of perception.

Third, this thesis provides computational level analysis (Marr, 1982) of multisensory perception and cross-modal transfer of knowledge. That is, our models specify a computational goal (e.g., cross-modal transfer of categorical knowledge), and then describe a computational model to achieve this goal optimally under the assumptions of the model for the given experimental task. The computational level analysis aspect of our work is particularly pronounced in Chapter 4, where we show that interesting aspects of human behavior can emerge from the optimal solution to the problem of multisensory perception. Importantly and not mutually exclusively, we also believe that the models described in this thesis could be starting points to identifying underlying process models. Particularly, we believe that the proposed representational formalisms (e.g., compositional “language”-like representations) should be investigated further as the actual data structures over which psychological processes are carried out in the brain.

Overall, this thesis proposes a computational framework to tackle the problem of interfacing perception and conception, by showing how to acquire abstract or conceptual representations on the basis of sensory input. The computational framework consists of two components: multisensory representations (e.g., compositional “language”-like representations) plus sensory-specific forward models. The models that instantiates this computational framework successfully address the acquisition, representation, and cross-modal transfer of psychologically important world variables.
2 Transfer of Object Category Knowledge Across Visual and Haptic Modalities: Experimental and Computational Studies

2.1 Introduction

When recording neural activity in the human medial temporal lobe, Quiroga, Kraskov, Koch, and Fried (2009) found individual neurons that explicitly encode multisensory percepts. For example, one neuron 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 actress 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. Clearly, our brains encode abstract representations of objects that are multisensory in the sense that these representations are activated by perceptual inputs, but these inputs span multiple sensory formats or modalities.
Why would our brains acquire abstract representations that are activated by inputs from a variety of sensory modalities? One possible answer to this question is that these representations facilitate the transfer of knowledge across modalities. Consider, for instance, a person that learns to categorize a set of objects based solely on tactile or haptic inputs. Would the person be able to categorize these same objects when the objects are viewed but not grasped? Would the person be able to view novel objects from the same categories and be able to categorize these?

Here, we report experimental and computational studies of the acquisition of multisensory representations of object category, and the role these representations play in the transfer of knowledge across visual and haptic modalities. Our work includes three contributions. First, our experiment used an unusual set of visual-haptic stimuli known as “Fribbles”. Fribbles are complex, 3-D objects with multiple parts and spatial relations among the parts (see Figure 2.1). Moreover, they have a categorical structure—that is, each Fribble is an exemplar from a category formed by perturbing a category prototype. Fribbles have previously been used in the study of visual object recognition (Hayward & Williams, 2000; Tarr, 2003; Williams, 1997). An innovation of our work is that we have fabricated a large set of Fribbles using a 3-D printing process and, thus, our Fribbles are physical objects which can be both seen and grasped. Based on this set of stimuli, we have created a data set, referred to as the See & Grasp data set, containing both visual and haptic features of the Fribbles. We are making this data set freely available on the world wide web with the hope that it will encourage quantitative research on computational models of visual-haptic perception.

Second, we conducted an experiment evaluating whether people can transfer knowledge of object category across visual and haptic modalities. Previous researchers have considered the transfer of knowledge of object identity across visual and haptic modalities (e.g., Lacey, Peters, & Sathian, 2007; Lawson, 2009;
Norman, Norman, Clayton, Lianekhammy, & Zielke, 2004). They have also compared similarity and categorization judgements based solely on visual input with those based solely on haptic input (Gaißert, Wallraven, & Bülthoff, 2008; Gaißert, Wallraven, & Bülthoff, 2010; Gaißert, Bülthoff, & Wallraven, 2011; Gaißert & Wallraven, 2012). To our knowledge, our experiment is the first focused on the transfer of object category knowledge across visual and haptic modalities.

Lastly, we developed a computational model, referred to as the MVH (Multisensory-Visual-Haptic) model, accounting for how multisensory representations of prototypical 3-D shape might be acquired, and of the role these representations might play in the transfer of category knowledge across visual and haptic modalities. Like some previous models in the literature (Biederman, 1987; Marr & Nishihara, 1978), the model makes use of part-based representations of prototypes. However, it goes beyond previous work by introducing a learning mechanism for the acquisition of these representations. Using its acquired multisensory representations along with sensory-specific forward models for predicting visual or haptic features from multisensory representations, the model transfers object category knowledge between visual and haptic modalities, thereby providing a qualitative account of our experimental data.

2.2 Previous Research on Visual-Haptic Object Perception

Previous research has shown that knowledge of object identity transfers (at least in part) across visual and haptic domains (e.g., Lacey, Peters, & Sathian, 2007; Lawson, 2009; Norman et al., 2004). For example, Lacey et al. (2007) trained subjects to identify objects either visually or haptically. Following training, subjects were tested on the same task using the untrained sensory modality. Subjects
showed excellent transfer to the novel modality when objects were presented at the same orientation as experienced during training, and still showed good transfer when objects were rotated to a new viewpoint.

Researchers have also compared people’s vision-only and haptic-only similarity judgements. For example, Gaißert and colleagues collected people’s unisensory similarity judgements for naturalistic objects resembling sea shells (Gaißert et al., 2008; Gaißert et al., 2010; Gaißert et al., 2011; Gaißert et al., 2012). Analyses based on multidimensional scaling showed that people’s vision-only and haptic-only similarity spaces were nearly identical. Gaißert and colleagues also examined people’s vision-only and haptic-only categorization judgements. Analyses showed that these categorizations were highly similar to each other, and that they were consistent with people’s similarity judgements (also see Haag, 2011).

Additional research has compared the acquisition of haptic concepts by blind individuals and sighted controls. Homa, Kahol, Tripathi, Bratton, & Panchanathan (2009) found that blind subjects learned the categories quickly and comparably with sighted subjects. Other research has studied transfer from haptics to vision in special populations, such as an individual blinded as a child or born with congenital cataracts, but with vision partially restored as an adult (Fine, Wade, Brewer, May, Goodman, et al., 2003; Held, 2009; Held, Ostrovsky, de Gelder, Gandhi, Ganesh, et al., 2011; Ostrovsky, Andalman, & Sinha, 2006). For example, Held et al. (2011) studied congenitally blind individuals born with dense bilateral cataracts. Following surgical removal of the cataracts, they were tested on a haptic-to-vision match-to-sample task in which an observer touched an object and selected an image that he or she thought depicted the same object. It was found that subjects performed poorly two days after surgery, but their performances improved significantly when tested five days after surgery.

Finally, behavioral and neural evidence support the idea that object features extracted by vision and by touch are integrated into multisensory object represen-
tations that are accessible to memory and higher-level cognition (e.g., Amedi, Jacobson, Hendler, Malach, & Zohary, 2002; Amedi, von Kriegstein, van Atteveldt, Beauchamp, & Naumer, 2005; Ballesteros, Gonzalez, Mayas, Garcia-Rodriguez, & Reales, 2009; Easton, Srinivas, & Greene, 1997; James, Humphrey, Gati, Servos, Menon, & Goodale, 2002; Lacey et al., 2007; Lacey, Tal, Amedi, & Sathian, 2009; Lawson, 2009; Norman et al., 2004; Pascual-Leone & Hamilton, 2001; Reales & Ballesteros, 1999; Tal & Amedi, 2009; Taylor, Moss, Stamatakis, & Tyler, 2006). For example, based on fMRI data, Taylor et al. (2006) argued that posterior superior temporal sulcus (pSTS) extracts pre-semantic, cross-modal 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 cross-modal repetition-suppression effects, indicating that these regions are involved in visual-haptic representation.

In summary, previous research strongly suggests the existence and use of multisensory representations of objects. This research leads to, but does not address, our research questions: Can people transfer categorical knowledge about objects across visual and haptic modalities? If so, what computations might underlie this behavior?

2.3 Fribbles and the See & Grasp Data Set

A key component of our research is the unusual visual-haptic stimuli that we used in both our experimental and computational studies. These stimuli are a subset of a larger set of stimuli known as “Fribbles”.\footnote{We thank M. Tarr for making the 3-D object files for Fribbles available on his web pages. We slightly modified these object files so that the connections among parts would be stronger when the objects are fabricated.} Fribbles have previously been used in the vision sciences to study visual object recognition (Hayward &
Figure 2.1: The top row shows computer-generated images of Fribbles which are rendered using the Fribbles’ 3-D object models. The bottom row shows photographs of the physical objects corresponding to these same Fribbles which were fabricated via a 3-D printing process using the same 3-D object models. Pairs of columns illustrate exemplars from different categories (e.g., columns 1-2 illustrate exemplars from Category A).

Williams, 2000; Tarr, 2003; Williams, 1997). Each Fribble is a complex, 3-D object with multiple parts. Our subset includes 40 Fribbles organized into 4 categories with 10 exemplars per category. Category prototypes differed in their parts and the spatial layout of these parts. Exemplars were created by perturbing a category prototype (both in terms of its parts and the spatial relations among these parts). An innovative aspect of our research is that we have obtained physical copies of Fribbles fabricated using an extremely high-resolution 3-D printing process. Consequently, subjects in our experiment were able to see, grasp, or both see and grasp these objects. Each Fribble is about 12 cm in length, 10 cm in width, and 8 cm in height. Figure 2.1 illustrates 8 Fribbles, 2 from each of 4 categories (see caption for explanation).

Our stimuli have several advantages. First, the objects that we use are complex and realistic, each with multiple parts and spatial relations. These stimuli are, thus, more ecologically valid than objects that are typically used in haptic or visual-haptic experiments. Second, our objects are organized into categories. This property allows us to study both object recognition and object categorization, as well as their interactions (Goldstone & Barsalou, 1998). Again, the categorical nature of our stimuli makes them highly realistic. Lastly, the visual and haptic
renderings of our objects are perfectly matched because they are both created from the same 3-D object models.

There does not currently exist a public data set containing both visual and haptic features of complex, realistic objects. As a result, quantitative computational models of visual-haptic interactions or even of haptic perception are nearly non-existent. We have created such a data set, referred to as the See & Grasp data set. Because we are making this set freely available on the world wide web, we believe that it will become a major resource to the cognitive science and computer science communities interested in perception.²

The data set contains 40 items corresponding to our 40 Fribbles. There are 3 entries associated with each item. One entry is the 3-D object model for a Fribble. The second entry is an image of a Fribble rendered from a canonical viewpoint so that the Fribble’s parts and spatial relations among the parts are clearly visible. (Using the 3-D object model, users can easily generate new images of a Fribble from any desired viewpoint.) The third entry is a way of representing a Fribble’s haptic features. It is a set of joint angles obtained from a grasp simulator known as “GraspIt!” (Miller & Allen, 2004). GraspIt! contains a simulator of a human hand. When forming the representation of a Fribble’s haptic features, the input to GraspIt! was the 3-D object model for the Fribble. Its output was a set of 16 joint angles of the fingers of a simulated human hand obtained when the simulated hand “grasped” the Fribble. Grasps—or closings of the fingers around a Fribble—were performed using GraspIt!’s AutoGrasp function. Each Fribble was grasped twice, once from its front and once from its rear, meaning that the haptic representation of a Fribble was a 32-dimensional vector (2 grasps × 16 joint angles per grasp). To be sure that Fribbles fit inside GraspIt!’s hand, their sizes were reduced by 67%.

Caveat: The field of cognitive science currently has an incomplete understand-

²The data set can be downloaded at the URL http://www.bcs.rochester.edu/people/robbie/jacobslab/dataset.html
The notion of “haptic features” (interested readers may want to see the pioneering work on this topic by Klatzky, Lederman, and their colleagues; e.g., Lederman & Klatzky, 1987). Consequently, our choice of joint angles as haptic features follows a common practice in the field of postural hand analysis (e.g., Santello, 1998; also see Thakur, Bastian, & Hsiao, 2008). Consistent with previous research (e.g., Santello, 1998), analyses of the features produced by GraspIt! (joint angles at the time of a stable grasp) reveal that these features contain much information about Fribbles’ shapes. For example, when feature vectors are clustered using a simple “k-means” clustering algorithm (Bishop, 2006), the discovered clusters correspond perfectly to the 4 categories of Fribbles comprising our stimuli.

2.4 Experiment

Questions about categorization and generalization are fundamental to cognitive science, yet many open questions about them remain, particularly in the context of multisensory perception. Important questions include: To what extent does knowledge of object categories gained through one modality transfer to another modality? Is the amount of transfer the same for familiar and novel objects? For example, if a person learns to visually categorize a set of objects, can the person categorize these same objects when the objects are grasped but not seen? Can the person grasp novel objects belonging to the same categories and correctly categorize them too? If so, then the person can be said to have transferred categorical knowledge across modalities.

2.4.1 Participants

Participants were 27 students (6 male and 21 female) from the University of Rochester who reported normal or corrected-to-normal visual and haptic percep-
tion. All participants were at least 18 years old (Min age = 20, Max age = 24, Mean age = 21.5, $SD = 0.96$). We obtained all participants’ written informed consent. Each experimental session lasted about an hour, and participants were paid $10. This study was approved by the University of Rochester Research Subjects Review Board.

\subsection*{2.4.2 Stimuli}

Our experiment made use of 40 Fribbles from the \textit{See \& Grasp} data set, 10 exemplars from each of 4 categories. Visual stimuli consisted of images of Fribbles rendered from a canonical viewpoint so that a Fribble’s parts and spatial relations among the parts were clearly visible (Figure 2.1, top row). Stimuli were presented on a 19-inch CRT computer monitor. Subjects sat approximately 60 cm from the monitor. When displayed on the monitor, visual stimuli spanned about 12 degrees in the horizontal dimension and 10.5 degrees in the vertical dimension. Visual displays were controlled using the Psychtoolbox extension of Matlab (Brainard, 1997; Pelli, 1997).

Participants received haptic inputs when they touched physical copies of Fribbles fabricated using a 3-D printing process (Figure 2.1, bottom row). Participants were blindfolded on trials in which they received haptic inputs. They were instructed to freely and bimanually explore the Fribbles.

\subsection*{2.4.3 Procedures}

Our experiment included three groups of eight participants each (three participants were excluded on the basis of Grubbs tests for outliers). Participants in Group V-H were initially trained to visually categorize 24 Fribbles, 6 exemplars from each of 4 categories. On a training trial, the image of a Fribble was displayed for 8 seconds. When the image disappeared, a participant indicated the category
that he or she believed that the depicted Fribble belonged to by pressing a key on the keyboard. An auditory sound provided feedback as to whether the response was correct or incorrect. Training consisted of 7 blocks where each block consisted of 24 trials (one trial for each Fribble). The presentation order of Fribbles was randomized in each block.

Following training, participants in Group V-H performed test trials. Participants were blindfolded during testing. On a test trial, a participant bimanually grasped and explored a Fribble for 8 seconds (auditory beeps demarcated the beginning and end of the 8-second period). The participant then verbally judged the category of the Fribble. This response was entered into the computer by an experimenter. Feedback about the correctness of the participant’s response was not provided. Participants performed 40 test trials using 10 exemplars from each of 4 categories (the presentation order was randomized). Of the 10 exemplars from a category, 6 were familiar (i.e., these were seen during training) and 4 were novel.

Participants in Group Vs-H were trained and tested identically to participants in Group V-H, except the duration of visual displays was 3 seconds (down from 8 seconds). The stimulus duration on haptic-only test trials remained 8 seconds. Visual-haptic experiments often use visual stimulus durations that are roughly half the length of haptic stimulus durations. The inclusion of Group Vs-H allows us to examine the effect of visual stimulus duration on cross-modal generalization.

Participants in Group H-V were trained and tested in a manner analogous to the training and testing of Group V-H, but training used the haptic modality and testing used the visual modality. That is, on a training trial, participants bimanually grasped and explored (but did not view) a Fribble, judged its category, and received auditory feedback about the correctness of their response. On a test trial, they viewed (but did not grasp) a Fribble and judged its category (without receiving feedback).
Figure 2.2: Learning curves for Groups V-H (mid gray), H-V (dark gray), and Vs-H (light gray) during training. The horizontal axis plots the training block number, and the vertical axis plots the average percent correct. Error bars show the standard errors of the means.

2.4.4 Results

The graph in Figure 2.2 shows the performances of Groups V-H, Vs-H, and H-V during training. This graph plots each group’s average percent correct as a function of the training block number (error bars indicate the standard errors of the means). All groups succeeded at learning, and Group V-H seems to have learned fastest. A mixed-design ANOVA confirmed that there is a significant main effect of group ($F = 44, Df = 2, MSE = 0.46, p < 0.001$) and block number ($F = 70, Df = 6, MSE = 0.74, p < 0.001$), as well as a significant interaction of these factors ($F = 4, Df = 12, MSE = 0.04, p < 0.001$).

Figure 2.3 shows the groups’ performances on the final training block (left panel) and during testing (right panel). For Groups V-H and H-V, the differences between each group’s final training and test performances were not significantly
Figure 2.3: (Left) Performances of Groups V-H, H-V, and Vs-H on the final training block. Dots indicate the performances of individual participants. (Right) Performances during testing.

different (based on two-tailed t-tests; Group V-H: $p = 0.93$; Group H-V: $p = 0.90$). For Group Vs-H, the difference between its final training and test performances was either not statistically significant or it was marginally significant ($p = 0.08$).³ In other words, Groups V-H and H-V showed complete cross-modal transfer during test, and Group Vs-H showed at least partial transfer. Furthermore, participants’ test performances with familiar objects (those seen or grasped during training) did not differ significantly from their performances with novel objects (Group V-H: $p = 0.70$; Group H-V: $p = 0.22$; Group Vs-H: $p = 0.81$).

We are interested in whether people show cross-modal generalization of object category knowledge. When they are trained to categorize objects using one sensory

---

³An anonymous reviewer pointed out the possibility of ceiling effects, which would violate the normality assumptions underlying the standard t-test. Consequently, we also conducted a Wilcoxon rank-sum test (Wilcoxon, 1945), a non-parametric counterpart of the standard t-test (and, thus, this test does not make any distributional assumptions). The results of this test are consistent with the results of the t-test (Group V-H, $p = 0.95$; Group H-V, $p = 0.95$; Group Vs-H, $p = 0.054$).
modality, can they categorize these objects when the objects are sensed through another modality? If so, can they also categorize novel objects from these same categories? Our experimental results indicate that the answers to these questions are “yes”. Our experiment also examined whether the extent of generalization from vision to haptics depends on the duration of visual stimulus presentation during training. Here, our results are inconclusive.

