Behavior-grounded multi-sensory object perception and exploration by a humanoid robot

by

Jivko Sinapov

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY

Co-majors: Computer Science; Human Computer Interaction

Program of Study Committee:
Alexander Stoytchev, Major Professor
Nicola Elia
Yan-Bin Jia
Jonathan Kelly
Jin Tian

Iowa State University
Ames, Iowa
2013

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

There are many people whose help, advice, and presence made this dissertation possible. First and foremost, I am grateful to have an exceptional advisor, Alexander Stoytchev. His guidance and patience went above and beyond what was required. I especially appreciate the freedom he provided for me to pursue my research topic and his deep interest in my professional development.

Sincere thanks also go out to the rest of my committee: Nicola Elia, Jin Tian, Yan-Bin Jia, and Jonathan Kelly. Each played a special role by providing valuable help, constructive critiques, and professional advice. The questions they posed to me during our discussions and my dissertation proposal were some of the most challenging and thoughtful questions that I have had to answer and for that I am very grateful. I feel lucky to have such a highly accomplished committee with a very diverse research expertise.

Next, I would also like to thank all of my fellow graduate and undergraduate research collaborators for their help in conducting experiments with our robot: Connor Schenck, Vladimir Sukhoy, Shane Griffith, Matt Miller, Taylor Bergquist, Liping Wu, David Johnston, Peter Wong, Mark Wiemer, Ugonna Ohiri, Ritika Sahai, and Kerrick Staley. Robotics requires a team effort and I could not have done this without their help. I would also like to thank Steven Lischer, Izaak Moody, Michael Steffen, Brad Smith, John O'Brien, and Frederick Thompson who helped prototype and build the robot.

Special thanks go out to Cornelia Caragea, Adrian Silvescu, and Vasant Honavar who were my first research collaborators. Their guidance throughout the initial stages of my graduate career was extremely helpful. I learned a lot about machine learning from them.

I would also like to acknowledge the people, institutions, and sponsors that helped fund my employment. I am lucky to have worked as a teaching assistant for some great professors, including Alex Stoytchev, Jonathan Kelly, Chris Harding, and Soma Chaudhuri. I am especially

grateful to Stephen Gilbert for providing me with a teaching opportunity in the summer of 2013. I would also like to thank the John Deere Co. for its support during my first few years here. I am also grateful to the National Science Foundation for funding the REU program at ISU, which allowed me to mentor and work with three bright undergraduates who were passionate about research. I also appreciate the continued support from the Virtual Reality Application Center, the Computer Science department, and the Human-Computer Interaction program. In addition to everyone at Iowa State, I would also like to thank Willow Garage and HRL Laboratories for providing me with research internships that allowed me to apply my expertise towards solving practical perception and manipulation tasks in robotics.

Finally, I would like to thank my mother, my father, and my sister for their love and support over these past eight years. I could not have done this without them.

ABSTRACT

Infants use exploratory behaviors to learn about the objects around them. Psychologists have theorized that behaviors such as touching, pressing, lifting, and dropping enable infants to form grounded object representations. For example, scratching an object can provide information about its roughness, while lifting it can provide information about its weight. In a sense, the exploratory behavior acts as a "question" to the object, which is subsequently "answered" by the sensory stimuli produced during the execution of the behavior. In contrast, most object representations used by robots today rely solely on computer vision or laser scan data, gathered through passive observation. Such disembodied approaches to robotic perception may be useful for recognizing an object using a 3D model database, but nevertheless, will fail to infer object properties that cannot be detected using vision alone. To bridge this gap, this dissertation introduces a framework for object perception and exploration in which the robot's representation of objects is grounded in its own sensorimotor experience with them. In this framework, an object is represented by sensorimotor contingencies that span a diverse set of exploratory behaviors and sensory modalities. The results from several large-scale experimental studies show that the behavior-grounded object representation enables a robot to solve a wide variety of tasks including recognition of objects based on the stimuli that they produce, object grouping and sorting, and learning category labels that describe objects and their properties.

CHAPTER 1. INTRODUCTION

1.1 Object Perception using Exploratory Behaviors

Infants learn the properties of objects through active exploration. Behaviors such as touching, pressing, lifting, and dropping play an important role in the acquisition of object knowledge (Rochat, 1989; Power, 2000). This type of exploration is crucial for solving a vast array of problems, including the formation and establishment of object representations (Meltzoff and Moore, 1998), recognition of objects based on the stimuli that they produce (Ruff, 1980), object grouping and ordering (Starkey, 1981; Spinozzi et al., 1999), as well as learning words that describe objects and their properties (Nelson, 1973; Bloom, 2000). Psychologists have theorized that humans acquire grounded object representations through the use of a number of manipulation behaviors, commonly referred to as *exploratory procedures* (see Lederman and Klatzky (1990)) or *exploratory behaviors* (see Gibson (1988); Power (2000)). For example, scratching an object can provide information about its roughness, while lifting it can provide information about its weight. In a sense, the exploratory behavior acts as a "question" to the object, which is subsequently "answered" by the sensory stimuli produced during the execution of the behavior.

In contrast, the object representations used by most robots are carefully designed by human programmers. For example, many state-of-the-art approaches to robotic manipulation are based on precise 3D object models that are typically not acquired by the robot itself. Given such representations, it is not surprising that most robots perceive objects using solely 2D and/or 3D vision sensors. While such representations allow the use of planning methods for manipulation, they still suffer from the limitation that many object properties cannot always be perceived through vision alone. For example, a robot that perceives an object using only vision cannot tell the difference between a full and an empty container, nor can it distinguish between soft

and hard objects that look the same.

Another major limitation of existing approaches to robotic object perception is that most existing architectures lack the ability to perform tasks that go beyond simply perceiving the object's identity and location. Indeed, most current methods are designed with the sole purpose of estimating an object's pose so that it can be used for grasping. While useful for pick-and-place tasks, these methods do not afford a robot the ability to form, acquire, and recognize object categories, or to infer pairwise object relations. Thus, while the state of the art in robotic manipulation may allow a robot to pick up objects from a table and throw them in a waste bin, there is still no clear way of training a robot to recognize which objects belong in the trash and which do not.

As an example, consider the task of cleaning up a kitchen. To solve this task, a robot must be able to detect, grasp and manipulate objects in order to move them from place to place. All of these are problems that have traditionally been addressed through the use of computer vision and 3D perception coupled with planning for grasping and manipulation. To fully solve the problem, however, a robot must also be able to perceive many object properties (e.g., material type), object categories (e.g., cups vs. plates) and relations between objects (e.g., smaller plates are placed on top of larger plates, not the other way around). Furthermore, robots that rely exclusively on passive sensory modalities for object perception (e.g., 2D and 3D computer vision) are missing important information about objects (e.g., how they feel, how they sound, how heavy they are, etc.) that we humans take for granted in our daily activities.

I propose that this gap between human and robot object perception may be bridged if robots are enabled to explore objects using a diverse set of exploratory behaviors coupled with a large number of sensory modalities. To address the limitations of the current state of the art, this dissertation describes a framework for object perception that enables a robot to acquire and use knowledge about objects that is grounded in its own behavioral interactions with them. Unlike passive approaches to object perception, in this framework, the robot perceives objects and their properties by applying a wide variety of exploratory behaviors on them and detects the perceptual stimuli produced by a variety of sensory modalities.

1.2 Research Questions

The main research question that is investigated in this dissertation is the following:

How can a robot use exploratory behaviors to acquire grounded representations that are useful for perceiving objects and their properties?

More specifically, this question is addressed by breaking it into the four subsidiary questions that are listed below.

1.2.1 How can a robot use its own behaviors to recognize objects?

Object recognition is typically addressed as a sub-problem of visual classification. Vision-based approaches to object recognition, however, are of little use when confronted with objects that are visually identical (e.g., a full bottle and an empty bottle that are opaque). Furthermore, studies in psychology have demonstrated that recognizing the identity of an object is a multi-modal process, mediated by physical interaction with the object (Ruff, 1980). This dissertation proposes to bridge that divide by developing and evaluating methods that would enable robots to recognize objects using multiple behaviors as well as multiple sensory modalities (e.g., touch, audio, etc.). The experimental results described in Chapters 4 and 5 show that a robot can indeed recognize the identity of objects using exploratory behaviors by training a collection of recognition models, where each classifier corresponds to a unique behavior-modality combination. The results also show that the use of a diverse set of behaviors significantly boosts object recognition rates as different behaviors capture different aspects of an object.

1.2.2 How can a robot use its own behaviors, coupled with non-visual sensory modalities, to group objects according to human-provided semantic labels?

Similarly to object recognition approaches, most methods for unsupervised and supervised object categorization are purely vision-based (Fergus et al., 2004; Ponce, 2006; Opelt et al., 2006). While vision-based approaches to object classification can be useful in a variety of

applications, the object classification models are rarely *learned* by the robot itself. Instead, such models are typically trained on pre-recorded 2D and 3D vision datasets, often collected using a different set of sensors from the ones used by the robot. Such disembodied approaches to robotic perception cannot handle object categories for which the underlying features are non-visual (e.g., classifying objects as either *soft* or *hard* requires touching them). Human beings are subject to the same limitations, and thus it is not surprising that we represent object properties using the data from multiple sensory modalities (Lynott and Connell, 2009). Therefore, the research described in this dissertation explores how human-provided semantic category labels can be associated with specific behaviors and sensory modalities. Chapter 6 describes an unsupervised method for object grouping that indeed shows that auditory and proprioceptive sensory feedback can be used to group objects according to their categories. Furthermore, Chapter 7 shows that the same categories can be explicitly learned by the robot using a labeled set of objects for which the categories are known.

1.2.3 How can robotic categorization of objects be scaled to a larger number of objects, behaviors, sensory modalities, and category types?

One major limitation of existing approaches to object classification in robotics is that they typically use a small number of objects and only one type of sensory modality, usually vision. To advance the state of the art, this dissertation shows that a robot's ability to assign category labels to objects can be scaled to a much larger number of objects, category labels, behaviors, and sensory modalities. More specifically, Chapter 8 demonstrates a category recognition framework that doubles the number of behaviors and objects as compared to our previous work and uses additional visual and non-visual sensory modalities.

Another major limitation of existing work is that virtually all models for object categorization used by robots today are only useful for assigning labels to single objects. In contrast, many semantic category labels – for example, *heavier than*, and *taller than* – have meaning only when they are applied on a pair of objects. Furthermore, other category labels, such as "vary by size," describe relationships within a set of objects. To handle this rich space of category labels, this dissertation describes a relational learning framework, described in Chap-

ter 9, which, unlike existing approaches to robotic object category recognition, can learn more complex category relations such as the ones mentioned above.

1.2.4 How can a robot use its own behaviors to individuate objects?

A related problem to object recognition is that of *object individuation*. Psychologists define the problem as that of deciding how many objects have been perceived, which is a necessary step for establishing a representation that is suitable for the task of object recognition (Kemp et al., 2009). In contrast, virtually all methods for object recognition and object categorization that are used by robots today assume that training data, labeled with the corresponding object identity, is available for each object in the robot's training set. In other words, such methods explicitly make the assumption that the object individuation task has already been solved. Providing labeled data, however, becomes increasingly more difficult as the number of objects increases. Furthermore, an autonomous robot is bound to encounter new objects that were not in its training set. In order for the robot to learn models that can recognize these novel objects, it must first be able to individuate them as separate objects. To address these challenges, this dissertation describes a behavior-grounded approach to object individuation that enables a robot to estimate how many objects it has interacted with and group its sensorimotor experiences with different objects according to the estimated object identities. This approach is described in Chapter 10.

1.3 Contributions

A long-standing goal for the robotics community is to enable robots to function autonomously in human environments such as our homes and offices (Kemp et al., 2007). Such unstructured environments present many challenges to robotic perception and manipulation due to the large number of objects that robots have to deal with in an intelligent manner. While many lines of research have explored specific manipulation problems (e.g., planning, obstacle avoidance, and grasping), robots still lack the basic abilities needed to intelligently acquire and use object knowledge that is grounded in their own sensorimotor experience.

To address these challenges, this dissertation describes a theoretical framework for acquiring and grounding object knowledge in a diverse set of robot behaviors and sensory modalities. This framework was evaluated on a variety of tasks, including object recognition, object classification, and object individuation. In doing so, the research described here advances the state of the art in the following ways:

1. It develops a behavior-grounded framework that enables a robot to recognize objects by performing exploratory behaviors on them (Chapters 4 and 5).

2. It develops feature extraction methods that can be applied on a wide variety of sensory feedback coming from different sensory modalities (Chapters 4 and 8).

3. It demonstrates that sensorimotor interaction can be used to group objects according to their physical properties and human-provided labels using both unsupervised (Chapter 6) and supervised machine learning methods (Chapters 7 and 8).

4. It develops a novel framework that enables a robot not only to assign labels to individual objects, but also to detect relational categories that describe how objects relate to each other (Chapter 9).

5. It demonstrates a solution to the object individuation problem that enables a robot to infer the number of objects that it interacted with and group its sensorimotor data according to the estimated object identities (Chapter 10).

Currently, most research in robotic manipulation focuses on enabling a robot to grasp an object and move it from one place to another. Such methods typically rely solely on vision-based techniques that are used to detect an object, recognize it using a 3D model database, and compute a pose for the robot's end effector that will presumably result in a stable grasp (Quigley et al., 2007; Srinivasa et al., 2009; Rasolzadeh et al., 2010; Rusu et al., 2008). More recent approaches to the same problem have focused on incorporating machine learning methods to detect graspable features in visual object data (Saxena et al., 2008; Erkan et al., 2010), and to learn manipulation skills such as pushing a button, or opening a drawer (Sukhoy et al., 2010; Dang and Allen, 2010).

In contrast, the research described here starts with the assumption that the robot is already capable of performing a number of behaviors on objects (e.g., grasping, lifting, shaking, etc.). Given this assumption, this dissertations advances the state of the art by enabling robots to solve a variety of object related tasks such as object recognition and object categorization. These abilities are quintessential for solving many household tasks. For example, to clean up a table, a robot will need not only pick and place manipulation skills, but also object classification skills in order to detect whether an object belongs in the waste bin or in the kitchen cupboard.

Furthermore, sensorimotor interaction with objects is a fundamental prerequisite for word learning. For example, the only way to learn the meaning of the words *soft* and *hard* is to physically interact with objects that are either soft or hard and to detect differences in at least one sensory stream that can be used to distinguish reliably between the two categories. The research described here advances the state in the art in that area by demonstrating that a robot can learn semantic object labels and relations that are represented in terms of the robot's own object-directed exploratory behaviors and their perceptual outcomes.

The experiments described in this dissertation were influenced by many observational and experimental studies reported in the developmental psychology literature. For example, the developmental progression of infant object knowledge, summarized in Section 2.1.5, served as an inspiration for some of the robotic experiments. Nevertheless, the research presented here is only inspired by relevant work in psychology and does not attempt to directly model how infants and humans learn about objects.

1.4 Overview

The rest of this dissertation is organized as follows. Chapter 2 provides a summary of the related work in robotics, as well as in psychology, cognitive science, and philosophy. Chapter 3 describes the upper-torso humanoid robot that was used to perform this research. Chapters 4 and 5 describe the behavior-grounded approach to object recognition proposed in this dissertation. Chapter 6 describes an unsupervised theoretical model that enables a robot to solve the *odd-one-out* task using multiple measures of object similarity grounded in the robot's sensorimotor repertoire. Chapter 7 describes a theoretical model in which those similarity measures are used to learn object categories in a supervised manner, while Chapter 8 demonstrates that the behavior-grounded representation can scale to an even larger number of objects, behaviors, and sensory modalities. Chapter 9 extends the behavior-grounded category recognition model to cover categories that describe object pairs and object groups. Chapter 10 shows that the representation is not only useful for recognizing object identities and object categories, but also can be used to individuate (i.e., *form* the object identities for) a set of novel objects. Finally, Chapter 11 provides a summary of this dissertation and also suggests several direct lines for future work.

CHAPTER 2. BACKGROUND AND RELATED WORK

2.1 Related Work in Philosophy, Psychology, and Cognitive Science

2.1.1 Object Concepts in Philosophy

The ability to form concepts from experience is an important hallmark of human intelligence. It enables us to solve a wide variety of problems, ranging from learning the names of objects in our infancy to formulating complex scientific theories in our adulthood. As such, concepts and their formation have been studied by scientists and philosophers for a long time. In the 17^{th} century, Locke (1690) postulated that category formation is the process of abstracting and inferring commonalities from a set of specific instances. Later on, Hume (1776) wrote that all ideas with meaning must be related to "livelier" cognitive and perceptual experiences (i.e., what he called *impressions*). More specifically, Hume argued that our ability to create mental concepts depends directly on our perception and experience:

> *"[A]ll this creative power of the mind amounts to no more than the faculty of compounding, transposing, augmenting, or diminishing the materials afforded us by the senses and experience. When we think of a golden mountain, we only join two consistent ideas, gold, and mountain, with which we were formerly acquainted" (Hume, 1776).*

In other words, complex ideas can be decomposed into simple ideas, and simple ideas, in turn, correspond to "impressions" from which they are derived. This notion, commonly referred to as Hume's *copy principle*, is what highlights Hume's brand of empiricism.

Influenced by Hume, Kant (1781) argued that all concepts fall within one of two categories: *a posteriori* or *a priori*. The first, a posteriori, refers to concepts that, as Hume argued, are

abstracted from experience and observation (i.e., empirical concepts). The second type, *a priori*, on the other hand, refers to concepts that are not directly abstracted from experience. For example, the statement "all balls are round" is an *a priori* concept, since, by definition, a ball is round. On the other hand, the statement "some balls are blue" is an *a posteriori* concept, which is derived from the observation that some balls are indeed blue.

One criticism of Hume's empiricism is that the *copy principle* can only be used to make the claim that an object is like another, but not that they are the same. The standard example of this paradox is the "mother" object – how does an infant know that the mother it sees one day is the same as the mother it saw the previous day? In the second half of the 20^{th} century, Quine attempted to resolve this paradox by postulating that humans rely on an innate bias that uses the observed spatio-temporal continuity of objects in order to distinguish items that have been previously observed from novel items that are merely similar to those encountered in the past (Quine, 1986).

More recently, Sloman made the distinction between two types of empiricism: *concept empiricism* and *knowledge empiricism*. Concept empiricism is the position that a concept may only be learned by experiencing examples of that concept. Under concept empiricism, the important question is, where do concepts come from? Knowledge empiricism, on the other hand, is concerned with knowledge rather than concepts and can be summarized as follows:

> *"All knowledge (about what is and is not true) has to be derived from and testable by sensory experience, possibly aided by experiments and scientific instruments." (Sloman, 2008).*

Unlike concept empiricism, knowledge empiricism allows for the possibility that some knowledge (although not empirical) may exist *a priori*. To illustrate this, Sloman and Chappell (2005) give an example with precocial species, which are born well developed and able to solve complex tasks. For example, a newly hatched chick not only can peck at food right away, but also can break out of the shell in the first place. Thus, they argued that the information structures that are responsible for problem solving are not devoid of semantic content. Further, they argued in favor of knowledge empiricism by noting that certain concepts are not abstracted away from

sensory experience, but instead are the result of powerful bootstrapping mechanisms:

> *"These mechanisms (a) acquire many discrete chunks of knowledge through play and exploratory behaviour which is not directly reinforced, and (b) combine such chunks in novel ways both in solving problems and in further play and exploration."*
> *(Sloman and Chappell, 2005).*

Overall, the influence of empiricism on developmental robotics is evident in several key principles described by Stoytchev (2009). For example, the *verification principle*, first proposed by Sutton (2001), states that *"An AI system can create and maintain knowledge only to the extent that it can verify that knowledge itself"*. In the context of robotics, this principle entails that a robot's knowledge about the world must be extracted from the robot's own experience, such that the knowledge can be verified through the robot's future sensorimotor interaction with the world.

The *embodiment principle* naturally follows, since without a body, there is no means of verification (Stoytchev, 2009). This principle is closely related to the empiricist notion that everything known about objects in the world is mediated by our senses. While this may seem obvious to philosophers, both the *embodiment principle* and the *verification principle* contrast sharply with the traditional robotics approach of representing objects as 3D models that are not extracted by the robot itself, but instead are provided by human programmers.

Along those lines, the principle of *grounding* postulates that a robot's knowledge of the world must contain (possibly probabilistic) representations that couple actions with perceptual outcomes (Stoytchev, 2009). Thus, a successful grounding of the robot's knowledge is one that enables a robot to verify what it knows by performing specific sequences of actions and detecting the results.

Guided by these principles, the research described here aims to develop novel methods, algorithms, and representations that would enable a robot to acquire grounded object knowledge and use that knowledge to solve a variety of cognitive tasks (e.g., object recognition and categorization). The next several sub-sections provide an overview of the relevant findings from psychology and cognitive science. In particular, the overview is designed to answer several key

questions: 1) *How do humans form object concepts and object categories in the first few years of life?* 2) *How do humans make use of multiple sensory channels for object perception?* and 3) *What role do exploratory behaviors play in the context of learning about objects?*

2.1.2 Object Individuation, Identification, and Categorization

Wilcox et al. (2006) define the problem of object individuation as that of determining whether two perceptual stimuli (e.g., visual images, sounds, or tactile signals) belong to the same object or not. Such an ability is a pre-requisite for representing the world in terms of distinct objects and the relations between them. The wider problem of object identification is defined by Kemp et al. (2009) as that of inferring how many distinct objects the environment contains, recognizing when the same object is encountered twice, and identifying whether a stimulus comes from a novel object. Studies in developmental psychology have shown that these skills are fundamental to establishing an internal object representation that can handle the large number of objects that humans encounter in their daily lives (Tremoulet et al., 2000; Krojgaard, 2004).

For this reason, a question of significant interest to developmental psychologists is how infants establish an object representation and subsequently use it to recognize the identities of objects. For example, a study by Tremoulet et al. (2000) showed that even at the age of 12-months, human infants are able to individuate objects using both shape and color information. The same study also found that while both object features were used for the task of figuring out how many objects are there, only the shape feature was used when recognizing the identity of an object that was previously individuated. Other studies have shown that when identifying objects, infants often make different judgments from adults based on the differences in the objects' features (see Wilcox and Baillargeon (1998)), indicating that at such an early age the biological circuits that allow the problem to be solved are still developing.

The ability to individuate objects has also been studied in human adults. As described by Kemp et al. (2009), in a typical scenario the human participant observes (or interacts with) objects one at a time, where the next object may or may not be a previously encountered one. Subsequently, participants may be asked to enumerate the objects that they have observed, or

match an object stimulus to one of the estimated object identities. For example, in a study with human adults, Kemp et al. (2009) showed that as the number of observed objects increases, the likelihood that a novel object will be classified as a previously observed object goes down. The same study also found that humans rely on prior information when solving identification problems. More specifically, to determine whether two perceptual stimuli originate from the same object, humans need prior experience in the form of pairs of perceptual stimuli for which the relationship is known (Kemp et al., 2009). In other words, prior experience with objects with known object identities is necessary in order to solve the object individuation task on a novel set of objects.

Therefore, it is not surprising that humans use a variety of cues, other than object features, when individuating objects (Kemp et al., 2009; Krojgaard, 2004). For instance, spatial cues can be used to individuate objects since observing two objects next to each other indicates that the two objects are not the same (Xu and Chun, 2009). Humans also use temporal cues, e.g., they assume that an object would remain the same object over the course of contiguous manipulation or observation (Becchio and Bertone, 2003). Most importantly, such spatial and temporal cues can inform the observer that the featural differences between the objects are not due to noisy observations, but due to the two objects being different (Kemp et al., 2009; Xu and Chun, 2009).

Developmental psychology also studies how infants and adults form object categories and relational concepts. An important finding is that certain experimental settings can elicit spontaneous sorting and grouping behaviors by infants (see Nelson (1973) and Starkey (1981) for examples). This suggests that even without any specific guidance, from an early age, humans are biased towards spontaneous categorization and grouping of objects. Starkey (1981) reports that both 9 and 12-month-old infants exhibit sorting behaviors when presented with a set of 8 objects, where the set contains 2 groups of four objects that are similar along some dimension (e.g., size, color, etc.).

Sorting and grouping behaviors have also been observed with non-human primates (Potì, 1997; Spinozzi et al., 1999). For example, Spinozzi et al. (1999) found that human-encultured Bonobos and Chimpanzees are capable of spontaneously partitioning a set of objects into two

categories. The authors also report that when chimpanzees partition a set of objects, they predominantly manipulate objects from only one of the two object classes. This procedure is consistent with the behavior of 3-year-old infants observed in a study by Spinozzi et al. (1999). Overall, these findings suggest that the ability to sort objects is fundamental to primate intelligence.

For humans in particular, object grouping skills are thought to be closely related to our language acquisition abilities. For example, Nelson (1973) argued that children form primitive conceptual categories that are later used when binding the meaning of a word. Similarly, based on a large volume of experimental research, Bloom (2000) argues that a large part of early language learning is about establishing a relation that maps language symbols (e.g., individual nouns) to already existing concepts that are formed independently of the language in question. An example of what this may look like is provided by Kemp et al. (2010) who write:

> *"Before learning her first few words, a child may already have formed a category that includes creatures like the furry pet kept by her parents; and learning the word 'cat' may be a matter of attaching a new label to this pre-existing category." (Kemp et al., 2010, p. 216)*

Not surprisingly, a large volume of research has focused on revealing how humans learn the names of categories (see Ashby and Maddox (2005) for a review). In this framework, the participants are typically presented with several examples from each object category and are subsequently asked to categorize a novel item. Researchers have postulated that humans use two different strategies (sometimes in combination) to learn categories from examples. The first strategy involves finding the common features of members of an individual category, while the second strategy consists of identifying the distinctive features among the non-members of that category (Hammer et al., 2009, 2010). Several experiments described by Hammer et al. (2009) have shown that adults can learn categories even when presented only with pairs of objects from different categories. Children between the ages of 6-9 years, however, could only learn the same categories when provided with object pairs in which the two objects come from the same category, indicating that the two strategies for solving this task have different developmental

trajectories (Hammer et al., 2009).

In addition to learning discrete categories, researchers have also examined how human adults and infants learn comparative relations such as "A is bigger than B" (Smith et al., 1986; Gentner and Namy, 2006). As with category learning, humans can learn such relations when presented with paired examples for which the relation is provided by the instructor or inferred by some other means. Thus, the robot in this work will be tested in a similar fashion – after initially interacting with the objects, the learned computational models will be evaluated using both discrete categorization as well as continuous ordering tasks.

While most related studies in psychology have focused on the visual sensory domain, Lederman (1982) argues that human perception of objects is an inherently multi-modal process, one in which humans perceive objects and form object concepts using a variety of sensory modalities (e.g., vision, touch, audio, etc.). In addition, perception of objects is not a passive process – instead, humans actively interact with objects through the use of what psychologists call exploratory behaviors and procedures (Lederman and Klatzky, 1990; Power, 2000). The next two subsections examine in detail how multiple sensory modalities and a rich behavioral repertoire enable humans to solve a wide array of problems, including object recognition and categorization.

2.1.3 Multi-Modal Object Perception

In the field of robotics, object recognition is almost exclusively considered to be a computer vision problem. Research in psychology and cognitive science, however, highlights the importance of sensory modalities other than vision for object recognition tasks. For example, Sapp et al. (2000) describe a study in which toddlers were presented with a sponge that was deceptively painted as a rock. As expected, the toddlers believed that the object was a rock until the moment they interacted with it (by touching it or picking it up). This and several other studies (Heller, 1992) illustrate that *proprioceptive* information (i.e., how objects feel when lifted or pushed) can be very useful when vision alone is insufficient. Studies have also shown that tactile exploratory behaviors are commonly used by infants when exploring a novel object (Ruff, 1984). For example, Stack and Tsonis (1999) have reported that, in the absence

of visual cues, 7-month-old infants use more efficient tactile exploratory strategies and can perform tactile surface recognition to some extent.

Natural sound is also an important source of cues about objects. The work of Gaver (1993) and Grassi (2005) has shown that even when a direct line of sight is not available, humans can extract the physical properties of objects from the sounds that they produce. The importance of everyday natural sounds is perhaps best summarized by Don Norman in his book "The Design of Everyday Things":

> *"[. . .] natural sound is as essential as visual information because sound tells us about things that we can't see, and it does so while our eyes are occupied elsewhere. Natural sounds reflect the complex interaction of natural objects: the way one part moves against another; the material of which the parts are made – hollow or solid, metal or wood, soft or hard, rough or smooth." (Norman, 1988, p. 103)*

According to Gaver (1993), the ecological approach to perception provides the insight that *listening* consists of perceiving the properties of the sound's source (e.g., bouncing ball, car engine, footsteps, etc.), rather than the properties of the sound itself (e.g., pitch, tone, etc.). These insights have been confirmed by multiple experimental studies. For example, Giordano and McAdams (2006) demonstrated that humans can accurately recognize an object's material (e.g., wood, glass, steel, or plexiglass) when listening to the sounds generated when the object is struck. Sound also allows us to perceive many physical properties of objects. Grassi (2005) showed that human subjects were able to provide reasonably good estimates for the size of a ball dropped on a plate by simply listening to the impact sound.

In addition to perceiving the physical properties of objects, non-visual sensory modalities are also useful for object individuation. Wilcox et al. (2006) describe several experiments documenting how infants use auditory information when figuring out whether two stimuli are produced by the same object or by two different objects. Their findings show that sounds that reveal the physical properties and the structure of objects (e.g., rattling sounds) are more useful for individuation than sounds that do not (e.g., tones produced by an electric keyboard).

In a follow-up study, Wilcox et al. (2007) conducted experiments that showed how prior

experience with an object in the tactile sensory domain can subsequently improve an infant's object individuation performance when using color alone. More specifically, their results revealed that combined tactile and visual exploration of objects increases the sensitivity to color differences of 10.5-month-old infants on an object individuation task. According to Wilcox et al. (2007), one possible explanation for this observation is that combined visual and tactile exploration of objects produces more detailed and robust object representations than the ones attained when using visual exploration alone. In fact, other research in psychology has shown that object exploration in a natural setting (as opposed to a research lab) is an inherently multi-modal process. Consider the simple act of touching an object. In Chapter 4 of "Tactual Perception: A Sourcebook", Lederman writes:

> *"Perceiving the texture of a surface by touch is a multi-modal task in which information from several different sensory channels is available. In addition to cutaneous and thermal input, kinesthetic, auditory, and visual cues may be used when texture is perceived by touching a surface. Texture perception by touch, therefore, offers an excellent opportunity to study both the integrated and the independent actions of sensory systems. Furthermore, it can be used to investigate many other traditional perceptual functions, such as lateralization, sensory dominance, and integration masking, figural aftereffects, and pattern recognition." (Lederman, 1982, p. 131)*

Indeed, Lynott and Connell (2009) have shown that humans require the use of two or more sensory modalities to accurately represent many object properties (e.g., texture, stiffness, and material type). This finding suggests that humans can integrate feedback from multiple channels of information in an efficient manner when perceiving objects. Ernst and Bulthof (2004) provide some details on how this is done based on an experimental study in which human participants were tasked with inferring an object's height using both proprioceptive and visual feedback. Their results suggest that humans use a weighted combination of the predictions of the two modalities, where the weights are proportional the estimated reliability of each modality (Ernst and Bulthof, 2004). The weighted combination ensures that a sensory

modality that is not useful in a given context will not dominate over other more reliable channels of information.

Inspired by these findings from psychology, this dissertation shows that a robot's ability to represent objects and perceive their properties may be greatly improved if the robot can experience the objects through a wide variety of sensory modalities. More specifically, this dissertation aims to show that many object properties can only be grounded successfully if the robot is allowed to use non-visual sensory feedback. Indeed, the studies described in this dissertation have already shown that a robot can recognize objects and their properties using auditory, proprioceptive, and tactile feedback, provided that the robot can estimate the reliability of each modality in different sensorimotor contexts (see Chapters 4 and 5).

2.1.4 Object Perception using Exploratory Behaviors

One way in which humans leverage information from different sensory modalities is through the use of what psychologists call *exploratory behaviors* (Power, 2000) or *exploratory procedures* (Lederman and Klatzky, 1990). In his book, "Play and Exploration in Children and Animals", Power writes:

> *"[. . .] exploratory behavior in infancy and childhood appears to serve an information-gathering function. Using a variety of methods, researchers have demonstrated that during exploration infants and young children extract at least short-term information about the characteristic of objects, including information about texture, hardness, weight, shape, size, and sound potential." (Power, 2000)*

Infants' use of exploratory behaviors when learning about objects is tightly connected to their ability to detect sensory events that occur over the course of object manipulation. Gibson (1988) concludes that our basic knowledge about how objects behave in the natural world is gathered through constant observation of how objects are affected by our own actions during play. In other words, when exploring an object, infants observe perceptual outcomes (e.g., sounds and movement patterns) that are subsequently used to form expectations about how an object behaves when a specific action is applied on in (Gibson, 1988).

Numerous experiments in psychology have investigated how such expectations are formed and how they are used to anticipate events in the future. For example, Hauf and Aschersleben (2008) have shown that 9-month-old infants can predict the occurrence of auditory and visual events that occur after pressing a button. The same line of research has even shown that exploratory behaviors may have a role in the early social development of infants. An experiment by Hauf et al. (2007) investigated infants' interest in the actions of others and showed that infants are more interested in watching another person manipulate an object if they themselves have had a chance to explore the object beforehand.

Other research has studied how exploratory behaviors enable infants to ground object properties in their own experience with objects. In a study by Paulus and Hauf (2011), 11-month-old infants were initially exposed to objects of two different materials, one heavy and one light, and after exploring the objects through manipulation, the infants showed preference for the lighter objects. Furthermore, at 13 months, the infants were able to associate the visual appearance of objects with their material type and used that knowledge to show preference towards novel objects made of the lighter material (Paulus and Hauf, 2011).

Combined, these studies show that the ability to apply exploratory behaviors on objects is fundamental to the development of motor, perceptual, and social skills in infancy. An important question is how systematic object exploration strategies emerge over the course of infant development. Power notes:

> *"[. . .] exploratory behaviors become more planful and systematic and are less driven by stimulus characteristics, with increasing child age. Moreover, the use of systematic exploratory strategies is associated with greater information yield." (Power, 2000)*

Thus, when infants first start exploring objects through actions, their behaviors tend to be random and seemingly without an intended purpose or plan. As the infant develops, however, exploration strategies become more systematic and show greater levels of intent. This progression is likely mediated by the acquisition of object knowledge, which serves to guide the application of specific exploratory behaviors intended to uncover specific object properties.

The research described in this dissertation is largely inspired by these findings from psychology. Therefore, the robot in this work explored objects using a wide variety of behaviors, many of them modeled after the ones performed by infants, toddlers, and young children (e.g., scratching, shaking, pushing, grasping, etc.). The studies described here have indeed shown that by using exploratory behaviors, a robot may recognize objects (see Chapters 4 and 5), solve the odd-one-out task (see Chapter 6), as well as assign category labels to novel objects (see Chapter 7).

2.1.5 The Development of Object Knowledge in Infancy

Infants begin to acquire knowledge about objects at a very early age. One of the first things that infants learn about objects is that objects are enduring across time and complete across space (Johnson et al., 2003; Spelke and Kinzler, 2007). How such object representations are formed is a key question in developmental psychology. In an attempt to answer it, Johnson et al. (2003) conducted several experiments that showed that 4 month olds can be trained to recognize that an object moving behind an occluder remains the same object if the trajectory can be observed without the occluding object. At the same age infants can not only track objects, but also they can predict their movement trajectories (Von Hofsten et al., 2007). The findings of Johnson et al. (2003) and Von Hofsten et al. (2007) suggest that infants learn object permanence representations using real-world experience that is derived from viewing different objects undergoing occlusion.

Infants reach another early developmental milestone when they acquire the ability to recognize objects that they have encountered in the past. At 3 months of age, infants are already able to visually recognize an object (Kraebel and Gerhardstein, 2006). More specifically, two experiments conducted by Kraebel and Gerhardstein (2006) showed that training experience consisting of multiple views of the same object can enable a 3-month-old infant to recognize that object in the future, even if it is placed in a previously unobserved orientation. By 5 months of age infants can recognize an object even if it is rotated around a novel axis, i.e., during training the object was rotated around one axis, but at test time it was rotated around another axis (Mash et al., 2007). Their results suggest that by the age of 5 months, infants are not merely

interpolating between the views observed during training, but instead are performing mental rotation when recognizing an object in a novel orientation. Psychologists suggest that such early object recognition skills are acquired by the infant through constant observation of how objects around them move and rotate (Wilcox and Baillargeon, 1998).

In addition to passive observation, the development of object concepts in early infancy has also been shown to be facilitated by the acquisition of motor skills. For example, a study with 4.5 to 7.5 month olds by Soska et al. (2010) showed that the infants' self-sitting and manual object manipulation skills could be used to predict the outcome of a visual object completion task. According to the authors, this result suggests that by the time infants are 4.5 month old, the ability to acquire and use 3D object representations is already tightly connected to motor skills that enable both visual and manual object exploration.

Around the same age, infants begin to form object categories, which initially are based on the objects' perceptual properties and similarities (Colombo et al., 1990; Eimas and Quinn, 1994; Rakison and Butterworth, 1998). Gradually, infants' categorization skills expand and enable the formation of categories based on more abstract object properties. For example, an experiment conducted by Luo et al. (2009) shows that at 5-6 months of age infants differentiate between inert and self-propelled objects and form different expectations for physical events for the two object categories. By 6 to 10 months of age, infants can learn categories of abstract properties such as the objects' function or their spatial relations (Casasola and Cohen, 2002; Casasola et al., 2003; Horst et al., 2005).

Another important developmental milestone of is the ability to use social cues (e.g., spoken words) when learning categories. It has been shown that the category learning performance of 6-month-old infants is sensitive to the presence of auditory words that serve as a label for each category (Fulkerson and Waxman, 2007). While there is some debate in the field regarding the strength of this effect at such an early age (Plunkett et al., 2008), other studies have shown that as the infant grows older the effects of labeling become much more pronounced. For example, the experiments of Plunkett et al. (2008) show that at 10 months of age, the presence of labels that are uncorrelated with the objects' categories inhibits category recognition performance. A possible way to interpret this is that by that age the infant has an expectation that verbal

labels used by adults to describe objects must be correlated with the objects' physical and functional properties.

The effects of labeling during categorization tasks become much more pronounced in the second year of life. For instance, a study by Booth and Waxman (2002) found that by 18 months of age category acquisition and generalization skills are significantly enhanced when training examples are accompanied by a verbal label. This effect was not observed at 14 months of age, which may indicate that a significant developmental shift may be occurring in between. Interestingly enough, providing a label along with a functional cue (e.g., during training, the experimenter performed a specific action with the object, where the action depended on the category) enhanced the categorization skills of both 14 and 18 month olds. The authors postulated that, for 14-month-old infants, demonstrating the function of each object category provides a core meaning for the associated label (Booth and Waxman, 2002). It has also been postulated that 18-month-old infants already know that names refer to some core functional meaning, and therefore, they are able to use names as cues for categorization even when they have not yet discovered what that meaning entails (Davidson and Gelman, 1990; Booth and Waxman, 2002).

Table 2.1 shows a summary of the developmental milestones associated with the gradual accumulation of object knowledge in infancy. Inspired by this developmental progression, the research described in this dissertation is motivated by two main skills that infants acquire in their first year and a half: 1) the ability to individuate and recognize objects, 2) the ability to learn labels that describe individual objects, as well as labels that are used to describe relations between objects (e.g., "bigger"). The next section gives an overview of the work in robotics that is most relevant to this dissertation.

Table 2.1 The Development of Object Knowledge in Infancy

Age	Knowledge or Skill	Source
4 mo	**Object Permanence:** infants can learn that an object moving behind an occluder remains the same object after it reappears. Infants can track an object's motion and predict its trajectory.	Johnson et al. (2003); Von Hofsten et al. (2007)
4 mo	**Categorization:** infants can learn object categories based on perceptual object properties (e.g., perceptual object similarity).	Colombo et al. (1990); Eimas and Quinn (1994); Rakison and Butterworth (1998)
3-5 mo	**Recognition:** infants can perform visual object recognition. Training under multiple orientations facilitates the ability to recognize the object in a novel orientation. By 5 months of age, infants recognize objects rotated around a novel axis.	Kraebel and Gerhardstein (2006); Mash et al. (2007)
5-6 mo	**Categorization:** infants form categories for inert and self-propelled objects. Infants form different expectations for the two categories.	Luo et al. (2009)
7 mo	**Recognition:** infants can discriminate textures by touch. In the absence of visual cues, infants use more efficient exploratory strategies.	Stack and Tsonis (1999)
6-10 mo	**Categorization:** infants begin to form categories based on abstract properties such as the objects' function and their spatial relations. Experiments suggest that the presence of verbal labels associated with categories facilitates category learning.	Casasola and Cohen (2002); Casasola et al. (2003); Horst et al. (2005); Fulkerson and Waxman (2007)
10-12 mo	**Individuation:** infants can individuate objects. Furthermore, naming objects enhances infant object individuation.	Van de Walle et al. (2000); Xu et al. (2005)
10-14 mo	**Categorization:** at 10 months of age, the effects of labeling become more pronounced. Labels that are uncorrelated with the objects' categories were shown to be detrimental to categorization. At 14 months, naming objects in a social setting was shown to enhance category acquisition.	Booth and Waxman (2002); Plunkett et al. (2008)

2.2 Related Work in Robotics

2.2.1 Behavior-Based Object Property Estimation

Enabling robots to manipulate objects in unstructured environments has been a long standing goal of robotics research (Kemp et al., 2007). A wide variety of methods, frameworks, and algorithms have been developed to estimate an object's 3D pose (i.e., position and orientation) and to subsequently compute a joint-space configuration that would enable the robot to grasp the object in order to solve a pick-and-place task.

In contrast, relatively little research has been conducted with the aim of enabling robots to use behaviors as a means of perceiving objects instead of simply changing the objects' states. One exception is a study by Krotkov (1995) in which a robot used behaviors in order to acquire information about objects and their properties. His experiment showed that a robot may estimate an object's mass and coefficient of sliding friction by striking the object with a wooden stick and subsequently observing the object's visual displacement. In a related study, Fitzpatrick et al. (2003) showed that by pushing an object and observing its visual displacement a robot can learn the rolling properties of the object. More recently, Katz and Brock (2008) have demonstrated a framework in which the robot tracks the displacements of individual object features in order to estimate the kinematics of jointed objects (e.g., scissors).

Other experiments have demonstrated that acoustic patterns can also be used to perceive object properties. For example, Krotkov et al. (1996) conducted experiments in which the task of the robot was to identify the material type (e.g., glass, wood, etc.) of different objects by probing them with its end effector. In that study, the robot used a hitting behavior to recognize five different materials: aluminum, brass, glass, wood, and plastic. The results indicate that the spectrogram of the detected sound can be used as a powerful representation for discriminating between the five materials. Subsequent work by Klatzky et al. (2000) showed that modeling frequency and decay parameters of sounds can also be used to build a sound model for each material.

Similarly, experiments by Richmond and Pai (2000) and Richmond (2000) have shown that modeling the spectrogram of the sounds using spectrogram averaging across multiple trials

allows a robot to detect different types of materials from contact sounds. A limitation of these studies is that the robot interacted with a very small number of objects. For example, in the work by Krotkov et al. (1996), the robot only explored one object from each of the five material types, and therefore it is impossible to evaluate whether the learned auditory models would generalize to different objects of the same material types.

Other properties that may be estimated through interacting with the objects include mass and moment of inertia. For example, Kubus and Wahl (2008) described a method for estimating the internal load parameters of an object using force-torque sensors in the robot's joints and an accelerometer in the robot's end effector. Since rigidly grasped objects can be treated as additional links of the robot, methods designed to estimate the robot's own kinematics and dynamics may also be applied in this setting (Atkeson et al., 1986; Hollerbach and Wampler, 1996; Nanayakkara et al., 1999; Krabbes and Döschner, 1999). In contrast, the research proposed here aims to use proprioceptive sensors as just one of many sensory channels used by the robot to learn multi-modal object representations.

The main limitation of the studies reviewed so far is that they typically use just one type of behavior and a limited number of sensory modalities (in most cases, just one). While such a limited sensorimotor repertoire may be sufficient to capture a specific object property, it would not scale up to capture multiple properties, especially ones that are not known to the human programmers in advance. In addition, the studies discussed so far typically use a very small number of objects (usually less than 10 and in some cases only one object per property value). Such experimental designs make it next to impossible to evaluate how well the learned representations generalize to new objects that were not included in the robot's training set. The research described in this dissertation addresses these limitations by using a large and diverse set of robot behaviors, sensory modalities, and object types.

2.2.2 Using Behaviors to Recognize Objects

In addition to estimating specific object properties, behaviors have also been used by robots to recognize the identity of objects. Traditionally, object recognition has been treated as a computer vision problem. Indeed, the majority of robots today can only recognize objects

using visual and/or 3D laser scan data (Quigley et al., 2007; Srinivasa et al., 2009; Rasolzadeh et al., 2010; Rusu et al., 2008). With a clear view of the target object, such systems can achieve high recognition rates, but suffer from several limitations. For example, using vision alone, a robot cannot distinguish between a heavy object and a light object that otherwise look the same. Furthermore, such a system would be of little use if the object is outside the robot's field of view (e.g., grasping an object that is inside a bag).

Several lines of research have attempted to address the limitations of the visual sensory modality by enabling robots to recognize objects using proprioceptive, auditory, and tactile sensory feedback. One of the first such examples is the work of Natale et al. (2004) in which proprioceptive data captured by the robot's hand was used to recognize objects. In their experiments, the robot grasped seven different objects and the resulting joint-angles of the fingers were fed as inputs to a Self-Organizing Map (SOM). The SOM subsequently allowed the robot to distinguish objects of different sizes, as well as objects of similar size but different rigidity. Another approach to proprioceptive object recognition consists of estimating physical properties such as the objects' mass and moment of inertia, and using that information to detect if a given object has been previously observed. Using this method, Kubus et al. (2007) performed an experiment in which a robot was able to recognize the identity of three different objects.

Other studies in non-visual recognition have investigated how robots can recognize surface textures using various forms of tactile feedback. Tanaka et al. (2003) developed an artificial finger that uses strain gauges and polyvinylidene fluoride (PVDF) foil to generate tactile feedback when sliding across a surface. In subsequent experiments, Tanaka et al. (2007) demonstrated how their sensor can detect roughness and temperature changes in the textures of six different fabrics. A similar sensor was developed by Hosoda et al. (2006). By applying two different exploratory behaviors – *pushing* and *rubbing* – their robot was able to distinguish between five different materials. A robotic finger with randomly distributed strain gauges and PVDF films was also proposed by Jamali and Sammut (2010). In their experiments, a Naive Bayes classifier trained with the Fourier coefficients of the sensor's output was used to recognize eight different surface textures. While these studies demonstrate the utility of tactile feedback for recognition

tasks, they typically consider such feedback in isolation and only make use of a limited number of behaviors (usually only one). In contrast, the research proposed here plans to investigate how the tactile sensory modality, coupled with scratching behaviors, can be used in conjunction with other channels of information to build a multi-sensory object representation.

Most of the studies in behavior-based object recognition reviewed so far typically assume that the robot can perform only one behavior on the objects that it explores. More recently, it has been demonstrated that robots can boost their recognition rates by applying multiple different behaviors on the test object. For example, Sinapov et al. (2009) proposed a framework for auditory object recognition using a set of five behaviors: grasp, shake, drop, push, and tap. Using auditory information alone, the robot was able to achieve a recognition rate of over 99% (measured with 36 different household objects). Such a high rate was possible only by applying all five behaviors on each test object and combining the outputs of the recognition models associated with specific behaviors. In subsequent studies, the same boosting effect was also observed when performing recognition using proprioceptive (Bergquist et al., 2009) as well as tactile feedback (Sinapov et al., 2011b). More recently, the feature extraction and similarity estimation methods proposed by Sinapov et al. (2009) were used by Rebguns et al. (2011) to solve an acoustic object recognition task with 10 objects, in which the robot used reinforcement learning to select which behaviors to apply in order to maximize recognition performance.

Another limitation of most current methods for recognizing objects using behaviors is that they typically use only a single sensory modality. In a recent study, we have shown that a robot may further improve its object recognition rate by not only performing multiple behaviors, but also by using multiple sensory modalities (Sinapov et al., 2011a). In that experiment, the robot explored 50 household objects using five different behaviors. Using both auditory and proprioceptive feedback, the robot was able to achieve a recognition rate of over 98%. The results also showed that increasing the number of sensory modalities boosts the object recognition rates similar to the boosting observed with increasing the number of behaviors. In another line of research, Gijsberts et al. (2010) describe a multi-modal object recognition approach that uses grasp affordance features that encode different ways in which an object can be grasped. Using a combination of the grasp affordance features and visual appearance

features, the robot was able to recognize 7 different objects.

A further limitation of most methods used by robots to recognize objects is that they start with a fixed object representation in which the robot's training data is labeled with one of a finite number of object identities (see Torres-Jara et al. (2005); Sinapov et al. (2009); Natale et al. (2004); Rasolzadeh et al. (2010); Bergquist et al. (2009); Rusu et al. (2008); Sinapov et al. (2011a); Marton et al. (2012) for a representative sample of such approaches). These methods implicitly make the assumption that the object individuation task (i.e., inferring how many unique objects have been observed) has already been solved. Providing labeled data, however, becomes increasingly more difficult as the number of objects increases.

In summary, while object recognition in robotics has traditionally been addressed as a visual classification problem, more recent lines of research have explored how the robot's own behaviors can be used to solve this task. Most approaches to date only use a single behavior and a single modality and are typically evaluated on a small set of objects. In addition, virtually all previous approaches assume that all of the training data is labeled with the correct object identity. This assumption, however, is impractical since it would be impossible for a human instructor to label the data for each individual object that a robot may possibly interact with in a home or an office. The research proposed here will address these limitations by developing methods that can scale up to a larger number of behaviors, sensory modalities, and objects. In addition, as described in Chapter 10, the robot in this research is not only tasked with recognizing the identity of a previously explored object, but is also tasked with solving the object individuation problem. Thus, this research relaxes the assumption that all perceptual experience with objects that is used to train the robot must be annotated with an object label.

2.2.3 Object Categorization

Most object categorization methods in robotics fall into one of two broad categories: 1) unsupervised methods, in which objects are categorized using unsupervised machine learning algorithms (e.g., k-Means, Hierarchical Clustering, etc.) and 2) supervised methods, in which a training set of objects is annotated with the correct labels and used to train a recognition model that can label new data points.

Several lines of research have demonstrated methods that enable robots to autonomously form internal object categories based on direct interaction with objects (Nakamura et al., 2007; Griffith et al., 2012; Dag et al., 2010; Sun et al., 2010b). For example, Griffith et al. (2012) showed how a robot can use the frequencies of auditory and visual events in order to distinguish between container and non-container objects. Dag et al. (2010) and Sinapov and Stoytchev (2008) have also shown that, through interaction with objects, robots can learn to categorize and relate objects based on the types of effects that they produce as a result of the robot's actions.

In contrast, supervised methods for object categorization attempt to establish a direct mapping between the robot's object representation and human-provided semantic category labels. A wide variety of computer vision methods have been developed that attempt to solve this problem using visual image features coupled with machine learning classifiers (Fergus et al., 2004; Ponce, 2006; Opelt et al., 2006). Several such methods have been developed for use by robots, almost all working exclusively in the visual domain (Lopes and Chauhan, 2007; Lai and Fox, 2009; Marton et al., 2009; Wohlkinger and Vincze, 2010; Leonardis and Fidler, 2011; Lai et al., 2011a). One advantage of visual object classifiers is that they can often be trained offline on large image datasets. Nevertheless, they cannot capture object properties that cannot always be perceived through vision alone (e.g., object compliance, object material, etc.). In other words, disembodied object category representations that are grounded solely in visual input cannot be used to capture object properties that require active interaction with an object. To address this limitation, the robot in this research grounded the semantic category labels of objects in its own sensorimotor experience with them, which is in stark contrast with approaches that rely purely on computer vision datasets.

Indeed, several studies have already demonstrated some ability of robots to assign category labels to objects based on interaction with them (Takamuku et al., 2007; Sinapov and Stoytchev, 2009; Araki et al., 2011; Chitta et al., 2011). For example, Takamuku et al. (2007) demonstrated that a robot can classify 9 different objects as either a rigid object, a paper object, or a plastic bottle using auditory and joint angle data obtained while the robot shook the objects. Araki et al. (2011) described a robot that learned to associate words describing an object (e.g., "cup")

with object clusters discovered using an unsupervised method. Sinapov and Stoytchev (2009) showed that by applying five different exploratory behaviors on 36 objects, a robot may learn to recognize their material type and whether they are full or empty, based on the auditory feedback produced by the objects.

More recently, we proposed a graph-based learning method that allows a robot to estimate the category label of an object based on pairwise object similarity relations estimated from different couplings of five exploratory behaviors and two sensory modalities (Sinapov and Stoytchev, 2011). In that experiment, the robot was able to classify 25 objects according to object categories such as plastic bottles, objects with contents, pop cans, etc. The accuracy was substantially better than chance, despite the fact that visual feedback was not used.

Despite all of these advances, current work on category recognition suffers from two broad limitations. First, most object category recognition approaches are entirely vision-based and as such, they would be unable to detect object properties that cannot be extracted using vision alone. While some research has focused on using different sensory modalities coupled with actions, most studies to date use a small number of behaviors (typically just one) and a small number of sensory modalities. To address this limitation, the research described here grounds human-provided category labels in a wide variety of robot behaviors and sensory modalities. Our results indicate that using a large number of different behaviors (10 in this case) coupled with the visual, auditory, and proprioceptive sensory modalities can enable a robot to recognize 20 different categories over a set of 100 different objects (Sinapov et al., 2012).

The second broad limitation of most existing approaches is that they only deal with human-provided semantic labels that can be expressed as a *unary* relations. For instance, any object category can be viewed as a collection of items that share some property (e.g., round, red, etc.). Many semantic labels, however, cannot be expressed with unary relations. For example, the label "taller than", can only be expressed as *binary* relation. Furthermore, in most learning tasks, the robot is only tasked with learning to detect the value of a given attribute (e.g., the color of an object). Such a robot would be able to classify a red ball as having the label "red," but would still be unable to detect that a *set* of objects vary by the attribute "color." To address these limitations, this document describes a relational approach to representing

semantic category labels that can handle many types of object relations beyond simple unary object categories (see Chapter 9).

2.3 Summary

Research in psychology and cognitive science has shown that humans ground object knowledge using a rich sensorimotor repertoire consisting of a diverse set of exploratory behaviors and sensory modalities. Our drive to explore objects by interacting with them during infancy is fundamental to our ability to perform complex manipulation and cognitive tasks in adulthood.

Despite these findings, most robots today do not explore objects the way humans do, but instead use object representations that are carefully designed by a human programmer. In addition, most paradigms in robotic perception focus almost exclusively on the visual sensory modality, which, while useful for some tasks, cannot capture many object properties that are relevant to the tasks that we want robots to perform in our homes and offices. To bridge this gap, this dissertation formulates a behavior-grounded approach to object perception and exploration that advances the state of the art in robotic recognition and categorization of everyday objects.

CHAPTER 3. EXPERIMENTAL PLATFORM

This chapter provides an overview of the robotic platform that was used to conduct the research described in this dissertation.

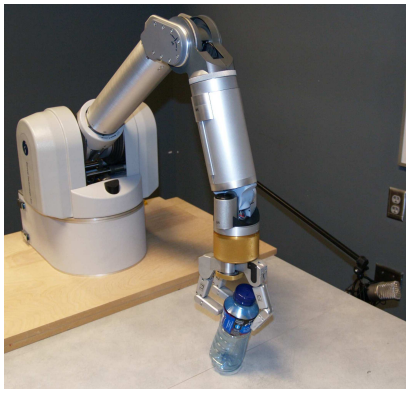
3.1 Robot

The experimental platform is an upper-torso humanoid robot. The robot's arms are two 7-DOF Barrett Whole Arm Manipulators (WAMs). Each arm is equipped with the 3-finger Barrett Hand as an end effector. The arms are controlled in real time from a Linux PC at 500 Hz over a CAN bus interface. The robot is shown in Figure 3.1.

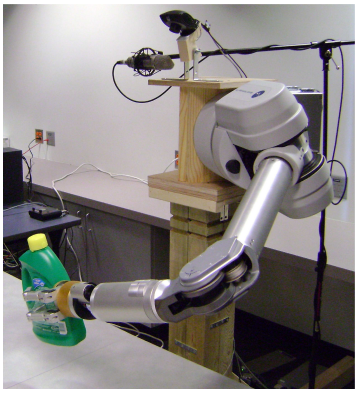
Figure 3.1 also shows how the robot's hardware configuration evolved over time. The first experiment with the robot, described in Sinapov et al. (2008), was conducted with a single Barrett WAM that was mounted horizontally on the table and a single Rode NT1-A microphone, as shown in Figure 3.1.a. The initial prototype for the upper-torso humanoid robot, shown in Figure 3.1.b was constructed by mounting the WAM on a wooden fixture. Finally, Figure 3.1.c shows the robot in its present state with two WAMs as arms.

The early prototypes did not include the actuated head, which was subsequently added to the robot. The head has two degrees of freedom in the neck, allowing it to pan and tilt. Each eye can pan independently and the two eyes can tilt simultaneously either up or down. Finally, the robot can also make facial expressions through the use of 4 servos that move the robot's mouth and eyebrows. Thus, the robot's head has a total of $2+2+1+4=9$ degrees of freedom.

The configuration of the robot's arms was designed to be similar to that of human arms so that the robot can manipulate objects placed in front of it in a human-like manner. Figure 3.2 shows several CAD drawings of the robot and its sphere of reach. The range of motion of each



a) June, 2008



a) September, 2008

b) April, 2011

Figure 3.1 Different stages of the design of the upper-torso humanoid robot used in the experiments conducted for this research.

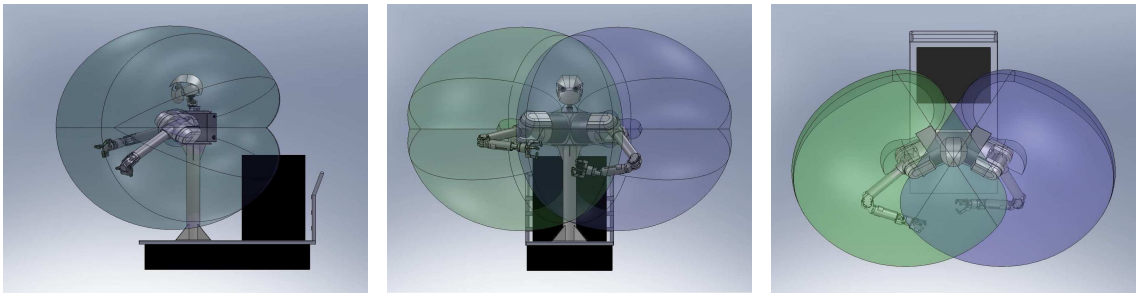
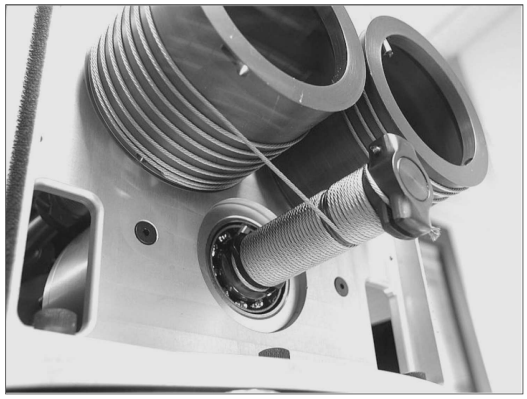


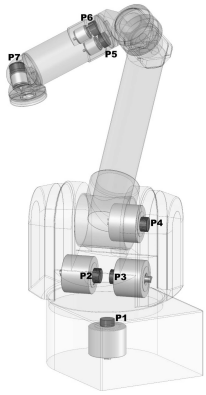
Figure 3.2 CAD drawings that illustrate the sphere of reach of each of the robot's arms. The intersection of the two hemispheres denotes the region in which bi-manual manipulation is possible. These images were drawn by Steven Lischer who helped design the robot's mounting fixture.

arm covers a hemisphere-shaped region of space. The intersection of the two regions denotes the space in which bi-manual manipulation is possible.

A distinct feature of the Barrett WAM is that it is backdrivable, allowing the joint controllers to detect the force applied to each joint and to apply joint torques at the same time. This feature is a direct result of the WAM's transmission system – while most robots use gears at each joint, the WAM uses a low-inertia and low-friction cable-and-cylinder drive, shown in Figure 3.3.a. The 7 joints of the WAM are controlled by miniature servo-controllers, also called motor pucks. Figure 3.3.b shows the layout of the pucks in the WAM. For more details on the WAM and its history, see Rooks (2006).



a) cable-and-cylinder drive



b) puck layout of the WAM

Figure 3.3 a) The cable-and-cylinder drive for one of the WAM's joints; b) The layout of the WAMs servo-controllers, also called motor pucks. Adapted from Rooks (2006).

The WAM's backdrivability allows experimenters to physically demonstrate a desired trajectory motion by moving the WAM with their hands. The ability to record motion trajectories allows for quick scripting of various exploratory behaviors (e.g., pushing, lifting, etc.) that the robot can perform on objects.

3.2 Sensors

3.2.1 Proprioception

The robot is equipped with a variety of sensors that enable it to perceive the properties of objects through a large number of sensory modalities. Each Barrett WAM has built-in sensors in the joints that measure joint angles and motor torques at 500 Hz. In addition to joint-torques, the strains and positions of the fingers of the Barrett Hand can also be measured. The newer hand design (BH8-280), which is placed on the robot's right arm, also provides the exact torques applied at each finger joint in real time. Finally, the robot's right hand is also equipped with a Force-Torque sensor at the wrist that measures 6-DOF forces and torques at the robot's end effector. Collectively, these sensory signals form the robot's proprioceptive or haptic sensory system.

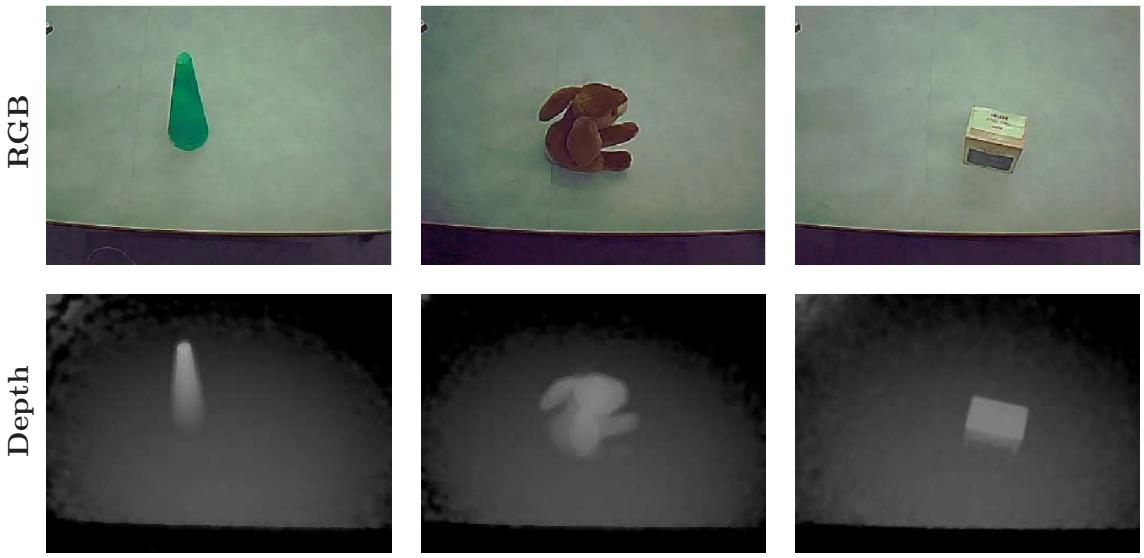


Figure 3.4 Example RGB images and their corresponding depth images taken by the 3DV Systems' ZCam that is mounted on the robot's head.

3.2.2 Vision

The robot's primary visual sensors consist of two Logitech webcams that capture 640×480 RGB images. Each of the webcams is mounted on a 2-DOF pan-tilt base unit, embedded in the robot's head, allowing the robot to control the gaze direction of each eye. The two pan axes are independent of each other, while the tilt axes are coupled.

The robot also has a ZCam, developed by 3DV Systems, which captures 640×480 RGB images as well as 320×280 depth images. As seen in Figure 3.1, the ZCam is mounted on top of the robot's head. Figure 3.4 shows example RGB and depth images capture by the robot's ZCam as the robot looks at different objects placed on the table in front of it. These images are part of the dataset described in Chapter 4.

In addition to the RGB webcam and the ZCam, the robot was recently equiped with a Microsoft Kinect sensor which captures RGB images as well as 3D point clouds. The sensor was mounted on the robots base and pointed towards the table used by the robot to interact with objects. Figiure 3.5 shows a sample image and its corresponding 3D point cloud captured by the sensor. The images are part of the dataset described in Chapter 9.

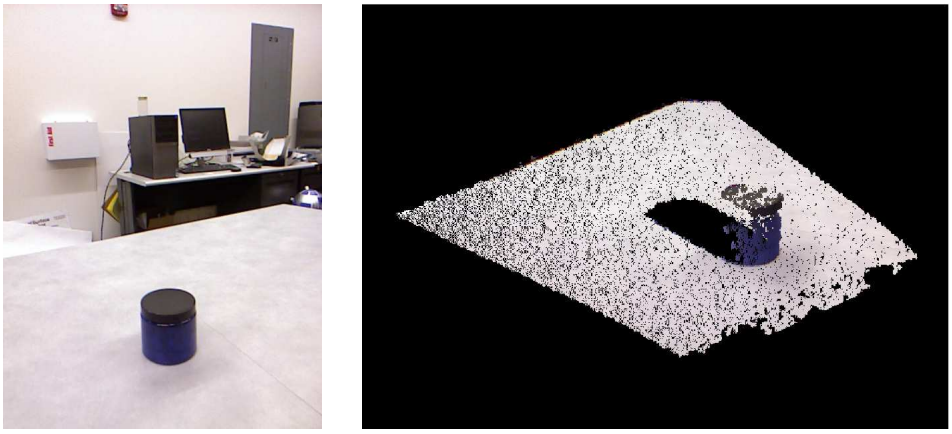
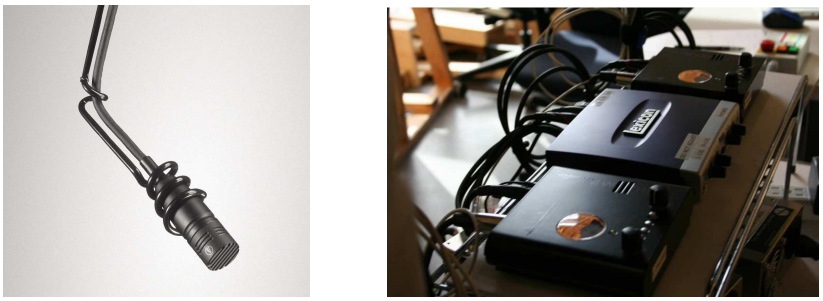


Figure 3.5 Example RGB image and its corresponding 3D point cloud captured by the robot's Microsoft Kinect sensor.