2.5 Preliminary Remarks Regarding the MVH Model

Our data show that participants transferred object category knowledge between visual and haptic modalities. How did they do this? To address this question, we propose a novel computational model, referred to as the MVH (Multisensory-Visual-Haptic) model, with several important properties. This model uses multisensory representations of prototypical 3-D shape. Like some previous models in the literature (Biederman, 1987; Marr & Nishihara, 1978), the model makes use of part-based representations of prototypes. However, it goes beyond previous work by solving the problem of learning these representations using a Bayesian inference algorithm. Because the representations are learned, the MVH model contributes to the growing literature on “grounded cognition” (Barsalou, 2008) by illustrating how high-level abstract representations (e.g., multisensory representations of 3-D prototypes) can be grounded in low-level perceptual features (e.g., image pixel values or joint angles of grasping hands). Using its multisensory representations of prototypes and sensory-specific forward models for predicting visual or haptic features from multisensory representations, the model transfers object category knowledge between visual and haptic modalities, thereby providing a qualitative account of our experimental data.
A complete specification of the MVH model requires a description of the model’s representations, a description of how these representations are learned, and a description of how the representations are used for the purpose of transferring object category knowledge across sensory modalities. This section discusses these aspects of the model in an intuitive manner. The next section provides the mathematical details underlying the model.

**Multisensory representations of prototypical shape:** Based on observed sensory features from either individual or multiple modalities, the model acquires latent or hidden representations of objects. These representations have three important properties.

First, the representations are multisensory, meaning they characterize properties of objects in a way that is independent of the individual modality or modalities through which those properties are sensed. Behavioral and neural data suggest the existence of multisensory representations, and also suggest that these representations underlie, at least in part, a variety of behaviors in visual-haptic environments (e.g., Amedi et al., 2002; Amedi et al., 2005; Ballesteros et al., 2009; Easton et al., 1997; James et al., 2002; Lacey et al., 2007; Lacey, Tal et al., 2009; Lawson, 2009; Norman et al., 2004; Pascual-Leone & Hamilton, 2001; Reales & Ballesteros, 1999; Tal & Amedi, 2009; Taylor et al., 2006).

Because the representations are multisensory, they can be used to predict or “imagine” sensory features from individual modalities. For example, given a multisensory representation of a particular Fribble, the model can predict what the Fribble would look like (perhaps a form of visual imagery) or predict the hand shape that would occur if the Fribble were grasped (perhaps a form of haptic imagery). A mapping from a multisensory 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 & Flanagan, 2009; Wolpert & Kawato, 1998). In cognitive science,
forward models are often mental or internal models. However, forward models exist in the external world too. For instance, a graphics software package is a vision-specific forward model because it maps a 3-D representation of an object to a prediction of an image of the object when viewed from a particular viewpoint. Similarly, the GraspIt! grasp simulator (described above) is a haptic-specific forward model because it maps a 3-D representation of an object to a prediction of the joint angles of the fingers of a hand when the hand grasps the object at a particular orientation.

Second, the representations characterize prototypical knowledge regarding the objects belonging to a category. A prototype is a summary representation of a category based on members’ most common feature values, average feature values, or ideal feature values. Prototype theories of categorization have been influential in the field of cognitive science for many years (see Minda and Smith, 2011, for a recent review).

Lastly, the representations characterize object shape via an object’s parts. Part-based representations of 3-D shape have been explored previously in the artificial intelligence and cognitive science literatures (e.g., Biederman, 1987; Marr & Nishihara, 1978). Our model draws on lessons learned from these earlier efforts. For example, our model uses shape primitives (cylinders as in Marr & Nishihara, 1978) to represent object parts, and uses spatial relations among parts to represent multi-part objects.

Learning process: Importantly, the MVH model’s representations are learned. 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, our model learns representations using a probabilistic or Bayesian inference algorithm.

Multisensory 3-D shape representations are characterized by several param-
eters in our model. These parameters include the number of object parts and the spatial configuration among parts. This information can be described by a network or graph in which nodes represent parts, and edges represent connections between parts. We use a prior distribution that favors spatial configurations in which relatively few parts have many connections and most parts have few connections (e.g., a power law distribution). For example, the prior distribution might assign a high probability to a shape with one main part (e.g., the trunk of a body) and other parts connected to this main part (e.g., the head, arms, and legs). In terms of networks or graphs, the prior favors shallow trees (e.g., a two-level network in which the root or parent node represents the trunk and child nodes represent the head, arms, and legs).

Each object part is represented by a shape primitive, namely a cylinder (Marr & Nishihara, 1978). Therefore, there are also parameters for the length, radius, and orientation of each part or cylinder. A uniform prior distribution is placed on these parameters.

Our description of the learning process also needs to include a likelihood function. Suppose that the model is attempting to acquire a multisensory representation of a category’s prototypical 3-D shape based on visual inputs. For each object belonging to the category, we assume that the model views the object from three orthogonal viewpoints (front, right, and top defined in a spatial reference frame), thereby receiving three images of the object. Given a multisensory 3-D shape representation and the three images, the value of the likelihood function is computed as follows. The model uses a vision-specific forward model to map from the shape representation to an image of the shape. This visual rendering process is repeated at each of the three viewpoints. (In simulation, rendering can be performed by a graphics library such as OpenGL.) The differences between the actual images of an object received by the model and the rendered images based on the multisensory 3-D shape representation are used to calculate a likelihood
value. Likelihood values are computed in an analogous way when the model attempts to acquire a multisensory representation based on haptic inputs. (In this case, haptic rendering can be performed by the GraspIt! simulator.)

Using Bayes' rule, prior probabilities and likelihood values are combined to form posterior probabilities over 3-D shape representations. The prototypical shape for each category of objects is the 3-D shape with the largest posterior probability.

**Transfer of object category knowledge across modalities:** Suppose that the MVH model is a participant from Group V-H in our experiment described above. During training, it acquired multisensory representations of prototypical 3-D shape, one representation for each category of Fribbles, based on visual inputs. Now, during testing, the model grasps, but does not view, a novel Fribble. Because the model has acquired multisensory representations of prototypical shapes, classifying this Fribble is straightforward. The model uses a haptic-specific forward model to haptically render each of the prototypes. It then calculates the differences between the actual haptic features received by the model when the Fribble is grasped and the rendered haptic features based on the prototypical shapes. The model's estimate is the category whose prototype is closest to this Fribble in "haptic feature" space. Classification is performed in an analogous way when the model is trained haptically and tested visually (i.e., when the model is a participant from Group H-V).

### 2.6 MVH (Multisensory-Visual-Haptic) Model

This section provides the mathematical details of the MVH model. We describe the model from the perspective of a participant from Group V-H in our experiment. During training, the model is provided with images of Fribbles along with the Fribbles' corresponding category labels. The model learns a multisensory
representation of each category’s prototypical 3-D shape on the basis of this information. The model is provided with Fribbles’ haptic features during testing, and it estimates the category to which each Fribble belongs.

For the purposes of statistical modeling, we present a “generative” model explaining our data set (our notation is summarized in Table 2.1). We assume that each data item—the visual and haptic features of a Fribble—was generated by the following steps:

1. At random, pick a category, denoted $C$.

2. Let $N_C$ denote the number of parts comprising a member of category $C$. For each part, pick a cylinder (i.e., pick values for the cylinder’s parameters, namely its length $l$, radius $r$, and orientation $o$). Let the part collection $\Omega_C$ denote the parameter values of all of an object’s parts (a vector with $N_C$ [number of parts] $\times$ 3 [number of parameters per part] elements).

3. Using a sequential procedure (see below), pick the spatial layout of a Fribble’s parts by selecting values for a directed tree graph $T_C$ and spatial configuration $S_C$. Tree $T_C$ has $N_C$ nodes where nodes correspond to parts and edges indicate which parts are connected (left side of Figure 2.4). Configuration $S_C$ indicates where parts are connected. As discussed below, parts can only connect at “docking locations.”

4. Parameters $\Omega_C$, $T_C$, and $S_C$ define a Fribble. That is, specific values for these parameters correspond to a specific Fribble. Given values for these parameters, visually project the corresponding Fribble onto an image plane to render its visual features, denoted $V_C$, and haptically project the Fribble to render its haptic features, denoted $H_C$ (right side of Figure 2.4). Note that $\Omega_C$, $T_C$, and $S_C$ define a specific Fribble but these parameters can also be used to define an ideal or prototypical Fribble.
<table>
<thead>
<tr>
<th>Variable</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C$</td>
<td>Category indicator; $C \in {1, 2, 3, 4}$</td>
</tr>
<tr>
<td>$N_C$</td>
<td>Number of parts in prototype for category $C$</td>
</tr>
<tr>
<td>$l_k$</td>
<td>Length of part $k$</td>
</tr>
<tr>
<td>$r_k$</td>
<td>Radius of part $k$</td>
</tr>
<tr>
<td>$o_k$</td>
<td>Orientation of part $k$</td>
</tr>
<tr>
<td>$\Omega_k$</td>
<td>Part $k$’s length, radius, and orientation; $\Omega_k = {l_k, r_k, o_k}$</td>
</tr>
<tr>
<td>$\Omega_C$</td>
<td>All parts in prototype for category $C$; $\Omega_C = {\Omega_1, \ldots, \Omega_{N_C}}$</td>
</tr>
<tr>
<td>$e_k$</td>
<td>Edge connecting $k^{th}$ parent-child pair; $k \in {1, \ldots, N_C - 1}$</td>
</tr>
<tr>
<td>$d_k$</td>
<td>Number of available docking locations for edge $e_k$; $k \in {1, \ldots, N_C - 1}$</td>
</tr>
<tr>
<td>$T_C$</td>
<td>Tree graph characterizing prototype in terms of parts (nodes) and spatial configurations among parts (edges)</td>
</tr>
<tr>
<td>$S_C$</td>
<td>Spatial configurations for all parts; each configuration indicates where (which docking locations) a child part connects to its parent part</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>Scaling parameter for prior distribution over $T_C$ and $S_C$</td>
</tr>
<tr>
<td>$V_C$</td>
<td>Visual features for category $C$ prototype obtained via visual forward model</td>
</tr>
<tr>
<td>$H_C$</td>
<td>Haptic features for category $C$ prototype obtained via haptic forward model</td>
</tr>
<tr>
<td>$V_1, \ldots, V_M$</td>
<td>Visual features for $M$ exemplars</td>
</tr>
<tr>
<td>$R$</td>
<td>Number of pixel-wise disagreements between visual features of exemplar and category prototype</td>
</tr>
</tbody>
</table>

Table 2.1: Summary of the variables in the MVH model.

Figure 2.4: Schematic illustrating steps 3 and 4 in the generative process for the MVH model. (Left) Tree graph characterizing a prototype in terms of its parts (nodes) and spatial configurations among parts (edges). Illustration to the right of the tree shows that the bottom-right edge represents the connection between a child part and its parent part. A subset of the “docking locations” on the parent part are illustrated with black and white dots. The child part is “docked” on a docking location at the top of the parent part. (Right) Haptic and visual features of a prototype are obtained through the use of haptic and visual forward models.
Suppose that, during training, the MVH model is provided with the visual features of \( M \) exemplars from category \( C \), denoted \( V_{C_1}, \ldots, V_{C_M} \). The posterior distribution of the latent variables \( \Omega_C, T_C, \) and \( S_C \) given this data can be computed via Bayes’ rule:

\[
p(\Omega_C, T_C, S_C|V_{C_1}, \ldots, V_{C_M}) \propto p(\Omega_C) \ p(T_C, S_C|\Omega_C) \ p(V_{C_1}, \ldots, V_{C_M}|\Omega_C, T_C, S_C).
\]

(2.1)

The values of \( \Omega_C, T_C, \) and \( S_C \) with the highest joint probability define the multisensory representation of category \( C \)’s prototypical 3-D shape. The right-hand side of Equation 2.1 has three terms which we describe in order. In the remainder, we drop the redundant category subscripts \( C \) for the sake of clarity.

**Prior distribution over part collection \( \Omega \):** Members of category \( C \) have \( N \) parts where each part is modeled as a cylinder with length \( l \), radius \( r \), and orientation \( o \). Let \( \Omega_k \) denote the portion of part collection \( \Omega \) corresponding to part \( k \), and let \( l_k, r_k, \) and \( o_k \) denote this part’s length, radius, and orientation, respectively. We use a prior distribution that assumes that parts are independent, meaning that \( p(\Omega) \) can be factored:

\[
p(\Omega) = \prod_k p(\Omega_k)
\]

(2.2)

and that a part’s parameters are independent, meaning that \( p(\Omega_k) \) can be factored:

\[
p(\Omega_k) = p(l_k) \ p(r_k) \ p(o_k).
\]

(2.3)

In our simulations reported below, we set \( p(l_k) \) to be a uniform distribution over integers in the range \([1, \ldots, 40]\), and set \( p(r_k) \) to be a uniform distribution over the range \([0.5, 1, 1.5, \ldots, 20]\) (arbitrary units). The orientation of a part was always parallel to one of the three axes in our spatial reference frame and, thus, \( p(o_k) \) was a uniform distribution over the set \( \{1, 2, 3\} \).
Prior distributions over tree $T$ and spatial configuration $S$: The spatial layout of a Fribble’s parts is parameterized by two variables. Directed tree graph $T$ contains $N$ nodes where each node corresponds to a part. Edges in the tree indicate which parts are connected. Spatial configuration $S$ indicates where the parts are connected.

We modeled part connections by assuming that parts could only connect at docking locations. The cylinder corresponding to a parent part (where parent-child relationships are given by tree $T$) was approximated by an elongated cube with 6 orthogonal planar surfaces. Each surface contained 18 equally spaced docking locations. At the connection between parent and child parts, a child could cover one or more docking locations on the parent depending on the child part’s size (as given by part collection $\Omega$).

Tree $T$ and spatial configuration $S$ are constructed in a sequential manner resembling a breadth-first search. We start with one part corresponding to the root node located at the highest level of $T$. Then a new node and edge are inserted in $T$ such that the new node is a child to the root node. The docking locations on the root or parent node that are covered by the new node must be selected. Next, a third node and edge are inserted such that the node is either a sibling or a child to the second node. Again, the docking locations on the parent node that are covered by the child node must be selected. Importantly, new nodes and edges are always added to $T$ such that nodes are inserted at a higher level (closer to the root) before nodes are added at a deeper level (further away from the root). This sequential procedure provides an ordering to the nodes and edges of $T$, and this ordering influences the values of $S$. For example, suppose Part 2 covers docking locations 1, 2, and 3 on Part 1. When adding Part 3 as a child to Part 1, Part 3 cannot connect to Part 1 at these same locations.

Suppose that at a particular moment in the sequential procedure, we are adding a new edge, denoted $e_k$, joining a new node to a parent node. Let $d_k$
denote the number of unoccupied docking locations on the parent node. Using this notation, we define the prior probability over $T$ and $S$ as:

$$p(T, S|\Omega) \propto \prod_{k=1}^{N_C-1} \exp(-\alpha d_k) \quad (2.4)$$

where $\alpha$ is a scaling parameter (we set $\alpha = 1$ in all our simulations). This prior prefers spatial layouts in which relatively few parts have many connections and most parts have few connections (e.g., a power law distribution). For example, the prior distribution might assign a high probability to a shape with one main part (e.g., the trunk of a body) and other parts connected to this main part (e.g., the head, arms, and legs connected to the trunk).

**Likelihood function $p(V_1, \ldots, V_M|\Omega, T, S)$:** The likelihood function measures how well the model accounts for the data. In our simulations, the model was provided with the visual features of $M$ exemplars from category $C$, denoted $V_1, \ldots, V_M$, during training. As described below, we used three images of each Fribble rendered at orthogonal viewpoints. In addition, pixel values were binary. In this case, $V_i$ is a binary vector of pixel values from all three images of the $i^{th}$ Fribble.

The likelihood function is computed in two stages. First, the prototypical 3-D shape defined by $\Omega$, $T$, and $S$ is visually rendered using the same three viewpoints as used to generate the training images of Fribbles (Figure 2.5). This can be accomplished by a vision-specific forward model. Next, the rendered images of the prototype are compared to the training images. Let $R$ denote the number of pixel-wise disagreements between the prototype images and the training images. Then

$$p(V_1, \ldots, V_M|\Omega, T, S) \propto \exp(-R) \quad (2.5)$$

defines the likelihood function.

**Inference:** Exact inference in the MVH model is computationally intractable.
Figure 2.5: Our inference algorithm consists mostly of Metropolis-Hastings (MH) steps. In each iteration $i$, a proposal prototype is drawn from the generative process. Then the visual forward model is used to obtain the visual features of the proposed prototype. These visual features are compared against the visual features of the training exemplars to compute a log-likelihood value. This value is used to evaluate the proposal with respect to the MH acceptance function and the current state of the chain. Details of our inference algorithm can be found in the Appendix.
Therefore, we developed an approximate Markov chain Monte Carlo inference algorithm that discovers good point estimates of parameters $\Omega$, $T$, and $S$. This algorithm is described in the Appendix.

2.7 Simulation Results

In the simulations reported here, we used a slightly modified version of the See & Grasp data set for the four categories used in the experiment. We used three images of each Fribble rendered from three orthogonal viewpoints—a top view, a front view, and a right view. In addition, we simplified the images by using low-resolution images (80 pixels $\times$ 80 pixels) and by converting pixel values to binary numbers using a thresholding scheme. Therefore, the visual representation of a Fribble was a 19200-dimensional binary vector (3 images $\times$ 80 pixels $\times$ 80 pixels).

As discussed above, the haptic representation of a Fribble was a 32-dimensional real-valued vector (2 grasps $\times$ 16 joint angles per grasp).

Four data items used in our simulations are illustrated in the four rows of Figure 2.6. Column 1 of each row shows a Fribble. Columns 2-4 show binary images of the Fribble from top, front, and right viewpoints. Columns 5-6 show the simulated hand grasping the Fribble, once from the Fribble’s front and once from its rear.

We simulated the MVH model from the perspective of a subject in Group V-H in our experiment. That is, the model was trained with visual inputs and tested with haptic inputs. As in our experiment, we trained the model with 24 Fribbles, 6 from each of 4 categories. The model was then tested with 40 Fribbles, 10 from each of 4 categories. Of the 10, 6 were familiar (these were seen during training) whereas 4 were novel. The simulation results are presented in two parts. We first examine the multisensory prototypes acquired by the MVH model. Can the model learn reasonable multisensory prototypes of object categories based solely
Figure 2.6: Four data items used in our simulations, one from each category of Fribbles. Column 1 of each row shows a Fribble. Columns 2-4 show binary images of the Fribble from top, front, and right viewpoints. Columns 5-6 show the simulated hand grasping the Fribble, once from the Fribble’s front and once from its rear.
Figure 2.7: (Top) 3 exemplars from each category. (Bottom) The multisensory prototypes learned by the model, visually rendered at a canonical projection.

on visual features? Next, we examine the generalization performances of the model when it was tested with haptic features. Can it correctly estimate the categories of Fribbles based on haptic inputs even though it has never previously touched a Fribble?