3.2.3 Audio

The robot is also equipped with microphones in order to detect auditory feedback that is produced by different objects as the robot interacts with them. The early prototype of the robot used a single Rode NT1-A microphone (see Figures 3.1.a and 3.1.b). That microphone had a cardioid polar pattern with an average self noise of 5 dB. The microphone's output was routed through an ART Tube MP Studio pre-amplifier. The pre-amplifier supplied 48 volt phantom power to the microphone and sufficient gain was used on the pre-amplifier to provide a suitable input level.

The later version of the robot (see Figure 3.1.c) was equipped with two Audio-Technica U853AW cardioid hanging microphones that were placed inside the robot's head. The output of each microphone was first routed through an ART Tube MP Studio Microphone pre-amplifier and was subsequently processed through a Lexicon Alpha bus-powered audio interface, which connects to the PC using USB. Sound input was recorded at 44.1 KHz using the Java Sound API over a 16-bit channel. Figure 3.6 shows an image of the type of microphone that was used as well as a picture of the audio system that was used to route the microphones' output to the PC.



a) U853AW Microphone b) Audio system

Figure 3.6 a) The Audio-Technica U853AW microphone; b) the two pre-amplifiers (ART Tube MP Studio Microphone pre-amplifiers) and the buspowered interface (a Lexicon Alpha bus-powered interface) that are used to route the microphones' output to the PC.

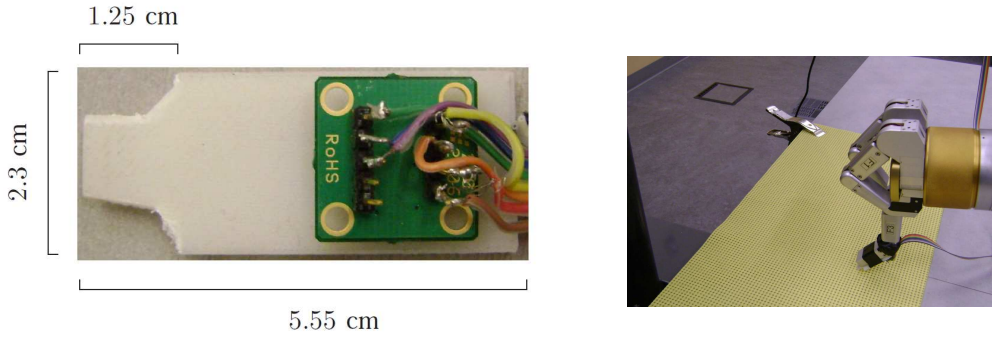


Figure 3.7 The artificial fingernail with the three-axis accelerometer sensor, shown by itself (left) and mounted on one of the robot's fingers (right). The thickness of the fingernail was 0.3175 cm (1/8th of an inch).

3.2.4 Tactile

To perceive object properties through touch, the robot has a vibrotactile sensor in one of its fingers, shown in Figure 3.7. The sensor consists of an artificial fingernail made of ABS plastic and the ADXL345 3-axis digital accelerometer mounted on the EVAL-ADXL345Z evaluation board. Both the accelerometer and the evaluation board were manufactured by Analog Devices. The accelerometer's output rate was 400.0 Hz using ten-bit resolution with a range of ± 2 g for each axis. The ADXL-345 accelerometer uses an on-board digital low-pass filter, but does not have any analog anti-aliasing filters.

The ABS plastic fingernail was designed with computer-aided design software and printed using a rapid prototyping 3-D printer. The EVAL-ADXL345Z accelerometer evaluation board was mounted on the fingernail, which, in turn, was attached to the middle finger (i.e., F3)

Table 3.1 Summary of the Robot's Sensory Modalities

Sensory Modality	Sensors
Vision	2 RGB Logitech Webcams 1 RGB-D ZCam 1 Microsoft Kinect
Proprioception	Joint-torque sensors (arm) Finger strains and torques (hand)
Tactile	3-axis fingertip accelerometer
Audio	2 Audio-Technica U853AW microphones

of the robot's left hand such that its tip protruded from the robot's finger. When the robot performed a scratching behavior, the vibrations of the fingernail were captured by the attached accelerometer. The accelerometer data were transferred to the PC over a universal serial bus (USB) at 400 Hz using the Arduino Duemilanove microcontroller. The sampling-frequency limitation was due to the limited serial port bandwidth of the Arduino board that was used to communicate with the accelerometer.

3.3 Summary

To summarize, the experimental platform is an upper-torso humanoid robot that has two Barrett WAMs as arms. Each WAM is backdrivable, which allows a human user to quickly program a new exploratory behavior by recording a new joint-space trajectory that can later be replayed by the robot. A table is placed in front of the robot so that it can reach and interact with objects placed on the table.

The robot has sensors that capture data from four different sensory modalities: vision, proprioception, audio, and touch. Table 3.1 lists the sensors for each modality. The aim of the research described here is to use all of the robot's sensors to acquire a rich multi-modal object representation. The following chapters describe the experiments that were conducted using the experimental setup described here. They also explicitly mention which subset of the sensory modalities listed in Table 3.1 were used in each experiment.

CHAPTER 4. BEHAVIOR-GROUNDED OBJECT RECOGNITION*

4.1 Introduction

Object exploration is one of the hallmarks of human and animal intelligence. As noted by Piaget (1952), infants perform a large set of exploratory behaviors such as grasping, shaking, dropping, and scratching on most objects they encounter. Such behaviors are commonly used to learn about objects and their physical properties (Lederman and Klatzky, 1987). Object exploration procedures have also been observed in a wide variety of animal species (Power, 2000). Some birds, for example, perform almost their entire behavioral repertoire when exploring an object for the first time (Lorenz, 1996).

Interactive object exploration is also an inherently multi-modal process. Lederman (1982) notes that surface texture can be perceived by sliding one's finger on the surface to obtain tactile sensations, but that behavior also produces auditory feedback, which can help to identify the texture. Indeed, many object properties can only be characterized using multiple modalities (Lynott and Connell, 2009). In contrast, most object recognition systems used in robotics today use almost exclusively computer vision techniques and thus rely on a single modality, see (Quigley et al., 2007; Srinivasa et al., 2009; Rusu et al., 2008; Rasolzadeh et al., 2010) for several examples.

To address the inherent limitations of the visual sensory modality, this dissertation introduces a novel behavior-grounded method for interactive recognition of household objects using the sensory feedback produced over the course of manipulating the object. While vision-based approaches typically use passive observation, the framework described here uses active interac-

*This chapter is based on the following paper: Sinapov, J., Bergquist, T., Schenck, C., Ohiri, U., Griffith, S. and Stoytchev, A., "Interactive Object Recognition Using Proprioceptive and Auditory Feedback", *The International Journal of Robotics Research*, 30(10), pp. 1250–1262, 2011.

tion to recognize the objects. The framework was tested with an upper-torso humanoid robot, which interacted with 50 different household objects. The robot recognized the objects by extracting features from its proprioceptive and auditory sensory streams, while applying five different exploratory behaviors on the objects: *lift*, *shake*, *drop*, *crush*, and *push*. The robot was evaluated on the task of object recognition given the feedback from either one or both of these sensory modalities.

The results show that both auditory and proprioceptive feedback, coupled with specific behaviors, contain information indicative of the object being manipulated. In addition, the robot was able to integrate feedback from multiple modalities and multiple behaviors performed on each test object, which resulted in recognition accuracy of over 98%. Further analysis of these results gives a strong indication that equipping robots with a diverse set of exploratory behaviors is necessary in order to scale up interactive recognition methods to a large number of objects.

4.2 Related Work

4.2.1 Psychology and Cognitive Science

The work presented in this chapter is directly inspired by research in psychology and cognitive science, which highlights the importance of sensory modalities other than vision for object recognition tasks. For example, Sapp et al. (2000) described a study in which toddlers were presented with a sponge that was deceptively painted as a rock. As expected, the toddlers believed that the object was a rock until the moment they interacted with it (by touching it or picking it up). This and several other studies illustrate that proprioceptive information about objects can be very useful when vision alone is insufficient (Heller, 1992).

Natural sound is also an important source of information. It allows us to perceive events and to recognize objects and their properties even when a direct line of sight is not available. The ecological approach to perception provides the insight that *listening* consists of perceiving the properties of the sound's source (e.g., bouncing ball, car engine, footsteps, etc.), rather than the properties of the sound itself (e.g., pitch, tone, etc.) (Gaver, 1993). Thus, the human

auditory system plays a crucial role in both understanding and representing object knowledge. Our hypothesis is that this association can be learned by coupling behaviors performed on objects with the sounds produced during these interactions.

These insights have been confirmed by multiple experimental studies. For example, Giordano and McAdams (2006) demonstrated that humans can accurately recognize an object's material (e.g., wood, glass, steel or plexiglass) when listening to the sounds generated when the object is struck. Sound also allows us to perceive many physical properties of objects. Grassi (2005) showed that human subjects were able to provide reasonably good estimates for the size of a ball dropped on plates by simply hearing the impact sound. Motivated by these and other examples, this chapter investigates a method that allows a robot to use sound as a source of information about objects in a similar manner.

4.2.2 Robotics

Traditionally, most object recognition systems used by robots have relied heavily on computer vision techniques (Quigley et al., 2007; Srinivasa et al., 2009; Rasolzadeh et al., 2010) and/or 3D laser scan data (Rusu et al., 2008). There has been relatively little previous work dealing exclusively with proprioceptive and auditory object recognition. One of the few examples is the work by Natale et al. (2004) in which a robot was able to recognize seven objects with the help of a Self-Organizing Map using proprioceptive data extracted from the robot's hand as it grasped an object.

Proprioceptive data has also been used to estimate an object's mass and moment of inertia (Kubus et al., 2007; Kubus and Wahl, 2008). Methods for estimating the dynamics of a robot's body could also be applied to estimate the mass of an object or some other properties (Atkeson et al., 1986; Hollerbach and Wampler, 1996; Nanayakkara et al., 1999; Krabbes and Döschner, 1999). In contrast, the research presented in this chapter explores how a general sequential representation for high-dimensional sensory data, coupled with standard machine learning algorithms, can be used by the robot to learn to recognize the objects that it manipulates. Thus, the method described here is not specific to proprioception, but can be applied to two (and possibly more) different modalities.

In other related work, Nakamura et al. (2007) describe a robot that uses proprioception along with visual and auditory information when interacting with objects. The robot used one modality to infer the outputs of another (e.g., whether an object would make noise when picked up after only looking at it). Metta and Fitzpatrick (2003) show that integrating proprioception with vision can bootstrap a robot's ability to manipulate objects.

Similarly, there has been some work on the use of auditory information for recognizing objects and their properties. One of the first studies in this area was conducted by Krotkov et al. (1996). Their robot was able to identify the material type (aluminum, brass, glass, wood, or plastic) of several objects by probing them with its end effector. Auditory-based material recognition has also been the topic of research conducted by Richmond and Pai (2000) and Richmond (2000), who described a platform for measuring contact sounds between a robot's end-effector and objects made of different materials. The robot was able to acquire acoustic models for four objects of different material types by repeatedly striking the objects at different positions.

Torres-Jara et al. (2005) demonstrated a robot that can perform acoustic-based object recognition using the sounds generated when tapping on the objects with its end effector. When tapping on a novel object, the spectrogram of the detected sound was matched to one that was already in the training set, which resulted in a prediction for the object's type. This allowed the robot to correctly recognize four different objects.

More recently, Sinapov et al. (2009) have shown that object recognition using auditory feedback can be scaled up to a larger number of objects - 36 - and extended to multiple robot behaviors (e.g., grasp, shake, tap, drop, push). The robot was able to recognize with high accuracy both the type of object and the type of interaction (i.e., exploratory behavior) using only the detected sound.

4.3 Theoretical Framework

4.3.1 Problem Formulation

Let N be the number of behaviors in the robot's repertoire, and let M be the number of sensory modalities (in our case, $N = 5$ and $M = 2$). Upon executing behavior i on a target object, the robot detects sensory stimuli $X_i^1, \ldots, X_i^M$, where each X_i^j is the sensory feedback from modality j. In the most general case, each stimulus can be represented either as a real-valued vector, or as a structured data point (e.g., a sequence or a graph).

The task of the robot is to recognize the target object by labeling it with the correct object label $o \in \mathcal{O}$, the set of all objects. To solve this problem, for each behavior, i, and each modality, j, the robot learns a model $\mathcal{M}_i^j$ that can estimate the object label probability $Pr(o|X_i^j)$. In other words, for each combination of behavior and modality, the robot learns a classifier that estimates the class label probability for each $o \in \mathcal{O}$. The following two sub-sections describe how the robot integrates stimuli from multiple modalities and multiple behaviors in order to further improve the accuracy of its predictions.

4.3.2 Combining Multiple Modalities

For each behavior i, the robot learns a model $\mathcal{M}_i$, which combines the class-label probabilities of the modality-specific models $\mathcal{M}_i^j$ (for $j = 1$ to M). Given sensory stimuli $X_i^1, \ldots, X_i^M$ detected while performing behavior i on a given object, the robot estimates the class-label probabilities for this object as:

$$Pr(o|X_i^1, \ldots, X_i^M) = \alpha \sum_{j=1}^{M} w_i^j Pr(o|X_i^j)$$

In other words, given the stimuli from the M available sensory modalities, the robot combines the class-label estimates of the modality-specific models $\mathcal{M}_i^j$ using a weighted combination rule. The coefficient α is a normalizing constant, which ensures that the probabilities sum up to 1.0. Each weight w_i^j corresponds to an estimate for the reliability of the model $\mathcal{M}_i^j$ (e.g., its accuracy).

It is worth noting that humans integrate information from multiple modalities in a similar way when performing the same task (Ernst and Bulthof, 2004). For example, when asked to infer an object property given proprioceptive and visual feedback, humans use a weighted combination of the predictions of the two modalities. Experimental results described by Ernst and Bulthof (2004) have shown that the weights are proportional to the estimated reliability of each modality. The weighted combination of predictions ensures that a sensory modality that is not useful in a given context will not dominate over other more reliable channels of information.

4.3.3 Combining Multiple Behaviors

To further improve the quality of its predictions, the robot uses not only multiple sensory modalities, but also applies multiple behaviors. After performing n distinct behaviors on the test object (where $n \leq N$), the robot detects sensory stimuli $[X_1^1, \ldots, X_1^M], \ldots, [X_n^1, \ldots, X_n^M]$. As in the case of combining multiple modalities, the robot uses a weighted combination rule and labels the test object with the class label $c \in \mathcal{C}$ that maximizes:

$$Pr(c|X_1^1, \ldots, X_1^M, \ldots, X_n^1, \ldots, X_n^M) = \alpha \sum_{i=1}^{n} \sum_{j=1}^{M} w_i^j Pr(c|X_i^j)$$

The following sub-sections describe the experimental setup, as well as the feature extraction and machine learning algorithms that were used to evaluate the object recognition framework described here.

4.4 Experimental Setup

4.4.1 Robot

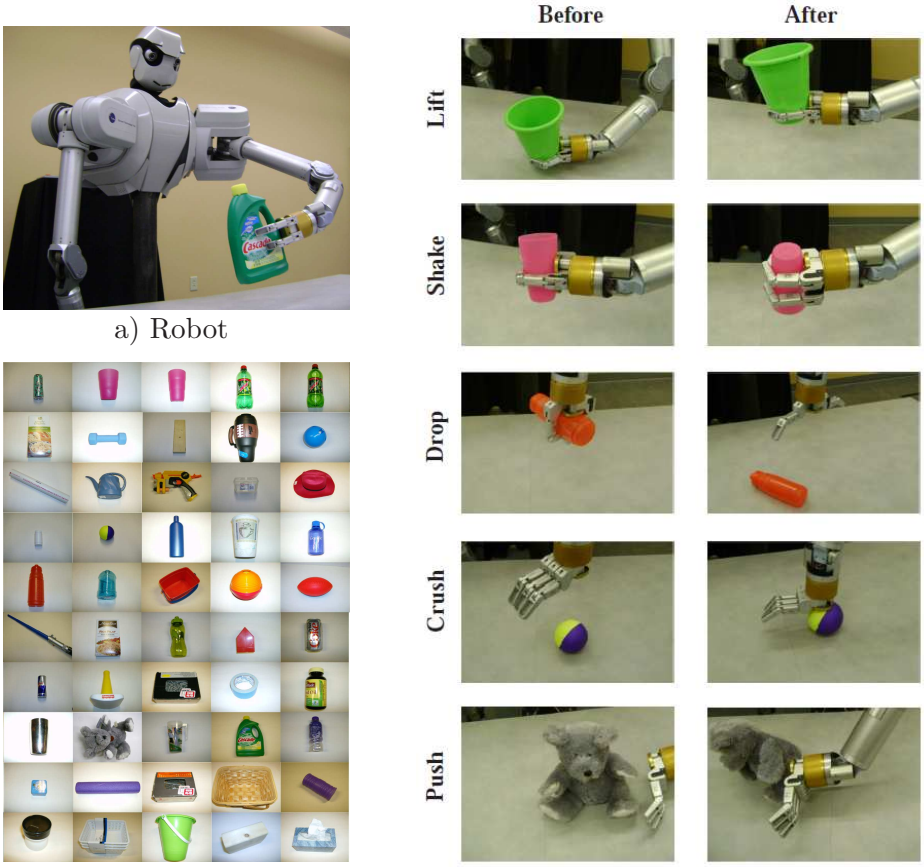
The robot used to evaluate the proposed recognition method was an upper-torso humanoid robot, with two 7-DOF Barrett WAMs for arms and two 3-finger Barrett Hands as end effectors (see Fig.4.1.a). The robot was controlled in real time from a Linux PC at 500 Hz over a CAN bus interface. The raw torque data was captured and recorded at 500Hz using the robot's low-level API. The robot's head was equipped with an Audio-Technica U853AW cardioid hanging microphone. The microphone's output was first routed through an ART Tube MP Studio Microphone pre-amplifier, and subsequently processed through a Lexicon Alpha bus-powered audio interface, which connects to the PC using USB. Sound input was recorded at 44.1 KHz using the Java Sound API over a 16-bit channel. Chapter 3 provides additional details about the robot.

4.4.2 Objects

The robot interacted with a set of objects, $\mathcal{O}$, consisting of 50 common household objects, including cups, bottles, and toys (see Fig. 4.1.b). The objects were made of various materials such as metal, plastic, paper, foam, and wood. Objects were selected using three criteria: 1) they must be graspable by the robot; 2) they must not break or permanently deform when the robot interacts with them; and 3) they must not damage the robot.

4.4.3 Behaviors

The set of behaviors, $\mathcal{B}$, consisted of five exploratory behaviors that the robot performed on each object: *lift*, *shake*, *drop*, *crush*, and *push*. The behaviors were performed with the robot's left arm, and encoded with the Barrett WAM API. Fig. 4.1.c shows *before* and *after* images for each of the five exploratory behaviors. Prior to the execution of each trial, each object was placed in roughly the same configuration (position and orientation). Due to human error, however, there was still some variation of the grasp contact points, as well as the contact



a) Robot

b) Objects

c) Behaviors

Figure 4.1 a) The upper-torso humanoid robot used in this experiment; b) The set of 50 household objects explored by the robot; c) The five exploratory behaviors that the robot performed on each object.

points with the object during the *push* and *crush* behaviors across multiple trials with the same object.

For each of the five interactions, the robot performed ten trials with each of the 50 objects for a total of $5 \times 10 \times 50 = 2500$ recorded interactions. For each trial, the raw proprioceptive and auditory data were recorded for the duration of each behavior. These data were later used to evaluate the behavior-grounded object recognition method proposed here.

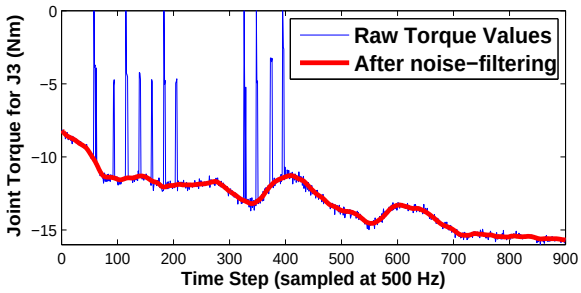


Figure 4.2 Joint torque values for J_3 as the robot lifts the dumbbell object. The thinner line shows the raw joint torques recorded using the robot's low-level API. The thicker line shows the filtered joint torques.

4.5 Feature Extraction and Learning Methodology

4.5.1 Proprioceptive Feature Extraction

The first step in the feature extraction routine was to noise filter the raw joint torque values of the left arm, which were recorded during each interaction. As can be seen in Fig. 4.2, the raw values were somewhat noisy, containing many spike readings. To handle this noise, the raw data was filtered using a filter of width 10, which checked for data points that lie more than 3 standard deviations away from the window median. Any such values were thrown out and replaced with the window median. The time series was then smoothed using a moving-average filter of size 10. The solid line in Fig. 4.2 shows the resulting smoothed torque values after the noise-filtering procedure was performed.

The proprioceptive feedback, P_i, from the i^{th} interaction was represented as a sequence of states in a Self-Organizing Map (SOM) (Kohonen, 2001), one of several ways to quantize data vectors. This representation was obtained as follows: let $T_i = [t_1^i, t_2^i, \ldots, t_{li}^i]$ be the noise-filtered joint torque values for some interaction i, such that each $t_j^i \in \mathbb{R}^7$ denotes the torque values for all 7 joints of the left arm at time step j. Given a set of joint torque records $\mathcal{T} = \{T_i\}_{i=1}^K$, collected over K interactions with different objects, a set of individual joint torque vectors was sampled at random and used as an input training data set for the SOM. In other words, the SOM was trained with seven-dimensional input vectors, $t_j^i \in \mathbb{R}^7$, where each data point denoted a particular record of joint torque values (for all 7 joints). To avoid overfitting and to speed up the training process, only 1/5 of the available input data points were used for training. The

Growing Hierarchical SOM toolbox was used to train a 6 by 6 SOM (i.e., 36 total nodes) using the default parameters[†] for a non-growing 2-D single layer map (Chan and Pampalk, 2002). Figure 4.4.a gives an overview of the training procedure while Figure 4.4.b shows how a torque record, T_i, can be mapped to a discrete sequence of states in the SOM.

4.5.2 Auditory Feature Extraction

Similarly, the auditory feedback from each interaction, A_i, was also represented as a sequence of states in another Self-Organizing Map. To do this, features from each sound were first extracted using the log-normalized Discrete Fourier Transform (DFT), using $2^5 + 1 = 33$ frequency bins with a window of 26.6 milliseconds, computed every 10.0 milliseconds. The SPHINX4 natural language processing library was used to compute the DFT (Lee et al., 1990). The spectrogram (see Fig.4.4.c for an example) encodes the intensity level of each frequency bin (vertical axis) at each given point in time (horizontal axis).

As in the case with proprioceptive data, a 6 by 6 SOM was trained on extracted column vectors from the set of DFT spectrograms detected by the robot (see Figure 4.4). In other words, the SOM was trained with input data points in $\mathbb{R}^{33}$ that represented the intensity levels for each of the 33 spectrogram frequency bins at a given point in time. Once the auditory SOM was trained, a column vector from any particular spectrogram could be efficiently mapped to a unique state in the SOM that has the highest activation value given the input vector. Thus, each sound was represented as a sequence, $A_i = a^i_1 a^i_2 \ldots a^i_{m^i}$, where each $a^i_k \in \Gamma_a$, Γ_a was the set of nodes in the auditory SOM, and m^i was the number of column vectors in the spectrogram (see Fig. 4.4).

[†]Planar SOM with Euclidean distance metric, learning rate $\lambda = 0.7$, and 5 training cycles. The size of the SOM (6 by 6) was heuristically chosen based on prior work by Sinapov et al. (2009) and was not tuned to maximize performance. Parameters governing the growth of the map did not affect the results because the training option for a non-growing map was used.

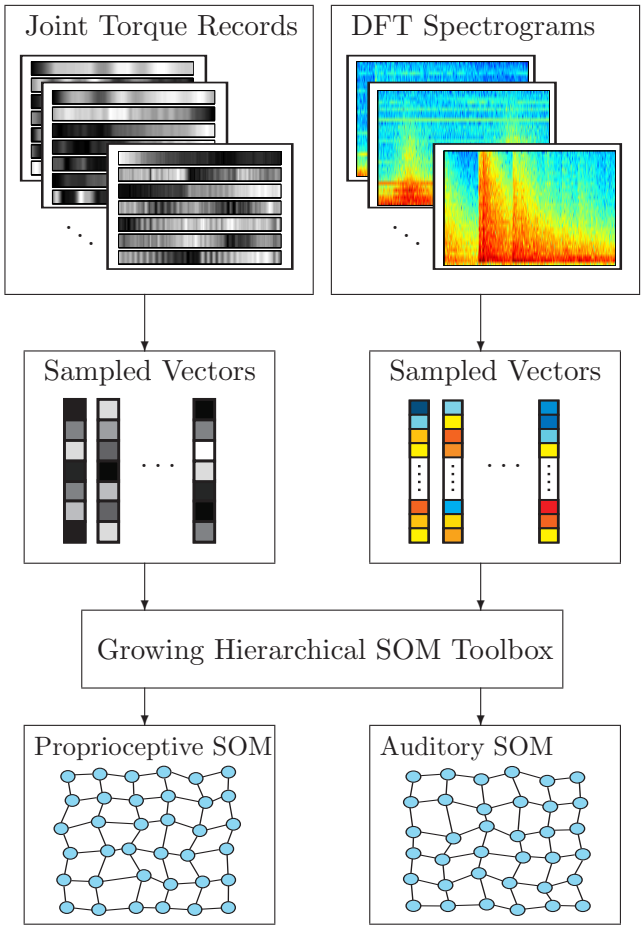


Figure 4.3 Illustration of the procedure used to train the proprioceptive and auditory Self Organizing Maps.

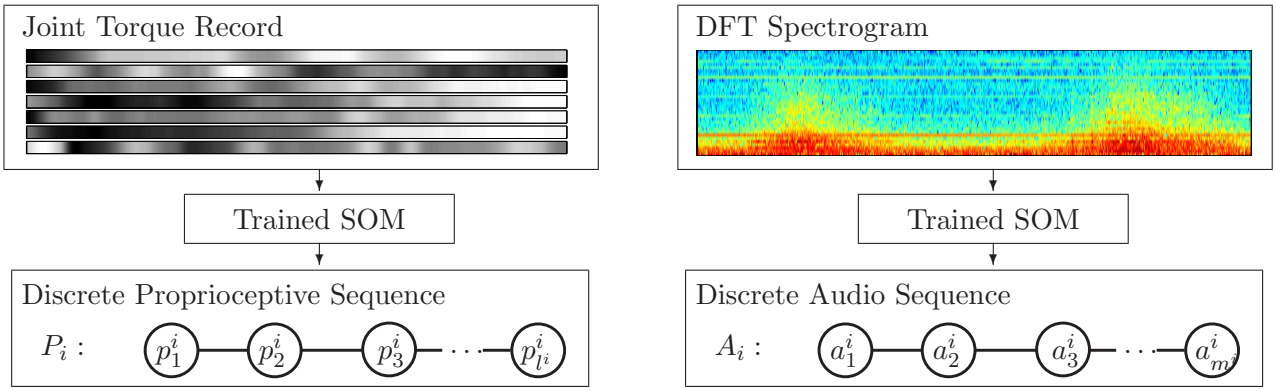


Figure 4.4 Illustration of the procedure used to turn high-dimensional proprioceptive (left) and auditory (right) sensorimotor feedback into discrete sequences using trained Self-Organizing Maps.

4.5.3 Machine Learning Classifier

To recall, in the behavior-grounded object recognition framework proposed here, the robot learns a recognition model $\mathcal{M}_i^j$ for each combination of behavior i and modality j, In the object recognition experiments conducted in this study, the robot used the k-NN classifier to implement the recognition model. The k-NN algorithm is a distance-based method, which does not build an explicit model of the training data (Aha et al., 1991; Atkeson et al., 1997). Instead, given a test data point, it simply finds the k closest neighbors and outputs a prediction, which is a smoothed average over those neighbors. In this study, k was set to 3. An estimate for the probability of each object, given the sequences, was computed by counting the class labels of the k neighbors. For instance, if two of the three neighbors had an object class label *plastic ball* then $Pr(O_i = \textit{plastic ball}) = \frac{2}{3}$. Similarly, if the class label of the remaining neighbor was *plastic cup*, then $Pr(O_i = \textit{plastic cup}) = \frac{1}{3}$. The value for k was chosen heuristically, such that it is both large enough to allow probabilistic interpretation of the model's output, and also small enough relative to the number of trials per object that were used to train each of the robot's behavior-grounded recognition models (e.g., 9 trials when performing 10-fold cross-validation).

The k-NN algorithm requires a distance measure to be used to compare the test data point to the training data points. In our case, each sensory feedback signal was encoded as a

sequence, where each token corresponds to the most highly activated node on a Self-Organizing Map. Since each data point in this study was represented as a sequence over a finite alphabet, the Needleman-Wunsch global alignment algorithm (Navarro, 2001; Needleman and Wunsch, 1970) was used to estimate the similarity between two sequences. While normally used for comparing biological or text sequences, the algorithm is applicable to other situations that require a distance measure between two strings. The algorithm requires a substitution cost to be defined over each pair of possible sequence tokens, e.g., the cost of substituting 'a' with 'b'. Since each token represents a node in a Self-Organizing Map, the cost for each pair of tokens was set to the Euclidean distance between their corresponding SOM nodes in the 2-D plane.

4.5.4 Evaluation

The performance of the recognition models coupled with each behavior-modality combination was evaluated using 10-fold cross validation, i.e., the full set of data points was split into ten folds corresponding to the ten trials performed with each object. During each of the ten iterations, nine of these folds are used for training the models and the remaining fold is used for evaluation. The recognition rate is reported in terms of accuracy, i.e.,

$$\%\ Accuracy = \frac{\#\ correct\ outputs}{\#\ total\ outputs} \times 100.$$

In addition, the performance was also measured as a function of the number of different exploratory behaviors that the robot performed on the test objects. To do so, the recognition performance was computed for various combinations of behaviors, ranging from 1, the default, to 5, the full set of behaviors. When using two, three and four interactions with the test object, all possible combinations of behaviors were evaluated and the average recognition rate was recorded. Whenever the robot was performing two or more exploratory behaviors on the test object, the predictions from the corresponding recognition models were combined, as described in Section 4.3.3.

Table 4.1 Object Recognition accuracy using k-NN model

Behavior	Audio	Proprioception	Combined
Lift	17.4 %	64.8 %	66.4 %
Shake	27.0 %	15.2 %	29.4 %
Drop	76.4 %	45.6 %	80.8 %
Crush	73.4 %	84.6 %	88.6 %
Push	63.8 %	15.4 %	65.0 %
Average	51.6 %	45.1 %	66.0 %

4.6 Results

4.6.1 Recognition Rates using a Single Interaction

The first set of results reports the recognition rates of the individual behavior-grounded recognition models, each of which is coupled with a specific behavior-modality combination. Table 4.1 shows the recognition rates for the object recognition dataset. The recognition rates are reported for each of the 10 behavior-modality combinations, as well as when using feedback from both sensory modalities.

As a reference, a chance predictor would be expected to achieve $(1/|\mathcal{O}|) \times 100 = 2.00\%$ accuracy (for $|\mathcal{O}| = 50$ different objects). Both the auditory and proprioceptive recognition models perform substantially better than chance, with the auditory model achieving slightly better accuracy on average. It is clear that the reliability of each modality is contingent on the type of behavior being performed on the object. For example, when the object is lifted, the proprioceptive model fares far better than the auditory model (since little sound is generated when an object is lifted). When performing the *push* behavior, on the other hand, the auditory modality dominates in performance.

Overall, the auditory stream is most informative when the object is dropped. The sound produced when the object hits the table implicitly captures many properties of the object: material type, size, and even shape. Proprioception, on the other hand, is most reliable when the object is crushed. The proprioceptive sequence implicitly captures the compliance and the

height of the object through the initial contact force and the timing of the first contact with the object. As expected, proprioception is also useful when lifting the object, since it implicitly captures the object's weight.

The results also show that combining the predictions from the two modalities improves the recognition accuracy for each of the five behaviors. This improvement is greatest for behaviors that yield reasonable performance for both modalities (e.g., *drop* and *crush*). However, even for behaviors where one of the modalities is far less reliable than the other (e.g., *lift*), there is still an improvement in object recognition accuracy. These results indicate that the use of multiple sensory modalities in object recognition models leads to greater robustness and higher overall accuracy.

4.6.2 Scalability with a Single Behavior

The second set of results looks at how the object recognition performance varies as the robot interacts with more and more objects. Most studies in robotics typically use a small number of objects. Presumably, it may be possible to achieve a high recognition accuracy when dealing with a small set of objects, but low recognition accuracy when the number of objects is increased. To test this hypothesis, the number of objects, n, was varied from 2 to 50 and for each n smaller than 50, the model was evaluated on 20 different randomly chosen object subsets of size n. For each subset, the accuracy of each of the five behavior-grounded models was recorded and used to compute the expected accuracy (and standard deviation) for each value of n.

Figure 4.5 shows the mean accuracies and standard deviations for all five behavior-grounded recognition models as a function of the number of objects in the data set, when using the weighted combination of the proprioceptive-auditory model outputs. With a small number of objects, the robot is able to achieve a high recognition rate. As the robot interacts with more and more objects, however, the recognition rate drops since a larger set of objects inherently contains objects with similar physical properties. The same trend can also be seen in Figure 4.6, which shows the mean accuracy rates for the three modality conditions, averaged across all behaviors. Therefore, robots that learn about objects should ultimately be evaluated on large

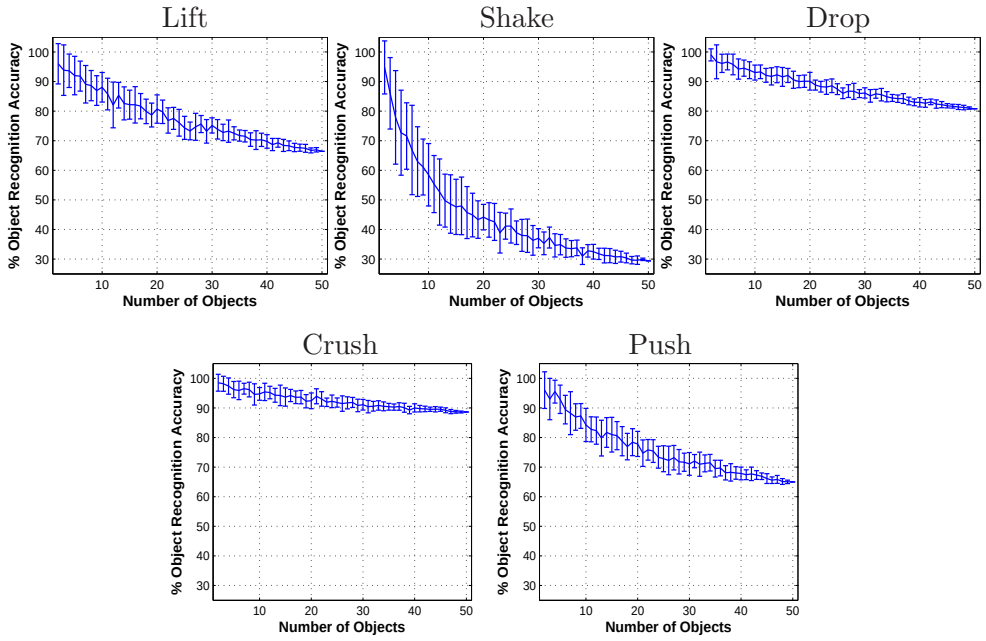


Figure 4.5 Recognition rates for the robot's behavior-grounded object recognition models (using both proprioceptive and auditory feedback) as a function of the number of objects, n, in the data set. For each value of n, 10-fold cross-validation was performed 20 different times, each with a different randomly selected object subset of size n. The solid lines indicate the resulting mean accuracy estimates while the error bars indicate the standard deviation of those estimates.

sets of objects in order to obtain more realistic and robust performance estimates.

4.6.3 Recognition Rates using Multiple Interactions

The next set of results examines how the recognition accuracy can be improved if the robot uses feedback generated from the execution of multiple different behaviors on the test object. Intuitively, it should be easier to recognize the test object if the robot lifts, shakes and then drops the object, than if it applies just a single behavior. To test this, the number of available interactions with the test object is varied from 1 (the default case, used to generate Table 4.1) to 5 (i.e., performing all five behaviors). When estimating the performance for 2, 3 and 4 interactions with the object, all possible combinations of behaviors are considered (e.g., for 2 interactions, there are 10 possible combinations), and the mean accuracy is reported. Model predictions from multiple interactions with the object are combined using the reliability weights estimated for each combination of behavior and modality, as previously described.

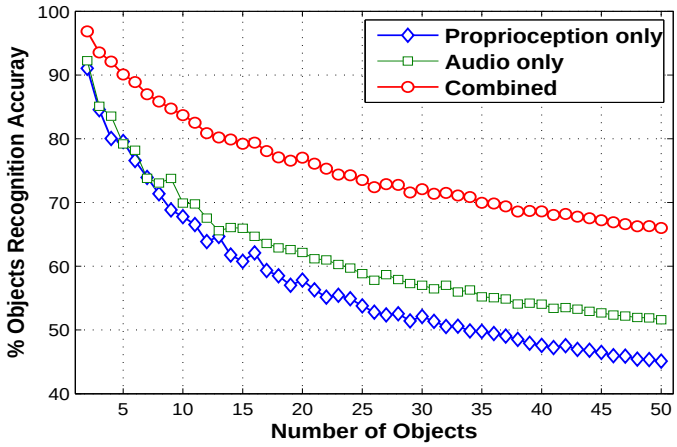


Figure 4.6 Average recognition accuracies from a single behavioral interaction as the number of objects, n, is varied from 2 to 50. For each value of n, 10-fold cross-validation was performed 20 different times, each with a different randomly selected object subset of size n.

Figure 4.7 shows the results of this experiment. Not surprisingly, the recognition accuracy improves dramatically as the robot interacts with the object using more and more behaviors – once all five behaviors are performed, it reaches 98.2%. This shows that *interactive* object recognition can provide highly accurate classification for a large set of objects, as long as the robot is allowed to perform several behavioral interactions with the object and combine their resulting predictions in an efficient manner.

A subsequent question to answer is whether the same type of recognition improvement can be achieved by performing the same behavior multiple times on the test object (as opposed to applying multiple different behaviors). An evaluation experiment was conducted in which the data set was split into 5 folds (each containing 2 trials with all five behaviors performed on each object) and 5-fold cross validation was performed. In other words, during each of the five iterations, the model was trained on 4 of the folds, and tested on the remaining one. For each of the five behaviors, the test set now contains two instances of the same behavior applied on each of the 50 objects. The test set also contains 4 instances for each of the $\binom{5}{2} = 10$ unique combinations of different behaviors (e.g., lift-shake) per object. After all five rounds of cross-validation, the individual accuracies of the five behaviors were estimated from the recorded model outputs when compared to the actual object IDs. The accuracies for each combination of exploratory behaviors were also estimated and stored in a 5×5 matrix. The diagonal

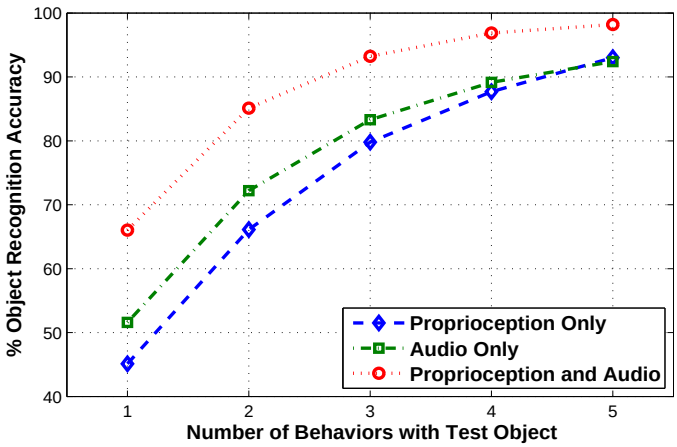


Figure 4.7 Average recognition accuracies from a single behavioral interaction as the number of objects, n, is varied from 2 to 50. For each value of n, 10-fold cross-validation was performed 20 different times, each with a different randomly selected object subset of size n.

entries of this symmetric matrix contain the 5 accuracy estimates obtained when performing the same behavior twice, while the 10 lower-diagonal entries contain the accuracy obtained when combining feedback from each of the 10 unique pairs of behaviors. These estimates were used to compute the improvement in recognition accuracy for different combinations of behaviors as described below.

Let $acc(\mathcal{M}^i, \mathcal{M}^j)$ be the estimated recognition accuracy when combining the outputs of recognition models $\mathcal{M}^i$ and $\mathcal{M}^j$ associated with behaviors B_i and B_j, and let $acc(\mathcal{M}^i)$ and $acc(\mathcal{M}^j)$ be their individual accuracies estimated when performing a single behavior execution on the test object. Given two behaviors B_i and B_j (which may be the same if $i = j$), the recognition improvement (RI_{ij}) obtained when applying the two behaviors sequentially on the test object can be measured relative to the recognition accuracy of the individual behaviors, i.e.,

$$RI_{ij} = acc(\mathcal{M}^i, \mathcal{M}^j) - \frac{acc(\mathcal{M}^i) + acc(\mathcal{M}^j)}{2}$$

With this formulation we can test whether combining feedback from two different behaviors results in greater recognition boost than combining feedback from two executions of the same behavior. The results of this evaluation, shown in Figure 4.8, confirm that the recognition improvement is higher when two different exploratory behaviors are applied on the test object,

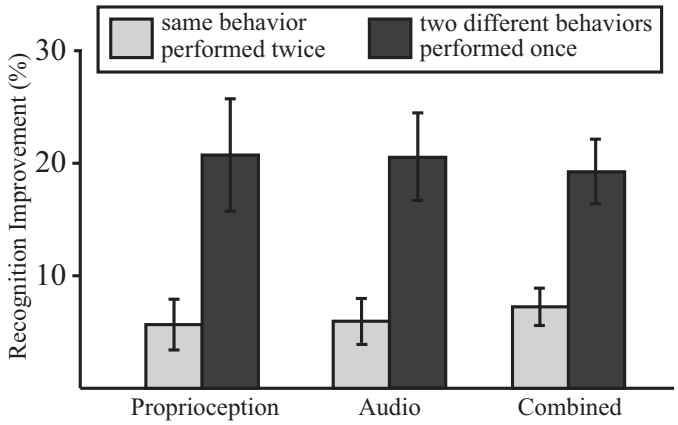


Figure 4.8 Object recognition improvement obtained by combining model outputs after two executions of the *same* behavior as well as two executions of *different* behaviors, estimated using 5-fold cross-validation. In all cases, the recognition improvement is higher when combining feedback from two distinct exploratory behaviors. When applying the same behavior twice, the standard deviation of the recognition improvement was estimated from 5 samples, one for each behavior. When applying two different behaviors, the standard deviation was estimated from 10 samples, one for each unique pair of behaviors.

as opposed to applying the same behavior twice. This result gives a strong indication that the diversity of the exploratory behaviors is more important than the number of times each behavior is executed when classifying an object.

4.7 Summary

This chapter introduced a framework for interactive object recognition, that can handle multiple exploratory behaviors and multiple sensory modalities. The proposed object recognition framework was evaluated using a large-scale experimental study, in which the robot interacted with 50 different objects using five exploratory behaviors (*lift*, *shake*, *drop*, *crush*, and *push*) and two sensory modalities (audio and proprioception). The feedback from the two sensory modalities, detected by the robot while interacting with an object, was represented as two sequences of the most highly activated nodes in two Self-Organizing Maps (one for each modality). Using global sequence comparison coupled with the k-Nearest Neighbors algorithm, the robot was able to recognize the explored object with accuracy substantially better than chance. The robot was also able to compute estimates for the reliability of each sensory modality and use them to improve its object recognition accuracy.

More importantly, after applying all 5 exploratory behaviors on the test object, the robot's recognition accuracy reached 98.2%, highlighting the importance of combining information extracted using multiple behaviors and multiple sensory modalities. These results give a strong indication that traditional vision-based object recognition systems can be further improved by the additional use of auditory and proprioceptive feedback. This is particularly important for objects that may not be easily recognized using vision alone (e.g., a heavy and a light object that look identical). Thus, active interaction (as opposed to passive observation) is a necessary component for resolving perceptual ambiguities about objects. Active object exploration is one of the hallmarks of human and animal intelligence (Power, 2000; Lorenz, 1996), which lends further credence to our approach to object recognition using exploratory behaviors.

There are several possible avenues for future work. First, other methods for dimensionality reduction (e.g., vector quantization, or Spatio-Temporal Isomap, as used by Peters et al. (2006)) can be applied in order to find meaningful patterns in the robot's proprioceptive and auditory sensory streams. Second, while the robot in our study was tested on an object recognition task, it is also possible to use auditory and proprioceptive feedback to detect certain physical properties of the object (e.g., its material type, whether it is hollow or solid, etc.). Some

preliminary results indicate that after applying all 5 behaviors on a novel object, the robot can detect its material type and other physical properties significantly better than chance (Sinapov and Stoytchev, 2009). Furthermore, the method for integrating information from proprioceptive and auditory feedback can be generalized to an arbitrary number of sensory modalities, allowing the robot to detect the reliability of each modality for each exploratory behavior. Integrating proprioceptive and tactile information from the robot's hand, as well as color and depth information from the robot's camera will allow the robot to further improve its ability to learn about common household objects. Robots that can interactively explore objects and make use of multiple sensory modalities will ultimately be better suited for working in human-inhabited environments.

CHAPTER 5. THE BOOSTING EFFECT OF EXPLORATORY BEHAVIORS AND SENSORY MODALITIES*

5.1 Introduction

From an early age, infants explore the objects around them through the use of exploratory behaviors such as grasping, shaking, dropping, and scratching (Piaget, 1952). Research in psychology has shown that the perceptual outcomes of these behaviors can be used to learn about objects and their physical properties (Lederman and Klatzky, 1987). Similar exploration procedures have also been observed in a wide variety of animal species, including primates and birds (Lorenz, 1996; Power, 2000).

Exploratory behaviors reveal information about an object by producing sensory feedback across a wide variety of sensory modalities. As noted by Lederman (1982), manual exploration of a surface texture not only generates tactile sensations but can also produce auditory feedback. These type of sensations allow humans to perceive a large number of object properties that cannot be detected through passive observation (Lynott and Connell, 2009). For example, the proprioceptive feedback produced when lifting an object can inform us of its weight, while the tactile feedback produced when scratching it can inform us of its roughness. In light of these findings, research in robotics has confirmed that the use of multiple exploratory behaviors and multiple sensory modalities improves interactive object recognition rates (Sinapov et al., 2009; Bergquist et al., 2009). But what causes this improvement?

This chapter addresses this question by analyzing previously published datasets from two different interactive recognition tasks: 1) object recognition using auditory and proprioceptive feedback; and 2) surface texture recognition using tactile and proprioceptive feedback. More

*This chapter is based on the following paper: Sinapov, J. and Stoytchev, A., "The Boosting Effect of Exploratory Behaviors", *In proceedings of the 24th National Conference on Artificial Intelligence (AAAI)*, 2010.

specifically, this chapter examines whether metrics designed to measure classifier diversity can be used to estimate the expected improvement of accuracy when combining information from multiple modalities or multiple behaviors. The results explain, for the first time, why using multiple exploratory behaviors and multiple sensory modalities leads to a boost in object recognition rates.

5.2 Related Work

The use of behaviors in robotics has a long history (Brooks, 1986; Arkin, 1987; Matarić, 1992). Initially, they were introduced as an attempt to simplify the control problem by splitting the robot's controller into tiny modules called behaviors (Brooks, 1986). At that time, the behavior-based approach outperformed other existing control methods, which quickly increased its popularity. Recently, the research focus has shifted from using behaviors for controlling the robot to using behaviors for extracting information about objects (Fitzpatrick et al., 2003; Stoytchev, 2005).

It was also realized that each behavior produces sensory signatures across one or more sensory modalities. This insight was used to improve the robot's knowledge about objects and their properties. For example, it was shown that integrating proprioception with vision can bootstrap a robot's ability to interact with objects (Fitzpatrick et al., 2003). Interaction with objects could also enable a robot to recognize them based on the sounds that they produce (Krotkov, 1995; Torres-Jara et al., 2005) or based on the proprioceptive data generated by the robot's hand as it grasps the objects (Natale et al., 2004). Other experimental results show that using multiple modalities leads to a boost in recognition performance (Saenko and Darrell, 2007; Morency et al., 2005).

Subsequent experiments have shown that robots can boost their object recognition rates by performing multiple exploratory behaviors as opposed to just one. This effect has been demonstrated with various sensory modalities, including audio (Sinapov et al., 2009), proprioception (Bergquist et al., 2009), and touch (Sinapov et al., 2011b; Hosoda et al., 2006). The source of this boosting effect, however, has not been adequately explained so far. The goal of this chapter is to provide a theoretical link between the boosting effect and exploratory behaviors.

5.3 Theoretical Framework

The theoretical framework described here uses the concept of classifier diversity to study the recognition improvement attained when a robot uses multiple exploratory behaviors and multiple sensory modalities. At first glance, the boosting effect appears similar to the classification improvement attained when using machine learning techniques such as bagging and boosting in conjunction with an ensemble of classifiers. Machine learning theory has attempted to explain the success of ensemble classifiers by introducing the concept of classifier diversity (Lam, 2000; Kuncheva and Whitaker, 2003). In this framework, combining predictions from diverse or complementary classifiers is thought to be directly related to the improvement in classification accuracy of the ensemble when compared to that of the individual base classifiers.

5.3.1 Problem Formulation

Let N be the number of behaviors in the robot's repertoire, and let M be the number of sensory modalities. Upon executing behavior i on a target object, the robot detects sensory stimuli $X_i^1, \ldots, X_i^M$, where each X_i^j is the sensory feedback from modality j. In the most general case, each stimulus can be represented either as a real-valued vector, or as a structured data point (e.g., a sequence or a graph).

The task of the robot is to recognize the target object by labeling it with the correct discrete label $c \in \mathcal{C}$. To solve this problem, for each behavior, i, and each modality, j, the robot learns a model $\mathcal{M}_i^j$ that can estimate the class label probability $Pr(c|X_i^j)$. In other words, for each combination of behavior and modality, the robot learns a classifier that estimates the class label probability for each $c \in \mathcal{C}$. The following two sub-sections describe how the robot integrates stimuli from multiple modalities and multiple behaviors in order to further improve the accuracy of its predictions.

5.3.2 Combining Multiple Modalities

For each behavior i, the robot learns a model $\mathcal{M}_i$, which combines the class-label probabilities of the modality-specific models $\mathcal{M}_i^j$ (for $j = 1$ to M). Given sensory stimuli $X_i^1, \ldots, X_i^M$

detected while performing behavior i on a given object, the robot estimates the class-label probabilities for this object as:

$$Pr(c|X_i^1, \ldots, X_i^M) = \alpha \sum_{j=1}^{M} w_i^j Pr(c|X_i^j).$$

In other words, given the stimuli from the M available sensory modalities, the robot combines the class-label estimates of the modality-specific models $\mathcal{M}_i^j$ using a weighted combination rule. The coefficient α is a normalizing constant, which ensures that the probabilities sum up to 1.0. Each weight w_i^j corresponds to an estimate for the reliability of the model $\mathcal{M}_i^j$ (e.g., its accuracy).

It is worth noting that humans integrate information from multiple modalities in a similar way when performing the same task (Ernst and Bulthof, 2004). For example, when asked to infer an object property given proprioceptive and visual feedback, humans use a weighted combination of the predictions of the two modalities. Experimental results have shown that the weights are proportional to the estimated reliability of each modality (Ernst and Bulthof, 2004). The weighted combination of predictions ensures that a sensory modality that is not useful in a given context will not dominate over other more reliable channels of information.

5.3.3 Combining Multiple Behaviors

To further improve the quality of its predictions, the robot uses not only multiple sensory modalities, but also applies multiple behaviors. After performing n distinct behaviors on the test object (where $n \leq N$), the robot detects sensory stimuli $[X_1^1, \ldots, X_1^M], \ldots, [X_n^1, \ldots, X_n^M]$. As in the case of combining multiple modalities, the robot uses a weighted combination rule and labels the test object with the class label $c \in \mathcal{C}$ that maximizes:

$$Pr(c|X_1^1, \ldots, X_1^M, \ldots, X_n^1, \ldots, X_n^M) = \alpha \sum_{i=1}^{n} \sum_{j=1}^{M} w_i^j Pr(c|X_i^j).$$

Intuitively, it is expected that by combining the predictions of the models $\mathcal{M}_i^j$ it is possible to achieve higher recognition accuracy than with any single model alone, especially if the weights w_i^j can be estimated accurately from the training dataset. This expected improvement is assumed to be directly related to the level of *diversity* between individual models (Lam, 2000;

Kuncheva and Whitaker, 2003). The next subsection describes several metrics for estimating model diversity that are commonly used in the machine learning literature.

5.3.4 Estimating Model Diversity

Combining predictive or recognition models (e.g., classifier ensembles, mixture of experts, etc.) is an established area of research within the machine learning community. A wide variety of metrics have been developed to measure the level of *diversity* among classifiers, with emphasis on establishing a relationship between diversity and accuracy (Kuncheva and Whitaker, 2003). Traditionally, such metrics have been used to compare classifiers that are trained on biased or re-weighted subsets of the original dataset. In contrast, each of the robot's recognition models $\mathcal{M}_c$ is trained and tested on data from a particular behavior-modality combination. Next, we show how several of the proposed metrics can be extended in order to measure the diversity of the robot's recognition models derived from the N exploratory behaviors and M sensory modalities.

Let $[X_1, \ldots, X_t]_k$ constitute the sensory feedback signals detected during the k^{th} interaction trial (where $k = 1$ to K) during which the robot sequentially performs all N behaviors on a test object. The output of a recognition model $\mathcal{M}_a$ can be represented as a $K-$dimensional binary vector $\mathbf{y_a} = [y_{1,a}, \ldots, y_{K,a}]^T$, such that $y_{k,a} = 1$ if the model $\mathcal{M}_a$ correctly labels the object present during trial k, and 0 otherwise. One strategy for measuring the pairwise diversity between two models $\mathcal{M}_a$ and $\mathcal{M}_b$ is to compare the corresponding vectors $\mathbf{y_a}$ and $\mathbf{y_b}$.

The first metric used in this study is the *disagreement measure*, which was previously used by Skalak (1996) to quantify the diversity between a base model and a complementary model. The disagreement measure is defined as:

$$DIS_{a,b} = \frac{N^{01} + N^{10}}{N^{11} + N^{10} + N^{01} + N^{00}}$$

where N^{pq} is the number of trials (out of K) for which $y_{k,a} = p$ and $y_{k,b} = q$ (see Table 5.1).

In other words, the disagreement measure is simply the ratio of the number of trials in which one model was correct and the other was wrong to the total number of trials. The measure is

Table 5.1 The relationship between a pair of recognition models $\mathcal{M}_a$ and $\mathcal{M}_b$ can be expressed using a 2 x 2 table, which shows how often their predictions coincide (N^{11} and N^{00}) and how often they disagree (N^{01} and N^{10}).

	$\mathcal{M}_a$ correct	$\mathcal{M}_a$ wrong
$\mathcal{M}_b$ correct	N^{11}	N^{10}
$\mathcal{M}_b$ wrong	N^{01}	N^{00}

always in the range of 0.0 to 1.0. Low values indicate that the predictions of the two models mostly agree (whether right or wrong).

The second metric used in this study is Yule's Q-Statistic (Yule, 1900; Kuncheva and Whitaker, 2003), which is defined for two models $\mathcal{M}_a$ and $\mathcal{M}_b$ as:

$$Q_{a,b} = \frac{N^{11}N^{00} - N^{01}N^{10}}{N^{11}N^{00} + N^{01}N^{10}}.$$

The Q-statistic ranges from -1.0 to 1.0. For statistically independent models, the expectation of $Q_{a,b}$ is 0 Kuncheva and Whitaker (2003). A high value of Q indicates that both models label objects either correctly or incorrectly during the same interaction trials, while a low value of Q indicates that the two models commit errors on different trials.

5.4 Experimental Setup

This section briefly describes the two previously published datasets from our lab, which were obtained from their authors (along with the corresponding source code) for the purposes of this study. For more details, please refer to the original papers.

5.4.1 Tactile Surface Recognition Dataset

In the first dataset, the task of the robot was to recognize surface textures by applying exploratory scratching behaviors on them (Sinapov et al., 2011b). The robot was programmed with five different exploratory behaviors, which constitute scratching trajectories performed at different speeds and in different directions. During each scratching interaction, the robot recorded the tactile feedback from an artificial fingernail with an embedded 3-axis accelerometer and the proprioceptive joint-torque feedback from all 7 joints. Twenty different surfaces were included in the experiments. The robot performed all five scratching behaviors on each surface ten different times for a total of 1000 behavioral interactions.

5.4.2 Interactive Object Recognition Dataset

In the second dataset, the task of the robot was to (interactively) recognize objects using only proprioceptive and auditory feedback (Sinapov et al., 2011a). The robot was programmed with five exploratory behaviors: *lift*, *shake*, *drop*, *crush*, and *push*. Each of these behaviors was applied ten times on fifty different objects, for a total of 2500 behavioral interactions. During each interaction, the robot recorded auditory feedback through a microphone and proprioceptive feedback in the form of joint-torque values.

5.4.3 Feature Extraction and Learning Algorithm

For all three modalities (auditory, tactile, and proprioceptive), the sensory stimuli X_i^j were encoded as a sequence of states in a Self-Organizing Map (SOM). A separate SOM was trained on input from each modality. Given a recorded audio signal, the Discrete Fourier Transform (DFT) was computed, which resulted in a matrix containing the intensity levels of each fre-

quency bin over time. This high-dimensional feedback, was transformed into a sequence over a discrete alphabet by mapping each column vector of the DFT matrix to a state in a trained SOM (see Sinapov et al. (2009) for details). Similarly, the DFT was computed for the tactile sensory feedback as described by Sinapov et al. (2011b), and subsequently mapped to a discrete sequence of activated states in a SOM. The proprioceptive feedback was also represented as a sequence by mapping each recorded joint-torque configuration to a state in a SOM, which was trained on proprioceptive data as described by Bergquist et al. (2009).

Each recognition model $\mathcal{M}_i^j$ was implemented as a k-Nearest Neighbor classifier with $k = 3$. The global pairwise sequence alignment score was used as the k-NN similarity function, which was computed for sequences of the same sensory modality.

Table 5.2 Surface Recognition from a Single Behavior

Behavior	Tactile	Proprioceptive	Combined
Lateral, fast	50.0 %	30.5 %	55.5 %
Lateral, medium	53.5 %	35.5 %	62.5 %
Lateral, slow	48.5 %	35.0 %	57.0 %
Medial, fast	42.0 %	48.5 %	57.0 %
Medial, slow	33.5 %	52.5 %	56.0 %
Average	45.5 %	40.4 %	57.6 %

5.5 Experiments and Results

5.5.1 Boosting Accuracy with Multiple Modalities

The first experiment explores whether the improvement attained when using multiple sensory modalities is related to the pairwise diversity metrics defined earlier. In this scenario, the robot is first evaluated on how well it can recognize the target object (or surface texture) from a single behavioral interaction with it. Table 5.2 shows the recognition rates for the surface texture recognition dataset when using either modality alone, as well as when the two modalities are combined. For comparison, the expected chance accuracy is $1/20 = 5.0\%$. For all 5 scratching behaviors, using both modalities always results in recognition rates substantially higher than the ones obtained with either modality alone.

Table 5.3 shows the results from the same experiment performed on the object recognition dataset. In this case, a chance predictor is expected to achieve $1/50 = 2.0\%$ accuracy. For this dataset, there is far greater variation in recognition rates across different behavior-modality combinations. It is also clear that the reliability of each modality is contingent on the type of behavior being performed on the object. For example, when the object is lifted, the proprioceptive model fares far better than the auditory model (since little sound is generated when an object is lifted). When the object is pushed by the robot, however, the auditory modality dominates in performance.

Table 5.3 Object Recognition from a Single Behavior

Behavior	Auditory	Proprioceptive	Combined
Lift	17.4 %	64.8 %	66.4 %
Shake	27.0 %	15.2 %	29.4 %
Drop	76.4 %	45.6 %	80.8 %
Crush	73.4 %	84.6 %	88.6 %
Push	63.8 %	15.4 %	65.0 %
Average	51.6 %	45.1 %	66.0 %

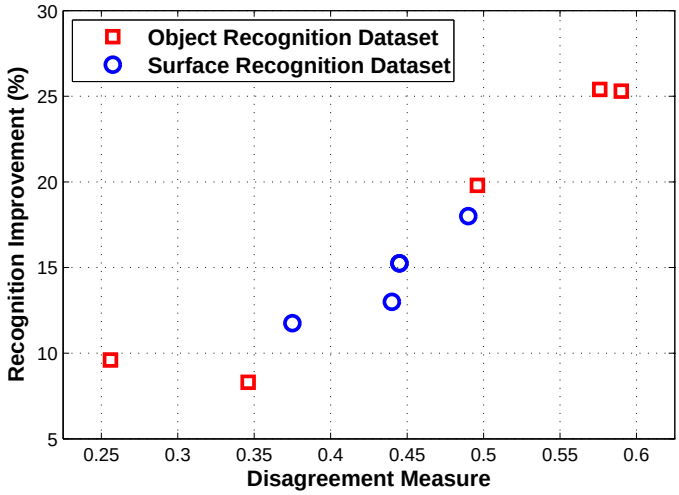


Figure 5.1 Pairwise disagreement measure vs. recognition improvement. Each point corresponds to one of the five behaviors in the two datasets. The horizontal axis shows the disagreement measure between the two modality-specific models, $\mathcal{M}_i^1$ and $\mathcal{M}_i^2$, for each behavior. The vertical axis shows the recognition improvement attained when both modalities are combined. In the surface recognition dataset, the points for two of the behaviors coincide.

For both datasets, combining modalities significantly improves recognition performance as compared to using either modality alone. But what is the source of this improvement? To answer this question, we can quantify the improvement in recognition accuracy and relate it to the diversity of the models. For each behavior i, let $acc(\mathcal{M}_i^j)$ be the % accuracy of the modality-specific recognition model $\mathcal{M}_i^j$ and let $acc(\mathcal{M}_i)$ be the % accuracy of the modality-combining model $\mathcal{M}_i$, a model that outputs a weighted combination of the outputs of the models $\mathcal{M}_i^1, \ldots, \mathcal{M}_i^M$. We define the *Recognition Improvement* (RI) for the i^{th} behavior as:

$$RI_i = acc(\mathcal{M}_i) - \frac{\sum_{j=1}^{M} acc(\mathcal{M}_i^j)}{M}.$$

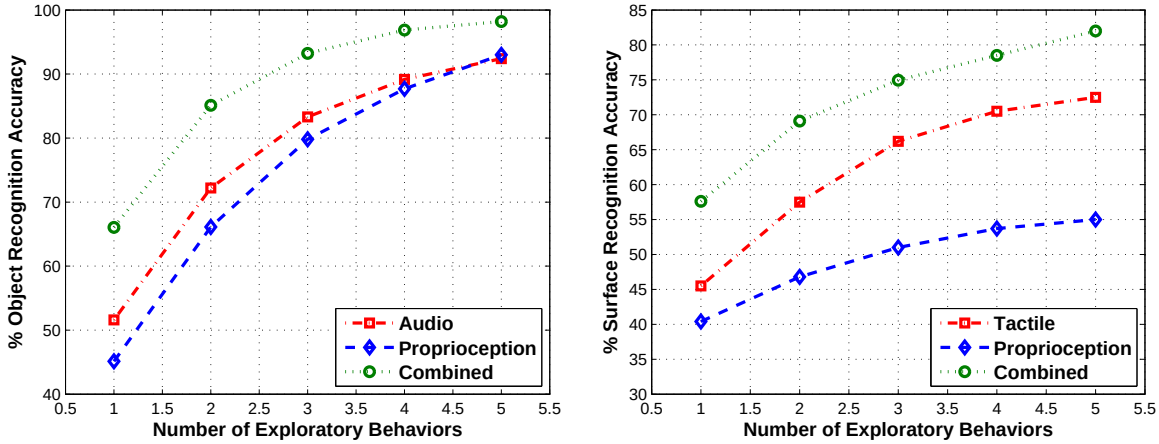


Figure 5.2 Recognition accuracy for the two datasets as the number of behaviors is varied from 1 (the default) to 5 (i.e., performing all five behaviors on the test object).

To see if there is a relationship between model diversity and recognition improvement, the disagreement metric was computed for each possible combination of modality-specific models. Figure 5.1 shows that for both datasets this relationship is approximately linear. As predicted by machine learning theory, high pairwise disagreement generally results in higher recognition improvement. This result shows that the concept of classifier diversity can indeed be applied to the robot's behavior-derived recognition models.

5.5.2 Boosting Accuracy with Multiple Behaviors

The next set of experiments examines the improvement in recognition rate achieved by performing multiple exploratory behaviors on the test object/surface. Figure 5.2 shows the recognition accuracy for both recognition tasks as the number of behaviors applied on the test object/surface is varied from 1 (the default, used to generate Tables 5.2 and 5.3) to 5 (i.e., performing all five behaviors). The results clearly show that the robot can significantly improve its recognition accuracy by applying multiple exploratory behaviors.

In the case of the surface task, the recognition rate increases at a faster pace when the predictions of the tactile models are combined, than when the predictions of the proprioceptive models are combined as shown in Figure 5.2. To understand the reasons why, we look at how this improvement is related to different measures of model diversity.

Given two distinct behaviors i and j, let $acc(\mathcal{M}_i, \mathcal{M}_j)$ be the estimated recognition accuracy attained by combining the predictions of the models $\mathcal{M}_i$ and $\mathcal{M}_j$ (which can be either modality-

specific models or modality-combining models). The recognition improvement for two behaviors i and j is defined as:

$$RI_{ij} = acc(\mathcal{M}_i, \mathcal{M}_j) - \frac{acc(\mathcal{M}_i) + acc(\mathcal{M}_j)}{2}.$$

Figure 5.3 plots the disagreement measure vs. the recognition improvement for the surface recognition dataset. Because there are 5 behaviors in that dataset, we can form 10 different pairs of behaviors for which the improvement in recognition accuracy can be calculated under three different conditions: touch only, proprioception only, or both. We can also calculate the diversity between any two behavioral models. The results show that the amount of disagreement is directly related to the expected improvement. On average, the pairwise disagreement for the tactile recognition models is higher than that for the proprioceptive models. This explains why the improvement attained by applying multiple behaviors is greater with the tactile sensory modality.

The same plot can also be calculated for the object recognition dataset. A comparison plot in Figure 5.4 shows the relationship between the disagreement measure and the classification improvement for both datasets. There is a linear relationship between the diversity metric and the observed boost in the recognition rate. As predicted by machine learning theory, higher diversity results in higher accuracy improvement. This result shows that the disagreement measure is a good indicator for the expected recognition improvement, a finding that generalizes to both datasets.