**Multisensory prototypes:** Figure 2.7 illustrates multisensory prototypical 3-D shapes learned by the model for each of the four categories. The top row illustrates three exemplars from each category. The prototypes learned by the model are illustrated in the bottom row. Although different simulations of the MVH model produced slightly different results, the prototypes shown in the figure are typical.

For all categories, the model learned 3-D, part-based, prototypical representations which are remarkably accurate. Prototypes characterized the major components of Fribbles—for example, prototypes consistently approximated the main bodies of Fribbles with large cylinders. Prototypes also characterized many of the subtle features of Fribbles—prototypes approximated Fribbles’ smaller appendages with smaller cylinders attached to the large cylinders. In other words, the number, positions, and orientations of prototypes’ cylinders, while not always perfect, were close approximations to the number, positions, and orientations of Fribbles’ body parts. The accuracies of the acquired prototypes are especially
impressive when one recalls that the model learned these prototypes from three binary images of each exemplar.

**Testing the model with haptic features:** Following visual training, we tested the MVH model by using it to classify 40 Fribbles based solely on their haptic features. Predictions for category membership were generated as follows. Using the GraspIt! simulator as a haptic-specific forward model, we obtained the haptic features of each of the four category prototypes. For each Fribble in the test set, we measured the Euclidean distance between the haptic features of the test item and the haptic features of each of the four prototypes. The item was classified based on the prototype it was closest to in “haptic feature” space. The model achieved perfect performance for all test items, both Fribbles seen during visual training as well as novel Fribbles.

To better understand why the model performed so well, we performed a Principal Component Analysis (PCA) using the haptic features of the 40 test items and the 4 prototypes. Based on the results of this analysis, we reduced the 32-dimensional haptic-feature space to two dimensions which accounted for 77% of the variance of the data. Figure 2.8 shows the projections of the haptic features of the test items and prototypes into this two-dimensional space. Clearly, exemplars for each category are tightly clustered in this space, and category prototypes lie close to exemplars from the same category.

In summary, the model acquired multisensory categorical representations in the form of prototypical 3-D componential shapes. The multisensory prototypes learned by the model preserved with high fidelity the typical shapes of category exemplars. In addition, the MVH model achieved excellent performance when, following visual training, it was tested with the haptic features of Fribbles. In some sense, this is surprising because the model had never previously touched a Fribble. Nonetheless, it was able to use its haptic-specific forward model to predict the haptic features of each category’s multisensory prototype. The model illustrates
the potential importance of multisensory prototypes and sensory-specific forward models for the transfer of object category knowledge across modalities, and thus for accounting for subjects’ performances in our experiment.

2.8 Discussion

In summary, this article has addressed people’s abilities to transfer object category knowledge across visual and haptic domains. Our work has made three contributions. First, by fabricating Fribbles (3-D, multi-part objects with a categorical structure), we developed (and are making freely available on the web) visual-haptic stimuli that are highly complex and realistic. Second, we conducted an experiment evaluating whether people transfer object category knowledge across visual and haptic domains. Our data clearly indicate that we do. Third, we developed a computational model that learns multisensory representations of prototypical 3-D
shape through the use of sensory-specific forward models that play important roles during both learning and transfer. The model provides an excellent qualitative account of aspects of our experimental data.

Many articles in the literature on multisensory perception emphasize the role of multisensory representations. Our work is unusual in its additional emphasis on sensory-specific forward models. We hypothesized that forward models allow people to make predictions of sensory features from multisensory representations. Future work will need to experimentally and theoretically evaluate the role of forward models in multisensory perception.

For instance, it would be interesting to know the extent that deliberate intent is needed for the use of forward models. Subjects in our experiment were told at the start of an experimental session that they would be trained with one sensory modality and tested with another. This knowledge may have encouraged subjects to attempt to use their forward models during training to facilitate performance during testing. Consider a subject in Group H-V. During training, the subject may have deliberately engaged in visual imagery in the belief that predicting visual features during haptic training would aid the transfer of knowledge from haptic to visual domains. In future experiments, subjects should not be given advance knowledge of testing with an untrained modality. If test performances are significantly poorer in this case, then this would suggest that the use of forward models requires deliberate intent. If test performances are unchanged, then this would suggest that the use of forward models is automatic.

Our emphasis on sensory-specific forward models also has theoretical implications for how we interpret existing data. As mentioned above, Held et al. (2011) studied congenitally blind individuals born with dense bilateral cataracts. Following surgical removal of the cataracts, these individuals were tested on a haptic-to-vision match-to-sample task in which an observer touched an object and selected an image that he or she thought depicted the same object. It was found
that subjects performed poorly two days after surgery, but their performances improved significantly when tested five days after surgery. Why did their performance improve in the interim? We speculate that they performed poorly two days post-surgery because they had poor vision-specific forward models. That is, they could not accurately predict the visual features of objects they had touched. Their performances improved after a few more days, we hypothesize, because the accuracy of their vision-specific forward models improved.

Whether or not this speculation is correct, our experimental results about cross-modal transfer and our theoretical results about the mechanisms that might underlie this transfer suggest a close interaction between multisensory representations and forward models when learning multisensory representations and when transferring knowledge across sensory domains. Consequently, we believe that the experimental and theoretical approaches advocated here provide new perspectives on crucial questions about multisensory perception, and new opportunities to study old and new questions.
3 Learning Multisensory Representations for Auditory-Visual Transfer of Sequence Category Knowledge: A Probabilistic

3.1 Introduction

Human cognition is robust, at least in part, because people mentally represent objects and events in a variety of ways, such as perceptual, motoric, and semantic representations. Even within perception, people represent objects and events in multiple ways. This fact is demonstrated by cross-modal transfer of knowledge. 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 learning is used during haptic testing, this finding suggests the existence of both visual and haptic representations of objects. Below, we report an experiment in which people were trained to either auditorily
or visually categorize sequences of events. When tested with sequences presented in a novel sensory modality, people were often able to categorize these sequences too. Because training and testing used different modalities, this result indicates that people had representations of event sequences in both modalities.

How do people transfer knowledge across sensory modalities? A plausible hypothesis, referred to here as the Multisensory Hypothesis, is that people use sensory-specific representations of objects and events to infer amodal or multisensory representations characterizing objects’ and events’ intrinsic properties. These representations facilitate cross-modal transfer of knowledge. To understand this hypothesis, it is important to recognize the distinction between objects’ and events’ intrinsic (or “deep”) properties and the sensory (or “surface”) features that these properties give rise to. For instance, the location of an event is a modality-independent intrinsic property. Visual and auditory features are modality-dependent sensory cues to the event’s location arising when the event is viewed or heard, respectively. To explain how the Multisensory Hypothesis accounts for cross-modal transfer, consider a sequence categorization task. For example, sequences of events moving in a clockwise direction belong to category $A$, whereas sequences moving in a counterclockwise direction belong to category $B$. When a person is trained to visually categorize event sequences, the person uses his or her visual representations to infer multisensory representations characterizing sequences’ intrinsic properties. When subsequently tested with an auditory sequence, the person judges its category based on whether it is more consistent with the intrinsic properties of sequences belonging to category $A$ or category $B$.

The Multisensory Hypothesis predicts that people acquire modality-independent representations of objects’ and events’ intrinsic properties. Converging neural, behavioral, and computational evidence suggests that this is the case. A striking example comes from Quiroga (2012) who argued that human brains contain “concept” cells which are involved in the representation of individual people or objects
regardless of the modality used to sense those people or objects. For instance, when recording in the human medial temporal lobe, he and his colleagues reported a neuron that selectively responded when a person viewed images of the television host Oprah Winfrey, viewed her written name, or heard her spoken name (Quiroga, Kraskov, Koch, & Fried 2009). These and similar findings indicate that our brains encode abstract representations that are amodal or multisensory in the sense that they are activated by perceptual inputs spanning multiple modalities.

Why focus on the Multisensory Hypothesis? The Multisensory Hypothesis is an appropriate focus because of its inherent interest and potential importance. Due to their abstract, modality-independent nature, multisensory representations are a form of conceptual representation. Currently, the field of Cognitive Science knows very little about how people acquire conceptual representations from sensory data, though this topic has garnered much interest in recent years (e.g., see the literature on “grounded cognition”; Barsalou, 2008). Furthermore, the study of multisensory perception is attracting much attention (Calvert, Spence, & Stein, 2004; Stein, 2012). It may be that recent advances in our understanding of multisensory perception shed new light on the Multisensory Hypothesis.

To our knowledge, no one has attempted to explicitly define and implement a model of cross-modal transfer based on the Multisensory Hypothesis. The sole exception is our earlier work where we showed how multisensory representations of object shape—consisting of representations of an object's parts and the spatial relations among these parts—can be acquired from visual or haptic features, and showed how these representations can facilitate transfer of object category knowledge across visual and haptic modalities (Yildirim & Jacobs, 2013). The work reported in this article builds on our earlier work. However, it studies a new domain, namely cross-modal transfer of sequence category knowledge across visual and auditory modalities. In addition, it uses a different modeling approach.

In cognitive modeling, one school of thought favors symbolic approaches, such
as approaches based on production rules or logic. Another school of thought favors statistical approaches, such as approaches based on connectionist networks or Bayesian inference. Advocates of these different schools of thought have different perspectives, and have often engaged in heated debates (McClelland & Patterson, 2002a, 2002b; Pinker & Ullman, 2002a, 2002b). Unfortunately, these debates have not led to a resolution as to which framework is best.

Our viewpoint is that both symbolic and statistical frameworks have important merits, and thus it may be best to pursue a hybrid approach taking advantage of each framework’s best aspects. This viewpoint is recently emerging in the Cognitive Science literature (e.g., Goodman, Tenenbaum, Feldman, & Griffiths, 2008; Piantadosi, Tenenbaum, & Goodman, 2012; Ullman, Goodman, & Tenenbaum, 2012). It is referred to as a probabilistic language of thought (pLOT) approach because it applies Bayesian inference to a representation consisting of symbolic primitives and combinatorial rules (Fodor, 1975). To date, the pLOT approach has been used almost exclusively in domains that are typically modeled using symbolic methods, such as human language and high-level cognition. A contribution of the work presented here is that we apply this approach to the study of human perception, an area whose study is dominated by statistical techniques. We believe that the pLOT approach can be advantageous for characterizing perceptual processes, particularly multisensory processes, including the acquisition of amodal, multisensory representations of objects and events from sensory data and their subsequent use.

3.2 Experiment

Previous experimental and theoretical studies examined people’s performances in tasks requiring them to learn about sequences. For example, researchers studied the learnability of sequences with different kinds of structural (e.g., Marko-
vian, non-Markovian, hierarchical) dependencies (e.g., Jordan, 1986; Elman, 1990; Cleeremans & McClelland, 1991; McCallum, 1996; Fiser & Aslin, 2002), and proposed different kinds of cognitive architectures to explain the observed behavioral patterns (see Gureckis & Love, 2010, for a critical review and comparison).

A subset of these researchers used sequences of spatial locations (e.g., Hunt & Aslin, 2001; Deroost & Soetens, 2006; Hunt & Aslin, 2010; Bo & Seidler, 2010). The serial reaction time task is frequently used in these studies. It has been found that people’s reaction times decline more quickly with a structured or highly predictable sequence than with a random or relatively unpredictable sequence (e.g., Hunt & Aslin, 2001).

Our experiment focuses on categorization of spatial sequences, and on generalization of sequence category knowledge to exemplars presented in an untrained sensory modality. The experiment made use of an innovative auditory-visual environment (see Figure 3.1A) whose major components are a vertically-oriented (and oriented perpendicular to a subject’s line of sight) planar surface covered with sheet metal, speakers, and light emitting diodes (LEDs). Each speaker and LED has a magnet attached to its back, meaning that each speaker and LED can be placed at any location on the vertical surface. Because a scrim (a curtain made from light gauzy material often used in theatre productions) covers the environment, the speakers and unlit LEDs are not visible by a subject. However, lit LEDs are visible to a subject due to the scrim’s translucent properties. This environment is very useful for auditory-visual experiments. It is a large-scale environment—when a subject is seated 60 cm from the vertical surface, speakers and LEDs can be placed over a region subtending nearly 90 degrees of visual angle. The environment is flexible because speakers and LEDs can be placed at any location on the vertical surface, and precise because each speaker and LED is controlled independently on a millisecond time scale.
Figure 3.1: **A.** Photos of the audio-visual environment. In the left photo, the speakers, LEDs, and electrical hardware are visible. In the right photo, a scrim conceals the environment, meaning that the speakers, unlit LEDs, and electrical hardware are not visible. **B.** A schematic of the 7 locations used in our experimental stimuli, and the speaker and LED at each location.

### 3.2.1 Participants

Participants were 21 students from the University of Rochester. All participants were at least 18 years old. We obtained all participants’ written informed consent. Each experimental session lasted less than an hour, and participants were paid $10. This study was approved by the University of Rochester Research Subjects Review Board.

### 3.2.2 Stimuli

Stimuli consisted of temporal sequences of spatial locations presented in the auditory-visual environment. There were 7 possible locations arranged on an imaginary circle of radius about 57 cm (see Figure 3.1B). Sequence lengths were sampled from a uniform distribution with minimum and maximum values of 6 and 15, respectively. When a sequence was presented auditorily, a location was indicated by a beep emitted by a small speaker. When a sequence was presented visually, a location was indicated by a flash of a white LED. Beeps or flashes lasted 200 ms, and pauses between beeps or flashes lasted 300 ms.
Sequences were exemplars from 4 possible categories. Fourteen exemplars from each category were generated. For each category, each of the 7 possible locations was used as a starting location twice.

Locations in exemplars from Category 1 change one unit in a clockwise direction at each time step, referred to as a [+1] pattern. Using the location indices in Figure 3.1B, exemplars from Category 1 include “456712”, “2345671234”, and “45671234”.

Exemplars from Category 2 are clockwise cycles of length 3, denoted [+1 +1 -2]. That is, the second location is one clockwise unit from the first location, the third location is one clockwise unit from the second location, and the fourth location is equal to the first location. This pattern repeats until the end of the sequence. Exemplars include “23423423”, “7127127”, and “456456456”.

For Category 3, exemplars are counterclockwise cycles of length 4, denoted [-1 -1 -1 +3]. Exemplars include “17651765176517”, “6543654”, and “54325432”.

Exemplars in Category 4 follow a [+2 -1] pattern. Exemplars include “72132435465”, “132435”, and “4657617213”.

3.2.3 Procedures

At the start of a trial, a red LED at the center of the auditory-visual environment was illuminated for 1000 ms. Participants were asked to fixate this LED. Next, a sequence was presented, either auditorily or visually. Following the sequence presentation, participants indicated the category to which they thought the sequence belonged by pressing a key on a keyboard. On training trials, auditory feedback indicated whether a response was correct or incorrect. Feedback was not provided on test trials.

Participants were seated approximately 50 cm from the auditory-visual display panel. However, because people’s auditory estimates of location are less
accurate than their visual estimates (Battaglia et al., 2003; Alais & Burr, 2004), and because localization of auditory events was difficult during preliminary studies, participants were encouraged to lean forward to be as close as possible to the panel while observing auditory events.

Two groups of 9 people each participated in the experiment (3 people were excluded because they performed at chance on training trials or because they did not complete the full set of training and test trials). Participants in Group A-V were initially trained to categorize 36 exemplars (9 exemplars from each of the 4 categories selected at random for each participant) presented auditorily. This auditory training stage consisted of blocks of 36 trials, where all exemplars were presented once and in randomized order in a block. At the end of a block, a message appeared on a computer screen informing a participant of his or her performance during that block. During training, a participant’s performance was monitored within a window of the most recent 36 trials. Training was terminated as soon as this performance exceeded 90%, or when the participant completed 7 blocks of training.

Following training, participants in Group A-V performed test trials. Test trials were identical to training trials except that sequences were presented visually, and participants did not receive feedback about the correctness of their responses. Participants performed 56 test trials (14 exemplars from each of 4 categories; 9 of the 14 exemplars were familiar [these sequences were presented during auditory training], whereas 6 exemplars were novel). Presentation order of the test sequences was randomized.

Participants in Group V-A followed the same procedures as participants in Group A-V except that the training and test modalities were switched. These participants underwent visual training and auditory testing.
3.2.4 Results

The left and right panels of Figure 3.2 show each participant’s learning curve during training for Groups V-A and A-V, respectively. The horizontal axis of each graph plots the training trial number, and the vertical axis plots the percent correct in the most recent 36 trials (for trials up to the 36\textsuperscript{th}, we assumed that a participant made 36 – $t$ incorrect responses where $t$ is the trial number). Participants in Group V-A were faster at learning the categories as demonstrated by their steeper learning curves which reached the training cut-off criteria of 90% earlier than the learning curves of participants in Group A-V. In contrast, not all participants in Group A-V achieved the training performance criteria. The performances of these participants tended to plateau at around the 100\textsuperscript{th} trial. These differences in the learning curves between the two groups are most likely due to the differing reliabilities of audition and vision for spatial localization. Clearly, however, all participants acquired significant knowledge of the sequence categories (chance performance is 25%).

The left panel in Figure 3.3 shows participants’ average performances on the final training block (i.e., the last 36 trials of training) and on the test block for both groups (error bars indicate standard errors of the means). The training performance of Group V-A reflects the fact that all participants achieved the training cut-off criteria of 90% correct. On auditory test trials, the performance of this group remained high (about 75%). The drop in performance from training to test is most likely due to the lower reliability of audition for spatial localization. The training performance of Group A-V was also good (slightly less than 75%), and its test performance was not significantly different than its training performance. Neither Group V-A’s nor Group A-V’s test performances differed on trials with familiar (i.e., previously observed during training) versus novel sequences.

The right panel in Figure 3.3 shows the final training block and the test per-
performances across the two groups when trials are sorted by the category of the sequence observed on the trial. For example, the leftmost bar in this panel shows the average performance on the final training block on trials that used sequences that are exemplars from Category 1. When examining the data in this manner, chance performance is 50% correct because a participant either correctly classified a sequence from, for instance, Category 1 or did not. This analysis allows us to examine the relative ease of correctly classifying exemplars from each category. When sorted by the categories, a Friedman test revealed that there was a statistically significant rank ordering of the categories across training ($p < 0.001$) and test ($p < 0.01$) blocks. From participants’ average performances, categories can be ordered with respect to their learnability (from highest to lowest) as follows: Category 1, Category 4, Category 2, and Category 3.

In summary, we are interested in people’s abilities to acquire and transfer knowledge of sequence categories. When categorical knowledge of spatial sequences is obtained through one sensory modality, can people transfer this knowledge to conditions in which sequences are observed through an untrained modality? Our experimental results indicate that the answer is yes.
Figure 3.3: (Left) Average performances on the final training block (last 36 trials of training) and on the test block for Groups V-A and A-V (error bars indicate standard errors of the means). (Right) Groups’ final training block and test performances when trials are sorted by the category of the sequence observed on the trial.