Figure 5.4 also shows the relationship between the Q-statistic and the recognition improvement for both datasets. The Q-statistic is approximately linearly related to the accuracy improvement in the surface recognition dataset, but there is no clear relationship in the object recognition dataset. This is indeed a surprising result, since the Q-statistic is typically the most common metric used for estimating the diversity between two classifier models and has been recommended by Kuncheva and Whitaker (2003) as a good metric for measuring classifier model diversity. Several factors might explain this apparent discrepancy. First, the individual classifier models in the experiments conducted by Kuncheva and Whitaker (2003) had approximately the same individual accuracies. The individual recognition models used in

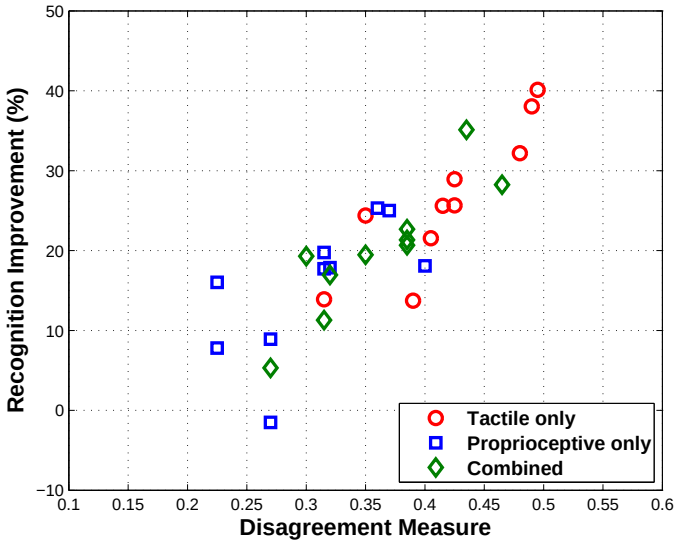


Figure 5.3 Pairwise disagreement measure vs. recognition improvement for the surface recognition dataset. For every unique combination of 2 behaviors (10 total for 5 behaviors), there are 3 points in the plot, one for each of the three conditions: touch, proprioception, or both. The horizontal axis shows the estimated disagreement measure between the two behavior-derived models, while the vertical axis shows the recognition improvement attained when applying both behaviors.

the interactive object recognition task, however, have very different accuracies (see Table 5.3). For example, performing the *shake* behavior results in 29.4% recognition rate, while the *drop* behavior achieves 80.8%. Second, it has been shown by Dietterich (2000) that different methods for building collections of classifiers can result in different relationship patterns between diversity and improvement. Typically, it is assumed that each classifier model in the ensemble is trained on some biased subset (or otherwise modified version) of the original training set. In contrast, the recognition models learned by the robot are constructed in a profoundly different manner - each of the robot's recognition models is trained and tested only on data from a particular behavior-modality combination. Despite these differences, the concept of classifier diversity was still found to be useful for explaining the improvement in recognition accuracy in the robot experiments.

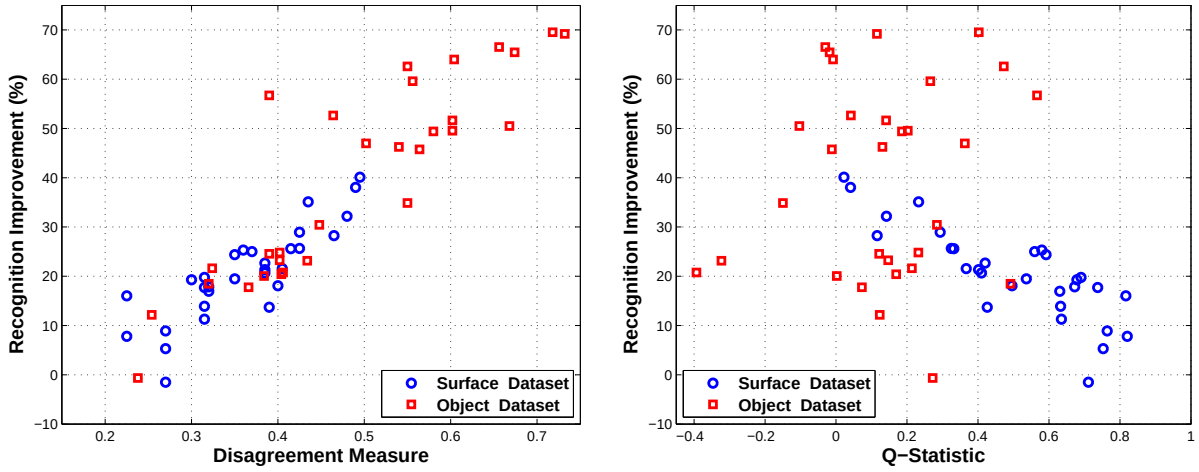


Figure 5.4 *Left:* Pairwise disagreement measure vs. recognition improvement for each of the 10 possible pairs of behaviors, under three different modality conditions (modality 1 only, modality 2 only, or combined) for both datasets. *Right:* Pairwise Q-statistic vs. recognition improvement for each of the 10 possible pairs of behaviors, under three different modality conditions (modality 1 only, modality 2 only, or combined) for both datasets.

5.6 Summary

Exploratory behaviors play an important role in the object exploration patterns of humans and animals (Lorenz, 1996; Power, 2000). When these behaviors are applied on objects they act like "questions" that the object "answers" by producing effects across multiple sensory modalities. When multiple behaviors are performed the identity of the object can be uniquely identified. Recent studies have shown that robots can also use exploratory behaviors to improve their object recognition rates. The reasons for this improvement, however, have not been adequately explained so far.

This chapter formulated a new metaphor to explain these results, namely, *behaviors are classifiers*. Thus, the behavioral repertoire of the robot can be viewed as an ensemble of classifiers, which can be boosted. The boosting effect generalizes not only to multiple exploratory behaviors, but also to multiple sensory modalities. Each new modality and each new behavior provides additional information that can be used to construct new classifiers.

Two large datasets with 50 objects and 20 surfaces were used to generate the results, which clearly show that the metrics designed to measure the diversity of classifiers can be applied

to measure the diversity of the behaviors in the robot's behavioral repertoire. In particular, the *disagreement measure* for two behavior-derived recognition models was found to be linearly related to the observed boost in recognition rate when both behaviors are applied. This is an important contribution as it establishes for the first time a link between empirical studies of exploratory behaviors in robotics and theoretical results on boosting in machine learning.

CHAPTER 6. THE ODD-ONE-OUT TASK: TOWARD AN INTELLIGENCE TEST FOR ROBOTS*

6.1 Introduction

The experiments described so far showed that a robot can ground its object recognition models in its own sensorimotor repertoire. Besides object recognition, however, there are many other tasks that a robot may be expected to solve. For example, detecting an item that does not belong in a given set is a standard problem in modern Intelligence Quotient (IQ) tests. This is known as the *odd one out* task, which is formulated as follows: given a set of items, the participant is asked to decide which one among them is most dissimilar from the rest. Variants of this task have been used extensively in a wide variety of disciplines to test for brain abnormalities (Buckley et al., 2001), learning disabilities (Roberson et al., 1999), and categorization abilities (Stephens and Navarro, 2008). It has also been used by Luria (1976) to probe the cultural and social foundations of cognition. Typically, the presented items vary along one dimension (e.g., size, shape, color), which, if identified by the participant, could be used to pick the most dissimilar item. In more complex settings, however, picking the odd object requires comparing along multiple sensory dimensions (Stephens and Navarro, 2008). This task has also been tried in the auditory domain with human participants (Snowling et al., 1994), which indicates that the general principles used to pick the odd item are not necessarily tied to the visual sensory modality.

The ubiquity of the odd one out task makes it an attractive candidate for an intelligence test in developmental robotics. The task has been used extensively to study how humans

*This chapter is based on the following paper: Sinapov, J. and Stoytchev, A., "The Odd One Out Task: Toward an Intelligence Test for Robots", *In proceedings of the 9th IEEE International Conference on Development and Learning (ICDL)*, pp. 126-131, 2010.

estimate object similarity and form object categories. Therefore, it may also be a valuable tool for conducting experiments with robots. Recent work in robotics has focused on detecting object similarities and forming object categories (Nolfi and Marocco, 2002; Natale et al., 2004; Nakamura et al., 2007; Takamuku et al., 2008; Sinapov et al., 2009; Griffith et al., 2009; Sun et al., 2010b), indicating that robots should, in principle, be capable of solving the odd one out task in a variety of settings.

This chapter proposes a framework that allows a robot to estimate the similarity between objects based on its prior interactive experience with them. A theoretical model is presented that uses the estimated object relations to solve the odd one out task by selecting the most dissimilar object from a given set. The experiments were conducted with an upper-torso humanoid robot, which interacted with fifty different objects by applying five types of exploratory behaviors (lift, shake, drop, crush, and push). Over the course of each interaction, the robot detected auditory and proprioceptive sensory feedback. The robot was able to estimate pairwise object similarity relations for each behavior-modality context, which were used to select the odd object in subsequent tests. The framework was repeatedly evaluated on six natural object categories (e.g., cups, bottles, pop cans, etc.). During each test, a group of three objects from the target category and one object from outside the category were presented. The robot's internal models were queried to pick the most dissimilar object. The results show that the estimated object relations were successful in capturing the properties of natural object categories, since the robot was able to solve the task with success rates substantially better than chance. This suggests that it may be possible to ground the semantic labels for many object categories in the robot's sensorimotor experience.

6.2 Related Work

Asking participants to pick the odd item from a set is a task that can provide valuable insights into how humans categorize objects. One of the early experiments that used this task was performed by Luria, who studied how social and cultural upbringing affect development (Luria, 1976). Uneducated Soviet peasants were shown images of four objects (e.g., hammer, saw, hatched, and wooden log) and asked to select the object that does not belong in the group. The goal of this test was to determine whether the participants grouped items together based on their semantic category (e.g., hand-held tools) or not.

Other researchers have used the odd one out task to study how humans measure perceptual similarity. Stephens and Navarro (2008) investigated how people establish similarity relations for three-dimensional models of animal-like objects called "greebles." During each trial, the participants were asked to pick the odd one out from a set of three greebles. The data was used to generate a pairwise matrix that specified the similarity for each pair of greebles, as determined by the participants. The study presented in this chapter solves the opposite problem: the robot first estimated pairwise measures of object similarity, and then used these measures to solve the odd one out task.

In another notable experiment, Roberson et al. (1999) used the odd one out task to study the relationship between perceptual similarity and object categorization. By examining a patient's performance on this task, they concluded that the mapping between a perceptual representation (e.g., color) and the corresponding category label (e.g., the name of the color) is not as transparent as previously thought. In relation to developmental robotics, this study suggests that the odd one out task may indeed be useful as a testbed for studying how well the robot's perceptual experience with an object matches the object's human-defined category label.

Robots that can estimate the similarity between objects and form meaningful object categories would be more useful in dynamic and unstructured environments. Related work in robotics has demonstrated that, through active interaction, robots can derive a measure of perceptual as well as functional object similarity (Nolfi and Marocco, 2002; Nakamura et al., 2007; Takamuku et al., 2008; Sun et al., 2010b). For example, Natale et al. (2004) used a

Self-Organizing Map to illustrate the haptic similarity between objects, obtained as their robot repeatedly grasped them and recorded tactile sensations. Sinapov et al. (2009) demonstrated that a robot can estimate the similarity between objects based on the sounds that the objects generate when different behaviors are performed on them.

Other related research has focused on categorizing objects in terms of their functional properties. The simulated robot in the experiments described by Sinapov et al. (2008) was able to establish how similar two tools are based on what the tools allow the robot to do. Modayil and Kuipers (2008) introduced a general framework that allows a robot to discover classes of objects, based on their detected percepts over the course of an interaction. Griffith et al. (2009) demonstrated that a robot can form the functional category of "containers" by repeatedly observing visual movement patterns of objects dropped in, or near, the container. Along with other published research, these results give a strong indication that, in the right setting, robots should be able to solve the odd one out task.

6.3 Experimental Setup

The experimental setup and the dataset used in this study are identical to the ones described in Chapter 4 and are only briefly summarized here. The robot was an upper-torso humanoid robot with two 7-DOF Barrett WAMs for arms and two 3-finger Barrett Hands as end effectors. The robot's head was equipped with a Audio-Technica U853AW cardioid microphone.

The robot performed five exploratory behaviors on each object: *lift*, *shake*, *drop*, *crush*, and *push* (shown in Figure 4.1 in Chapter 4). The behaviors were encoded with the Barrett WAM API and performed with the robot's left arm. The raw proprioceptive data (i.e., joint torques of the left arm) and the raw audio were recorded for the duration of each behavior (start to end).

The proprioceptive and auditory feedback for each behavioral interaction were represented as discrete sequences, where each sequence element corresponded to the most highly activated state in an 6-by-6 Self-Organizing Map (SOM) (Kohonen, 2001). One SOM was trained for each modality, as described by Bergquist et al. (2009) and Sinapov et al. (2009). For example, given a specific joint-torque configuration (i.e., a vector in $\mathbb{R}^7$), the data point is fed as input to the proprioceptive SOM and the index of the most highly activated state in the map is appended as the next token in the proprioceptive sequence for that behavioral interaction. Similarly, given a Discrete Fourier Transform (DFT) of a recorded sound, each column vector of the DFT is given as input to the auditory SOM and the index of the most highly activated state is added as the next token in the auditory feedback sequence. The proprioceptive SOM was trained with sample joint-torque configurations experienced by the robot, while the auditory SOM was trained with a set of column vectors extracted from the recorded DFTs. This procedure is described in much more detail by Bergquist et al. (2009) for proprioception and by Sinapov et al. (2009) for audio.

After each behavioral interaction is performed, the robot records two sequences, $X_{prop} = p_1 p_2 \dots p_k$ and $X_{audio} = a_1 a_2 \dots a_l$. The two sequences are not necessarily of the same length, since proprioception and audio are sampled at different frequencies. Finally, the robot needs a metric that can establish the similarity between two sequences from the same sensory modality.

Figure 6.1 The six object categories, along with the remaining 25 objects, used in this study. An object may belong to more than one category - e.g., the three pop cans also belong to the set of *metal objects*. One of the pop bottles was full during the experiments and is not included in the *empty bottles* set.

As described by Bergquist et al. (2009) and Sinapov et al. (2009), the global alignment similarity function was used, which is a common choice for comparing discrete sequences.

The robot interacted with a set of objects, $\mathcal{O}$, consisting of 50 common household objects, including cups, bottles, and toys (see Figure 6.1). The figure also shows the object categories formed by the objects, which were used in the evaluation of the robot's performance on the odd one out task. During each test, 3 objects from a given category (e.g., pop cans) and 1 from outside the category were chosen. The robot's model was queried to select the object that, according to the robot's internal representation, does not belong in the group. The next section describes the theoretical model used by the robot to solve this task.

6.4 Methodology

This section describes the three stages used to solve the odd one out task. First, the robot interacts with the set of all objects, $\mathcal{O}$, by performing each of its exploratory behaviors on every object while recording the detected proprioceptive and auditory feedback. Second, after the interaction stage is over, the robot estimates a pairwise $|\mathcal{O}| \times |\mathcal{O}|$ object similarity matrix, $\mathbf{W}$, such that W_{ij} denotes the similarity between objects i and j. Finally, when presented with 4 (or in the general case, K) different objects, the robot uses the similarity matrix $\mathbf{W}$ to select the one object that does not belong. The next subsections describe these three stages in more detail.

6.4.1 Interacting with Objects

Let $\mathcal{B} = \{$*lift*, *shake*, *drop*, *crush*, *push*$\}$ be the set of N exploratory behaviors of the robot and let M be the number of sensory modalities (in our case, $N = 5$, and $M = 2$). Each behavior-modality combination (e.g., *lift-proprioception*) determines a context, which we will denote by $c \in \mathcal{C}$, where $\mathcal{C}$ is the set of all contexts. In our experiments, the size of $\mathcal{C}$ was $|\mathcal{C}| = N \times M = 5 \times 2 = 10$.

Given a context $c \in \mathcal{C}$, and an object $i \in \mathcal{O}$, let $\mathcal{X}_c^i = [X_1, \ldots, X_D]$ be the set of sensory feedback sequences detected while interacting with object i in context c. Each behavior was performed 10 times on each object, hence $|\mathcal{X}_c^i| = 10$. As described below, the robot estimates the similarity between objects using the sets $\mathcal{X}_c^i$ for all modality-behavior contexts $c \in \mathcal{C}$ and all objects $i \in \mathcal{O}$.

6.4.2 Estimating the Similarity between Objects

Next, the robot estimates an $|\mathcal{O}| \times |\mathcal{O}|$ pairwise object similarity matrix $\mathbf{W}$ such that each entry W_{ij} denotes how similar objects i and j are. The similarity matrix is calculated in two steps: 1) for each of the 10 contexts $c \in \mathcal{C}$, estimate an object similarity matrix $\mathbf{W}^c$; and 2) combine the 10 estimated similarity matrices $\mathbf{W}^c$ into a single consensus similarity matrix $\mathbf{W}$.

Let $\mathcal{X}_c^i$ and $\mathcal{X}_c^j$ be two sets containing the sensory feedback sequences detected in context

c with objects i and j, respectively. In our experiments, each set contained 10 such sequences, recorded while performing the same behavior ten times with each object. Let $sim(X_a, X_b)$ be the global alignment similarity function that measures the similarity between two sequences $X_a \in \mathcal{X}_c^i$ and $X_b \in \mathcal{X}_c^j$. For context $c \in \mathcal{C}$, the similarity between two objects i and j can be defined as the expected pairwise similarity of two sequences X_a and X_b:

$$W_{ij}^c = \mathbf{E}[sim(X_a, X_b) | X_a \in \mathcal{X}_c^i, X_b \in \mathcal{X}_c^j].$$

The expected value is estimated as follows:

$$\frac{1}{|\mathcal{X}_c^i| \times |\mathcal{X}_c^j|} \sum_{X_a \in \mathcal{X}_c^i} \sum_{X_b \in \mathcal{X}_c^j} sim(X_a, X_b).$$

In other words, the entry W_{ij}^c is estimated by calculating the average similarity of all possible pairs of sensory feedback sequences in the two sets $\mathcal{X}_c^i$ and $\mathcal{X}_c^j$. Let $\mathbf{W}^c \in \mathbb{R}^{|\mathcal{O}| \times |\mathcal{O}|}$ be the resulting pairwise object similarity matrix for behavior-modality combination c. The matrices $\mathbf{W}^c$ for all contexts are used to construct a single consensus similarity matrix, $\mathbf{W}$, using a weighted combination:

$$W_{ij} = \sum_{c \in \mathcal{C}} \alpha_c \times W_{ij}^c$$

where α_c is the weight assigned to context c (i.e., the consensus object similarity matrix $\mathbf{W}$ is a linear combination of the similarity matrices $\mathbf{W}^c$ for all contexts $c \in \mathcal{C}$).

Two different weighting schemes were used to calculate $\mathbf{W}$. In the first, the weights are uniform, i.e., $\alpha_c = \frac{1}{|\mathcal{C}|}$. In the second, it is assumed that the robot can estimate how useful each behavior-modality context is for the task of object recognition. A context that enables the robot to better distinguish between objects is deemed more useful and assigned a higher weight. Let a_c be the object recognition accuracy achieved in context c, estimated by performing 10-fold cross validation on all data recorded in that context and evaluating a classifier that estimates the object identity given the sensory feedback sequence as input. Once these accuracies are estimated, the weights α_c are computed such that $\alpha_c \propto a_c$ and $\sum_{c \in \mathcal{C}} \alpha_c = 1.0$. The classifier used in this stage was the k-Nearest Neighbor classifier with k set to 3, using the global alignment similarity function to rank neighbors. The classifier, the similarity function

$sim(X_a, X_b)$, and the cross-validation setup were identical to the ones used in the experiments described in Chapter 4.

6.4.3 Detecting the Odd Object

Given an object similarity matrix (either a context-specific matrix $\mathbf{W}^c$ or a consensus matrix $\mathbf{W}$), the robot's model is queried to select the most dissimilar object from a test set $\mathcal{T}$ of K objects, where $\mathcal{T} \subset \mathcal{O}$. For example, if presented with three pop cans and a cowboy hat, we expect the hat to be selected as the object that does not belong in that group. The robot's model selects the odd object i such that the pairwise object similarity within the remaining group of $K-1$ objects is maximized, while the similarity between the selected object i and the remaining $K-1$ objects is minimized.

Given a set of objects, $\mathcal{T}$, the odd object is selected as the object i that maximizes the following objective function:

$$q(\mathcal{T}, i) = \alpha_1 \sum_{j \in \mathcal{T}/i} \sum_{k \in \mathcal{T}/i} W_{jk} - \alpha_2 \sum_{j \in \mathcal{T}/i} W_{ij}.$$

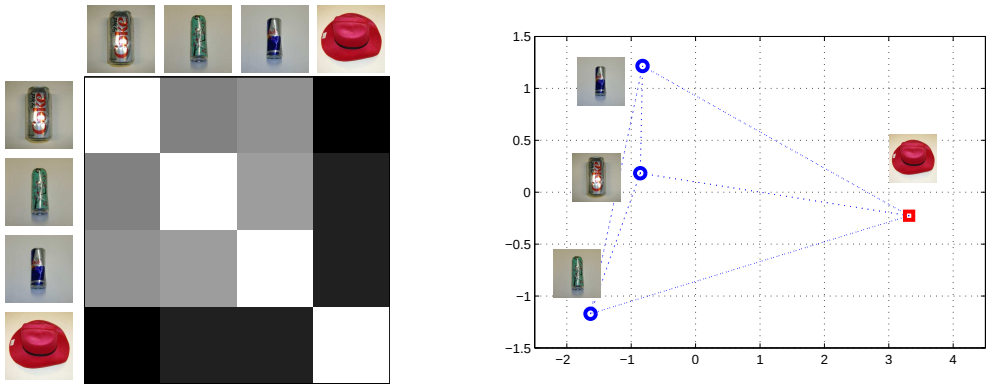
The first term captures the pairwise object similarity between the remaining objects in $\mathcal{T}$ (i.e., after i is removed from $\mathcal{T}$). The second term captures the similarity between the selected object i and the remaining $K-1$ objects in $\mathcal{T}$. It is worth noting that the objective function is based on the general normalized-cut criterion, which is commonly used in the machine learning community for clustering data points whose similarity is specified by an affinity matrix (von Luxburg (2007)). The constants α_1 and α_2 are normalizing weights, which ensure that the objective function is not biased towards either one of the two terms. In our case, the weights were set to:

$$\alpha_1 = \frac{1}{(|\mathcal{T}|-1) \times (|\mathcal{T}|-1)},$$

$$\alpha_2 = \frac{1}{|\mathcal{T}|-1}.$$

6.4.4 Evaluation

The framework was evaluated as follows. Given a target category (e.g., metal objects), three objects from the category and one from outside the category were chosen at random. The robot's model was then queried to pick the odd object. If the selected object matched the object from outside the category, then the solution was deemed a success. This process was repeated for all possible combinations of three objects from the category and one object from outside the category. For example, consider the *metal objects* category, which has 5 objects (see Figure 6.1). There are $\binom{5}{3} = 10$ possible ways to select three objects out of five. For each of these, there are $50 - 5 = 45$ ways to select a fourth object from the dataset that does not belong to that category. Thus, a total of $10 \times 45 = 450$ odd one out tests were performed with that category. The extensive evaluations of these tests were performed off-line after the robot interacted on-line with all 50 objects.



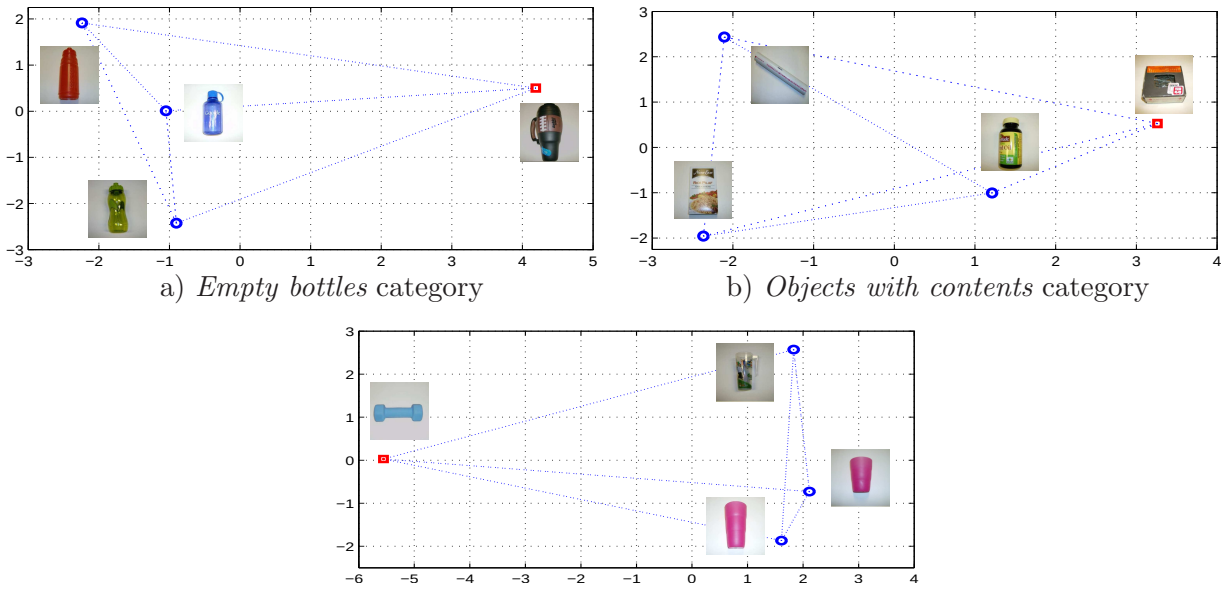
a) Pairwise object similarity matrix b) 2D ISOMAP embedding

Figure 6.2 An example odd one out task. Four objects are presented: three pop cans and a cowboy hat. As expected, the hat is selected by the robot's model as the object that does not belong in that group. a) The pairwise object similarity matrix for the four objects (a sub-matrix of the unweighted consensus similarity matrix **W**). White color indicates high similarity, while black color indicates low similarity. b) A 2D embedding of the pairwise similarity matrix, produced by converting it into a distance matrix and applying the ISOMAP method for non-linear dimensionality reduction. This visualization clearly shows that the cowboy hat is the object in the group that is most distant from the remaining three. The distance between points in the 2D embedding approximates the distance in the matrix used as an input to the ISOMAP algorithm.

6.5 Results

6.5.1 Example

Figure 6.2 shows an example task in which the robot is presented with the three pop cans along with the cowboy hat, and is asked to select the object that does not belong in this group. Figure 6.2.a shows images of the objects and the pairwise object similarity for these four objects (i.e., a sub-matrix of the uniformly-weighted consensus similarity matrix **W**). As expected, the matrix shows that the three pop cans are far more similar to each other, than they are to the cowboy hat. To better visualize the similarity relationships between the four objects, the similarity matrix is embedded onto the 2D plane by first converting it into a distance matrix and then applying the ISOMAP method for dimensionality reduction (Tenenbaum et al., 2000). Figure 6.2.b shows the resulting graph. The distance between two nodes in the graph is an approximation of their distance specified in the input to the ISOMAP algorithm. The hat



a) *Empty bottles* category

b) *Objects with contents* category

c) *Cups* category

Figure 6.3 Examples and solutions of the odd one out task with different object categories. See the text for more details.

object, maximizes the objective function defined earlier.

Figure 6.3 shows three more example tasks, including one in which the robot's model makes a mistake. The object selected by the robot's model is denoted by a red square glyph, while the remaining objects are denoted by blue circle glyphs. Figure 6.3.a shows an example task in which the robot's model is queried to pick the odd object out of three different types of bottles and a mug. The visualization shows that the mug is clearly the most different object. Figure 6.3.b shows a test in which the four objects presented to the robot include three that have contents inside of them (a box of rice, a bottle with pills, and a box with screws) and one that does not (a PVC pipe). The robot's model selects the box with screws as the most different object, which is an incorrect choice, according to the human-labeled object category (i.e., *objects with contents*). Finally, Figure 6.3.c shows a test in which the dumbbell is correctly selected as being different from the three plastic cups.

6.5.2 Success Rates Per Object Category

The performance rates for all six object categories are shown in Table 6.1, averaged over all possible instantiations of the odd one out task for each category. Rates are shown for both the uniform weighting scheme as well as the weighting scheme in which contexts are weighted

Table 6.1 Success Rates per Task Category

Category	Uniform Combination	Weighted Combination	Best Context
Pop Cans	100.00 %	100.00 %	100.00 %
Plastic cups	76.59 %	87.23 %	97.87 %
Metal objects	70.00 %	80.00 %	95.33 %
Empty bottles	62.96 %	66.47 %	63.97 %
Soft objects	50.44 %	67.78 %	97.33 %
Objects w/ contents	45.34 %	49.89 %	66.71 %

based on their usefulness for distinguishing between objects. In addition, for each category, the individual context that results in the highest success rate is determined and the resulting success rate is reported. The idea behind this test is that certain behavior-modality contexts may be better suited for detecting certain object categories than others. The expected success rate when randomly selecting the odd object is 25% (i.e., randomly choosing 1 out of 4).

The results show that the robot's unsupervised model is substantially better than chance for all six object categories. The weighted combination scheme performs better than the uniform combination. The best results are achieved with the *pop cans* category, for which the robot was able to select the object that is not a pop can in 100% of the tests. These results indicate that the robot's behavioral and perceptual repertoire was able to capture the common properties that define pop cans (e.g., material type, specific sounds they generate, weight, etc.). The worst performance is for the *objects with contents* category. The only thing that these objects have in common is that they make noise when shaken (i.e., only 1 of 10 contexts, *shake-audio*, may be able to capture that). The robot's model, however, is completely unsupervised and does not know that the object similarity matrix extracted in the *shake-audio* context is the most relevant for this category type.

The last column of Table 6.1 shows that for most object categories, there exists a specific behavior-modality context that results in a success rate that is higher than the one achieved when using the consensus similarity matrix. For example, when using only the object similarity matrix extracted from the *shake-audio* context, the success rate for the *objects with contents*

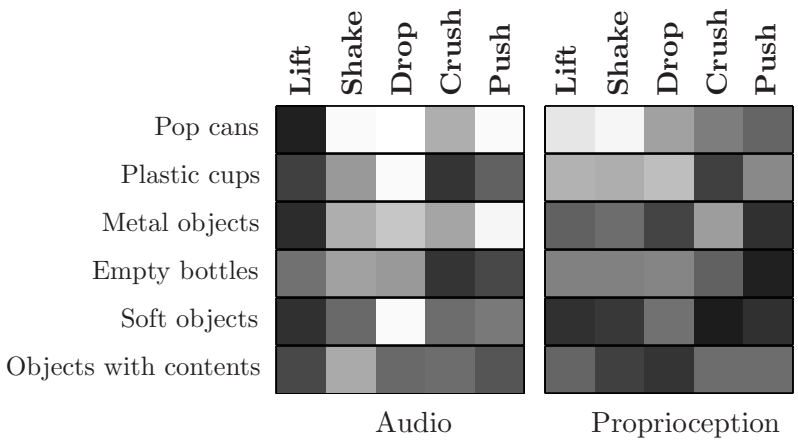


Figure 6.4 Success rates for the odd one out task, shown for each category, and for each behavior-modality context. Light color indicates high success rates, while dark color indicates low success rates.

category jumps to 66.71%. The *empty bottles* category was an exception - in that case, using the weighted consensus similarity matrix results in a higher success rate than with any individual context-specific similarity matrix.

Figure 6.4 visualizes the success rates for each category when using each context-specific similarity matrix $\mathbf{W}^c$. Light color indicates high success rates. The results show that the properties of different categories are best captured by different behaviors and modalities. For example, the *plastic cups* category is best captured by the *drop-audio* behavior-modality context. This context is also very useful when the robot is evaluated on the *soft objects* category, likely because the robot detects an absence of a loud noise when these objects are dropped on the table. As expected, the *objects with contents* category is best captured by the *shake-audio* behavior-modality combination, since the contents make distinct sounds when the objects are shaken.

6.6 Summary

This chapter introduced an interactive framework and a theoretical model that allow a robot to solve the *odd one out* task by estimating the similarity relations between objects in different behavior-modality contexts. The experimental evaluation showed that the robot's choice for the odd object was consistent with human-defined object categories, with success rates varying from 45% to 100%, depending on the category. Certain behavior-modality combinations produced object similarity relations that were able to better capture the target category. These results show that sensorimotor interaction can capture many of the physical properties of objects that define an object category.

One limitation of the model presented here is that the objective function for deciding which of the objects does not belong in the group was pre-defined. Future work can address this by incorporating some amount of human supervision into the overall framework. If the robot knows whether its choice for the odd object is right or wrong, it could potentially estimate which behavior-modality combinations are most suitable for capturing the properties of a target object category. This information can also be used to estimate specific weights for each context in order to learn a new object similarity relation that better captures a given human-defined category. Pursuing this line of research would allow robots to solve a variety of additional tasks, including sorting and ordering objects.

CHAPTER 7. OBJECT CATEGORY RECOGNITION USING BEHAVIOR-GROUNDED RELATIONAL LEARNING*

7.1 Introduction

Learning to classify objects into categories is a fundamental milestone in human development. Such an ability is crucial for robots that have to operate in human environments where object categorization skills are required for solving many practical tasks (e.g., sorting objects in order to clean a room or unload a dishwasher). Not surprisingly, there has been much recent progress in enabling robots to robustly recognize and categorize objects, using both supervised and unsupervised machine learning methods.