3.3 Overview of the Model’s Components

Our experiment suggests that participants transferred sequence category knowledge across auditory and visual modalities. How did they do this? To address this question, we propose a computational model accounting for our experimental results. The model includes a multisensory representation of each sequence category. A multisensory representation characterizes the intrinsic properties of a category in a modality-independent manner (Yildirim & Jacobs, 2013). The model also includes sensory-specific forward models. Because sensory-specific forward models map multisensory representations to sensory data, they can be thought of as implementing a type of mental imagery (Miall & Wolpert, 1996; Ito, 2008; Tian & Poeppel, 2010; Yildirim & Jacobs, 2013). This section provides an overview of the model’s components. The next section describes learning and cross-modal transfer by the model in the context of our experiment.
3.3.1 Multisensory Representations of Sequence Categories

Given auditory exemplars, visual exemplars, or both, the model learns a multisensory representation of a category. Behavioral and neural data suggest the existence of multisensory representations, and also suggest that these representations underlie, at least in part, a variety of behaviors in auditory-visual environments (e.g., Calvert et al., 1997; Pascual-Leone & Hamilton, 2001; Pekkola, Ojanen, Autti, Jääskeläinen, et al., 2005; Tanabe, Honda, & Sadato, 2005; de Gelder & Vroomen, 2000; von Kriegstein & Giraud, 2006; Lehmann & Murray, 2005; Quiroga et al., 2009; Liang, Mouraux, Hu, & Iannetti, 2013).

We characterize multisensory representations as computer programs for generating or predicting exemplars from a category. This approach builds on earlier work by Piantadosi, Tenenbaum, & Goodman (2012) who used computer programs to characterize people’s mental representations of numerical concepts.

When designing the computational model, our main focus was not on developing new insights regarding human sequence learning. Although this is an important topic, many researchers already study this topic (e.g., Jordan, 1986; Elman, 1990; Cleeremans & McClelland, 1991; Gureckis & Love, 2010; McCallum, 1996; Fiser & Aslin, 2002). Rather, our goal was to understand how multisensory representations can be learned from sensory data, and to understand how multisensory representations can facilitate transfer of knowledge across sensory modalities. Because our model needs to represent sequences, it necessarily resembles previously existing models that also represent sequences. Of particular interest is the fact that our model shares important features with an early model of sequence learning by Simon and Kotovsky (1963). Although our model and their model have different goals, the two models use “programming languages” with similar symbolic operators to represent sequences. Indeed, it is only a moderate stretch to say that our model might be seen as a revised version of their model, modified to
accept and process sensory data and updated to use modern learning (Bayesian inference) techniques.

We describe the programming language used by our model by explaining several sample programs. Consider the program in Panel A of Figure 3.4. This program generates sequences in which locations change one unit in a clockwise direction at each time step (Category 1 from the experiment reported above). The first line of the program, denoted L1, randomly initializes a spatial cursor, denoted k. The spatial cursor is a variable that keeps track of the current spatial location. The \texttt{init} function randomly sets the cursor to a random integer between 1 and 7 (recall that there are 7 possible locations). Line L2, \texttt{next(k)}, moves the cursor one unit in a clockwise direction. Line L3, \texttt{go to L2}, states that the next line to be executed is L2, thereby creating a loop. Putting aside the fact that the program creates infinite sequences of locations (see below), the reader should intuitively understand that this program is consistent with sequences such as “456712” and “23456”, but inconsistent with sequences such as “124” and “765”.

Next, consider the program in Panel B. It uses the same primitives as the previous program, but it composes them in a different way. This program generates sequences of length 3 in which the second location is one clockwise unit from the first location, and the third location is two clockwise units from the second location (e.g., sequences such as “124” and “457”). Importantly, this program uses recursion (see line L3).

The program in Panel C generates sequences that alternate between two neighboring locations, first moving one clockwise unit, then returning to the original location by moving one counterclockwise unit. Movement of the spatial cursor by one counterclockwise unit is achieved using the command \texttt{prev(k)} (line L3). This program is consistent with sequences such as “454545” and “1212121”.

The program in Panel D generates sequences with counterclockwise cycles of length 4. It is consistent with sequences such as “654365” and “32173217”.
This program illustrates an additional feature of the init function that was not illustrated by earlier programs. The first time that init is called, it sets the spatial cursor to a random location. It then stores this location. Subsequent calls to init set the cursor to the stored location.

Based on these programs, the reader should have a good understanding of the nature of the model’s programming language. Programs contain line numbers, a spatial cursor, and init, next, prev, and go to commands. Programs are capable of looping and of recursion. Clearly, these elements provide the model with a rich, expressive language for characterizing sequence categories.

### 3.3.2 Sensory-Specific Forward Models

Because multisensory representations are modality-independent, sensory-specific forward models are needed to relate the representations to sensory data. An auditory-specific (vision-specific) forward model maps an exemplar to a prediction of the auditory (visual) features that an observer would perceive when the exemplar is auditorily (visually) rendered. Because of the simple nature of our
stimuli—beeps and flashes—our forward models are relatively simple.\textsuperscript{1} In particular, the location of an observed beep or flash is predicted to be equal to the location of the actual beep or flash plus some additive noise sampled from a circular Gaussian or von Mises distribution (recall that locations lie on a circle). The key feature of these forward models are their noise distributions. Because vision is a more precise cue to spatial location than audition (Battaglia et al., 2003; Alais & Burr, 2004), the vision-specific forward model used a noise distribution with a significantly higher precision ($\kappa_V = 4.0$ roughly corresponding to a variance of $17^\circ$) than the noise distribution used by the audition-specific forward model ($\kappa_A = 2.5$ corresponding to a variance of $32^\circ$). The values of $\kappa_A$ and $\kappa_V$ were chosen on the basis of a trial-and-error search for values that allowed the model’s predictions to match our experimental data. This occurred whenever vision was a reasonably more precise cue to spatial location than audition—that is, the model’s performance was highly robust to the exact values chosen.

### 3.4 Learning and Cross-Modal Transfer

Having introduced the multisensory representations and sensory-specific forward models, we now describe how the model learns and transfers knowledge across auditory and visual modalities. We do so in the context of the experiment described above.

The model learns multisensory representations of sequence categories based upon its sensory input. As described above, the hypothesis space of category representations (i.e., the space of possible computer programs) is large. How should the model evaluate different hypotheses during learning? Here, we cast

\textsuperscript{1}In other cases, sensory-specific models can be complex. For example, Yildirim & Jacobs (2013) considered visual and haptic cues to object shape. In this case, the vision-specific forward model was a graphics library (e.g., OpenGL) and the haptics-specific forward model was a simulator of a human hand.
this problem as an instance of Bayesian inference.

For ease of exposition, we describe the model from the standpoint of a participant in Group V-A. Let \( V = \{ \vec{v}_1, \ldots, \vec{v}_N \} \) denote \( N \) visual sequences from one of the four categories (we exclude subscripts indexing categories to avoid unnecessary notation). Each \( \vec{v}_i \) is a vector of \( K_i \) spatial locations, where \( K_i \) is the length of the \( i \)th visual sequence. Thus, we write \( \vec{v}_i = [v_{i1}, \ldots, v_{iK_i}]^T \), and let \( v_{ij} \) denote the visual observation at the \( j \)th time step in sequence \( \vec{v}_i \).

The model learns multisensory representations from sensory data as follows. Cognitive models often assume that sensory data are the products of a generative process. In the context of our experiment, a visual sequence is generated when a multisensory representation for a sequence category produces a sequence of locations and this sequence is visually rendered. To learn about multisensory representations, this generative process can be inverted via Bayes’ rule:

\[
P(R|V) \propto P(R) p(V|R) = P(R) \prod_{i=1}^{N} \prod_{j=1}^{K_i} p(v_{ij}|R) \quad (3.1)
\]

where \( P(R) \) is the prior probability of multisensory representation \( R \), and \( p(V|R) \) is the likelihood function arising from the vision-specific forward model. We consider each of these quantities—the prior and the likelihood function—in turn.

As described above, multisensory representations are computer programs. For the purpose of assigning prior probabilities to programs, we characterize these programs using the probabilistic context-free grammar (PCFG) in Figure 3.5 (this general approach is adopted from Piantadosi, Tenenbaum, & Goodman, 2012). A particular program, \( R \), can be generated from the start symbol \( S \) by a derivation, a sequence of productions in the PCFG 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 com-
### Production rule

<table>
<thead>
<tr>
<th>Production rule</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>( S \rightarrow \text{init}(\text{L}) )</td>
<td>1.0</td>
</tr>
<tr>
<td>( U \rightarrow \text{L} )</td>
<td>0.25</td>
</tr>
<tr>
<td>( U \rightarrow \text{O} )</td>
<td>0.25</td>
</tr>
<tr>
<td>( U \rightarrow \text{L} )</td>
<td>0.25</td>
</tr>
<tr>
<td>( U \rightarrow \text{O} )</td>
<td>0.25</td>
</tr>
<tr>
<td>( \text{O} \rightarrow \text{go to [one of the earlier lines} )</td>
<td></td>
</tr>
<tr>
<td>( \text{in the current program} )</td>
<td>1.0</td>
</tr>
<tr>
<td>( \text{use equal probabilities]} )</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Figure 3.5: PCFG for multisensory representations of sequence categories.

A complete derivation is the product of the probabilities for these choices. In principle, the prior probability of a program should be defined as the sum of the probabilities of its possible derivations. However, derivations using our grammar are unique due to the structure of the grammar (there is, at most, only one non-terminal on the right-hand side of a production rule) and the consistent order in which we expand the non-terminals.

This prior probability distribution is similar (but not identical) to the prior distribution used by Goodman, Tenenbaum, Feldman, & Griffiths (2008). An advantage of this distribution is that it favors “simple” programs, meaning programs with short derivations. Consequently, it can be regarded as a type of Occam’s Razor.

The likelihood of a visual sequence, \( p(\vec{v}_i|R) \), was estimated as follows. The initial observed location, \( v_{i1} \), is an imperfect cue to the actual starting location of a sequence due to sensory noise. To deal with this uncertainty, we used the
vision-specific forward model to select the three most probable locations based upon the value of $v_{i1}$, and averaged the likelihood scores over these locations:

$$p(\vec{v}_i|R) = \frac{1}{3} \sum_{l_1 \in L} p(\vec{v}_i|R, l_1)$$  \hspace{1cm} (3.2)$$

where $l_1 \in L$ indexes the three most probable locations, and $p(\vec{v}_i|R, l_1)$ is the likelihood score of sequence $\vec{v}_i$ based on multisensory representation $R$ assuming that the sequence started at location $l_1$.\(^2\)

To compute $p(\vec{v}_i|R, l_1)$, we used program $R$ to generate a sequence. The initial location of this sequence was set to $l_1$. If the program was not capable of generating a sequence whose length is at least as long as $K_i$—the length of visual sequence $\vec{v}_i$—then the likelihood score was set to 0 (e.g., the program in Panel B of Figure 3.4 only generates sequences of length 3). Otherwise the program was used to generate a sequence of length $K_i$. Let $l_j$ denote the $j^{th}$ element of this sequence. The likelihood score of $p(\vec{v}_i|R, l_1)$ is computed using the vision-specific forward model as follows:

$$p(\vec{v}_i|R, l_1) = K_i \prod_{j=1}^{K_i} VM(v_{ij}|l_j, \kappa_V)$$  \hspace{1cm} (3.3)$$

where $VM(\cdot|l_j, \kappa_V)$ is the univariate von Mises probability density function with mean $l_j$ and precision $\kappa_V$. To simulate participants from Group A-V, the model is identical except that visual sequences $V = \{\vec{v}_1, \ldots, \vec{v}_N\}$ are replaced with auditory sequences $A = \{\vec{a}_1, \ldots, \vec{a}_N\}$, and visual precision $\kappa_V$ in Equation 3.3 is replaced with auditory precision $\kappa_A$.

Ideally, we would insert the prior distribution and likelihood function into Bayes’ rule (Equation 3.1) to compute the posterior distribution over multisensory representations. Unfortunately, computing the posterior distribution in this man-

\(^2\)We also considered a different likelihood function in which, instead of summing over possible initial locations, we searched for the initial location that maximized the likelihood score of $\vec{v}_i$ with respect to $R$. Our simulation results were qualitatively indistinguishable between these two alternatives.
ner is intractable, and thus we performed numerical simulations to search the space of multisensory representations. Specifically, we used a tree-based Monte Carlo Markov chain (MCMC) algorithm—a type of Metropolis-Hastings algorithm—based upon the algorithm in Goodman, Tenenbaum, Feldman, & Griffiths (2008).

The algorithm was initialized with a random multisensory representation by drawing a random derivation from the PCFG. This random representation was used as the current hypothesized program, also known as the current state of the Markov chain, at iteration 1. At each subsequent iteration, a proposal program was compared against the current hypothesized program. Proposals were generated as follows. A derivation of a program can be represented by a tree in which internal nodes represent non-terminals and leaf nodes represent terminals. A proposal program was formed by randomly perturbing the current program. A node from the derivation tree of the current program was randomly selected. The subtree below this node was deleted. Non-terminals in the remaining tree were then expanded using random choices of productions from the PCFG. Finally, a choice was made between the proposal and current program based on the Metropolis-Hastings acceptance function. The proposal was accepted, and thus became the new current program, with probability equal to the minimum of 1 and:

$$\frac{p(V|R') |R'|}{p(V|R) |R|}$$  

(3.4)

where $R'$ is the proposal, $|R'|$ is the number of non-terminals in the derivation of $R'$, $R$ is the current program, and $|R|$ is the number of non-terminals in the derivation of $R$.\(^3\) This process of randomly generating a proposal and stochastically choosing between the proposal and the current program was repeated for many iterations. The current programs from the final iterations (presumably following convergence of the algorithm) are samples from the posterior distribution.

---

\(^3\)As shown by Goodman et al. (2008), including the number of non-terminals terms, $|R|$ and $|R'|$, in the acceptance function ensures the detailed balance condition of the MCMC algorithm.
over multisensory representations.

3.5 Simulation Results

We trained the model in a manner analogous to the way that participants in our experiment were trained. Recall that the experiment had 2 groups (Groups A-V and V-A), with 9 participants per group. Each participant was trained with exemplars from 4 categories. Correspondingly, our model included 2 groups of simulations, one group for auditory training and the other for visual training, with 9 participant-level simulations per group. Each participant-level simulation consisted of 4 category-level simulations. A category-level simulation used the same sensory modality as its corresponding participant, and the same number of exemplars from a category as was observed by this participant during the experiment. For example, consider a category-level simulation corresponding to Category 1, Participant 1, Group A-V. This simulation was conducted using the same number of exemplars from Category 1 as were heard by Participant 1 in Group A-V during the experiment. To mimic sensory noise in our simulations, each location in an exemplar (i.e., a sequence of locations) was perturbed by adding a random number drawn from a von Mises distribution to the location. This was accomplished using the vision-specific or auditory-specific forward models described above.

Each category-level simulation was run for 150,000 iterations of the MCMC algorithm. Samples from the first 100,000 iterations were excluded as burn-in. Samples from the remaining 50,000 iterations were thinned to a set of 5,000 samples to reduce autocorrelations between samples. This set of 5,000 samples is referred to as the category-level simulation’s posterior sample.
3.5.1 Posterior distributions

Figure 3.6 illustrates our results. The first and second rows correspond to Groups A-V and V-A, respectively. The four columns correspond to categories 1-4. Each graph shows the posterior probabilities based upon the category-level simulations for a given group and category (there are 9 participant-level simulations per group, and thus there are 9 category-level simulations for a given group and category). The horizontal axis of a graph gives a program identification number. Each program appearing in the posterior sample was assigned a unique ID based on the rank of its posterior probability (the program with the largest probability was numbered 1, the program with the next largest probability was numbered 2, and so on). The vertical axis gives the posterior probability of a program in the posterior sample. For a given group and category, the probability distribution over programs was calculated as follows. For each category-level simulation, we first calculated each program’s unnormalized posterior score—defined as the product of a program’s likelihood score and its prior—and then normalized these scores. The normalized scores are a posterior probability distribution over programs for a given category-level simulation. These scores were then averaged across category-level simulations to arrive at the final estimate of a program’s posterior probability. The entropy (denoted $H$ and measured in bits), an information-theoretic measure of uncertainty (Cover & Thomas, 1991), of each posterior distribution is shown in each graph.

There are several important features of these data. First, posterior distributions are peaked around a single program. Although the hypothesis space of possible programs is infinite (i.e., the probabilistic context-free grammar can generate an infinite number of programs), our results indicate that only a small number of these programs have significant posterior probability for each group and category.

Second, the model shows perfect modality invariance. The bottom row of
Figure 3.6: First and second rows show posterior probabilities based upon the category-level simulations. The horizontal axis of each graph gives a program identification number, and the vertical axis gives the posterior probability of a program in the posterior sample. The entropy ($H$) of each posterior distribution is shown in each graph. The bottom row shows the program with the highest posterior probability (i.e., the MAP estimate) for each category when samples are combined across all participant-level simulations. The model shows modality invariance as evidenced by the fact that MAP estimates were identical for simulations of visual and auditory training groups.
Figure 3.6 shows the program with the greatest posterior probability for each category when samples are combined across all participant-level simulations. These programs are the model’s maximum a posteriori (MAP) estimates of the multisensory representations. Critically, the MAP estimates for each category are identical for simulations of visual and auditory training groups. That is, the model learns the same program regardless of the sensory modality used to perceive a category’s training exemplars. Moreover, for all categories, the MAP estimate is correct, meaning that it is identical to the actual program used to generate the exemplars. This result indicates that our model is very effective at learning multisensory representations from sensory data.

Lastly, the entropies (i.e., uncertainties) of the posterior distributions follow an interesting pattern. As expected, entropies are higher for auditory training than for visual training. More surprisingly, entropies are lowest for Category 1, highest for Category 3, and have intermediate values for Categories 2 and 4. This result suggests that learning about Category 1 should be easiest, learning about Category 3 should be hardest, and learning about Categories 2 and 4 should have intermediate levels of difficulty, consistent with our experimental data (see the right graph in Figure 3.3).

3.5.2 Training and Test Categorization Performances

We computed the model’s categorization performances as follows. Consider the model’s performance during visual training. For the moment, we focus on one participant-level simulation. Recall that a participant-level simulation consists of 4 category-level simulations. For each of these category-level simulations, we calculated the MAP estimate of a multisensory representation (i.e., for each category, we found the program with the highest frequency in the posterior sample). These MAP representations may be regarded as (point estimates of) the category representations acquired by a participant-level simulation. We used them
to classify individual training exemplars. Given an exemplar, we computed a posterior score for each MAP representation based solely on that exemplar. The computation of this score used the prior distribution and likelihood function described above (Section 3.4). The categorization response was taken to be the category corresponding to the MAP representation with highest posterior score. This process was repeated for each of the 36 visual exemplars (9 exemplars × 4 categories) used in a participant-level simulation. Analogous computations were performed during visual testing and during auditory training and testing (as in the experiment, testing included 14 exemplars × 4 categories).