There are two main limitations of current approaches to object category recognition. First, most methods rely exclusively on computer vision or laser scan data, gathered through passive observation (Quigley et al., 2007; Rusu et al., 2008; Srinivasa et al., 2009; Endres et al., 2009). Given a clear view of the object, such methods can achieve high classification accuracy. Nevertheless, experiments in psychology have shown that many object properties (e.g., material type, weight, etc.) can only be perceived through the use of auditory, proprioceptive, and other non-visual sensory modalities (Lynott and Connell, 2009). For example, using vision alone, a robot cannot distinguish between an empty bottle and a full bottle that otherwise look the same.

Another major limitation of current approaches to object classification is that they typically fail to exploit relational information that specifies how similar two objects are in a given context. Instead, objects are usually classified based on static visual features alone. Recent results from

*This chapter is based on the following paper: Sinapov, J. and Stoytchev, A., "Object Category Recognition by a Humanoid Robot Using Behavior-Grounded Relational Learning", *In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 184-190, 2011.

the machine learning community, however, have shown that by exploiting relations that link objects (e.g., citations link academic papers, hyperlinks connect web pages, etc.) it is possible to further increase the classification accuracy (see Getoor and Diehl (2005) for a literature survey).

To address these limitations, this chapter proposes a behavior-grounded approach for classifying objects into categories that estimates and uses object similarity measures grounded in raw sensorimotor interactions. Rather than trying to classify objects through passive observations, our robot actively interacts with them by applying five different exploratory behaviors. Over the course of each interaction, the robot detects auditory feedback captured by a microphone and proprioceptive feedback captured by joint torque sensors in the robot's arm. The sensorimotor data is used to estimate multiple pairwise measures of object similarity, each corresponding to a unique coupling between an exploratory behavior and a sensory modality. A graph-based recognition model is trained by extracting features from the estimated similarity relations, allowing the robot to recognize the category memberships of novel objects based on the objects' similarity to the set of familiar objects.

The framework was evaluated on an upper-torso humanoid robot with two large sets of objects. The results show that the model was able to recognize human-provided object categories significantly better than chance. The results also make a strong case that robots should interact with objects using a rich behavioral repertoire and many sensory modalities in order to better ground object categories in sensorimotor experience.

7.2 Experimental Setup

The experimental setup used in this study is identical to the one described in the previous chapter. To summarize, the robot was an upper-torso humanoid robot with two 7-DOF Barrett WAMs for arms, each with a 3-finger Barrett Hand as an end effector. The robot's head was equipped with an Audio-Technica U853AW cardioid microphone that was used to capture auditory feedback. Joint torque sensors in each joint were used to capture proprioceptive feedback at 500 Hz using the robot's low-level API.

The robot explored objects by applying five exploratory behaviors on them: *lift*, *shake*, *drop*,

Figure 7.1 The six object categories. An object may belong to multiple categories, e.g., the three pop cans also belong to the set of metal objects.

crush, and *push*. All behaviors were encoded with the Barrett WAM API and performed with the left arm. During the execution of each behavior, the raw proprioceptive data (i.e., joint torques of the left arm) and the raw audio were recorded from start to end. The proprioceptive and auditory feedback for each behavioral interaction were represented as discrete sequences, by reducing the dimensionality of the raw sensory input using Self-Organizing Maps (SOMs) (Kohonen, 2001). The feature extraction routines that were used are identical to the ones described in Chapter 4.

The object categories used to test the category recognition model were identical to the six categories described in the previous chapter, and shown here in Figure 7.1. The original data set was collected for a different purpose and had an additional 25 objects that are not included here because they did not fall into any of the 6 object categories, nor did they form any other object categories that we could identify. As described in Section 7.4.5, the proposed model was also evaluated on another data set from an earlier experiment with a different set of objects, which was originally used for the task of acoustic object recognition, as described in Sinapov et al. (2009).

7.3 Theoretical Model

This section describes how a robot can classify objects into object categories using relational machine learning methods. The approach consists of 3 broad stages: 1) interaction stage – the robot explores the objects by applying its set of exploratory behaviors on them; 2) similarity estimation stage – the robot estimates multiple pairwise measures of similarity between the objects, each corresponding to a specific coupling between an exploratory behavior and a sensory modality; and 3) category learning stage – relational features are extracted from the similarity relations and used to train recognition models that can estimate the category memberships of novel objects.

7.3.1 Interacting with Objects

During the first stage, the robot interacts with the set of objects $\mathcal{O}$ using a set $\mathcal{B}$ of N exploratory behaviors. For the experimental setup described so far, $\mathcal{B} = \{$*lift*, *shake*, *drop*, *crush*, *push*$\}$ and $N = 5$. During the execution of each behavior, feedback from M sensory modalities is recorded (in our case $M = 2$). Each unique behavior-modality combination (e.g., *drop-auditory*) specifies a sensorimotor context $c \in \mathcal{C}$, where $\mathcal{C}$ is the set of all contexts (in our case $|\mathcal{C}| = 10|$).

Let $\mathcal{X}_c^i = [X_1, \ldots, X_D]$ be the set of sensory feedback sequences detected while interacting D times with object o_i in context c. In our experiments, the robot performed each behavior 10 times on each object, thus $|\mathcal{X}_c^i| = 10$. The next subsection describes how the sets $\mathcal{X}_c^i$ can be used to estimate multiple pairwise similarity measures for all objects in the set $\mathcal{O}$ and all modality-behavior contexts $c \in \mathcal{C}$.

7.3.2 Estimating the Similarity Between Objects

After the interaction stage, the robot estimates pairwise object similarity matrices $\mathbf{W}^c \in \mathbb{R}^{|\mathcal{O}|\times|\mathcal{O}|}$ for all behavior-modality contexts $c \in \mathcal{C}$. Each entry $W_{ij}^c \in \mathbb{R}$ denotes how similar objects o_i and o_j are in sensorimotor context c.

Intuitively, if two objects produce similar sensory feedback sequences when a particular

behavior is applied on them, then they should be considered similar in that context. Given two objects o_i and o_j, let $\mathcal{X}_c^i$ and $\mathcal{X}_c^j$ be the two sets containing the sensory feedback sequences detected with these objects in context c. Let $sim(X_a, X_b)$ be the global alignment similarity function that measures the similarity between two sequences from the same modality. The pairwise object similarity between objects o_i and o_j can then be defined as the expected pairwise similarity of two sequences $X_a \in \mathcal{X}_c^i$ and $X_b \in \mathcal{X}_c^j$:

$$W_{ij}^c = \mathbf{E}[sim(X_a, X_b) | X_a \in \mathcal{X}_c^i, X_b \in \mathcal{X}_c^j],$$

where the expected value is estimated from available data as:

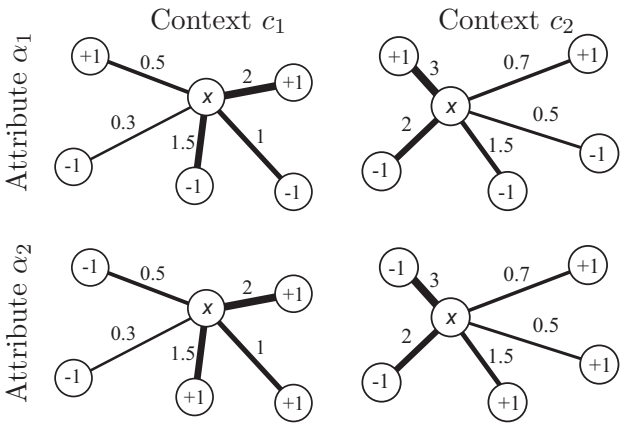
$$\frac{1}{|\mathcal{X}_c^i| \times |\mathcal{X}_c^j|} \sum_{X_a \in \mathcal{X}_c^i} \sum_{X_b \in \mathcal{X}_c^j} sim(X_a, X_b).$$

In other words, given a context c and objects o_i and o_j, the entry W_{ij}^c is computed by calculating the average similarity for all possible pairs of sensory feedback sequences detected with the two objects.

7.3.3 Object Category Recognition using Relational Features

During the third and final stage, the robot learns a set of relational classifiers that can estimate the category memberships of a novel object using the entries of the similarity matrices $\mathbf{W}^c$ and the category labels of familiar objects. Let $\mathcal{A} = [\alpha_1, \ldots, \alpha_K]$ be the set of attributes (or category memberships) used to label the familiar objects, each corresponding to a particular object category (e.g., *PopCans* or *PlasticCups*). Let the function $label(o_i, \alpha) \rightarrow \{-1, +1\}$ specify whether object o_i belongs to category α $(+1)$ or not (-1). In our experiments, there were six object category attributes ($K = 6$). Figure 7.1 shows the category memberships of the objects.

Given a set of objects with known attribute labels, the task of the robot is to learn a classification model that can be used to estimate the labels (either -1 or $+1$) of novel objects for all attributes in $\mathcal{A}$. This task is solved in two steps: 1) for each object, extract relational features from the similarity graphs defined by $\mathbf{W}^c$ for all sensorimotor contexts $c \in \mathcal{C}$; and



$f^{\alpha_1}_{x,c_1} = \frac{0.5+2}{2} = 1.25$ $f^{\alpha_1}_{x,c_2} = \frac{3+0.7}{2} = 1.85$

$f^{\bar{\alpha_1}}_{x,c_1} = \frac{0.3+1.5+1}{3} = 0.933$ $f^{\bar{\alpha_1}}_{x,c_2} = \frac{2+1.5+0.5}{3} = 1.333$

$f^{\alpha_2}_{x,c_1} = \frac{2+1+1.5}{3} = 1.5$ $f^{\alpha_2}_{x,c_2} = \frac{0.7+0.5+1.5}{3} = 0.9$

$f^{\bar{\alpha_2}}_{x,c_1} = \frac{0.5+0.3}{2} = 0.4$ $f^{\bar{\alpha_2}}_{x,c_2} = \frac{3+2}{2} = 2.5$

Figure 7.2 A simple example of relational feature extraction. In this case, there are two contexts (c_1 and c_2) and two attributes (α_1 and α_2). There are five familiar objects with known labels (either -1 or $+1$) for both attributes and one unlabeled novel object (denoted with x). The edges correspond to the similarity between the novel object and the familiar ones (the edges between familiar objects are not shown). To represent the novel object, for each combination of a context c and an attribute α, two features are extracted, $f^{\alpha}_{x,c}$ and $f^{\bar{\alpha}}_{x,c}$. The first feature is simply the average similarity in context c between the novel object and familiar objects that are members of the category α. The second feature is calculated in a similar way but for the objects that do not belong to the category. There are 8 features in this example.

2) for each attribute α, train a recognition model M_α that can estimate the class label of an unlabeled object, given the extracted relational features for that object.

Let $\mathcal{O}_\alpha$ be the set of labeled objects for which $label(o_i, \alpha) = +1$, and let $\mathcal{O}_{\bar{\alpha}}$ be the remaining set of labeled objects that do not belong to category α, such that $\mathcal{O}_\alpha \cap \mathcal{O}_{\bar{\alpha}} = \emptyset$. Given an unlabeled object o_i, a context c, and an attribute α, we can then extract two features, $f^{\alpha}_{i,c} \in \mathbb{R}$ and $f^{\bar{\alpha}}_{i,c} \in \mathbb{R}$, which are defined as:

$$f^{\alpha}_{i,c} = \mathbf{E}[W^c_{ij} | o_j \in \mathcal{O}_\alpha],$$

$$f^{\bar{\alpha}}_{i,c} = \mathbf{E}[W^c_{ij} | o_j \in \mathcal{O}_{\bar{\alpha}}].$$

In other words, $f^{\alpha}_{i,c}$ specifies the expected similarity in behavior-modality context c between object o_i and all objects o_j for which $label(o_j, \alpha) = +1$, while $f^{\bar{\alpha}}_{i,c}$ specifies the same, but for all objects o_j that do not belong to category α. These expectations are estimated from the available data:

$$f^{\alpha}_{i,c} = \mathbf{E}[W^c_{ij} | o_j \in \mathcal{O}_\alpha] \cong \frac{1}{|\mathcal{O}_\alpha|} \sum_{o_j \in \mathcal{O}_\alpha} W^c_{ij},$$

$$f^{\bar{\alpha}}_{i,c} = \mathbf{E}[W^c_{ij} | o_j \in \mathcal{O}_{\bar{\alpha}}] \cong \frac{1}{|\mathcal{O}_{\bar{\alpha}}|} \sum_{o_j \in \mathcal{O}_{\bar{\alpha}}} W^c_{ij}.$$

Figure 7.2 shows an example of relational feature extraction with 2 contexts, 2 binary attributes, and 6 objects. Five of the objects are familiar (with known labels of either -1 or $+1$) and one is a novel object (denoted by x). The links correspond to the similarity between the novel object and the familiar ones (the thicker the link, the more similar the objects). In this example, 8 relational features are extracted. In the experimental setup described earlier, there were 10 contexts, 6 binary attributes, and 2 relational features for each context-attribute combination. Thus, each object was represented by a $10 \times 6 \times 2 = 120$ dimensional feature vector $\mathbf{f}_i \in \mathbb{R}^{10 \times 6 \times 2}$. For each attribute $\alpha \in \mathcal{A}$, a separate recognition model M_α (i.e., a classifier) is trained such that $M_\alpha(\mathbf{f}_i) \rightarrow label(o_i, \alpha)$. Three different machine learning methods were evaluated as implementations of the recognition models M_α: Support Vector Machine (SVM), k-Nearest Neighbors (k-NN), and Decision Tree (C4.5).[†]

[†]The WEKA machine learning library, which provides implementations of k-NN, SVM, and C4.5, was used (Witten and Frank, 2005). For SVM, the default polynomial kernel function with exponent set to 2.0 was used. For k-NN, the value of k was set to 3. To handle the unbalanced nature of the training sets (i.e., most data points have a class label of -1), an ensemble classifier approach was adopted, in which 20 different classifiers (all of the same type) were each trained on a randomly re-sampled version of the training set with equal number of positive and negative examples. The outputs of the individual classifiers in the ensemble were combined using uniform weights.

Table 7.1 Interpreting κ coefficient values, as proposed by Landis and Koch (1977).

κ	Strength of Agreement
$0.81 - 1.00$	Almost Perfect
$0.61 - 0.80$	Substantial
$0.41 - 0.60$	Moderate
$0.21 - 0.40$	Fair
$0.01 - 0.20$	Slight
≤ 0.0	Poor

7.4 Results

7.4.1 Evaluation

The recognition models M_α were evaluated using *object-based* cross-validation. During each round of evaluation, the robot's six category recognition models were trained with the known labels for $|\mathcal{O}| - 1$ objects and evaluated on the remaining one object. For the purposes of training, the relational features used to represent each object were estimated using only the labels of the $|\mathcal{O}| - 1$ objects in the training set. For each evaluation round, the output of each model M_α was logged and compared against the ground truth (i.e., human-provided labels). The end result of this classification procedure was one 2×2 confusion matrix for each individual attribute, which specified how many of the model's predictions were true positives, true negatives, false positives, and false negatives.

Because for many attributes most objects have a label of -1, reporting the raw accuracy may be misleading. For example, given the attribute $PopCans$, only 3 out of the 25 objects have a label of $+1$. Thus, a classifier that always predicts -1 can achieve 88% accuracy, and yet this performance is no better than chance. Therefore, the performance of the recognition models is reported in terms of Cohen's kappa coefficient (Cohen, 1960), a statistic that compares the classifier accuracy against chance accuracy, which is defined as:

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)},$$

where $Pr(a)$ is the probability of correct classification by the classifier and $Pr(e)$ is the prob-

Table 7.2 Kappa Statistics for classifiers M_α obtained with k-NN, Decision Tree, and SVM machine learning algorithms.

Category	k-NN	Decision Tree	SVM
Pop Cans	0.834	0.692	**0.834**
Plastic Cups	0.097	0.408	**0.412**
Metal Objects	**0.821**	0.667	0.750
Empty Bottles	0.337	0.072	**0.481**
Objects w/ Contents	0.547	0.669	**0.753**
Soft Objects	0.197	**0.858**	0.750

ability of correct classification by chance. For example, if the evaluation resulted in 3 true positives, 21 true negatives, 1 false positive, and 0 false negatives, then $Pr(a) = \frac{3+21}{25} = 0.96$, $Pr(e) = \frac{3+0}{25} \times \frac{3+1}{25} + \frac{21+1}{25} \times \frac{21+0}{25} = 0.7584$, and thus, $\kappa = 0.834$, which indicates almost perfect classification. On the other hand, a trivial classifier that always outputs -1 as the class label, results in $Pr(a) = Pr(e) = \frac{22}{25}$ and $\kappa = 0$. Table 7.1 shows how to interpret κ values as proposed by Landis and Koch (1977).

7.4.2 Object Category Classification Rates

The first experiment measures the performance of the classifiers M_α for all attributes α in terms of the kappa coefficient. Table 7.2 shows the resulting classification performance for the three different machine learning algorithms that were used to implement M_α. In nearly all cases, the performance is substantially better than chance (i.e., κ greater than 0.0). This result indicates that the relational features contain information that is useful for estimating the categories of novel objects, despite the fact that visual feedback was not provided to the classifier model. Furthermore, the classification rates highlight the importance of auditory and proprioceptive feedback for grounding complex object categories in raw sensorimotor experience.

It is also important to look at the type of errors made by the robot's classification model. For example, for the *PopCans* attribute, both k-NN and SVM make only one mistake, by incorrectly labeling the small metal cup as a pop can. This is not surprising, considering that

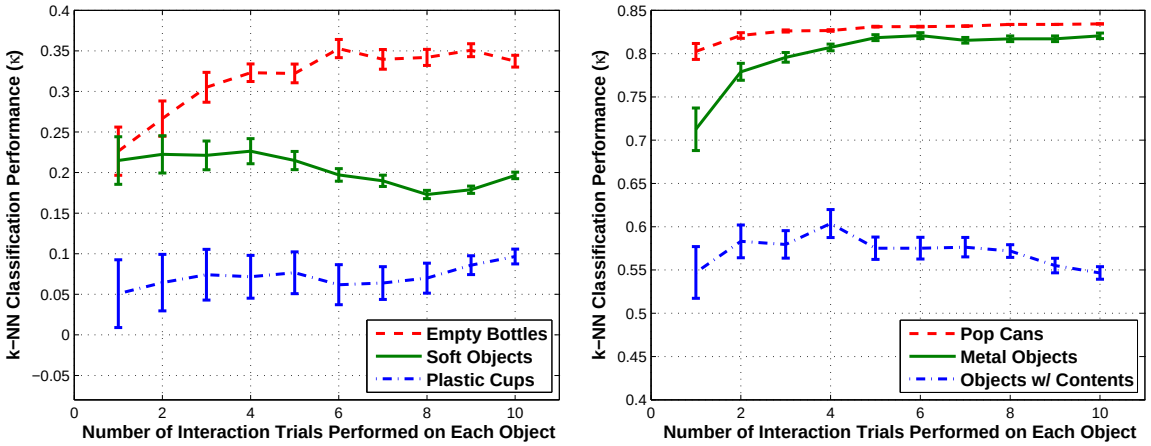


Figure 7.3 Classification performance of the k-NN category recognition model as a function of the number of interaction trials used to estimate the object similarity matrices $\mathbf{W}^c$.

the metal cup produces similar sounds to the pop cans as these objects share the same material type. Similarly, the hard plastic bottles are often misclassified as belonging to the *PlasticCups* category, due to both material and weight similarities.

7.4.3 Classification Performance vs. Amount of Interaction

The second experiment aims to see how much experience with the objects is necessary for the classification performance to converge. To find out, the number of trials used to estimate the similarity matrices $\mathbf{W}^c$ was varied from 1 to 10. Because there are multiple ways to choose which trials should be used, the evaluation was repeated 200 times at each level. The mean and the variance of the kappa statistic were recorded for each level and for each of the 6 attributes. Due to the large number of evaluations, only the k-NN algorithm was chosen because of its relatively fast runtime performance with 25 objects.

Figure 7.3 shows the results of this experiment. There is a slight to moderate improvement in the classification performance for several of the categories as the robot performs more interaction trials with the objects. For all six categories, there is a notable decrease in the variance of the classification performance as the robot gains more experience with the objects. For some of the object categories (e.g., *PopCans*), the model's performance is nearly the same, regardless of how many trials are used to estimate the object similarity relations.

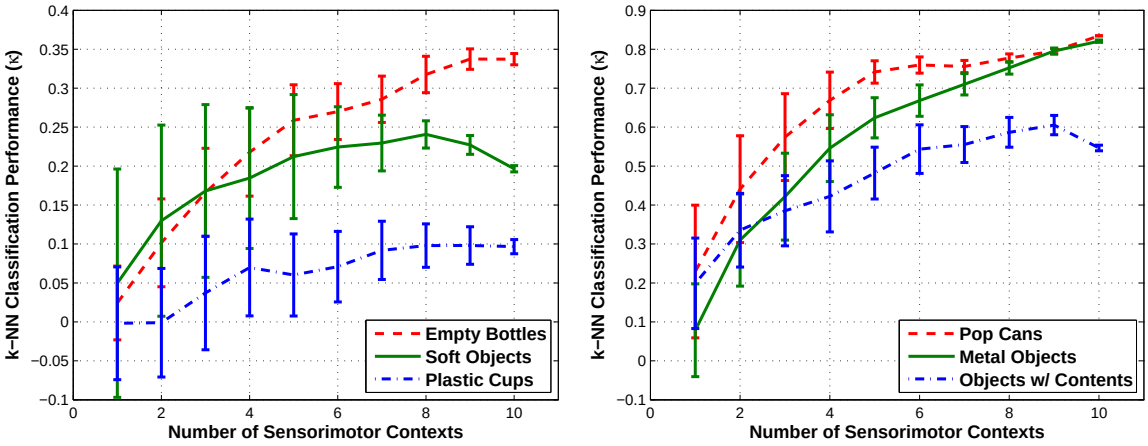


Figure 7.4 Classification performance of the k-NN category recognition algorithm as a function of the number of sensorimotor contexts available to the relational recognition model.

7.4.4 The Role of Exploratory Behaviors and Sensory Modalities

The next experiment measures the model's performance as a function of the number of available behavior-modality contexts. This is done by varying the number of object similarity relations $\mathbf{W}^c$ used to extract relational features from 1 (i.e., the robot has only one behavior and perceives only one sensory modality) to 10 (i.e., the results shown in Table 7.2). Since there are multiple ways to select a subset of contexts, the evaluation was repeated 200 times at each level with a different random seed.

Figure 7.4 visualizes the results of this experiment. As expected, the model's performance tends to increase as the model uses object similarity relations extracted from more contexts. More importantly, when compared with the results of the previous experiment, Figure 7.4 shows that the number of different behaviors and sensory modalities available to the robot is far more important than the number of interaction trials performed on each object. In other words, the performance improves much faster when more sensorimotor contexts are added, than when more trials are added. Thus, the diversity of experience with objects counts more than the sheer amount of experience. This result makes a strong case that robots should interact with objects using a rich behavioral repertoire and a large number of sensory modalities. It also complements our previous work (Sinapov and Stoytchev, 2010a), which has shown that exploratory behaviors act as classifiers that can be boosted. Research by Gibson (1988) and Power (2000) has indeed shown that animals and humans use multiple exploratory behaviors

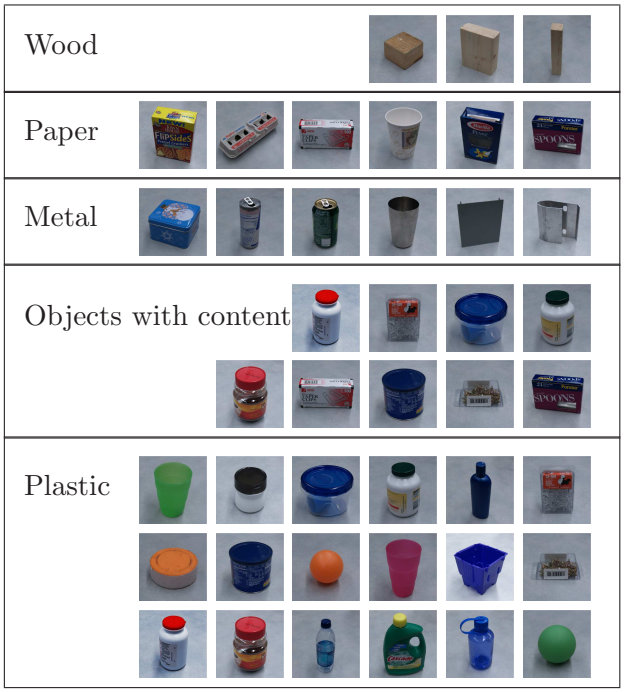


Figure 7.5 The objects from the second data set and their corresponding categories, which were used to further validate the method presented in this chapter. Some objects belong to multiple categories. Three of the objects in that data set do not belong to any of the five categories and are not shown here.

and multiple sensory modalities to both learn and represent the properties of objects.

7.4.5 Validation on a Second Data Set

Finally, the proposed method was evaluated on another data set, which was previously used for the tasks of acoustic object recognition (Sinapov et al., 2009) and categorization (Sinapov and Stoytchev, 2009). In this experiment, the robot performed five exploratory behaviors (grasp, shake, drop, push and tap) on 36 household objects (see Figure 7.5). The auditory data from each trial was recorded and converted into a discrete sequence using the method introduced by Sinapov et al. (2009). Since there is only one sensory modality, only 5 object similarity matrices $\mathbf{W}^c$ were estimated, one for each exploratory behavior. The objects were labeled according to five attributes: *Plastic*, *Paper*, *Metal*, *Wood*, and *Contents*. The first 4 refer to the objects' material type while the last indicates whether or not the object has contents inside of it (e.g., a full pill bottle). A detailed description of how each object was labeled is available in (Sinapov and Stoytchev, 2009).

Table 7.3 Classification Performance in terms of the *kappa* statistic (κ) on the data set from Sinapov et al. (2009).

Category	k-NN	Decision Tree	SVM
Plastic	**0.328**	0.100	**0.328**
Paper	0.110	**0.420**	0.178
Metal	**0.684**	0.641	0.625
Wood	0.262	0.222	**0.302**
Contents	0.633	0.892	**1.000**

Table 7.3 shows the results of this experiment. Overall, the classification model performs significantly better than chance for most of the object category attributes, despite the fact that the object similarity matrices were estimated using only auditory data (i.e., no proprioceptive measure of similarity between the objects was available). The validation experiment shows that the proposed relational learning model can be used by a robot to detect the labels of novel objects in a wide variety of settings. In other words, the model is not bound to specific objects, exploratory behaviors, or sensory modalities.

7.5 Summary

This chapter presented a novel relational (i.e., graph-based) learning framework that can enable a robot to recognize the categories of novel objects by relating them to familiar objects. In contrast to traditional object classification methods that directly map visual object features to categories, the model presented here makes use of relational information that specifies how similar two objects are in a variety of sensorimotor contexts. An important feature of our framework is its ability to simultaneously handle multiple robot behaviors, sensory modalities, and object attributes.

The results presented here were obtained by evaluating our method on two large-scale experimental data sets and have several important implications for research in robotics. First, the robot was able to achieve high object classification accuracy, despite the fact that visual feedback was not used as an input to the robot's model. This finding highlights the importance of non-visual sensory modalities for robotic perception of objects. Second, as the robot was able to experience the objects in more and more sensorimotor contexts, the model's performance increased dramatically. This result shows that the level of diversity of sensorimotor experience with objects is crucial for learning meaningful object representations through behavior-grounded exploration.

There are several directions for future research. First, while the model presented here uses dense object similarity matrices, sparse representations could be explored in order to scale up the framework to a much larger number of objects. Second, the relational object representation enables the use of semi-supervised graph-based learning methods, which have the added advantage of requiring only a few labeled objects (see Zhu et al. (2005) for a review). Finally, the duration of the object exploration stage can be reduced, while still maintaining good classification performance, by adapting active learning methods to operate on graph-based representations.

CHAPTER 8. GROUNDING SEMANTIC CATEGORIES IN BEHAVIORAL INTERACTIONS: EXPERIMENTS WITH 100 OBJECTS*

8.1 Introduction

Object categories are all around us - our homes and offices contain a vast multitude of objects that can be organized according to a diverse set of criteria ranging from form to function. A robot operating in human environments would undoubtedly have to assign category labels to novel objects because it is simply infeasible to preprogram it with knowledge about every individual object that it might encounter. For example, to clean a kitchen table, a robot has to recognize semantic object category labels such as silverware, dish, or trash before performing an appropriate action.

The ability to learn and utilize object category memberships is an important aspect of human intelligence and has been extensively studied in psychology (Ashby and Maddox, 2005). A large number of experimental and observational studies have revealed that object category learning is also linked to our ability to acquire words (Fulkerson and Waxman, 2007; Plunkett et al., 2008). Researchers have postulated that, with a few labeled examples, humans at various stages of development are able to identify common features that define category memberships as well as distinctive features that relate members and non-members of a target category (Hammer et al., 2009, 2010). Other lines of research have highlighted the importance of object exploration (Gibson, 1988; Power, 2000), which is important for learning object categories since many object properties cannot always be detected by passive observation (Ernst and Bulthof, 2004; Lynott and Connell, 2009).

*This chapter is based on the following paper: Sinapov, J., Schenck, C., Staley, K., Sukhoy, V. and Stoytchev, A., "Grounding Semantic Categories in Behavioral Interactions: Experiments with 100 Objects", *Robotics and Autonomous Systems*, (in press), 2012.

Figure 8.1 The humanoid robot used in our experiments, along with the 100 objects that it explored.

Recently, several research groups have started to explore how robots can learn object category labels that can be generalized to novel objects (Lopes and Chauhan, 2007; Griffith et al., 2009; Marton et al., 2009; Sinapov and Stoytchev, 2011; Leonardis and Fidler, 2011). Most studies have examined the problem exclusively in the visual domain or have used a relatively small number of objects and categories. To address these limitations, this chapter describes an approach to object categorization that enables a robot to acquire a large number of category labels from a large set of objects. This is achieved with the use of multiple behavioral interactions and multiple sensory modalities. To test our method, the robot in our experiment (see Figure 8.1) explored 100 different objects classified into 20 distinct object categories using 10 different interactions (e.g., grasp, lift, tap, etc.) making this one of the largest object sets that a robot has physically interacted with.

Using features extracted from the visual, auditory, and proprioceptive sensory modalities, coupled with a machine learning classifier, the robot was able to achieve high recognition rates on a variety of household object categories (e.g., balls, cups, pop cans, etc.). The robot's model was also able to identify which sensory modalities and behaviors are best for recognizing each category label. In addition, the robot was able to actively select the exploratory behavior that it should try next when classifying an object, which resulted in faster convergence of the model's accuracy rates when compared to random behavior selection. Finally, the model was

evaluated on whether it can detect if a novel object does not belong to any of the categories present in the robot's training set.

8.2 Related Work

Most object categorization methods in robotics fall into one of two broad categories: 1) unsupervised methods, in which objects are categorized using unsupervised machine learning algorithms (e.g., k-Means, Hierarchical Clustering, etc.) and 2) supervised methods, in which a labeled set of objects is used to train a recognition model that can label new data points. Several lines of research have demonstrated methods that enable robots to autonomously form internal object categories based on direct interaction with objects (Nakamura et al., 2007; Griffith et al., 2009; Dag et al., 2010; Sun et al., 2010b). For example, Griffith et al. (2009) showed how a robot can use the frequencies with which certain events occur in order to distinguish between container and non-container objects in an unsupervised manner. Dag et al. (2010) and Sinapov and Stoytchev (2008) have also shown that robots can categorize and relate objects based on the type of effects that they produce when an action is performed on them.

In contrast, the focus of this chapter is on supervised methods for object categorization, which attempt to establish a direct mapping between the robot's object representation and human-provided semantic category labels. A wide variety of computer vision methods have been developed that attempt to solve the problem using visual image features coupled with machine learning classifiers (Fergus et al., 2004; Ponce, 2006; Opelt et al., 2006). Several such methods have been developed for use by robots, almost all exclusively working in the visual domain (Lopes and Chauhan, 2007; Lai and Fox, 2009; Marton et al., 2009; Wohlkinger and Vincze, 2010; Leonardis and Fidler, 2011; Lai et al., 2011a). One advantage of visual object classifiers is that they can often be trained offline on large image datasets. Nevertheless, they cannot capture object properties that cannot always be perceived through vision alone (e.g., object compliance, object material, etc.). In other words, disembodied object category representations that are grounded solely in visual input cannot be used to capture object properties that require active interaction with an object. Thus, even the best visual classifier is guaranteed to fail on certain object classification tasks. For example, Lai et al. (2011b) report

that using state-of-the-art RGB and depth features for classifying 300 objects into 51 categories results in 85.4% accuracy, which demonstrates that there is still a lot of information about object categories that cannot be captured using disembodied vision-based systems. Furthermore, it has been argued that embodied perception is not only desirable, but also required for achieving intelligent autonomous behavior by a robotic system (Vernon, 2008). Therefore, to address the limitation of disembodied systems, our robot grounded the semantic category labels of objects in its own sensorimotor experience with them, which is in stark contrast with approaches that rely purely on computer vision datasets.

The importance of non-visual sensory modalities for robotic object perception has been recognized by several lines of research, which have shown that robots can recognize objects using auditory (Torres-Jara et al., 2005; Sinapov et al., 2009; Rebguns et al., 2011), tactile (Sinapov et al., 2011b; Saal et al., 2010), and proprioceptive (Natale et al., 2004; Bergquist et al., 2009) sensory modalities. For example, Natale et al. (2004) showed that proprioceptive information obtained from the robot's hand when grasping an object can be used to successfully recognize the identity of the object. Similarly, Bergquist et al. (2009) performed an experiment in which a robot was able to recognize a large number of objects using proprioceptive feedback from the robot's arm as it manipulated them. Other research has also shown that auditory features (e.g., sounds generated as the robot's end effector makes contact with an object) can also be useful for recognizing a previously explored object (Torres-Jara et al., 2005; Sinapov et al., 2009). Most recently, a study by Sinapov et al. (2011a) demonstrated that a robot can achieve high object recognition rates when tested on a large set of 50 objects by integrating auditory and proprioceptive feedback detected over the course of exploring the objects. In contrast to this previous work, the study described here demonstrates that behavior-grounded object perception can also be used by a robot to both learn and recognize human-provided semantic category labels for novel objects.

Several studies have already demonstrated some ability of robots to assign category labels to objects based on interaction with them. For example, Takamuku et al. (2007) demonstrated that a robot can classify 9 different objects as either a rigid object, a paper object, or a plastic bottle using auditory and joint angle data obtained when the robot shakes the objects. An

experiment by Chitta et al. (2011) has shown that tactile feedback produced during grasping can be useful for categorizing cans and bottles as either full or empty. In another study, Sinapov and Stoytchev (2009) showed that by applying five different exploratory behaviors on 36 objects, a robot may learn to recognize their material type and whether they are full or empty, based on the auditory feedback produced by the objects.

In previous work, we proposed a graph-based learning method that allows a robot to estimate the category label of an object based on pairwise object similarity relations estimated from different couplings of five exploratory behaviors and two sensory modalities (Sinapov and Stoytchev, 2011). In that experiment, the robot was able to classify 25 objects according to object categories such as plastic bottles, objects with contents, pop cans, etc. The accuracy was substantially better than chance, despite the fact that visual feedback was not used.

To further improve category recognition rates, the study presented in this chapter describes a method that scales to a much larger number of exploratory behaviors, sensory modalities, and objects than any previously published experiments in which robots have perceived objects by interacting with them. More specifically, in addition to doubling the number of objects, this study also doubles the number of behaviors and more than triples the number of sensorimotor contexts as compared to our previous work (Sinapov et al., 2011a) (which only focused on object recognition rather than category recognition). In addition, we also show that by using prior information in the form of confusion rates for all categories, the robot can actively select which behavior to apply next when classifying a novel object.

8.3 Experimental Platform

8.3.1 Robot and Sensors

The experiments were performed with the upper-torso humanoid robot shown in Fig. 8.1. The robot has as its actuators two 7-DOF Barrett Whole Arm Manipulators (WAMs), each with an attached 3-finger Barrett Hand. Each WAM has built-in sensors that measure joint angles and torques at 500 Hz. An Audio-Technica U853AW cardioid microphone mounted in the robot's head was used to capture auditory feedback at the standard 16-bit/44.1 kHz

Figure 8.2 The 100 objects explored by the robot, grouped in 20 object categories. From left to right and from top to bottom: 1) containers with different types of contents, 2) plastic bottles, 3) metal objects, 4) containers that vary by weight, 5) egg-coloring cups (vary only by color), 6) pop cans, 7) tin boxes (empty), 8) wicker baskets, 9) foam noodles, 10) medicine pill bottles, 11) pasta boxes (full), 12) big stuffed animals, 13) balls, 14) food cans, 15) cups (vary by material), 16) small stuffed animals, 17) easter eggs (vary by material), 18) styrofoam cones, 19) PVC pipes, and 20) wooden blocks.

resolution and rate over a single channel. The robot's right eye (a Logitech webcam) captured 640 by 480 images that were used for visual feature extraction.

8.3.2 Objects

The robot explored 100 different household objects, which, to the best of our knowledge, is currently the largest number of objects explored by a robot over the course of a single experiment. The 100 objects were selected from 20 object categories, each containing 5 objects that vary along certain dimensions while remaining constant along others. For example, the 5 *PVC pipes* vary by width and weight, but have the same shape, color, and material type. Figure 8.2 shows all objects and object categories that were used in the experiments.

8.3.3 Exploratory Behaviors

The robot was equipped with 10 behaviors: *look*, *grasp*, *lift*, *hold*, *shake*, *drop*, *tap*, *poke*, *push*, and *press*. The *look* behavior consisted of simply taking an RGB snapshot of the object on the table (see Fig. 8.3). All other behaviors were encoded as trajectories in joint-space that

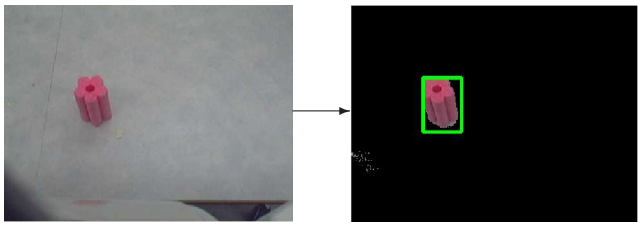


Figure 8.3 Illustration of the visual object detection routine. The position of the bounding box around the object was used by the robot to apply the *grasp* and *tap* behaviors in the correct location (the remaining behaviors either assumed a fixed object position, or the robot was already holding the object). Features for visual object category recognition were extracted from the pixels corresponding to the object as described in Section 8.4.3

were executed using Barrett's default PID controller (see Fig. 8.4). The only exceptions were the *grasp* and *tap* behaviors, which varied depending on the visually detected initial position of the object[1]. It is worth mentioning that the proposed method for learning object categories is independent of how the behaviors are encoded.

8.3.4 Data Collection

The robot interacted with the objects in a series of exploration trials. During each trial, an object was placed on the table by the experimenter and the robot performed all of its 10 exploratory behaviors on the object. The object was then switched with another object from the same category. This was repeated until the robot had explored each object from that category five times. If the objects within a given category could be placed in an order (e.g., by height or by weight), then they were explored in a sequence that is random with respect to the attribute by which they could be sorted. This process was repeated for all twenty categories. In the end, the robot had performed all 10 behaviors 5 times on each of the 100 objects, resulting in $10 \times 5 \times 100 = 5000$ behavior executions.

[1]Visual object detection was performed by estimating a background model of the table when there were no objects placed on it and using this model to fit a bounding box to the largest non-background connected component, which was assumed to be the object. Motor models for the *grasp* and *tap* behaviors were trained by repeatedly placing objects in various positions on the table and demonstrating initial and final joint angles for these behaviors by manually moving the robot's backdrivable arm. To synthesize these behaviors during object exploration, the robot used the three demonstrations closest to the current location of the object to compute average initial and final joint-space positions. The arm was then moved to these positions using the default PID controller.

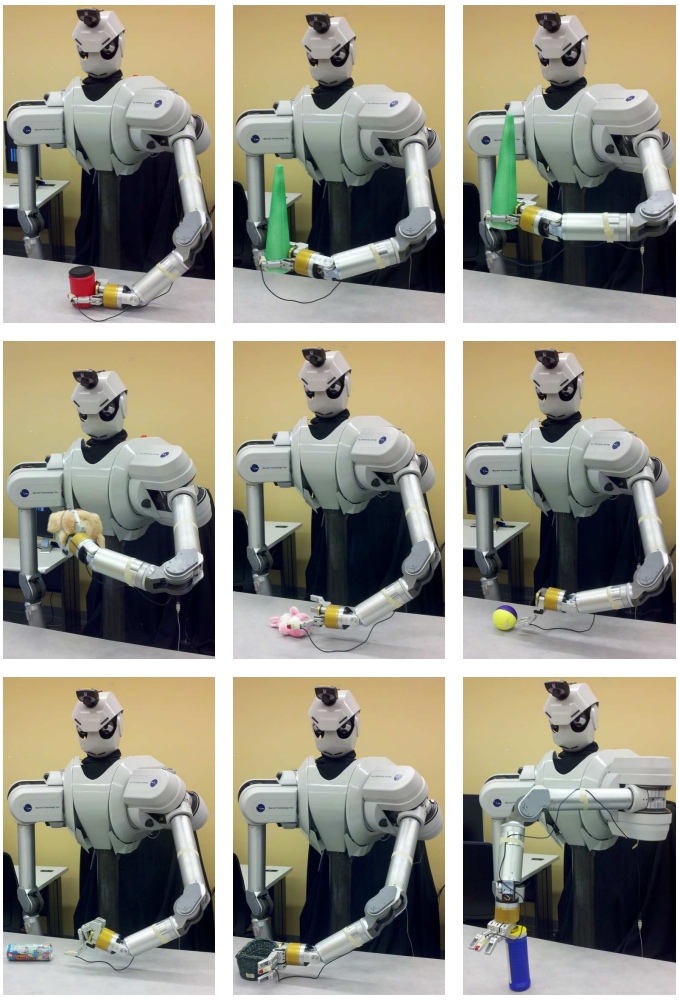


Figure 8.4 The exploratory behaviors that the robot performed on all objects shown in Fig. 8.2. From top to bottom and from left to right: 1) grasp, 2) lift, 3) hold, 4) shake, 5) drop, 6) tap, 7) poke, 8) push, and 9) press. The *look* behavior is described in Fig. 8.3.

While performing each behavior, the robot recorded proprioceptive, auditory, and visual sensory feedback. The next section describes the feature extraction routines that were used to compute features from the recorded sensory input streams.

8.4 Feature Extraction

8.4.1 Proprioceptive Feature Extraction

For each of the nine interactive behaviors shown in Fig. 8.4, proprioceptive features were extracted from the recorded joint torques from all 7 joints of the robot's left arm. The torques

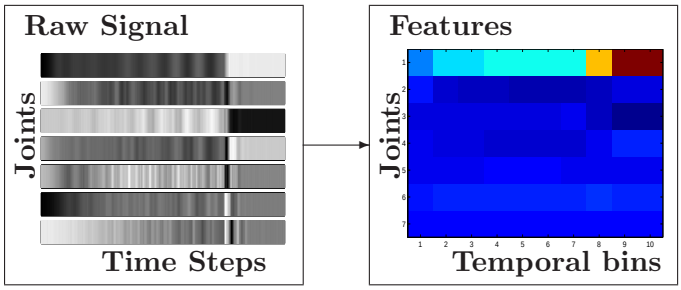


Figure 8.5 Illustration of the proprioceptive feature extraction routine. The input signal is sampled during the execution of a behavior at 500 Hz and consists of the raw torque values for each of the robot's seven joints. Features are extracted by discretizing time (horizontal axis) into 10 temporal bins, resulting in a $7 \times 10 = 70$ dimensional feature vector.

were recorded at 500Hz. The joint-torque record from each interaction was represented as a $\mathbb{R}^{n \times 7}$ vector, where n is the number of temporal samples recorded for each of the 7 joints. Histogram features were extracted from each joint-torque record by discretizing the series of torque values for each joint into 10 temporal bins. This resulted in lower-dimensional datapoints $\mathbf{x} \in \mathbb{R}^{10 \times 7}$, which were subsequently used for the tasks of training and applying the robot's category recognition model. Figure 8.5 shows an example of this feature extraction process.

8.4.2 Auditory Feature Extraction

Auditory features were extracted using the log-normalized Discrete Fourier Transform (DFT), which was computed for each detected sound using $2^7 + 1 = 129$ frequency bins. The SPHINX4 natural language processing library package was used to compute the DFT for each sound (Lee et al., 1990). The DFT encoded the detected intensity for all 129 frequency bins over time, but it was highly-dimensional and thus could not be used directly as an input to the machine learning algorithm. Therefore, given a DFT matrix of a detected sound, a 2D histogram was computed by discretizing time into k_t bins and frequencies into k_f bins. The value for each bin in the histogram was set to the average of the values in the DFT matrix that fell into it. In all experiments, both k_t and k_f were set to 10. Thus, each sound was represented by a feature vector x, where $x \in \mathbb{R}^{10 \times 10}$. Figure 8.6 shows an example of this feature extraction routine.

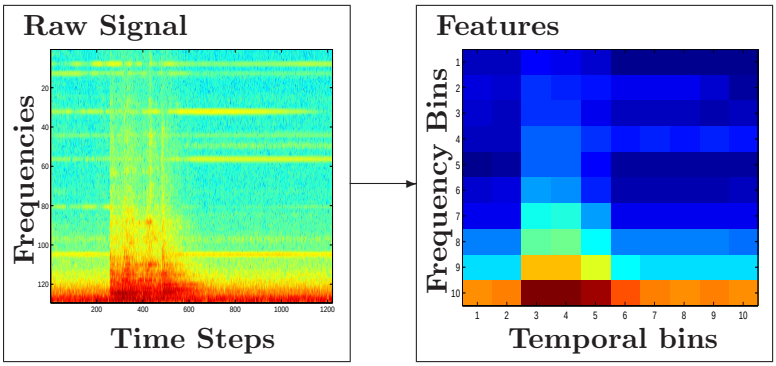


Figure 8.6 Illustration of the auditory feature extraction procedure. The input consists of the discrete Fourier transform spectrogram of the audio wave recorded while a behavior is executed. The spectrogram encodes the intensity of 129 frequency bins and was calculated using a raised cosine window of 25.625ms computed every 10.0ms. To reduce the dimensionality of the signal both the time and the frequencies were discretized into 10 bins, resulting in a $10 \times 10 = 100$ dimensional feature vector.

8.4.3 Visual Feature Extraction

Three types of visual features were extracted from the output of the robot's RGB camera:

8.4.3.1 Color

During the execution of the *look* behavior, the recorded RGB image of the object was used to compute an $8 \times 8 \times 8$ (i.e., 512-dimensional) color histogram in RGB space with uniformly spaced bins. For each image, the object was segmented from the background to ensure that only pixels that correspond to the object are used in the computation of the histogram.

8.4.3.2 Optical Flow

During the execution of all interactive behaviors, the stream of images captured by the robot's camera was used to extract optical flow features. To do so, the dense optical flow was first computed using the algorithm and MATLAB implementation proposed by Sun et al. (2010a). More specifically, given an image from the raw video stream, for each pixel, the algorithm computes a two-dimensional real-valued vector (u, v) encoding the direction of motion (i.e., the vector's angle) as well as the magnitude of the motion (i.e., the vector's norm). The region of interest was set to include the whole image and captured motion produced both by the robot's arm and by the object. Figure 8.7 illustrates this procedure. The optical flow data

is very dense and cannot directly be used as an input to a machine learning algorithm. To overcome this, *weighted angular histogram* features were extracted from the sequence of optical flow images by binning the angles into 10 equally spaced bins. More specifically, the norms of all optical flow vectors with angles ranging from 0 to $2\pi/10$ are added to bin number 1, the norms of all vectors with angles in the range of $2\pi/10$ to $2 \times 2\pi/10$ are added to bin number 2 and so forth.

8.4.3.3 SURF

The Speeded-Up Robust Features (SURF) proposed by Bay et al. (2008) were computed for all images captured by the robot's camera during the execution of each of the 10 behaviors. Figure 8.7 shows the detected SURF interest points for several images over the course of executing the *poke* behavior.

For the *look* behavior, the region of interest was set to the bounding box containing the segmented object from the background. For the remaining 9 behaviors, the SURF features were computed over a region of interest covering the entire table. Each SURF descriptor was represented as a 128-dimensional feature vector encoding the distribution of the first order Haar wavelet responses within the interest point neighborhood.

The detected SURF descriptors were quantized using the X-Means algorithm, an extension of k-Means that attempts to estimate the number of clusters using the Bayesian Information Criterion (see Pelleg and Moore (2000) for details). The quantization was learned using only 0.5% (or approximately 35,000)of the feature descriptors detected from all individual images captured by the robot's camera. The X-Means algorithm found 200 clusters that were interpreted as a dictionary of visual "words". Given a set of SURF descriptors detected over the course of executing a behavior on an object, a 200-dimensional feature vector was computed encoding a histogram of the SURF descriptors over the words in the dictionary.[2]

[2]Experiments were also conducted with larger visual word dictionaries, but no benefit to classification performance was observed.

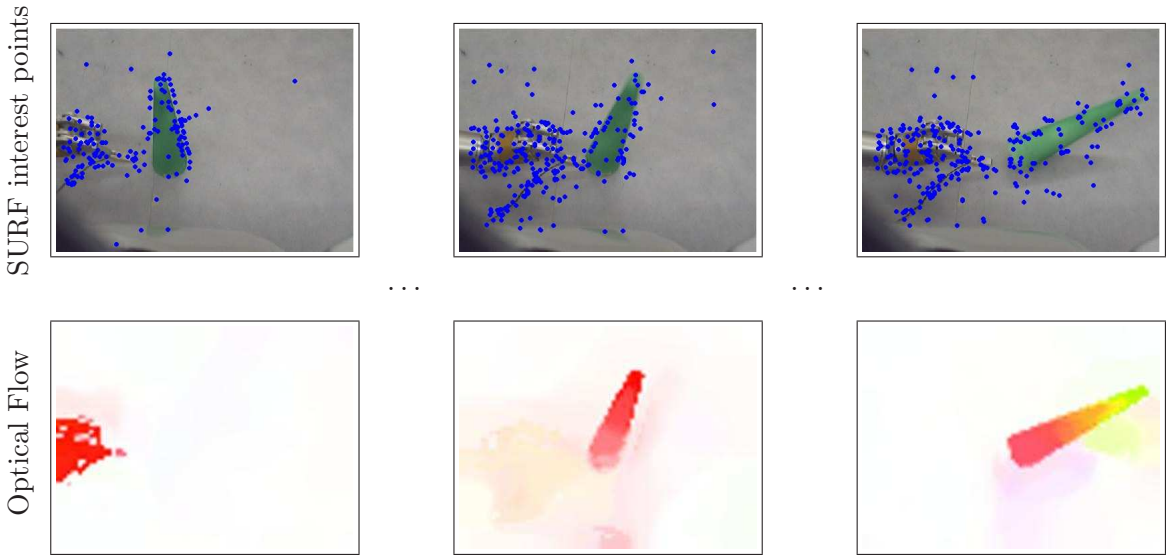


Figure 8.7 Illustration of the SURF features and the optical flow detected through the robot's camera during the execution of the *poke* behavior on one of the objects from the *styrofoam cones* category. The left column shows the raw camera images with the detected SURF interest points, while the right column shows the corresponding optical flow images. For each pixel in the optical flow images, the hue encodes the angle of the optical flow vector (u, v) for that pixel, while the intensity encodes the vector's norm.

8.4.4 Hand Proprioception Feature Extraction

The final configuration of the fingers at the end of the *grasp* behavior was also recorded. This resulted in a 3-dimensional feature vector, where each value indicates the end joint position for each of the three fingers of the Barrett Hand (BH-260). The final position of each finger was always in the range of 0 (fully open) to 20000 (fully closed). The spread of the fingers (joint number 4) was held fixed during the execution of each grasp.

8.4.5 Summary

To summarize, the robot perceived the objects using 6 different types of features: 1) auditory, 2) proprioceptive (arm), 3) proprioceptive (hand), 4) color, 5) optical flow, and 6) SURF. The auditory, proprioceptive and optical flow features were extracted from the robot's sensorimotor data recorded while performing each of the 9 interactive behaviors on the objects. Color features, on the other hand, were extracted from the static images of the object taken

Table 8.1 The 39 Sensorimotor Contexts used by the Robot

	Audio	Proprioception		Vision		
	Discrete Fourier Transform	Joint Torques	Finger Positions	Optical Flow	SURF Points	Color Histogram
look					X	X
grasp	X	X	X	X	X	
lift	X	X		X	X	
hold	X	X		X	X	
shake	X	X		X	X	
drop	X	X		X	X	
tap	X	X		X	X	
push	X	X		X	X	
poke	X	X		X	X	
press	X	X		X	X	

by the robot's camera during the execution of the *look* behavior. Finally, SURF features were extracted from both static images captured during the *look* behavior as well as the image sequences from the remaining 9 behaviors. The next section describes how these features are used for recognizing the category of an object.

8.5 Theoretical Model

8.5.1 Notation

Let $\mathcal{B}$ be the set of exploratory behaviors and let $\mathcal{C}$ be the set of sensorimotor contexts such that each context $c \in \mathcal{C}$ refers to a combination of a behavior and a sensory modality (e.g., *drop-audio*, *look-color*, etc.). In our case, 9 behaviors (all except *look*) produced 3 types of feedback: auditory, optical flow, and proprioceptive feedback from the robot's arm. SURF features were extracted during all 10 behaviors. In addition, color features were extracted during the *look* behavior. Finally, the *grasp* behavior also produced proprioceptive feedback from the robot's hand. Thus, the total number of sensorimotor contexts in our experiments was $9 \times 3 + 10 + 1 + 1 = 39$. In other words, $|\mathcal{C}|$ =39. The 39 sensorimotor contexts are visualized in Table 8.1, where each context corresponds to a combination of a behavior and sensory signal.

Let $\mathcal{O}$ be the set of all 100 objects. During the data collection, the robot was repeatedly presented with an object $o \in \mathcal{O}$ and subsequently applied all of its exploratory behaviors on the object, which constituted one trial. Thus, during the i^{th} exploration trial, the robot observed features x_i^c for each behavior-modality context c. The following subsections describe how these features can be used to solve the object category recognition task.

8.5.2 Problem Formulation

Each object in our dataset was labeled as belonging to one of the 20 categories shown in Figure 8.2. Let the function $label(o) \rightarrow y$ be a labeling function that outputs a label $y \in \mathcal{Y}$ given an object o, where $\mathcal{Y}$ is the full set of 20 category labels ($|\mathcal{Y}| = 20$). The task of the robot is to learn a category recognition model that outputs the correct category label y, given sensory feedback signals detected while interacting with object o using a set of behaviors $\mathcal{B}$.

8.5.3 Category Recognition Model

To solve this problem, for each sensorimotor context $c \in \mathcal{C}$, a category recognition model M_c is trained on input datapoints of the form $[x_i^c, y]$ where x_i^c is a feature vector detected in context c during trial i, while exploring an object with label y. The recognition model is tasked with estimating the category label probability for each class label, i.e., $Pr(\hat{y} = y|x_i^c)$ for all labels $y \in \mathcal{Y}$. In this work, two different machine learning algorithms were evaluated: k-Nearest Neighbors (k-NN) and Support Vector Machine (SVM).

8.5.3.1 K-Nearest Neighbor

The first algorithm, k-Nearest Neighbors (k-NN), falls within the family of *lazy learning* or *memory-based learning* algorithms (Aha et al., 1991; Atkeson et al., 1997) and does not build an explicit model of the data. Instead, it simply stores all data points and their category labels and only uses them when the model is queried to label a test data point.

To label a test data point, k-NN finds its k closest neighbors in the training set. The Euclidean distance function (i.e., L2-norm) was used to calculate the distances between the test data point and the training samples when computing the set of k closest neighbors. The

parameter k was heuristically set to 3. Probability estimates were computed by counting the category labels of the 3-neighbors. For example, if two of those neighbors have a class label "ball", then $Pr(\hat{y} = ball) = 2/3$. All experiments were conducted using the implementation of k-NN included in the WEKA machine learning library (Witten and Frank, 2005).

8.5.3.2 Support Vector Machine

The second machine learning algorithm, Support Vector Machine (SVM), falls in the family of *discriminative* models (Vapnik, 1998). Let $(\mathbf{x}_i, y_i)_{i=1,\ldots,l}$ be a set of labeled inputs, where $\mathbf{x}_i \in \mathbb{R}^n$ and $y_i \in \{-1, +1\}$ (i.e., a binary classification problem). The goal of the SVM algorithm is to learn a linear function $f(\mathbf{x}) =< \mathbf{x}, \mathbf{w} > +b$, $\mathbf{w} \in \mathbb{R}^n$ and $b \in \mathbb{R}$, that can accurately classify test data points. To do this, the SVM algorithm solves a dual quadratic optimization problem, in which $\mathbf{w}$ and b are optimized so that the margin of separation between the two classes is maximized (Vapnik, 1998).

A good linear decision function $f(\mathbf{x})$ in the n-dimensional input space, however, does not always exist and therefore the labeled inputs are typically mapped into a (possibly) higher-dimensional feature space, e.g., $\mathbf{x}_i \rightarrow \Phi(\mathbf{x}_i)$, where a good linear decision function can be found. The mapping can be defined implicitly with a kernel function $K(\mathbf{x}_i, \mathbf{x}_j) =< \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) >$ that replaces the dot product $< \mathbf{x}_i, \mathbf{x}_j >$ in the dual quadratic optimization problem (see Vapnik (1998); Burges (1998) for details). Intuitively, the kernel function can be interpreted as a measure of similarity between two data points.

In this work, several kernel functions were used. The first is the polynomial kernel function. Given two input feature vectors $\mathbf{x}_i$ and $\mathbf{x}_j \in \mathbb{R}^n$, the polynomial kernel function is defined as:

$$K_{poly}(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^{\mathbf{T}} \mathbf{x}_j + 1.0)^p.$$

While the polynomial kernel function is one of the most commonly used ones in the literature, it is not appropriate for all types of data. Thus, two other kernel functions were also used, one designed to work on data points encoding a histogram (Chapelle et al., 1999) and another designed to handle data points that represent matrices rather than flat feature vectors (Zhou, 2004). Let $\mathbf{x}_i$ and $\mathbf{x}_j$ be two histograms such that $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{N}_0^n$, where $\mathbb{N}_0$ is the set of

all non-negative integers. To handle histogram inputs, Chapelle et al. (1999) propose the use of a non-Gaussian RBF kernel function:

$$K_{hist}(\mathbf{x}_i, \mathbf{x}_j) = e^{-\rho d_{a,b}(\mathbf{x}_i, \mathbf{x}_j)}$$

where

$$d_{a,b}(\mathbf{x}_i, \mathbf{x}_j) = \sum_k |x_{ik}^a - x_{jk}^a|^b.$$

If $a = 1$ and $b = 2$, this function corresponds to the commonly-used Gaussian RBF kernel. As Chapelle et al. (1999) note, lowering b amounts to assuming that the data are generated by a mixture of distributions that are heavy-tailed when compared to the Gaussian distribution. Based on the experiments described by Chapelle et al. (1999), in this work the parameters a and b were set to 1.0 and 0.5 respectively, while ρ was set to 0.1 (similar classification performance was observed as long as ρ was between 0.005 and 0.25). The K_{hist} kernel function was used by the SVMs trained on optical flow histogram features, the SVMs trained on SURF histogram features, as well as the SVM trained to recognize the category of an object using its color histogram features.

Finally, since the auditory features correspond to a matrix (see Figure 8.6), the auditory SVMs were trained using the trace kernel function designed to handle matrices (Zhou, 2004). Given two $n \times m$ matrices $\mathbf{X}_i$ and $\mathbf{X}_j$, the trace kernel function can be defined as:

$$K_{trace}(\mathbf{X}_i, \mathbf{X}_j) = tr(\mathbf{X}_i^{\mathbf{T}} \mathbf{X}_j)^p.$$

In summary, three different kernel functions were used in this work: K_{poly}, K_{hist}, and K_{trace}. The SVMs trained on optical flow angular histogram features and the SVM trained on color histogram features all used the K_{hist} kernel function. The SVMs trained on auditory features used the K_{trace} kernel functions. All other SVMs used the polynomial function, K_{poly}. The exponent p in K_{trace} and K_{poly} was set to 2.0.

To generalize the binary SVM classifier to the multi-class problem of category recognition, the pair-wise coupling method proposed by Hastie and Tibshirani (1998) was applied in this work. Finally, to obtain probabilistic estimates from the SVM classifiers, Logistic regression

models were fit to the outputs of the SVMs as described by Witten and Frank (2005). The next subsection describes how the outputs from the context-specific category recognition classifiers were combined.

8.5.4 Combining Model Outputs

The outputs of several context-specific category recognition models can be combined in order to achieve better performance. The robot's experience with a given object o in multiple sensorimotor contexts during trial i can be represented by the set of features $\mathcal{X}_i = \{x_i^{c_1}, \ldots, x_i^{c_N}\}$, where each feature corresponds to the detected signal from a unique behavior-modality combination and N is the number of sensorimotor contexts ($N \leq |\mathcal{C}|$). The outputs of the individual models can be combined using the uniform combination rule:

$$Pr(\hat{y} = y|\mathcal{X}_i) = \alpha \sum_{x_i^c \in \mathcal{X}_i} Pr(\hat{y} = y|x_i^c),$$

where α is a normalization constant ensuring that the probabilities sum up to 1.0. By varying the number of elements in the input set $\mathcal{X}_i$, this formulation allows us to evaluate how the category recognition performance improves as the robot uses multiple sources of information.[3]

8.5.5 Active Behavior Selection

In practice, it would be highly desirable for a robot to minimize its object exploration time when classifying new objects. To address this challenge, the model in this work selected which behaviors to apply next based on prior information in the form of the confusion matrices associated with each behavior. More specifically, for a given behavior $b \in \mathcal{B}$, let $\mathbf{C}^b \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{Y}|}$ be a confusion matrix such that each entry C_{ij}^b encodes how many times an object from category y_i was classified as belonging to category y_j.

Given a probabilistic estimate for an object's category, the confusion matrices associated with the robot's behaviors can be used to guide subsequent exploration. For example, suppose

[3]Other combination rules that were explored include the product combination rule, a weighted combination rule, a majority vote rule, as well as a meta-learning approach in which the outputs of the individual classifiers were fed as input to a meta-learning classifier. The classification performance of these other rules was either nearly identical or slightly inferior to the rule used in this work.For a detailed review of different classifier combination schemes, see (Lam and Suen, 1995; Lam, 2000).

that after performing the *look* behavior, the robot's estimates for the object's category labels are $Pr(\hat{y} = \text{“}egg\text{”}) = 0.6$, $Pr(\hat{y} = \text{“}ball\text{”}) = 0.4$ and 0 for all others. Given this information, it may be possible to speed up exploration time if the next behavior that the robot chooses to apply is the one that confuses the "egg" and "ball" categories the least.

More specifically, for an exploratory behavior $b \in \mathcal{B}$, let $Pr_b(\hat{y} = y_i | y = y_j)$ be the probability of mis-classifying an object from category y_j as an object from category y_i when applying behavior b. Thus, the degree of confusion between categories y_i and y_j for behavior b can be defined as:

$$C_{ij}^b = \frac{Pr(\hat{y} = y_i | y = y_j) + Pr(\hat{y} = y_j | y = y_i)}{2}.$$

The estimates for the confusion between categories are used by the robot to guide exploration as follows. Let $\hat{\mathbf{p}} \in \mathbb{R}^{|\mathcal{Y}|}$ be the robot's current probabilistic estimate for the object's category labels such that $\hat{p}_i$ is the probability that the object's category is y_i. Let $\mathcal{B}_r$ be the remaining set of behaviors to choose from (i.e., the behaviors not performed so far on the test object). In this setting, the next behavior to be applied is selected using the following procedure:

1. Compute the set $\mathcal{Y}_K \subset \mathcal{Y}$ such that it contains the K most likely object categories according to $\hat{\mathbf{p}}$.

2. Pick the next behavior b_{next} with an associated confusion matrix that is least likely to confuse the categories within the set $\mathcal{Y}_K$, i.e.,

$$b_{next} = \arg\min_{b \in \mathcal{B}_r} \sum_{y_i \in \mathcal{Y}_K} \sum_{y_j \in \mathcal{Y}_K / y_i} C_{ij}^b.$$

3. Update the estimate $\hat{\mathbf{p}}$ using the classifiers associated with the sensorimotor contexts of b_{next}.

4. Remove b_{next} from $\mathcal{B}_r$. If $|\mathcal{B}_r| \geq 1$, go back to step 1).

Rather than setting a static value for the threshold K, this value is determined on-line given the current estimate $\hat{\mathbf{p}}$ such that the likelihoods of the K most likely categories sum up to at

least ω. For example, if there are only three categories, A, B, and C, with likelihood estimates 0.5, 0.4, and 0.1 respectively, and $\omega = 0.65$, then only the first two, A and B, will be included in $\mathcal{Y}_K$ since they are the two most likely categories and $0.5 + 0.4 > 0.65$. In our experiments, the value for the threshold ω was set to 0.65. The results remained similar provided that ω was between 0.5 and 0.8, with performance diminishing outside that range.

8.5.6 Detecting Outlier Categories

One limitation of the theoretical model presented so far is that it cannot handle objects that do not belong to any of the categories specified during training. This is an important problem because a robot operating in a human environment is guaranteed to encounter an object from a category that it has never been exposed to before. To handle such situations, this section describes a method that can enable the model to detect whether an object belongs to a known category or not.

The problem can be formulated as follows. Let o_{test} be a test object whose category label is unknown (it may be either from a known category or from an unfamiliar category). Let $\hat{y} \in \mathcal{Y}$ be the estimated category assigned to the object by the trained category recognition model. Finally, let the set $O_{\hat{y}} = \{o_1^{\hat{y}}, \ldots, o_n^{\hat{y}}\}$ contain the known objects from category $\hat{y}$. Given the object o_{test} and the set $O_{\hat{y}}$, the task is to detect whether or not o_{test} is from a novel category or not. In the machine learning literature, this problem is known as *outlier detection* (see Hodge and Austin (2004) for a review). While there are many approaches to this problem, most typically assume a flat feature vector representation for the data, as well as large amounts of data points. Therefore, in this work, the method for detecting the presence of novel categories is based on an approach, described by Sinapov and Stoytchev (2010b) (and also in Chapter 6), that was specifically designed to deal with a small number of objects that have been physically explored by a robot.

The original method can be summarized as follows. Let $\mathbf{W} \in \mathbb{R}^{N \times N}$ be an affinity matrix encoding the similarity relations among a set $\mathcal{D}$ of N objects (i.e., data points). The outlier object is then selected as the object o_i that maximizes the following objective function:

$$q(\mathcal{D}, o_i) = \alpha_1 \sum_{j \in \mathcal{D}/o_i} \sum_{k \in \mathcal{D}/o_i} W_{jk} - \alpha_2 \sum_{j \in \mathcal{D}/o_i} W_{ij}.$$

The first term captures the pairwise similarity between the remaining objects in $\mathcal{D}$ (i.e., after i is removed from $\mathcal{D}$) while the second term captures the similarity between the selected object i and the remaining $|\mathcal{D}| - 1$ objects in $\mathcal{D}$. The constants α_1 and α_2 are normalizing weights, which ensure that the function is not biased towards either one of the two terms. Thus, the weights were set to:

$$\alpha_1 = \frac{1}{(|\mathcal{D}| - 1) \times (|\mathcal{D}| - 1)}, \; \alpha_2 = \frac{1}{|\mathcal{D}| - 1}.$$

As reported by Sinapov and Stoytchev (2010b), given a set of physical objects explored by the robot, the proposed method is useful for detecting the object in the set that does not belong to the category. For example, given 3 pop cans and 1 hat, and a matrix encoding the similarity between the objects as measured by the sensorimotor features detected with the objects, the hat is selected as the odd object.