Figure 3.7 shows the model’s categorization performances. The left panel illustrates the training and test results. On average, the participant-level simulations corresponding to Group V-A performed at more than 95% correct on visual training exemplars. When tested with auditory test exemplars, the simulations showed excellent cross-modal transfer, performing at nearly 85% correct (recall that chance performance is 25%). Participant-level simulations corresponding to Group A-V performed at 85% correct on auditory training items, and more than 90% correct on visual test items, meaning that this group too showed excellent cross-modal transfer. We emphasize the match between our experimental (left panel of Figure 3.3) and modeling (left panel of Figure 3.7) results.

The right panel of Figure 3.7 shows the participant-level simulations’ training and test performances when trials are sorted by sequence category. This type of analysis was discussed above in the context of our experimental data (right panel of Figure 3.3), and is useful because it allows us to examine the relative ease of correctly classifying exemplars from each category. Recall that experimental participants performed best with exemplars from Category 1, worst with exemplars from Category 3, and at intermediate levels with exemplars from Categories 2 and 4. Does our model show this same rank ordering of category difficulty? The answer is yes. Based upon the participant-level simulations’ average performances,
Figure 3.7: Modeling results presented in the same format as our experimental results (Figure 3.3). (Left) Average performances of the participant-level simulations on the training and test exemplars for simulations corresponding to Groups V-A and A-V (error bars indicate standard errors of the means). (Right) Participant-level simulations’ training and test performances when trials are sorted by the category of the sequence observed on the trial.
Figure 3.8: Results from the model in which multisensory MAP representations are evaluated based on a likelihood score, not a posterior score.

the rank ordering of the category difficulties parallel the behavioral results.4 A Friedman test revealed a statistically significant rank ordering of the categories based upon training ($p < 0.01$) and test ($p < 10^{-4}$) block responses.

### 3.5.3 Role of the prior in accounting for participants’ performances

When considering our model, we would like to evaluate the role of the prior probability distribution. To do so, we could attempt to develop an alternative model that does not include a prior (i.e, a model in which all multisensory representations are equally probable). In our case, however, this would be difficult to do, particularly when training the model to acquire multisensory representations. As described above, our model learns multisensory representations via an MCMC

---

4Rather than use MAP estimates of multisensory representations, an alternative strategy is to randomly sample from the posterior distribution over representations. This alternative strategy provides very similar performances to those shown in Figure 3.7. This result was expected because posterior distributions tend to have small variances.
algorithm in which proposals are generated at each iteration by sampling from
the prior distribution. This is where a significant problem arises. If we did not
have a prior distribution, then where would proposals come from? Recall that
sampling from the prior occurs through the use of the production rules in the
PCFG. Even if we could devise a scheme in which all derivations from the PCFG
are equally probable (it is not clear that we could), then this would likely lead to
other challenges, such as creating MCMC algorithms that converge.

The model uses the prior distribution both when acquiring multisensory represen-
tations and when using these representations to categorize an exemplar. Given
that it may be impossible to ignore the prior during the acquisition stage, we
decided to focus on what would happen if we ignored the prior during the catego-
ration stage. When categorizing an exemplar, our model calculated a posterior
score for each multisensory MAP representation, and labeled the exemplar based
on the representation with the highest score. We wondered what would happen
if we did not calculate a posterior score but, rather, calculated a likelihood score.
That is, what if MAP representations were evaluated based solely on the likelihood
function, ignoring the prior distribution?

Figure 3.8 shows the results. Clearly, the model’s results are now less similar
to the experimental results. For example, differences between training and test
performances are now greatly reduced. Furthermore, the rank ordering of cate-
gory difficulties observed in the behavioral data and in the original model’s data is
+missing from this model’s data. Perhaps the most striking aspect of this model’s
data is that the performances are so high. That is, ignoring the prior distribution
when evaluating multisensory MAP representations leads to better performances.
At the same time, it also makes the model’s performances less similar to experi-
mental participants’ performances, suggesting that people, like our model, might
also have a bias toward “simpler programs”. Future work will need to address
this hypothesis.
3.6 Discussion

In summary, our goal has been to use the Multisensory Hypothesis to better understand how people acquire and use multisensory representations to facilitate transfer of knowledge across sensory modalities. We conducted an experiment evaluating whether people transfer sequence category knowledge across auditory and visual domains. Our experimental data clearly indicate that we do. We then developed a computational model accounting for our experimental results. To our knowledge, this is among the first formulations of the Multisensory Hypothesis that has been explicitly defined and implemented (also see Yildirim & Jacobs, 2013). Because our model demonstrates how the acquisition and use of amodal, multisensory representations can underlie cross-modal transfer of knowledge, and because our model accounts for subjects’ performances, our work lends credence to the Multisensory Hypothesis. Overall, our work suggests that people automatically extract and represent objects’ and events’ intrinsic properties, and use these properties to process and understand the same (and similar) objects and events when they are perceived through novel sensory modalities.

*pLOT approach beyond higher-level cognition:* Multisensory representations lie at the core of our computational model. An unusual aspect of the model is that these representations are characterized as computer programs, and programs are learned via Bayesian inference. As discussed above, our work contributes to the emerging pLOT perspective. Symbolic and statistical approaches to cognitive modeling often have complementary strengths and weaknesses. A strength of symbolic approaches is their representational expressiveness which comes from their use of highly structured, compositional data structures. However, symbolic approaches are often “brittle” (i.e., they often fail in uncertain environments) and often have limited learning capabilities. In contrast, statistical approaches tend to be robust in the sense that they often work well despite uncertainty. In addi-
tion, these approaches can excel at inference and learning, especially when using new computational techniques developed in the past 25 years (e.g., new Monte Carlo sampling methods or variational approximations). However, statistical approaches often require highly structured prior distributions or likelihood functions to work well (Tenenbaum, Kemp, Griffiths, & Goodman, 2011). By combining the strengths of symbolic and statistical approaches, the pLOT perspective may offer a unifying framework for thinking about many aspects of human cognition. Our work extends the small (but hopefully growing) literature on the pLOT modeling approach. To our knowledge, our model is among the first pLOT models to address the domain of human perception.

**Alternative models and the Direct Association Hypothesis:** The computational model described here is an appealing implementation of the Multisensory Hypothesis. Nonetheless, the specific details of the model may, ultimately, be less important than its general framework. The key elements of this framework are a vocabulary for expressing multisensory representations, and sensory-specific forward models which relate multisensory representations to sensory signals. In principle, one could implement these elements in other ways, perhaps using a purely statistical approach, a purely symbolic approach, or an alternative hybrid approach. A challenge for researchers interested in pursuing other implementations will be to account for the patterns in our experimental data (e.g., subjects showed cross-modal transfer, subjects showed differing levels of performance for different sequence categories).

Can models that do not make use of multisensory representations account for cross-modal transfer? Consider the Direct Association Hypothesis. According to this hypothesis, cross-modal transfer occurs because sensory representations from different modalities are directly associated with each other. For example, when a person is visually trained to categorize event sequences, the person may map visual sequences to corresponding auditory sequences, perhaps a type of auditory
imagery. In this case, the person will form two sensory representations on each training trial, a visual representation formed on the basis of a perceived visual sequence and an auditory representation formed on the basis of the imagined auditory sequence. Cross-modal transfer occurs when the person applies the same category label to both sensory representations.

We are not aware of any implemented models based on the Direct Association Hypothesis. It is, therefore, hard to evaluate this hypothesis. However, we suspect that the Direct Association Hypothesis will be difficult or impossible to implement without some multisensory aspect to it. For example, it is not clear how visual signals could be mapped to auditory signals using only unisensory operations and representations. An artificial neural network, for instance, would require hidden units to learn about the statistical relations among visual and auditory signals. The training signals underlying such learning would, by necessity, be multisensory in nature. Moreover, an analysis of the receptive and projective fields acquired by the hidden units would reveal that these units are forms of multisensory representations. Future work will need to consider possible implementations of the Direct Association Hypothesis and whether the Direct Association and Multisensory Hypotheses are truly distinct. If they are, note that these hypotheses are not mutually exclusive and, ultimately, models implementing aspects of both hypotheses may provide the best accounts of human behavior.

Extensions of the current model: Our model is an instance of a “single cause” model because it assumes that visual and auditory signals arise from a single source (i.e., a sequence of events). In the real world, however, visual and auditory signals sometimes arise from the same source and other times arise from different sources. People are able to learn if different sensory signals should be attributed to the same or different underlying causes. Future extensions of the current model will need to learn this too (Körding et al., 2007).

Future extensions will also need to consider how the model can be scaled to
larger, more realistic scenarios. In more realistic settings, richer sets of representational primitives will be needed, as well as more sophisticated forward models. We are encouraged by the fact that researchers are developing advanced software for perceptual (e.g., visual, auditory) rendering, for simulating the kinematics and dynamics of robots, and for simulating dynamic interactions among objects. Cognitive scientists can build larger, more realistic models by using these software packages as forward models in their models of human perception, motor control, and intuitive physics (see Battaglia, Hamrick, & Tenenbaum, 2013; Yildirim & Jacobs, 2013; ).

Computational and representation/algorithm levels of analysis: Cognitive models are often classified based on whether they contribute to Marr’s (1982) computational or representation/algorithm level of analysis. We believe that our model currently makes a contribution at the computational level and may, in the future, make a contribution at the representation/algorithm level. At the computational level, our model defines optimal performance on our experimental task (given the assumptions of the model; see Jacobs & Kruschke, 2011). Therefore, it can be used as a benchmark to evaluate subjects’ performances. Subjects performed correctly on about 70-75% of final training and test trials. In addition, they performed best on exemplars from Category 1, worst on exemplars from Category 3, and at intermediate levels on exemplars from Categories 2 and 4. Are these performances good or bad? By comparing subjects’ performances with those of the computational model, we see that subjects’ performances are similar to those of the model, though subjects are moderately less proficient. This indicates that subjects performed well, but that there was still room for improvements in these performances. The gap between subjects’ performances and the model’s performances may have been due to our training procedures. Future work will need to investigate this issue.

Our model can potentially be used as a starting point for a new model in-
tended to faithfully mimic people’s psychological operations and representations underlying cross-modal transfer, thereby making a contribution at the representation/algorithm level of analysis. As discussed above, the symbolic operators in our model’s programming language are similar to those proposed by Simon and Kotovsky (1963). These authors clearly believed that they were modeling people’s mental representations. If so, then our model can also be regarded as a realistic hypothesis about people’s representations. However, our Bayesian inference algorithm is not psychologically plausible, primarily due to its simultaneous use of all data items at each iteration. Interestingly, researchers have argued that inference algorithms of the type used here can often be replaced with closely related on-line algorithms (known as particle filter or sequential Monte Carlo algorithms) that are psychologically plausible because they use only one data item at a time (Sanborn, Griffiths, & Navarro, 2010; Griffiths, Vul, & Sanborn, 2012). Future work will need to study the benefits of extending our model in this manner to create a so-called “rational process model” (Sanborn, Griffiths, & Navarro, 2010; Griffiths, Vul, & Sanborn, 2012).
4 A Rational Analysis of the Acquisition of Multisensory Representations

4.1 Introduction

Much of the history of perceptual science can be characterized as a “sense by sense” approach in which each sensory modality is studied in isolation. For example, visual scientists study behavioral, cognitive, and neural aspects of visual perception, whereas auditory scientists study behavioral, cognitive, and neural aspects of auditory perception. However, it is an undeniable, but often overlooked, fact that perception is fundamentally multisensory. We learn about our environments by seeing, hearing, touching, tasting, and smelling these environments. Moreover, evolution has designed our brains so that our senses work in concert. Objects and events can be detected rapidly, identified correctly, and responded to appropriately because our brains use information derived from different sensory channels cooperatively. This fundamental property of human perception is a key reason that human intelligence is so robust.

To date, most computational work on multisensory perception focuses on the problem of sensory integration: how to combine information from two or more
sensory modalities to maximize performance on a task (e.g., Abidi & Gonzalez, 1992; Alais & Burr, 2004; Battaglia, Jacobs, & Aslin, 2003; Clark & Yuille, 1990; Gepshtein & Banks, 2003). Ernst and Banks (2002), for example, examined how people estimate the height of an object based on visual and haptic inputs. They compared people’s judgements with the predictions of a statistically optimal rule based on maximum likelihood estimation theory (the Fuzzy Logical Model of Perception of Massaro, 1998, provides a similar framework for sensory integration). According to this rule, an observer first forms two estimates of object height, one based on the visual input and the other based on the haptic input. Next, the observer forms an estimate of object height based on both sets of inputs by taking a weighted average of the estimates based on the individual inputs.

Although sensory integration is an important aspect of multisensory perception, it is only one of many aspects of multisensory perception. Here, we take a broader approach to the study of multisensory perception. We are interested in providing a single, unified account of a wide variety of multisensory phenomenon. We are not interested in providing an account solely of sensory integration, or of providing multiple accounts, one for each phenomenon that needs to be explained. Consequently, we emphasize the need to understand multisensory representations that can underlie many types of multisensory behaviors.

Specifically, we focus on the problem of learning multisensory, or amodal, representations from unisensory data. This focus distinguishes our work from both the maximum likelihood approach to sensory integration described above and from the Fuzzy Logical Model of Perception, neither of which is concerned with the acquisition of complex multisensory representations. The representations we seek are task-independent. That is, they are not acquired to facilitate performance on a particular task. Instead, they are acquired to facilitate performances on many different tasks. When performing a particular task, the acquired representations can serve as key components of a larger system that also includes components
that are task-specific. Our emphasis on task-independent multisensory representations is not intended to deny the existence of task-specific representations. To the contrary, we think that both task-independent and task-specific representations will play important roles in comprehensive theories of perception. Rather, our emphasis on task-independent representations allows us to focus on general theoretical principles with broad explanatory power.

The computational model that we propose uses a probabilistic framework, specifically a Bayesian framework, for studying the acquisition of multisensory representations. Bayesian modeling has become increasingly important in the field of cognitive science (e.g., Anderson, 1990; Barlow, 1959; Chater & Oaksford, 2008; Geisler, 2004; Green & Swets, 1966; Griffiths, Kemp, & Tenenbaum, 2008; Kahneman, Slovic, & Tversky, 1982; Knill & Richards, 1996; Marr, 1982; Oaksford & Chater, 1999; Todorov, 2004). A common observation of cognitive scientists is that we live in an uncertain world, and rational behavior depends on the ability to process information effectively despite ambiguity or uncertainty. Cognitive scientists, therefore, need methods for characterizing information and the uncertainty in that information. Fortunately, such methods are available: probability theory provides a calculus for representing and manipulating uncertain information. To us, an advantage of Bayesian models relative to many other types of models is that they are probabilistic.

Probability theory doesn’t provide just any calculus for representing and manipulating uncertain information, it provides an optimal calculus (Cox, 1961). Consequently, an advantage of Bayesian modeling is that it gives cognitive scientists a tool for defining rationality. Via Bayes’ rule, Bayesian models optimally combine information based on prior beliefs with information based on observations or data. Via Bayesian decision theory, Bayesian models can use these combinations to choose actions that maximize task performance. Due to these optimality properties, Bayesian models perform a task as well as the task can be performed
(given the assumptions built into the model), meaning that the performance of a Bayesian model on a task defines rational behavior for that task (again, based on the model’s assumptions).

Our model is motivated by the need for a computational theory (Marr, 1982) or rational analysis (Anderson, 1990) of the learning problem facing humans in complex multisensory environments. This approach emphasizes that important properties of cognition can be understood by characterizing the computational demands of a natural environment, and theorizing that evolution has shaped the brain so that it efficiently meets these demands. By studying multisensory behavior from this perspective, one can examine the properties of multisensory perception that emerge as part of the optimal solution to the problem of learning to perceive and represent multisensory environments. One can then compare human performances in multisensory settings to the properties predicted by the optimal model, and explain why these properties exist in human behavior. Cognitive scientists are increasingly using optimal models to study human cognition. Analyses based on optimal performance are referred to as rational analyses, ideal observer analyses, or ideal actor analyses in the literatures on cognition, perception, and motor control, respectively (e.g., Anderson, 1990; Geisler, 2004; Todorov, 2004).

This paper is organized as follows. Section 2 describes four hypotheses about multisensory perception that have appeared in the scientific literature. We highlight these hypotheses because they play important roles in our computer simulations. Sections 3 and 4 provide preliminary remarks regarding the proposed model and a detailed description of the model, respectively. Sections 5-7 provide simulation results on three data sets: a synthetic data set, a real-world visual-auditory data set, and a real-world visual-haptic data set. The final section provides a general discussion and concluding remarks.
4.2 Four Hypotheses About Multisensory Perception

Cognitive neuroscientists have recently studied at least four important hypotheses about multisensory perception. First, researchers have conjectured that multisensory representations are essential because sensory integration provides significant statistical advantages (e.g., Abidi & Gonzalez, 1992; Alais & Burr, 2004; Battaglia, Jacobs, & Aslin, 2003; Clark & Yuille, 1990; Ernst & Banks, 2002; Gepshtein & Banks, 2003).

For example, sensory integration can ameliorate the effects of noise contained in representations based on single modalities. Multisensory representations are, therefore, able to convey more accurate and reliable information than the unisensory representations from which they are derived. Consider an observer that sees and touches a surface slanted in depth. Suppose that the observer’s slant estimates based on the visual cue and on the haptic cue are each corrupted by sensory noise with some variance. It is easily shown that the statistically optimal estimate of surface slant obtained by combining information from both cues has a lower variance, and is thus more reliable, than estimates based on either cue alone.

Evidence that the brain is able to combine sensory information in such a manner has been obtained by several researchers. Ernst and Banks (2002) found that people’s estimates of object height based on both visual and haptic information were more reliable (i.e., had smaller variances) than their estimates based on either visual or haptic information alone. Alais and Burr (2004) accounted for the “ventriloquist effect” by showing that people combine visual and auditory information in a statistically optimal manner when judging the location of events. Fetsch, DeAngelis, and Angelaki (2010) reviewed evidence that people integrate visual and vestibular heading cues in a manner consistent with optimal integration theory. Wozny, Beierholm, and Shams (2008) showed that people combine
visual, auditory, and tactile information optimally when performing a numerosity judgment task.

A second hypothesis studied by cognitive neuroscientists is that training in multisensory environments leads to better learning than training in unisensory environments, and that this advantage holds even when testing is performed in unisensory conditions. Shams and Seitz (2008) argued that unisensory training is less efficient than multisensory training because it is unnatural, and thus fails to tap into multisensory learning mechanisms that have evolved to produce optimal behavior in the naturally-occurring multisensory environment.

Several investigators have reported experimental results supporting this hypothesis. Seitz, Kim, and Shams (2006) trained subjects to perform a motion detection task based on visual input or based on both visual and auditory input. When tested with visual input alone, subjects trained in the multisensory environment performed significantly better, both on the first day of learning and across all ten experimental sessions. Von Kriegstein and Giraud (2006) trained subjects to recognize voices based on auditory signals or based on auditory signals coupled with visual images of faces. Subjects trained in the multisensory setting performed significantly better even when test trials contained only auditory information. Lehmann and Murray (2005) asked subjects to indicate when an image appeared that had previously been presented during the experiment. Even though sound was never presented when an image recurred, images that previously appeared with congruent sounds (e.g., an image of a bell and the sound ‘dong’) were recognized better than those that had previously been presented either without sound or with incongruent sounds (e.g., an image of a bell and the sound ‘meow’).

A third hypothesis recently proposed by cognitive neuroscientists is that our neural representations of environmental events are often modality invariant, meaning they are the same (or at least similar) regardless of the sensory modalities through which we perceive those events. From a computational perspective,
modality invariance is a desirable property. It will be easier to recognize, reason, and learn about an event if the event has the same representation regardless of the sensory modality through which it is perceived.

Evidence consistent with this hypothesis has been obtained by several researchers. Konkle, Wang, Hayward, and Moore (2009) found that motion aftereffects transfer between vision and touch, suggesting the existence of shared representations of motion. Amedi et al. (2001) showed that a neural region known as the lateral occipital complex (LOC) shows similar patterns of activation regardless of whether an object is seen or touched. Quiroga, Kraskov, Koch, and Fried (2009) found single neurons in human brains that respond selectively to the same individual regardless of whether an observer sees a photograph of the individual, reads the name of the individual, or hears the individual’s voice.

Lastly, researchers have hypothesized that representations based on different modalities are associated with each other. Consider an observer that sees, but does not hear, an environmental event. A visual representation of that event will be active in the observer’s brain, and this representation will often predict or activate an auditory representation of the event even though the event is not heard. From a computational viewpoint, predictions of one modality’s sensory representations based on another modality’s representations might be useful top-down information leading to faster or more accurate processing in the first modality. These predictions may also help the observer efficiently allocate attention in this modality.

Data consistent with this hypothesis has appeared in the literature. Calvert et al. (1997) reported that viewing facial movements associated with speech (lipreading) leads to activation of auditory cortex in the absence of auditory speech sounds (see also Pekkola et al., 2005; Tanabe, Honda, & Sadato, 2005). Zhou and Fuster (2000) found sustained activity in primary somatosensory cortex during a delay period after presentation of a visual stimulus previously associated with a tac-
tile stimulus. Sathian et al. (1997) found that primary visual cortex (area V1) is active when observers perform a tactile discrimination task involving oriented gratings. Zangaladze et al. (1999) argued that V1 is crucial for tactile discrimination because disruption of V1 activation using transcranial magnetic stimulation impairs performance on this tactile task.

In summary, this section has reviewed four hypotheses\(^1\) about multisensory perception that will be addressed by the computer simulations reported below:

- Multisensory representations are essential because sensory integration provides significant statistical advantages;
- Training in multisensory environments leads to better learning than training in unisensory environments, even when testing is performed in unisensory conditions;
- Neural representations of events are often modality invariant, meaning they are the same (or at least similar) regardless of the sensory modalities through which the events are perceived; and
- Representations based on different modalities are associated with each other.

### 4.3 Preliminary Remarks Regarding the Proposed Model

To date, there has been relatively little research on how people acquire multisensory representations. An important hypothesis is that multisensory representations are not necessarily logically distinct because Hypothesis 2 (training in multisensory environments leads to better learning than training in unisensory environments) is a special case of Hypothesis 1 (sensory integration provides significant statistical advantages), and because Hypothesis 3 (neural representations of events are often modality invariant) is a special case of Hypothesis 4 (representations based on different modalities are associated with each other).

\(^{1}\)An anonymous reviewer noted that the four hypotheses are not necessarily logically distinct because Hypothesis 2 (training in multisensory environments leads to better learning than training in unisensory environments) is a special case of Hypothesis 1 (sensory integrations provides significant statistical advantages), and because Hypothesis 3 (neural representations of events are often modality invariant) is a special case of Hypothesis 4 (representations based on different modalities are associated with each other).
tations, especially representations of complex perceptual events, develop slowly. Unisensory representations develop first, and multisensory representations are acquired later based on statistical correlations among the unisensory representations (Alais, Newell, & Mamassian, 2010). At least in part, this hypothesis is motivated by neuroscientific evidence obtained in physiological investigations of the superior colliculus (SC), a region found in mammalian brains in which sensory integration has often been studied. For example, Wallace and Stein (2001) found that some superior colliculus (SC) neurons in newborn monkeys respond to signals from multiple sensory modalities, but that these neurons do not synthesize and represent complex multisensory events until later in life. These authors wrote, “These data, coupled with those from cat, suggest that the capacity to synthesize multisensory information does not simply appear in SC neurons at a prescribed maturational stage but rather develops only after substantial experience with cross-modal cues.”

To this main hypothesis, we add the conjecture that the acquisition of multisensory representations is often accomplished in an unsupervised or task-independent manner. This conjecture is consistent with (but not proven by) recent observations by Lacey, Hall, and Sathian (2010) who found that the performances of subjects performing a shape discrimination task were impaired by changes to objects’ task-irrelevant surface properties such as surface texture. The authors concluded that our multisensory representations integrate shape and modality-independent surface properties. If so, this result implies that task-irrelevant features, such as surface texture, are represented in multisensory object representations, thereby suggesting that these representations are acquired in an unsupervised or task-independent manner.

At an intuitive level, the idea that unisensory representations develop first and multisensory representations develop later (in an unsupervised manner) based on statistical correlations among the unisensory features is reasonable and appealing. Nonetheless, its lack of detail makes it difficult to rigorously understand, evaluate,
and extend. What predictions, if any, does this hypothesis make about multisensory perception? Does it predict that sensory integration provides significant functional advantages? Does it predict that multisensory training will be better than unisensory training even when testing is conducted in unisensory conditions? Does it predict the existence of modality-invariant multisensory representations? Does it predict the existence of associations among unisensory representations?

Guided by the main hypothesis and conjecture, we propose a model, referred to as the multisensory perception model, of the acquisition of multisensory representations. This model represents a novel approach to modeling the acquisition of multisensory representations, complementary to previous models that explored how acquisition might be algorithmically or neurally implemented. In the traditions of “ideal observer analysis” (Barlow, 1959; Geisler, 2004) or “rational analysis” (Anderson, 1990; Chater & Oaksford, 1999; Marr, 1982), we consider the abstract computational problem of learning multisensory representations that explain unisensory features in a generative or statistical manner described below, and show that several aspects of multisensory perception emerge as part of the optimal solution to this learning problem. Although the artificial intelligence, cognitive science, and computational neuroscience literatures contain several computational models (including Bayesian models) of sensory integration (e.g., Abidi & Gonzalez, 1992; Alvarado, Rowland, Stanford, & Stein, 2008; Anastasio, Patton, & Belkacem-Boussaid, 2000; Ernst & Banks, 2002; Hershey & Movellan, 1999; Landy, Maloney, Johnston, & Young, 1995; Massaro, Cohen, Campbell, & Rodriguez, 2001; Pouget, Deneve, & Duhamel, 2002), our model is, to our knowledge, the first attempt focusing on the problem of learning complex multisensory representations from a rational perspective (Anderson, 1990; Chater & Oaksford, 1999; Marr, 1982).

The model can be regarded as an implementation of the main hypothesis and conjecture described above for the purpose of addressing open questions about
their meanings, implications, and potential extensions. It is a model of the acquisition of multisensory representations that learns multisensory representations by learning about the statistical correlations among unisensory features. Moreover, it learns in an unsupervised manner.

In unsupervised learning, the data provided to a learner are unlabeled. The goal of the learner is to discover patterns and structure within the data set. There is a dichotomy in the cognitive science and machine learning literatures between parametric and nonparametric unsupervised learning methods. A parametric method uses a fixed representation that does not grow structurally as more data are observed. In contrast, nonparametric methods use representations that are allowed to grow structurally. These methods are advantageous when the goal is to impose as few assumptions as possible. For example, Dirichlet process mixture models are nonparametric models used to cluster data items (Ferguson, 1973; Neal, 2000; Rasmussen, 2000). However, unlike their parametric counterparts such as conventional Gaussian mixture models, they do not make assumptions about the number of clusters from which a set of data items are drawn. Similarly, Indian buffet processes are nonparametric models used to learn latent or hidden features underlying a set of observable variables (Griffiths & Ghahramani, 2005, 2006). Unlike their parametric counterparts such as factor analysis models, they do not make assumptions about the number of latent or hidden features best characterizing the observable variables. In this sense, Bayesian nonparametric techniques are said to “let the data speak for themselves” (Blei, Griffiths, & Jordan, 2010).

While both parametric and nonparametric methods have important roles to play in the study of human cognition, Bayesian nonparametric methods are appealing for our purposes because they have both the advantages of probabilistic methods, due to their foundations in Bayesian statistics, and the advantages of flexible representations, due to their nonparametric nature. With respect to representational flexibility, we regard the Bayesian nonparametric approach as an
important advance over conventional parametric approaches in which a researcher sets, for instance, the number of latent variables by hand, often in an ad hoc or unprincipled manner (Blei, Griffiths, & Jordan, 2010). How can a researcher be sure that the number of latent features should, for example, be exactly 10? Shouldn’t the number of latent features be determined by the structure of the task or data set? We also regard the Bayesian nonparametric approach as an advance over modeling approaches that define a set of models, each with a different number of latent features, for instance, and perform “model comparison” to select the best model (see Navarro, Griffiths, Steyvers, and Lee, 2006, for additional discussion of the relationships between Bayesian nonparametric approaches and approaches based on model comparison). Typical model comparison techniques are computationally expensive and, thus, only practical for comparing small numbers of models. How should a researcher pick a small number of models to consider? The Bayesian nonparametric approach eliminates (or at least ameliorates) the problems associated with model comparison.

The proposed model is an extension of a Bayesian nonparametric method known as the Indian buffet process (Griffiths & Ghahramani, 2005, 2006). It “explains” the representations arising from individual sensory modalities through the use of a set of latent variables representing multisensory information. Importantly, the number of latent variables is not fixed. Instead, this number is treated as a random variable whose probability distribution is estimated based on the unisensory data. Because the size of the latent multisensory representation is estimated from the observed unisensory data, nonparametric statistical methods are required for inference. As noted by Austerweil and Griffiths (2009), Bayesian nonparametric methods are particularly appropriate for the study of perceptual learning. It is known that people do not use a fixed number of perceptual features.

As a rule, this statement is correct. However, there are exceptions to this rule. That is, there exist parametric methods which infer the dimensionality of representations. The interested reader should see Green and Richardson (2001) and Rasmussen and Ghahramani (2001).
tures (Goldstone, 1998). Instead, people create new features, at least in part, by combining or differentiating existing features. Goldstone (1998) referred to these processes as unitization and differentiation, respectively. Hence, any plausible model of perceptual learning must allow the number of perceptual features to adapt.

The Indian buffet process does not make assumptions about the exact number of latent features underlying a finite set of observed objects, although it does make other assumptions about these features (Griffiths & Ghahramani, 2005, 2006). It assumes that the latent features are binary. Thus, an object either does or does not possess a feature. It also assumes that latent features are statistically independent, meaning that knowledge that an object possesses one feature does not provide information about whether it possesses other features. Lastly, it assumes that the latent features are a finite subset of an unbounded or infinite set of features.

### 4.4 Multisensory Perception Model

We describe the proposed model here in the context of a visual-auditory environment, though we note that the model is equally applicable to other sensory modalities and to any number of modalities (it can also be used when multiple cues arise within a single modality, such as visual stereo, visual motion, and visual shading cues). A coarse schematic of the model is illustrated on the left side of Figure 4.1. It contains three sets of nodes or variables corresponding to visual features, auditory features, and multisensory features. The visual and auditory features are statistically dependent. However, they are conditionally independent given values for the multisensory features. The values of the visual features are observed when an object is viewed. When an object is not viewed, the visual features are latent, and their distributions can be inferred. Similarly, the values of
the auditory features are observed when an object is heard. Otherwise, the auditory features are latent, and their distributions can be inferred. The multisensory features are always latent variables. Whereas the numbers of visual and auditory features are fixed, the number of multisensory features is not. Consistent with the nonparametric approach, this number is a random variable whose distribution is inferred from the data.

Formally, the model is a straightforward extension of the Indian buffet process (Griffiths & Ghahramani, 2005, 2006). A detailed graphical representation of the model is shown on the right side of Figure 4.1. An important goal of the model is to find a set of latent multisensory features, denoted $Z$, “explaining” a set of observed visual and auditory features, denoted $X_V$ and $X_A$, respectively. Assume that a learner both sees and hears a number of objects. Let $Z$ be a binary multisensory feature ownership matrix, where $Z_{ij} = 1$ indicates that object $i$ possesses multisensory feature $j$. Let $X_V$ and $X_A$ be real-valued visual and auditory feature matrices, respectively (e.g., $X_{Vij}$ is the value of visual feature $j$ for object $i$). The problem of inferring $Z$ from $X_V$ and $X_A$ can be solved via Bayes’ rule:

$$p(Z|X_V, X_A) = \frac{p(X_V|Z) p(X_A|Z) p(Z)}{\sum_{Z'} p(X_V|Z') p(X_A|Z') p(Z')} \quad (4.1)$$
where \( p(Z) \) is the prior probability of the multisensory feature ownership matrix, and \( p(X_V|Z) \) and \( p(X_A|Z) \) are the likelihoods of the observed visual and auditory feature matrices, respectively, given the multisensory features. We now describe the prior and likelihood distributions.

The multisensory feature ownership matrix is assigned a Bayesian nonparametric prior distribution known as the Indian buffet process (Griffiths & Ghahramani, 2005, 2006). It can be interpreted as a probability distribution over feature ownership matrices with an unbounded (infinite) number of features. The distribution is written as:

\[
p(Z) = \frac{\alpha^K}{\prod_{h=1}^{N-1} K_h!} \exp\{-\alpha H_N\} \prod_{k=1}^{K} \frac{(N - m_k)!(m_k - 1)!}{N!}
\]

(4.2)

where \( N \) is the number of objects, \( K \) is the number of multisensory features, \( K_h \) is the number of features with history \( h \) (the history of a feature is the matrix column for that feature interpreted as a binary number), \( H_N \) is the \( N \)th harmonic number, \( m_k \) is the number of objects with feature \( k \), and \( \alpha \) is a variable influencing the number of features. (We did not set \( \alpha \) to a fixed value in our simulations described below. To make the model more flexible, \( \alpha \) was assigned a vague prior distribution, and its posterior distribution was inferred.)

The visual and auditory likelihoods are each based on a linear-Gaussian model. Let \( z_i \) be the multisensory feature values for object \( i \), and let \( x_{i\beta} \) be the feature values for object \( i \) where \( \beta \) is set to either \( V \) or \( A \) depending on whether we are referring to visual or auditory features. Then \( x_{i\beta} \) is drawn from a Gaussian distribution whose mean is a linear function of the multisensory features, \( z_i W_\beta \), and whose covariance matrix equals \( \sigma_{X_\beta}^2 I \), where \( W_\beta \) is a weight matrix (the weight matrices themselves are drawn from zero-mean Gaussian distributions with
covariance $\sigma^2_{W\beta} I$). Given these assumptions, the likelihood for a feature matrix is:

$$p(X_{\beta}|Z, W_{\beta}, \sigma^2_{X\beta}) = \frac{1}{(2\pi \sigma^2_{X\beta})^{ND_{\beta}/2}} \exp\left\{ -\frac{1}{2\sigma^2_{X\beta}} \text{tr}\left( (X_{\beta} - ZW_{\beta})^T (X_{\beta} - ZW_{\beta}) \right) \right\}$$

(4.3)

where $D_{\beta}$ is the dimensionality of $X_{\beta}$, and $\text{tr}(\cdot)$ denotes the trace operator.

Exact inference in the model is computationally intractable and, thus, approximate inference must be performed using Markov chain Monte Carlo (MCMC) sampling methods (Gelman et al., 1995; Gilks, Richardson, & Spiegelhalter, 1996). In our simulations, we used the MCMC sampling algorithms of Griffiths and Ghahramani (2005).

### 4.5 Synthetic Data Set

This section applies the multisensory perception model to a synthetic data set. Our goal is to illustrate for the reader that the model performs sensory integration in a sensible manner. Toward this goal, we consider an experiment by de Gelder and Vroomen (2000). This experiment is useful for our purposes because it demonstrates important properties of human sensory integration in a straightforward and clear fashion.

De Gelder and Vroomen (2000) studied how people integrate visual and auditory information when judging the emotional content of facial expressions. They created a continuum of images of a person’s face ranging from a happy face at one end to a sad face at the other end. They also created a continuum of voices ranging from a happy voice to a sad voice. On each trial, subjects saw a face and heard a voice, and judged whether the facial expression was happy or sad. The results show that the voice influenced subjects’ categorizations. When a happy voice was paired with a facial expression, subjects were more likely to judge the face as happy than when the face was paired with a neutral or sad voice. Similarly,
subjects were more likely to judge a face as sad when it was paired with a sad voice than when it was paired with a neutral or happy voice. The authors referred to this result as an “emotional McGurk effect” (de Gelder & Vroomen, 2003).

Does the multisensory perception model show an analogous effect? To address this question, we created a synthetic set of stimuli which is meant to be similar in spirit to the stimuli used by de Gelder and Vroomen (2000). We defined a line in a two-dimensional space, and defined nine categories along this line. The category prototypes were located at \([4, 4], [3, 3], \ldots, [-4, -4]\). A normal distribution was centered at each prototype, where a distribution’s covariance matrix was \(\sigma^2 I\) (\(\sigma^2 = 0.3\)). Thirty samples were collected from each distribution. Twenty were used during training, and ten were used during testing.

For ease of explanation, we take some terminological liberties. First, we assign emotions to categories. Categories whose associated distributions are centered at positive coordinates are referred to as “happy”, the category whose associated distribution is centered at \([0, 0]\) is referred to as “neutral”, and categories whose associated distributions are centered at negative coordinates are referred to as “sad”. In addition, the multisensory perception model had two sets of input variables, which we will refer to as “visual” and “auditory” features.

During training, input variables were consistent, meaning that visual and auditory features were sampled from the same distribution or category. A single MCMC chain was simulated. The chain was run for 500 iterations. The first 400 iterations were discarded as burn-in. In simulations of additional chains, we found that the results reported here are typical.

Our simulations included three test conditions. In these conditions, the model was exposed to data items potentially containing sensory conflicts. In all conditions, visual features could come from any of the nine categories. In the first test condition, auditory features always came from a happy category (the category centered at \([1, 1]\)). In the second test condition, auditory features always came
from the neutral category. Auditory features always came from the sad category (the category centered at \([-1, -1]\)) in the third test condition.

Evaluating the model’s performance on test items presents unique challenges. Although it is reasonable to sample variables’ values, and thus estimate variables’ distributions, on the basis of training items, models are not meant to learn from test items. Consequently, we could not run the MCMC sampler on the model using the test items to evaluate the model’s categorization performance. Doing so would erase the distinction between training and test data items.

Instead, we proceeded as follows. Consider the latent feature representations obtained on iteration \(i\) of the MCMC sampler when the model was trained on the training data. There is one such representation for each training item. These are the latent representations with non-zero probability based solely on iteration \(i\). Let \(\mathcal{L}_i\) denote this set of representations. For each test item in each test condition, we searched \(\mathcal{L}_i\) to find the latent representation that was most probable given the item. The item was then categorized on the basis of this latent representation. Specifically, the item was assigned to category \(j\) if its most probable latent representation was associated with category \(j\) on training iteration \(i\). Categorization performances were averaged across iterations.

The results are shown in Figure 4.2. The horizontal axis plots the category from which a test item’s visual features were sampled, and the vertical axis plots the probability that the model classified the test item as sad. Data points indicated by diamonds correspond to the first test condition (auditory features were sampled from the happy category centered at \([1, 1]\)), points indicated by circles correspond to the second test condition (auditory features were sampled from the neutral category centered at \([0, 0]\)), and points indicated by squares correspond to the third condition (auditory features were sampled from the sad category centered at \([-1, -1]\)).

These results demonstrate that the multisensory perception model integrates
information from two sources in a sensible manner. The model was more likely to judge a test item as happy when visual features were paired with happy auditory features than when paired with neutral or sad auditory features. Similarly, the model was more likely to judge a test item as sad when visual features were paired with sad auditory features than when paired with neutral or happy auditory features. In this sense, the model replicates the “emotional McGurk effect”.

4.6 Visual-Auditory Data Set

We applied the multisensory perception model to a real-world visual-auditory data set known as Tulips1 (Movellan, 1995). Twelve people (9 adult males, 3 adult females) were videotaped while uttering the first four digits of English twice.

In each video frame, the image of a speaker’s mouth was processed to extract 6 visual features: the width and height of the outer corners of the mouth, the width and height of the inner corners of the mouth, and the heights of the upper and lower lips. The auditory signal corresponding to a frame was processed to extract 26 features: 12 cepstral coefficients (these are the coefficients of the Fourier transform representation of the log magnitude spectrum), 1 log-power, 12 cepstral coefficient derivatives, and 1 log-power derivative. Because speech utterances had different durations, we sampled 6 frames for each utterance spanning the entire duration of the utterance in a uniform manner. In summary, each data item contained values for 36 visual features (6 frames × 6 visual features per frame) and 156 auditory features (6 frames × 26 auditory features per frame).

Training and test sets were created as follows. For the first eight speakers, one utterance of each digit was used for training and the other utterance was used for testing. For the remaining speakers, both utterances were used for training. Thus, the training set contained 16 data items for each digit, and the test set contained 8 data items for each digit.
Figure 4.2: The performance of the multisensory perception model on the first synthetic data set. The horizontal axis plots the category from which a test item’s “visual” features were sampled, and the vertical axis plots the probability that the model classified the test item as “sad”. Data points indicated by diamonds correspond to the first test condition (auditory features were sampled from a “happy” category), points indicated by circles correspond to the second test condition (auditory features were sampled from the neutral category), and points indicated by squares correspond to the third condition (auditory features were sampled from a “sad” category).
To understand better the performance of the multisensory perception model, we also consider the performances of two other models. The vision-only model is identical to the multisensory model except that it contains only two sets of variables corresponding to visual and latent features. When applied to the Tulips1 data set, it received only the visual features. Similarly, the audition-only model contains only two sets of variables corresponding to auditory and latent features. It received only the auditory features from the data set.

A single MCMC chain of each model was simulated. The chain was run for 5000 iterations, where the first 3000 iterations were discarded as burn-in. To reduce correlations among variables at nearby iterations, the remaining iterations were thinned to every 10\textsuperscript{th} iteration (i.e., only variable values at every 10\textsuperscript{th} iteration were retained). Thus, the results below are based on 200 iterations. Simulations of additional chains produced nearly identical results as those reported here.

**Posterior distributions over latent features:** Recall that the number of latent features in each model is not fixed a priori. Instead, it is a random variable whose distribution is inferred from the training data. The three graphs in Figure 4.3 show the distributions of the numbers of latent features in the vision-only, audition-only, and multisensory models. The vision-only model used relatively few latent features, the audition-only model used more latent features, and the multisensory model used the most latent features. This result confirms that the models are highly flexible. Their nonparametric nature allows them to adapt their representational capacities based on the complexities of their data sets. Interestingly, the multisensory model learned a compact set of latent features: the number of features it acquired was always less than the sum of the number of features acquired by the vision-only and audition-only models.

**Categorization performances:** We evaluated each model’s ability to categorize the speech utterances as instances of one of the first four digits in English based upon its latent feature representations. At each iteration of an MCMC
Figure 4.3: The distributions of the numbers of latent features in the vision-only (left), audition-only (middle), and multisensory (right) perception models.

chain, a model sampled a latent feature representation for each data item in the training set. Using these representations, we performed k-means clustering with four cluster centers. We then performed an exhaustive search of assignments of clusters to English digits (e.g., cluster $A \rightarrow$ digit 3, cluster $B \rightarrow$ digit 1, etc.) to find the assignment producing the best categorization performance. Performances were averaged across iterations of a chain.

The results are shown in the leftmost graph of Figure 4.4. The horizontal axis gives the model, and the vertical axis plots the percent of data items in the training set that were correctly classified (error bars indicate the standard deviations of these percents across iterations of an MCMC chain). As expected, the vision-only model showed the worst performance, the audition-only model showed better performance, and the multisensory model showed the best performance (based on two-tailed t-tests, the differences between performances of the multisensory model and each of the other models are statistically significant at the $p < 0.05$ level).

It is possible that the multisensory model showed the best performance solely due to the fact that it received both visual and auditory features and, thus, received a richer set of inputs than either the vision-only or audition-only models.
Figure 4.4: Categorization performances of the vision-only, audition-only, multisensory, and mixed models on the training set (left) and on the test set (right). The horizontal axis of each graph gives the model, and the vertical axis plots the percent of data items correctly classified (error bars indicate the standard deviations of these percents across iterations of an MCMC chain).

To evaluate this possibility, we simulated a model, referred to as a ‘mixed’ model, that resembled the multisensory model in the sense that it received both visual and auditory features. However, for the mixed model, these features were not segregated into separate input streams. Instead, the mixed model contained a set of latent features that received inputs from a set of undifferentiated perceptual features, namely a concatenation of the visual and auditory features. The results for the mixed model on the training set are also shown in the leftmost graph of Figure 4.4. The mixed model showed significantly poorer performance than the multisensory model, thus suggesting that the multisensory model benefited from independently accounting for visual and auditory features (albeit at the expense of additional variables).

This analysis was repeated using the data items in the test set. For a given model, let $\mathcal{L}_i$ denote the latent feature representations obtained on iteration $i$ of the MCMC sampler when the model was trained on the training data. Recall that there is one such representation for each training item, and these are the latent representations with non-zero probability based solely on iteration $i$. For each data item in the test set, we searched $\mathcal{L}_i$ to find a latent representation that
was most probable given the item. This was repeated for every item in the test set. Using these representations, the analysis of the test set is identical to the analysis of the training set described above: latent representations were clustered using k-means clustering, and a search of assignments of clusters to digits was performed to find the assignment producing the best categorization performance. Performances were averaged across iterations.

The results are shown in the rightmost graph of Figure 4.4. Again, the multisensory model showed the best performance.

In summary, the multisensory perception model showed the best categorization performance on both training and test data sets. We conclude that its superior performance is due to both its rich set of inputs (it receives both visual and auditory features) and due to its internal structure (visual and auditory features are segregated perceptual streams). Clearly, this model received the statistical benefits of sensory integration.

**Multisensory versus unisensory training:** Above, we reviewed experimental evidence that training in multisensory environments leads to better learning than training in unisensory environments, and that this advantage occurs even when testing is conducted under unisensory conditions. Does the multisensory perception model predict the superiority of multisensory training?

To study this question, we compared the categorization performances of the multisensory and vision-only models when visual features were the only inputs to these models, and compared the performances of the multisensory and audition-only models when auditory features were the only inputs. For ease of explanation, we first consider the multisensory and vision-only models when visual features were the only inputs.

As above, let $\mathcal{L}_i$ denote the set of multisensory feature representations obtained on iteration $i$ of the MCMC sampler when the model was trained on the training data. Again, these are the latent or multisensory representations with non-zero
probability based solely on iteration $i$. For each data item in the training set, we calculated the most probable multisensory representation based solely on an item’s visual features. A set of most probable representations is formed when all training items are taken into account. Using this set, we categorized the data items in the same manner as above: multisensory representations were clustered using k-means clustering, and a search of assignments of clusters to digits was performed to find the assignment producing the best categorization performance.

For the vision-only model, categorization of training items was performed in a similar manner. At each iteration of an MCMC chain, the vision-only model sampled a latent feature representation for each data item in the training set. Categorization via k-means clustering was performed on the basis of these representations.

The results are shown on the left-side of the left graph in Figure 4.5. The horizontal axis labels the model, and the vertical axis plots the proportion of data items a model categorized correctly (error bars plot the standard deviations of these proportions across iterations of an MCMC chain). Clearly, the multisensory perception model outperformed the vision-only model, even though both models were evaluated solely on the basis of training items’ visual features.

For data items in the test set, the categorization performances of the multisensory and vision-only models were calculated in analogous ways. For each model, the most probable latent representations were found at each iteration based on the items’ visual features. Categorization was performed on the basis of these representations. The results are shown on the right-side of the left graph in Figure 4.5. Again, the multisensory perception model performed better than the vision-only model.

Analogous computations were carried out to compute the categorization performances of the multisensory and audition-only models when auditory features were the only inputs to these models. The results are shown in the right graph of
Figure 4.5: (Left) Categorization performances of the multisensory and vision-only models on the training (left-side) and test (right-side) data items when visual features were the only inputs to these models. The horizontal axis labels the model, and the vertical axis plots the proportion of data items a model categorized correctly (error bars plot the standard deviations of these proportions across iterations of an MCMC chain). (Right) Categorization performances of the multisensory and audition-only models on the training and test data items when auditory features were the only inputs to these models.

Figure 4.5. For training items, the multisensory model outperformed the audition-only model. For test items, however, the two models showed similar categorization performances.

In summary, the multisensory perception model performed better than the unisensory models. This result was obtained even though the multisensory model was trained in multisensory conditions but evaluated in unisensory conditions. The result was strongest when models were evaluated on the basis of visual features only. This is expected because auditory features contain more information about the identity of speech utterances than visual features, and thus visual-auditory training should have its largest impact when evaluation is conducted solely with visual features. Clearly, the multisensory perception model predicts that training in multisensory environments should lead to better learning than training in unisensory environments, and that this advantage should occur even when evaluation is conducted under unisensory conditions.
**Modality invariance**: As discussed above, neural representations of objects are often modality invariant (Amedi et al., 2001; Konkle et al., 2009; Quiroga et al., 2009). That is, the same (or at least similar) neural representations arise regardless of the modality through which an object is sensed. Does the multisensory perception model show this same property?

We investigated this question as follows. Once again, let \( L_i \) denote the set of multisensory feature representations obtained on iteration \( i \) of the MCMC sampler when the model was trained on the training data. These are the multisensory representations with non-zero probability based solely on iteration \( i \). For each data item in the training set, we calculated the probability distribution of the multisensory representation given an item’s visual features, and the distribution of the multisensory representation given an item’s auditory features where \( L_i \) was the set of possible multisensory representations. When all training items are taken into account, these distributions are denoted \( p(Z|X_V) \) and \( p(Z|X_A) \), respectively. We then calculated the Battacharyya distance\(^3\) between \( p(Z|X_V) \) and \( p(Z|X_A) \). On every iteration, this distance was zero.

We repeated this analysis using the data items in the test set. Again, we computed \( p(Z|X_V) \) and \( p(Z|X_A) \) where \( X_V \) and \( X_A \) refer to the visual and auditory features of test items, and where \( L_i \) is the set of possible multisensory representations. The Battacharyya distances between \( p(Z|X_V) \) and \( p(Z|X_A) \) are always small values—the distribution of these distances has values of 1.51, 1.55, and 1.68 as its 25\(^{\text{th}}\), 50\(^{\text{th}}\), and 75\(^{\text{th}}\) percentiles, respectively. By way of comparison, we also computed the distance between \( p(Z|X_A) \) and a uniform distribution over multisensory representations. The distribution of these distances has values of 3.49, 7.83, and 19.04 as its 25\(^{\text{th}}\), 50\(^{\text{th}}\), and 75\(^{\text{th}}\) percentiles.

\(^3\)The Battacharyya distance is a statistical metric measuring the similarity of two discrete or continuous probability distributions. For discrete distributions \( p \) and \( q \) defined over the domain \( X \), it equals \(-\ln\left(\sum_{x \in X} \sqrt{p(x) q(x)}\right)\). We also considered the Kullback-Leibler distance but use of this metric led to numerical instabilities.
In summary, both training and test sets suggest that the multisensory perception model did indeed acquire modality-invariant representations. Its latent multisensory features had the same or similar distributions regardless of whether a speech utterance was seen or heard.

**Predicting sensory representations in missing modalities:** Above, we reviewed evidence of activity in people’s auditory cortices when they viewed speech utterances but did not hear those utterances (Calvert et al., 1997; Pekkola et al., 2005). This result is consistent with the hypothesis that sensory representations in one modality can predict or activate representations in other modalities. Does the multisensory perception model support the ability to use sensory representations in one modality to predict representations in other modalities?

This question was studied using the data items in the test set. Let \( V \) and \( A \) denote the sets of visual and auditory feature representations for the data items in the training set. Once again, let \( L_i \) denote the set of multisensory representations obtained on iteration \( i \) of the MCMC sampler when the model was trained on the training data. For each test item, we computed the probability distribution of an auditory representation given a test item’s visual features. This was accomplished by first calculating a conditional joint distribution over both multisensory and auditory representations, and then by marginalizing over the multisensory representations where the set of possible auditory and multisensory representations were given by \( A \) and \( L_i \). Analogous computations were carried out to compute the distribution of a visual representation given an item’s auditory features.

Representative results are shown in Figure 4.6. Four test items (items 1, 12, 24, and 28) were selected at random with the constraint that one item corresponded to each spoken digit. The four graphs in the top row of the figure show the distributions of the visual representations given the auditory features of the test items. More precisely, the graphs show that when presented with the auditory
features corresponding to one of the digits, the model’s distribution of visual representations was tightly peaked at a representation corresponding to the same digit. The four graphs in the bottom row show analogous results for distributions of auditory representations given test items’ visual features.

In summary, the multisensory perception model learns to associate unisensory representations from different modalities. It successfully predicts representations from missing modalities based on features from observed modalities.

4.7 Visual-Haptic Data Set

This section reports the results of applying the multisensory perception model to an experimental data set collected by Cooke, Jäkel, Wallraven, and Bülthoff (2007). As discussed below, a motivation for this application is that it permits us to interpret the multisensory features acquired by the model.

Cooke et al. (2007) created novel, three-dimensional objects that varied para-
metrically in shape (macrogeometry) and texture (microgeometry). Subjects performed a similarity rating task in which they repeatedly rated the similarity of pairs of objects. One group of subjects based their similarity judgments on their visual percepts, another group based their judgments on their haptic percepts, and a third group based their judgments on both visual and haptic percepts. Similarity ratings were analyzed using multidimensional scaling (MDS). This analysis revealed that people’s similarity judgments were strongly dominated by shape when objects were seen. However, shape and texture were roughly equally important in determining similarity when objects were touched or when they were both seen and touched. The authors concluded that the perceptual modality used to interact with objects affects the mental representations used for similarity judgments and categorization.

Here, we apply the multisensory perception model to the experimental data collected by Cooke et al. (2007). Doing so, however, requires a small modification of the model. In the simulations reported above, the observable data used to train a model consisted of unisensory perceptual features. The data collected by Cooke et al. does not consist of perceptual features. Rather it consists of pairwise similarity ratings. The changes needed to modify the multisensory perception model so that it could be trained with similarity data were described by Navarro and Griffiths (2008) in the context of the Indian buffet process. In brief, Navarro and Griffiths adapted the Indian buffet process so that it can be interpreted as a Bayesian formulation of additive clustering, a technique for inferring the features of a set of stimuli from their similarities (Shepard and Arabie, 1979). Importantly for our purposes, the formulation uses similarity ratings to infer one or more latent or hidden features that “explain” these ratings. Associated with each feature is a non-negative weight quantifying the saliency of that feature.\footnote{We followed the simulation procedures of Navarro and Griffiths (2008) with the exception that we used a uniform distribution over the interval [0, 1] as a prior distribution on the weights instead of a gamma distribution.}
We trained four models using the experimental data of Cooke et al. (2007), a vision-only model, a haptic-only model, and two multisensory models. The vision-only model had two sets of variables. These corresponded to latent features and subjects’ average similarity ratings when objects were seen. The haptic-only model also had two sets of variables. These corresponded to latent features and subjects’ average ratings when objects were touched. The first multisensory model (Multisensory-1) had two sets of variables corresponding to latent features and subjects' average ratings when objects were both seen and touched. In contrast, the second multisensory model (Multisensory-2) did not receive the data from the multisensory experimental condition. Instead, it received the data from the two unisensory conditions. This model had three sets of variables corresponding to latent features, subjects’ ratings when objects were seen, and subjects’ ratings when objects were touched. Although the model received only the data from unisensory conditions, we predicted that it would infer latent features resembling those of the model receiving data from the multisensory experimental condition.

A single MCMC chain of each model was simulated. The chain was run for 21000 iterations. The first 1000 iterations were discarded as burn-in. To reduce correlations among variables at nearby iterations, the remaining iterations were thinned to every $10^{th}$ iteration (i.e., only variable values at every $10^{th}$ iteration were retained). Thus, the results below are based on 2000 iterations.

The results are shown in Figure 4.7. Each panel illustrates the 25 objects used in the experiment of Cooke et al. (2007). Objects vary systematically in texture properties along the horizontal axis, and vary in shape properties along the vertical axis. The four panels correspond to the four models that we implemented. Each panel illustrates properties of the latent features acquired by its corresponding model. Let $\mathcal{L}_i$ denote the set of latent features acquired by a model on iteration $i$ of the MCMC sampler when the model was trained on the training data. In Figure 4.7, we only consider “stable” features, meaning those that existed in $\mathcal{L}_i$. 
for at least 25% of the iterations.

The upper-left panel plots the results for the vision-only model. This model acquired two stable latent features. One feature was possessed by objects 1-15. These objects occupy the bottom three rows of the panel (region enclosed by the solid line). The other feature was possessed by objects 12-25, occupying nearly all of the top three rows (region enclosed by the dotted line). In other words, these features were “on” for different values of the shape parameter (vertical axis). The first feature was on when the shape parameter had a small value, and the second feature was on when the parameter had a large value. Clearly, subjects’ similarity ratings when objects were seen were based on shape properties. The weight values associated with each of the features are given at the bottom of the panel.

The upper-right panel plots the results for the haptic-only model. In this case, four latent features were acquired. Two of the features can be characterized as “shape” features. They are identical to the features acquired by the vision-only model. One feature was on when the shape parameter had a small value, and the other feature was on when the parameter had a large value. The remaining two features can be characterized as “texture” features. One feature was on when the texture parameter (horizontal axis) had a small value, and the other feature was on when the parameter had a large value. In contrast to the vision-only case, subjects’ similarity ratings when objects were touched seem to have been based on both shape and texture properties.

As illustrated in the bottom-left panel, the results for the Multisensory-1 model are nearly identical to those of the haptic-only model. Four latent features were acquired. Two of the features can be characterized as shape features, and the remaining two can be characterized as texture features. When objects were both seen and touched, subjects’ similarity ratings were based on both shape and texture properties.

The bottom-right panel shows the results for the Multisensory-2 model. Recall
Figure 4.7: Each panel illustrates the 25 objects used in the experiment of Cooke et al. (2007). Objects vary systematically in texture properties along the horizontal axis, and vary in shape properties along the vertical axis. The four panels correspond to the four models that were implemented. Each panel illustrates properties of the latent features acquired by its corresponding model. See text for further explanation. (Figure of Cooke et al., 2007, adapted with permission from Elsevier.)
that the Multisensory-1 model received one set of multisensory similarity ratings as input whereas the Multisensory-2 model received two sets of unisensory similarity ratings. Despite this difference, the two models acquired nearly identical latent features. The Multisensory-2 model acquired four latent features, two shape features and two texture features. (Because the Multisensory-2 model received two sets of similarity ratings, it had two weights associated with each latent feature. The values of these weights were averaged, and these average values are shown at the bottom of the panel.)

In summary, the features acquired by the models studied here are in agreement with the analyses of Cooke et al. (2007). Subjects’ similarity ratings seem to be based on shape properties when objects are viewed, and based on both shape and texture properties when objects are touched or seen and touched. Moreover, the multisensory perception model (Multisensory-2) acquired multisensory latent features from two unisensory data sets that were nearly identical to the features acquired directly from a single multisensory data set (Multisensory-1). This result suggests that the model acquired multisensory representations that closely resemble those used by human subjects when judging the similarity of objects in a multisensory environment.

4.8 Discussion

How do people learn multisensory representations, and what consequences do these representations have for perceptual performance? We addressed this question by performing a rational analysis of the problem of learning multisensory representations that optimally explain unisensory features detected by individual modalities (Anderson, 1990; Marr, 1982). This analysis indicates that several aspects of multisensory perception emerge as part of the optimal solution to the problem of learning to represent complex multisensory environments. On the basis of our
results, we argue that multisensory representations should: (i) be adaptable and compact; (ii) integrate information from multiple modalities in such a way that leads to superior performances in, for example, categorization tasks; (iii) lead to better learning than representations acquired in unisensory environments, even when testing is conducted in unisensory conditions; (iv) be modality invariant; and (v) support the prediction of “missing” sensory representations in modalities when the input to those modalities is absent.

An assumption of the research reported here is that multisensory representations do indeed exist in people’s minds and brains. The question of whether behavioral phenomenon are best accounted for by postulating the existence of a multisensory representation versus the existence of multiple modality-specific representations has been actively studied. In regard to interactions between visually and haptically derived representations of objects, Lacey, Campbell, and Sathian (2007) concluded that the weight of evidence suggests the existence of a multisensory representation. We concur with their conclusion, and speculate that multisensory representations are commonplace in biological organisms.

A second assumption of our work is that learning involves modifications to multisensory representations. There are alternatives to this possibility, namely that learning involves modifications to unisensory representations or to connections among unisensory representations. As pointed out by Shams and Seitz (2008), these possibilities are not mutually exclusive. It may be that learning involves all three types of changes. If so, then the work presented here is incomplete due to its exclusive focus on the acquisition of multisensory representations.

We regard the rational analysis performed here as a useful early step toward understanding the acquisition of multisensory representations. Future work will need to address a number of open issues. A prediction of our work is that there is a connection between data complexity and memory code length, and that Bayesian nonparametric models provide a novel method for accounting for this connec-
An unusual property of the proposed model is that it adapts the size of its multisensory representation based on the complexity of the unisensory data it receives. This can be understood using intuitions from information theory (Cover & Thomas, 1991). We expect that unisensory data that are highly regular (e.g., highly predictable or low entropy) are mentally represented or “explained” with a compact code, such as a small number of multisensory features. In contrast, unisensory data with fewer statistical regularities should be represented by longer codes, such as a large number of multisensory features. Future work will need to test these predictions.

Future work should also include computational extensions to the proposed model. The model currently learns in an unsupervised manner, but it could be extended to a supervised setting. If, in addition to explaining a data item’s unisensory features, the multisensory perception model was required to also explain an item’s target response, then the model would be applicable to supervised situations. For example, a model could receive as input each item’s visual features, auditory features, and category label. The model would learn a multisensory representation that explains all three inputs. Such a model would possess several interesting properties because the multisensory representations acquired by the model would be shaped by the target responses. If the target responses were category labels, then the model would be constrained to acquire similar multisensory representations for data items belonging to the same category. That is, the representations would be task-specific, not task-independent. Furthermore, when the target responses are not observed, the model could compute a probability distribution over possible target responses given a test item’s unisensory features. Thus the model could act as a classifier, regressor, or associative memory.

As a second possible extension, the multisensory perception model learns multisensory features at a single level of abstraction, but this representation could be extended to a hierarchy. Courville, Eck, and Bengio (2009) showed how an Indian
buffet process can be extended to two or more layers of latent variables. Each layer defines a distribution over the latent variables of the layer below via a noisy-or mechanism (Neapolitan, 2004). They showed that this hierarchical model often outperforms a single-level Indian buffet process in the sense that it achieves higher likelihoods on test data items and more compact latent representations. Similarly, Adams, Wallach, and Ghahramani (2010) developed a hierarchical model, referred to as the Cascading Indian Buffet Process, that learns a layered, directed belief network of latent variables that explain the observable variables. An interesting feature of this model is that it provides a prior distribution on the structure of belief networks that is unbounded in both depth and width, yet allows tractable inference. Applying the hierarchical extensions of either Courville et al. (2009) or Adams et al. (2010) to the multisensory perception model should be straightforward, and may provide insights into multisensory features at multiple levels of abstraction.

Finally, the model currently does not include a notion of time. Future work might consider a model version that is sensitive to temporal dependencies in sensory data (Williamson, Orbanz, & Ghahramani, 2010) or a version whose inferential processes are time-dependent (as in sequential Monte Carlo methods). These modifications would allow the model to potentially replicate the time-course of development and learning in multisensory perception, thereby expanding the range of experimental data to which the model could be applied.
5 Discussion

In this thesis, we studied how sensory inputs can give rise to conceptual, abstract representations by casting this long standing puzzle of cognitive science in the domain of multisensory perception. Our approach was to explore the Multisensory Hypothesis, which states that people automatically extract and represent objects’ and events’ intrinsic properties, and use these properties to process and understand the same (and similar) objects and events when they are perceived through novel sensory modalities. We reported behavioral and computational studies.

The behavioral experiments examined people’s abilities to transfer their world knowledge across modalities —an aspect of human behavior that stands out as an unattained challenge for existing computational models of human cognition. Our results indicated that people indeed successfully transferred their knowledge to untrained sensory modalities with no additional training required.

Our computational studies provided (to our knowledge) the first implementations of the Multisensory Hypothesis. They met the challenge that our behavioral experiments posed by illustrating how modality-independent representations could be acquired on the basis of sensory data, and could mediate transfer of knowledge across modalities. The key idea in our computational framework is that modality-independent, conceptual representations can be inferred from sensory data by a
process that inverts predictive models, also known as forward models. In characterizing the modality-independent representations, we made use of both statistical (Bayesian nonparametrics, Chapter 4) and hybrid —symbols plus probabilities— approaches (pLOT approach, Chapters 2 and 3). For example, in Chapter 2, a modality-independent “language”-like representation was used for representing objects, such as a set of possible object parts and their possible spatial relations.

As for the forward models, our models either acquired them jointly with the modality-independent representations (Chapter 4), or it used currently existing software packages (e.g., graphics engines, grasping simulator) that implement sensory-specific forward models. For example, the model in Chapter 2 includes a haptic forward model (GraspIt!) that maps an object representation to a prediction of the haptic features that would arise if the object were grasped (perhaps a form of haptic imagery).

Given these two components —the modality-independent representations and the forward models— and the sensory data (e.g., pixel images of an object), the models use Bayesian inference to find the modality-independent representations (e.g., object representations) that give rise to sensory data potentially of different modalities (e.g., a set of visual features observed when an object is viewed, a set of haptic features observed when an object is grasped, or both).

The pLOT approach is very promising as it combines the strengths of symbolic (very powerful representational formalisms) and statistical (very flexible representations) approaches. However, the pLOT approach is often exclusively applied for the study of higher-level cognition. Here, we extend the pLOT approach to the domain of human perception. Our use of this approach in perceptual domains raises the possibility that compositional, symbolic representations may also underlie human perception. We argue that it is possible that the principles of mental computation responsible for seemingly different functions (e.g., perception vs. language) may share important commonalities. The recent findings indicating that
the visual cortices of blind individuals process syntactic structures for language comprehension (Bedny et al., 2011), suggesting that the neural substrate for vision can also support linguistic processes, may need to be re-interpreted from the viewpoint advocated in this thesis (e.g., perhaps vision operates over compositional structures similar to language).

The insights that this thesis provides about human perception and cognition raise deep questions into the nature of mental computation. Below we discuss these questions from the point of each of the components of our computational framework.

Even though explaining the results of the cross-modal experiments challenge our models to a certain extent (particularly Chapter 3), the power of the representational formalisms that we utilize in these models (e.g., potentially Turing complete computer programs) demands further behavioral scrutiny. One research avenue to take is to devise behavioral tasks to collect more quantifiable and different kinds of data (e.g., similarity data, learning curves, reaction times), and use these data to explore the space of representations or to justify a particular format of modality-independent representations. We have already begun to pursue this research direction. In a cross-modal similarity judgement task, we found that our pLOT based 3-D object shape model better fits people’s ratings when compared to other models (Erdogan, Yildirim, & Jacobs, under review). Converging evidence can help the field make stronger cases for the psychological reality and neural signatures of the compositional/combinatorial representations for perception.

Another aspect of our computational framework, that is its use of forward models, also raises fundamental questions about the brain and the behavior. The sensory-specific forward models described here (particularly Chapter 2) are sophisticated simulation engines (e.g., a grasping simulator or a rendering engine). We are intrigued to seek evidence that our brains might be performing simulations in a similar manner. In fact, Battaglia, Hamrick, & Tenenbaum (2013) argued
that people’s physical scene understanding can be best described as a relatively full physics simulation engine with noisy input. Their model outperforms a battery of possible heuristics that in the past were argued to be crucial to physical scene understanding in fitting participants’ behavior over a wide range of experimental tasks. Encouraged by this evidence and the results reported in this thesis, we advocate newer paradigms to investigate the psychological reality and neural signatures of simulation based forward models in the brain.
References


Kosslyn, S. M., Thompson, W. L., & Ganis, G. (2006). *The case for mental imagery*. Oxford University Press.


Appendix for Chapter 2: An MCMC Algorithm for the MVH Model

Exact inference in the MVH model is computationally intractable. Therefore, we developed an approximate Markov chain Monte Carlo inference algorithm. The input to the algorithm is a set of images of Fribbles which belong to the same category. The output is an approximation of the prototypical 3-D shape for that category (i.e., estimated values for the prototype’s part collection $\Omega$, tree $T$, and spatial configuration $S$). This appendix describes the algorithm.

Pseudocode for the inference algorithm is provided in Algorithms 1-4. Algorithm 1 is the main loop, whereas Algorithms 2-4 are subroutines. The key idea underlying the algorithm is as follows. The algorithm is not provided with the number of parts belonging to the prototype. Therefore, it forms the prototype by sequentially adding parts one at a time until the prototype contains a maximum number of parts (15). It then prunes parts using a hypothesis-testing procedure. Readers should be able to understand the algorithm from the pseudocode. Additional comments regarding the algorithm are listed here:

- As described in Algorithm 1, parts are added to the prototype one at a time. For each part, multiple (50) iterations are used to generate a final proposal, and a set of final proposals is created by repeating this iterative process multiple (10) times. We concatenate the old state of the MCMC chain to the proposals. One member of the set is accepted. The chain is terminated if the accepted member is the old prototype instead of a proposal.
• When searching for a new part to add to the prototype, proposals (i.e., values for $\Omega$, $T$, and $S$) are sampled from broad distributions. Proposals are either accepted or rejected on the basis of how well they account for the visual data (as given by values of the likelihood function).

• When computing the likelihood function based on values of $\Omega$, $T$, and $S$, it is important to check that these values do not specify a spatial layout in which two parts occupy the same docking locations. If this constraint is violated, then set the likelihood function to minus infinity.

• Given a tree $T$ with $N$ nodes, there are only $N$ ways that a new node can be added such that the resulting structure is also a tree. Therefore, exhaustive search of the space of trees with $N + 1$ nodes is computationally tractable.

• In Algorithm 3, a new spatial configuration $S$ is sampled from a mixture distribution for the new node. What is meant is that a new center docking location is sampled from this distribution, and this location is concatenated to previous center locations (for previously added parts) to form a new spatial configuration.

• After running Algorithm 1, the prototype contains a maximum (15) number of parts. Next, parts are pruned using a sequential hypothesis-testing procedure adapted from Feldman and Singh (2006). Going from the last part that was added to the first, each part is subjected to a Bayesian posterior ratio test of significance. If the posterior probability of the prototype with $n$ parts is less than its probability with $n - 1$ parts, then the $n^{th}$ part is pruned. Otherwise, the $n^{th}$ part is retained and the pruning procedure is terminated.
Algorithm 1 Main loop for sequentially adding parts to prototype

for part = 1 to 15 do
  // Compute log likelihood for state of MCMC chain with part – 1 parts.
  // Log likelihood is negative infinity when part = 1.
  old_log_likeli = log_likelihood(T, S, Ω)
  // Generate 10 full proposals with part parts.
  for counter = 1 to 10 do
    // Initialize Ω for new part randomly.
    r ~ Uniform(1, 40)
    l ~ Uniform(1, 40)
    o ~ Uniform(1, 3)
    // Do 50 iterations of sampling to generate one full proposal of T, S, and Ω.
    for iteration = 1 to 50 do
      if part = 1 then
        // With only one part, there is no spatial layout among parts (no T and S).
        // Only goal is to sample r, l, and o for first part.
        Algorithm 4
      else
        // There is at least one existing part. Now want to add a new part.
        // Sample new T by adding a node and edge to old T, and sample new S
        // by adding a new center docking location to old S.
        Algorithm 2
        // Retain T but resample S.
        Algorithm 3
        // Sample r, l, and o for new part.
        Algorithm 4
      end if
    end for
    full_proposal(counter) = log_likelihood(T, S, Ω)
  end for
  full_proposal(11) = old_log_likeli
  W = normalized full_proposal array such that values are non-negative, sum to 1
  Sample T, S, and Ω according to discrete distribution given by W
  Break if old state [corresponding to full_proposal(11)] is accepted.
end for
Algorithm 2 Sample $T$ and $S$ given $\Omega$

// Possible values for new $T$ formed by adding a node and edge to old $T$.
// Sample new $S$ by adding a center docking location to old $S$.

old_log_likeli = log_likelihood($T, S, \Omega$)

for each possible value of new $T$ (where $i$ indexes this value) do
  for $j = 1$ to 10 do
    $S \sim$ Uniform(1, 108)
    $log_likeli(i, j) = log_likelihood(T, S, \Omega)$
  end for
end for

Concatenate old_log_likeli to end of log_likeli array

$W = \text{normalized} \ log_likeli \ array \ such \ that \ values \ are \ non-negative, \ sum \ to \ 1$

Sample $T$ and $S$ according to discrete distribution given by $W$
Algorithm 3 Resample $S$ given $T$, $S$, and $\Omega$

old $S = S$

old log likeli = log_likelihood($T, old S, \Omega$)

for counter = 1 to 10 do

new $S \sim [0.5 \times \text{Uniform}(old S - 3, old S + 3)] + [0.5 \times \text{Uniform}(1, 108)]$

new log likeli = log_likelihood($T, new S, \Omega$)

// Metropolis-Hastings step

if Uniform(0, 1) < exp(min(0, new log likeli - old log likeli)) then

$S = new S$

break

end if

end for
Algorithm 4 Sample Ω (i.e., r, l, and o) given T and S

//Sample orientation o.
old_log_likeli = log_likelihood(T, S, r, l, o)
new_o ~ Uniform(1, 3)
new_log_likeli = log_likelihood(T, S, r, l, new_o)
//Metropolis-Hastings step
if Uniform(0, 1) < exp(min(0, new_log_likeli − old_log_likeli)) then
    o = new_o
end if
old_log_likeli = log_likelihood(T, S, r, l, o)
//Sample r and l.
for counter = 1 to 10 do
    new_r ~ [0.7 × Uniform(r − 2, r + 2)] + [0.3 × Uniform(1, 40)]
    new_l ~ [0.7 × Uniform(l − 2, l + 2)] + [0.3 × Uniform(1, 40)]
    new_log_likeli = log_likelihood(T, S, new_r, new_l, o)
    //Metropolis-Hastings step
    if Uniform(0, 1) < exp(min(0, new_log_likeli − old_log_likeli)) then
        r = new_r
        l = new_l
        break
    end if
end for