Given the object o_{test}, its estimated category label $\hat{y}$, the set of objects $O_{\hat{y}}$, and a similarity matrix $\mathbf{W}_c$ associated with sensorimotor context $c \in \mathcal{C}$, the method described in Chapter 6 is adapted for outlier category detection using the following procedure:

- Let $o_{odd} = \arg\max_i q(O_{\hat{y}} \cup \{o_{test}\}, o_i)$.
- If $o_{odd} \neq o_{test}$, then classify the object o_{test} as belonging to the *familiar* category $\hat{y}$.
- If $o_{odd} = o_{test}$ and $q(O_{\hat{y}} \cup \{o_{test}\}, o_{test}) > \epsilon^c_{\hat{y}}$, then classify object o_{test} as belonging to a *novel* category. Else, classify o_{test} as an object from a known category, i.e., accept the estimated category label $\hat{y}$.

The threshold $\epsilon^c_{\hat{y}}$ is a parameter specific to the category $\hat{y}$ and context c, and is estimated from the available training data. This is done by repeatedly running the odd-one-out task on all groups of objects from the same category in the training set and recording the highest observed outlier score, $q^{c,\hat{y}}_{max}$. Thus, $\epsilon^c_{\hat{y}}$ is set to $r \times q^{c,\hat{y}}_{max}$, where $r \in \mathbb{R}$. For example, when $r = 1.0$, for an object to be considered an outlier, it has to have a higher odd-one-out score than the highest observed score for an object that belongs to the category.

Each entry in the matrices $\mathbf{W}_c$ is computed by estimating the expected similarity between the feature vectors detected with each pair of objects in context c. Given two feature vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ from the same context c, the similarity function that was used can be expressed as $e^{-\rho d_{L2}(\mathbf{x}_i,\mathbf{x}_j)}$ where d_{L2} is the L2 norm. The parameter ρ was heuristically set to 0.1, which is the same value as the one used in the definition of the SVM kernel function K_{hist} defined in Section 8.5.3.2.

The procedure for detecting the presence of an unfamiliar category takes as input just one similarity matrix $\mathbf{W}_c$, tied to a specific sensorimotor context. To use multiple sensorimotor contexts, the procedure is applied with several different matrices (one per sensorimotor context) and if more than half of the time the object o_{test} is detected as one from a novel category, then it is classified as such. In the experiments described in the next section, nine contexts were used for this task. This set of contexts was selected such that for each estimated category label $\hat{y}$, it contained the nine best contexts for recognizing category $\hat{y}$, as estimated by performing cross-validation on the training data.

8.5.7 Evaluation

8.5.7.1 Category Recognition

The robot's category recognition models were evaluated using *object-based cross-validation* as follows. During each round of evaluation, the robot's context specific models were trained on data from 4 objects from each category (a total of 80 objects) and evaluated on data from the remaining 20 objects. This process was repeated five times, such that each object was included four times in the training set and once in the testing set. Since the robot explored each object over 5 trials, during the training stage each context-specific classifier was trained on $80 \times 5 = 400$ data points and evaluated on the remaining $20 \times 5 = 100$. For the purposes of this evaluation, outlier category detection was turned off to ensure that the classifiers are trained and tested using all available datapoints. Two metrics were used to quantify the category recognition performance. The first metric was accuracy, defined as:

$$\%\ \text{Accuracy} = \frac{\#\ \text{correct classifications}}{\#\ \text{total classifications}} \times 100.$$

The second metric was the *f-Measure*, which is defined as the harmonic mean between the precision and recall for a given category label. It can be computed as follows:

$$\text{f-Measure} = 2 \times \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}.$$

The f-Measure is always in the range of 0.0 to 1.0. For a given category, a high-value of the f-Measure indicates that the category is easy to recognize, while a low value shows that the category is difficult to recognize.

In addition to evaluating the performance of the individual classifiers, the model's accuracy rates were also computed as the number of sensorimotor contexts available to the robot was varied from 1 to 39, and as the number of behaviors applied on the test object was varied from 1 to 10. For the latter case, both the random and the active behavior selection strategy were evaluated.

8.5.7.2 Outlier Category Detection

To evaluate the method for detecting the presence of novel categories, the initial set of categories was split into two groups of 10. The robot's category recognition models were subsequently trained with 4 out of 5 objects with the known category labels. In other words, the test set in this case contains data from 5 novel objects from each of the 10 novel categories as well as data from 1 novel object for each of the 10 familiar categories. The estimated category label for each object in the test set was computed using the trained category recognition model. Subsequently, the procedure described in Section 8.5.6 was used to decide whether to accept the category label or classify the object as one belonging to a category that was not present in the training set.

This test was repeated 20 times with different random seeds that determine how the set of categories is split into two groups. The results are reported in terms of *true positive rate*, i.e., the proportion of objects from novel categories that are classified as such, and *false positive rate*, i.e., the proportion of objects from familiar categories that are mistakenly classified as novel.

Table 8.2 Category Recognition Accuracy (%) using the 'Look' Behavior

	Color Histogram	SURF	All
k-NN	47.3	33.7	50.7
SVM	58.9	58.8	67.7

8.6 Results

8.6.1 Category Recognition using a Single Behavior

The first experiment evaluated the performance of the robot's recognition models for each of the 39 possible sensorimotor contexts. Tables 8.2 and 8.3 show the accuracy rates for every viable combination of behavior and sensory modality.[4] The results show that nearly every sensorimotor context contains information useful for category recognition. For comparison, a model that randomly assigns an object category label is expected to achieve only 5.0% accuracy as the number of object categories is 20. On average, SVM performs substantially better than k-NN for most sensorimotor contexts.

As expected, certain behaviors work better with certain modalities. For example, the proprioceptive features detected during the *lift* behavior are more useful for object category recognition than the auditory features produced by the same object. One unexpected result is that auditory features produced by relatively silent behaviors such as *lift* and *hold* produce recognition accuracies better than chance. One possible explanation is that certain objects with contents inside of them (e.g., pasta boxes) still produce some auditory feedback that is indicative of the object's category. In addition, the sounds produced by the robot's motors while lifting and holding objects depend on the weight of the objects (i.e., heavier objects require larger torques). Another important result is that the SURF features detected over the course of manipulating the object are more useful for recognition than the features detected from the static *look* behavior. One possible explanation is that when performing a behavior, the object is observed from more than just one side and for a longer time frame, indicating

[4]For the *grasp* behavior and the proprioceptive sensory modality, the outputs of the *arm* and *hand* proprioceptive recognition models were combined and the resulting accuracy is reported. Individually, the arm proprioceptive model achieved accuracy of 36.27%, while the hand proprioceptive model achieved 21.84% when using the k-NN algorithms. With SVM, the rates were 36.7% and 21.5%, respectively.

Table 8.3 Category Recognition Accuracy(%) using a Single Behavior

	Behavior	Audio	Proprio-ception	Optical Flow	SURF	All
k-NN	grasp	30.9	38.9	13.6	48.3	64.0
	lift	34.1	37.1	5.0	54.3	62.4
	hold	20.4	24.5	5.0	39.5	43.6
	shake	42.7	39.1	25.0	69.3	71.2
	drop	45.7	18.8	16.0	40.5	59.0
	tap	51.9	29.1	20.4	61.9	72.2
	push	64.2	58.6	22.8	65.0	84.8
	poke	48.5	53.1	18.8	57.7	76.0
	press	46.7	66.1	24.0	59.7	69.6
SVM	grasp	45.7	38.7	12.2	57.1	65.2
	lift	48.1	63.7	5.0	65.9	79.0
	hold	30.2	43.9	5.0	58.1	67.0
	shake	49.3	57.7	32.8	75.6	76.8
	drop	47.9	34.9	17.2	57.9	71.0
	tap	63.3	50.7	26.0	77.3	82.4
	push	72.8	69.6	26.4	76.8	88.8
	poke	65.9	63.9	17.8	74.7	85.4
	press	62.7	69.7	32.4	69.7	77.4

that even if a robot uses only vision-based sensors to perceive objects, active interaction with them can still further improve the classification accuracy.

To visualize the errors made by the robot's collection of recognition models, the 39 confusion matrices associated with the 39 sensorimotor contexts were summed up, producing the matrix $\mathbf{M} \in \mathbb{Z}^{20\times 20}$ in which each entry M_{ij} encodes how many times category i was confused with category j. A second, symmetric matrix $\mathbf{M}^{sym}$ was then computed such that $M_{ij}^{sym} = M_{ij} + M_{ji}$. The matrix $\mathbf{M}^{sym}$ was then used to produce a taxonomy of the categories by recursively applying the normalized-cut algorithm (Shi and Malik, 2000). The result is shown in Figure 8.8. Categories that are likely to be confused by at least some of the classifiers in the ensemble are close within the taxonomy while categories that are easy to distinguish are further apart. While the taxonomy is not expected to match how a human would organize the categories, it still shows how perceptually similar they are from the robot's point of view.

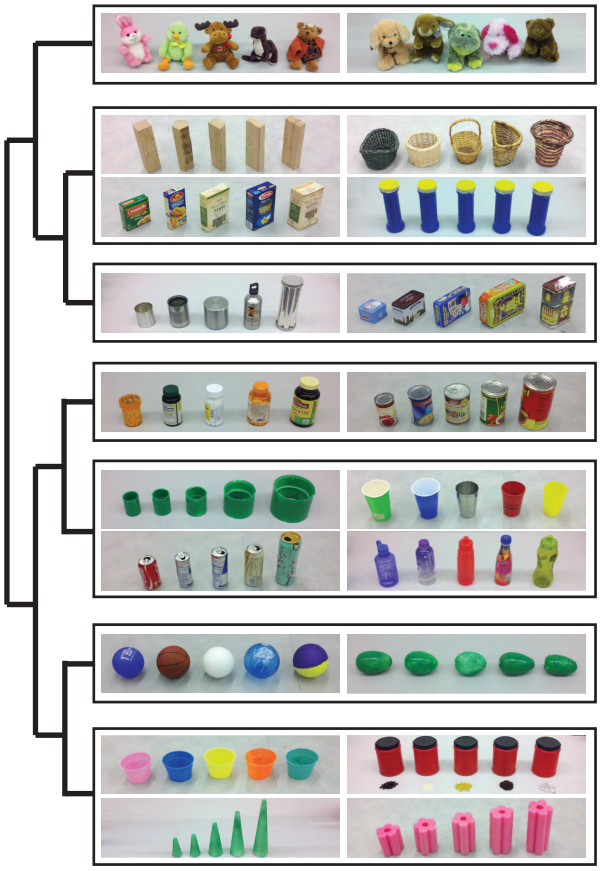


Figure 8.8 A hierarchical clustering of the 20 categories based on the confusion matrix encoding how often each pair of categories is confused by the robot's context-specific category recognition models.

8.6.2 Category Recognition from Multiple Sensorimotor Contexts

The next experiment evaluated whether the robot's category recognition performance could be improved by combining the outputs of individual recognition models trained on data from specific behavior-modality combinations. As before, the models were trained with known labels for 4 out of the 5 objects in each category and evaluated on the remaining set. In this case, however, the evaluation was performed by varying the number of sensorimotor contexts that were used for classifying a novel object from 1 to 39 (see Section 8.5.4 for details on how the outputs from multiple context-specific recognition models are combined). Due to the large number of tests that need to be performed for this experiment, only k-NN was evaluated with a variable number of sensorimotor contexts available to the robot.

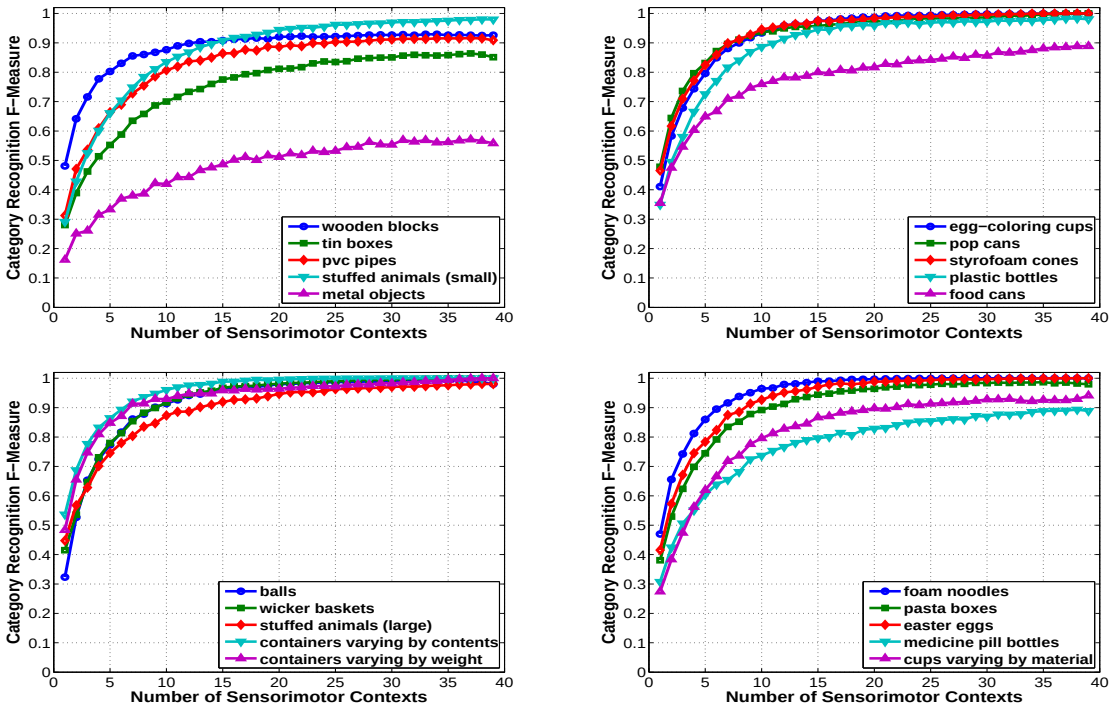


Figure 8.9 Category recognition rates as a function of the number of sensorimotor contexts from which features are extracted. The results of this experiment show that the f-measure increases dramatically as the robot experiences the objects using more behaviors and more sensory modalities.

Figure 8.9 shows the categorization performance for each of the 20 object categories as the number of contexts is varied from 1 to 39. As the robot is allowed to experience objects in more sensorimotor contexts its ability to classify them into categories increases. Most object categories (14 out of 20) can be recognized almost perfectly (i.e., f-measure greater than 0.9) if all sources of information are used. When all 39 sensorimotor contexts are used, k-NN achieved 94.6% category recognition accuracy. The SVM algorithm was also evaluated when using all 39 contexts, resulting in 97% accuracy.

Table 8.4 shows the specific precision and recall rates for all 20 categories when using all 39 contexts. The object category that was most difficult to recognize was the *metal objects* category, for which the f-measure was only 0.57. Objects from this category were most often mis-classified as belonging to the *tin boxes* category, which was likely due to the fact that both of these categories consisted of objects that were made of metal. This illustrates that for a large set of objects it may be difficult to specify perfectly disjoint category assignments. In

Table 8.4 Precision and recall rates for all 20 categories using all sensorimotor contexts.

	k-NN		SVM	
Category	Precision	Recall	Precision	Recall
wicker baskets	1.0	1.0	1.0	1.0
containers (vary by weight)	0.93	1.0	1.0	1.0
small stuffed animals	1.0	0.96	1.0	0.96
large stuffed animals	0.96	1.0	0.96	1.0
metal objects	0.67	0.48	0.64	0.76
wooden blocks	0.86	1.0	0.96	1.0
pasta boxes	1.0	0.96	1.0	1.0
tin boxes	0.91	0.8	0.77	0.96
PVC pipes	0.82	1.0	1.0	0.96
cups	0.89	0.96	1.0	1.0
pop cans	1.0	1.0	1.0	1.0
plastic bottles	0.96	1.0	1.0	1.0
food cans	1.0	0.8	0.96	0.92
medicine pill bottles	0.83	0.96	0.89	1.0
containers (vary by contents)	1.0	1.0	1.0	1.0
styrofoam cones	1.0	1.0	1.0	1.0
foam noodles	1.0	1.0	1.0	1.0
egg-coloring cups	1.0	1.0	1.0	1.0
easter eggs	1.0	1.0	1.0	1.0
balls	1.0	1.0	1.0	1.0

future work, we plan to address this by devising a category recognition method that can handle objects that may belong to multiple categories.

8.6.3 Identifying Task-Relevant Sensorimotor Contexts

The previous experiment showed that the robot can improve its category recognition performance by using information from all available sensorimotor contexts as opposed to just one. Nevertheless, this may not result in optimal recognition rates as certain contexts may produce features that are irrelevant for a given object category, thus making the learning task more difficult. To address this issue, in the next set of experiments the robot was tasked with estimating the most useful sensorimotor contexts for recognizing a given category.

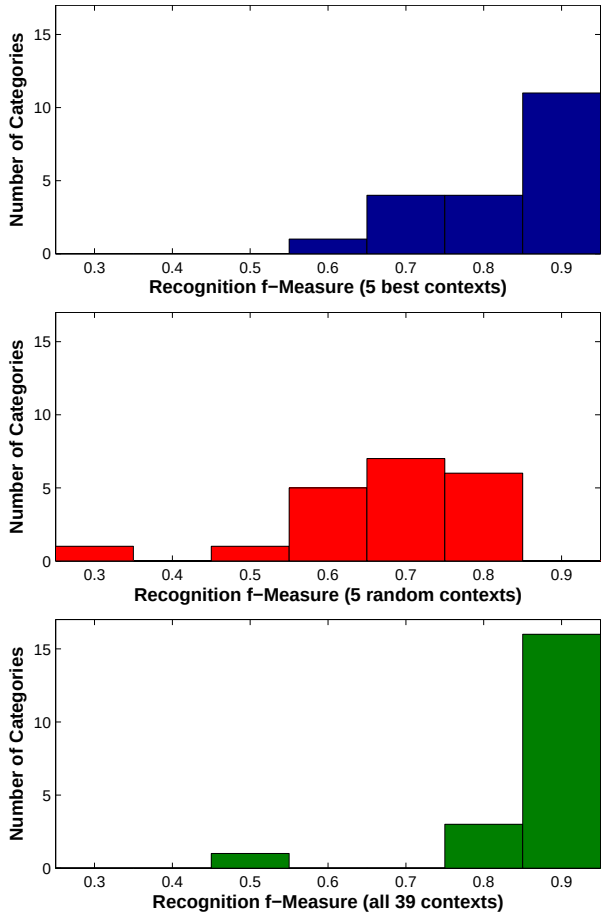


Figure 8.10 Histograms of individual f-Measures per object category under three different conditions: (top) when using the 5 best contexts for each category; (middle) when using 5 random contexts; and (bottom) when using all 39 sensorimotor contexts. The results show that by identifying which 5 sensorimotor contexts work best for a given category the robot's model can improve its recognition when compared to any random combination of the same number of contexts.

To do so, during the training stage, the model performed internal cross-validation on the training data for each possible context-category combination, and the resulting f-Measure was recorded. At test time, for each category, the three contexts with the highest f-Measures were used for detecting whether a novel object was a member of that category or not. Note that the set of 5 best contexts for each category is not necessarily the best *combination* of five contexts.

Figure 8.10 shows histograms of the category recognition rates (f-Measure) for three different conditions: 1) using the 5 best sensorimotor contexts (top); 2) using 5 random sensorimotor contexts (middle); and 3) using all 39 sensorimotor contexts (bottom). The results show that the robot is able to identify a group of five task-relevant sensorimotor contexts that can be used

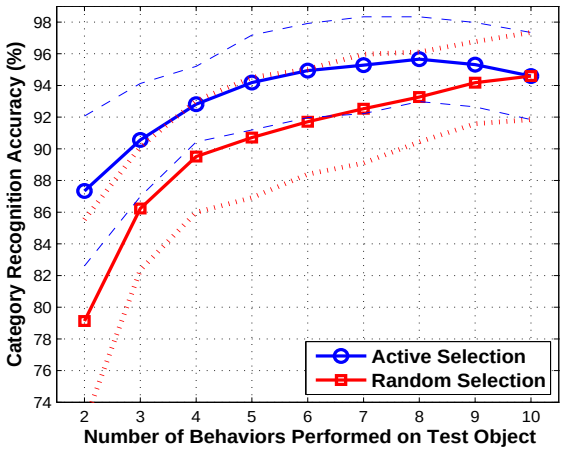


Figure 8.11 Category recognition rates with k-NN classifier as a function of the number of behaviors applied on the test object under two different conditions: random behavior selection and active behavior selection (see Section 8.5.5). For each condition, the evaluation was performed using 5 different train-test splits. For each of the five splits, the evaluation was performed using each of the 10 behaviors as an initial state. Thus, the means and the standard deviations were computed from samples of size 50.

to detect specific categories with performance comparable to that of using all 39 sensorimotor contexts. In other words, for each category, there exists a set of 5 contexts for which the performance is close to that achieved when using all sensorimotor contexts. Thus, if the robot is tasked with finding objects from a specific category, it could do this more efficiently by only applying the behaviors that are included in these 5 sensorimotor contexts.

It is important to note that the best sensorimotor features will be different for different categories. For example, the best sensorimotor context for the *blue containers* category was the *look-color* behavior-modality combination since the objects in that category vary by weight but are identical in color. The same combination, however, was not very useful for categories with objects that vary by color. The *egg coloring cups* category, for example, was easiest to recognize in the *press-proprioception* sensorimotor context since that context implicitly captures some of the objects' geometry and compliance (the objects in that category were identical in shape, height, and material type). For certain categories, auditory feedback was most useful for recognition. For example, the single best context for the *wooden blocks* category was *tap-audio* since wooden objects produce a distinct sound when tapped by the robot's fingers.

8.6.4 Active Behavior Selection

In practice, it may also be useful to know how many behaviors need to be performed to achieve a desired accuracy rate. To obtain this result, the number of behaviors performed at test time is varied from 2 to 10 under two different conditions: random behavior selection and active behavior selection (see Section 8.5.5). When evaluating the performance for active behavior selection, the first behavior is always chosen at random.

Figure 8.11 shows the result of this test in which both models converge to 94.6% when using all 10 behaviors. However, when randomly selecting the next behavior, the performance of the model crosses the 94% threshold after the 8^{th} behavior. On the other hand, the active behavior selection strategy converges to the same rate after only the 4^{th} behavior, i.e., the exploration time during testing is reduced by half. An interesting observation is that the active behavior selection strategy can achieve higher performance with slightly less than all 10 exploratory behaviors. A possible explanation for this is that under the active strategy, the last one or two behaviors that remain are the behaviors that are least accurate for the category of the test object, and thus their output acts as noise in the final combination.

8.6.5 Detecting Outlier Categories

In the last set of experiments, the robot's model was tasked with inferring whether a novel object belongs to a category that is not present in the robot's training set of categories. Figure 8.12 shows a sample case in which the category of the test object (*easter eggs*) is not actually present in the robot's training set. Initially, the category recognition model incorrectly classified the test object as a ball, most likely because the egg has many similar properties as the balls (e.g., shape, size, etc.). Next, the procedure for detecting the odd-one-out object (described in Section 8.5.6) was applied, and in this case, the egg was selected as the outlier. As a result, the estimated category label (*balls*) for the test object was rejected and instead, the object was classified as belonging to a novel category. The figure shows an ISOMAP embedding (Tenenbaum et al., 2000) of the matrix encoding the pair-wise distances between all five objects, as computed in the *press-proprioception* sensorimotor context. As can be seen from the figure,

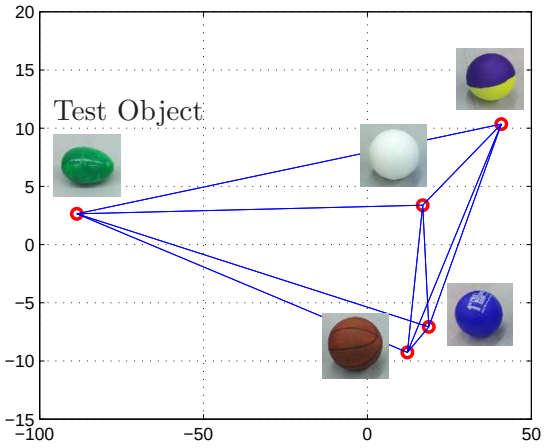


Figure 8.12 A sample case of outlier category detection. In this example, the category *easter eggs* is not present in the robot's training set. Initially, the test object (one of the eggs) is classified as belonging to the *balls* object category by the robot's recognition model. The graph represents a 2-dimensional ISOMAP embedding of a context-specific distance matrix between the 5 objects, i.e., the four known balls and the egg, which is the test object. The sensorimotor context in this example was *press-proprioception*. The distance matrix is converted to a similarity matrix and the procedure outlined in Section 8.5.6 is applied to detect whether the test object should indeed be classified as a ball, or whether it should be considered as one belonging to a novel category. In this case, the method correctly detects that the egg should be considered as belonging to a category not present in the robot's training set.

the four balls form a tight cluster in this context and the egg is easily identified as the odd-one-out.

Figure 8.13 shows the results after the entire evaluation, for different values of the constant r, which determines the necessary threshold that must be exceeded before the object is classified as belonging to an unfamiliar category. The results are reported in terms of true positive rate (the proportion of objects from novel categories classified as such) and false positive rate (the proportion of objects from familiar categories that are mistakenly classified as novel ones). When r is in the range of 1.5 to 2.0, most objects from novel categories can be detected as such, while only a small number of objects from familiar categories are falsely classified as novel.

A large portion of the mistakes made by the model involved the *metal objects* category. For example, when the *pop can* category was not present in the training set, objects from it were classified as belonging to the *metal objects* category. Since a pop can is made of metal,

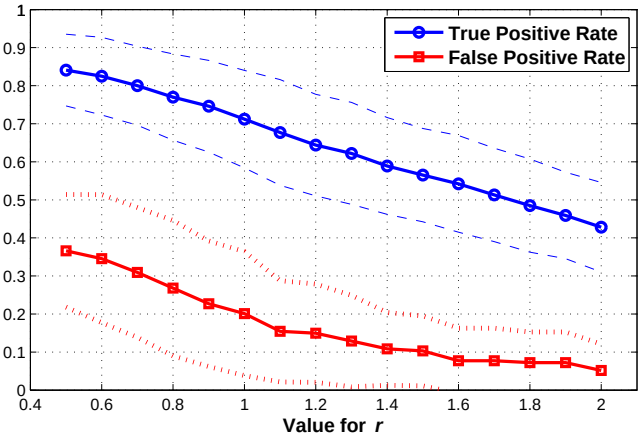


Figure 8.13 Evaluation of the robot's model for detecting the presence of unknown categories. The results are reported in terms of true positive rate (i.e., the proportion of objects from novel categories classified as such), and false positive rate (i.e., the proportion of objects from familiar categories that are mistakenly classified as novel ones). The model is evaluated for different values of the constant r, which determines the threshold that needs to be exceeded for an object to be classified as belonging to an outlier category.

the odd-one-out method was not able to clearly separate it from the known metal objects. In other words, many of the mistakes reflect the fact that specifying a perfectly disjoint object categorization for a large set of objects is nearly impossible.

8.7 Summary and Future Work

The ability to classify objects into categories is a pre-requisite for intelligent manipulation in human environments. To solve a wide variety of household tasks – from sorting objects on a table, to cleaning a kitchen, to taking out the trash – a robot must be able to recognize the semantic category labels of novel objects in its environment. This chapter addressed the problem of object category recognition by presenting an approach that enables a robot to acquire a rich sensorimotor experience with objects and subsequently use visual, auditory, and proprioceptive features to label them. Using simple sensorimotor features coupled with the k-NN and SVM classifiers, the category recognition model was able to scale up to a large number of objects with a diverse set of category labels. Our method was tested using a large-scale experiment in which the robot repeatedly interacted with 100 different objects from 20 object categories using 10 different behaviors (e.g., looking at the object, grasping it, shaking it,

tapping it, etc.). The high recognition rates achieved by the robot (e.g., 97% using SVM) show that perceiving objects using a diverse set of behaviors and sensory modalities is crucial for scaling up object category recognition to a large number of objects and object categories. The model was also able to identify task-relevant sensorimotor contexts for a given categorization task, which allow a robot to learn what specific behaviors and sensory modalities are best for recognizing a specific category label in a novel object. Most importantly, by actively selecting which behavior to apply next, the model was able to reduce by half the exploration time required for classifying a new object. Finally, the robot's model was extended to detect if the test object does not belong to any of the known categories.

There are several direct lines for future work that can further improve the robot's categorization skills. First, a limitation of the current system is that many of the features used to train the classifiers are not invariant with respect to many aspects of the environment that were fixed in the lab setting (e.g., background audio noise, etc.). While much work in the computer vision literature has focused on identifying and computing features that are invariant with respect to scale, orientation, and illumination, it is still an open research question how to do the same for other sensory channels such as audio and proprioception. In addition, some level of invariance to changes in the environment can also be attained by employing machine learning methods that assume that the input data is sampled from a non-stationary distribution (see Sugiyama and Kawanabe (2012) for a review).

Second, it would be highly desirable to relax the assumption that all objects in the robot's training set have corresponding category labels since it may be infeasible to provide such category assignments for all objects that a robot interacts with. This problem can be addressed by using semi-supervised learning methods (Zhu et al., 2005; Zhou et al., 2007) that can make use of both labeled and unlabeled data. Furthermore, since real world objects typically belong to more than one category, it may be desirable to employ a multi-label classification paradigm (see Tsoumakas and Katakis (2007) for a review). This can be achieved by either transforming the multi-label problem into a set of standard classification tasks (e.g., the method proposed by Boutell et al. (2004)) or by employing machine learning algorithms that are directly adapted to the multi-label data representation (e.g., the multi-label AdaBoost method proposed by

Schapire and Singer (2000)).

Finally, while in this chapter the robot was able to perform all of its behaviors on all 100 objects, this may not be feasible if the number of objects is scaled up to 1000 or more. Instead of exhaustively exploring the objects, a robot dealing with such a large number of objects would need to apply behaviors in a way that minimizes exploration time but maximizes the relevant information extracted from the objects. One way to address this problem is to apply models of intrinsic curiosity and motivation (Oudeyer and Kaplan, 2007) to behavior-grounded object exploration. Along those lines, advanced methods for classifier selection (e.g., Gao and Koller (2011)) could also be explored to further reduce the number of interactions required to correctly classify an object.

CHAPTER 9. LEARNING RELATIONAL OBJECT CATEGORIES USING BEHAVIORAL EXPLORATION AND MULTIMODAL PERCEPTION *

9.1 Introduction

The ability to learn and use object categories is an important aspect of human intelligence and has been extensively studied in psychology (see Ashby and Maddox (2005) for a review). Researchers have postulated that, with a few labeled examples, humans at various stages of development are able to identify common features that define category memberships as well as distinctive features that relate members and non-members of a target category (Hammer et al., 2009, 2010). Other lines of research have highlighted the importance of active object exploration for learning object categories (Gibson, 1988; Power, 2000). Studies have also demonstrated that many object properties cannot always be detected by passive observation alone (see Ernst and Bulthof (2004) and Lynott and Connell (2009)).

Recently, several research groups have started to explore how robots can learn object category labels that can be generalized to novel objects (Lopes and Chauhan, 2007; Griffith et al., 2012; Marton et al., 2009; Sinapov and Stoytchev, 2011; Leonardis and Fidler, 2011). Most studies have examined the problem exclusively in the visual domain or have used a relatively small number of objects and categories. Using vision alone, however, would preclude a robot from perceiving the tactile, auditory, and proprioceptive properties of the objects, and thus could severely limit the space of categories that may be learned. On the other hand, if only a small number of objects is used, then there is the potential to severely over-estimate the performance of the classification method (see Sinapov et al. (2011a) for a discussion).

*This chapter is based on the following paper: Sinapov, J., Schenck, C. and Stoytchev, A., "Learning Relational Object Categories Using Behavioral Exploration and Multimodal Perception", (*Under Review*).

A broader limitation of most existing approaches is that they only address human-provided semantic labels that can be expressed as *unary* relations. For instance, an object category can be viewed as a collection of items that share some property (e.g., color, shape, or weight). Many human-provided semantic labels, however, cannot be expressed as unary relations. For example, the label "taller than" can only be expressed as a *binary* relation between two objects. Another limitation is that, in most learning tasks, the robot is only trained to detect the value of a given attribute (e.g., the color of an object). Such a robot would be able to classify a red ball as having the label "red," but it would not be able to detect that a *set* of objects vary by (or are constant in) the attribute "color." To address these limitations, this chapter proposes a relational approach to representing category labels that can handle many types of object relations, not just unary relations.

9.2 Related Work

Supervised methods for object categorization attempt to establish a direct mapping between the robot's object representation and human-provided semantic category labels. A wide variety of computer vision methods have been developed that attempt to solve this problem using visual image features coupled with machine learning classifiers (Fergus et al., 2004; Ponce, 2006; Opelt et al., 2006). Several such methods have been developed for use by robots, almost all working exclusively in the visual domain (Lopes and Chauhan, 2007; Lai and Fox, 2009; Marton et al., 2009; Wohlkinger and Vincze, 2010; Leonardis and Fidler, 2011; Lai et al., 2011a).

Other studies have also demonstrated the ability of robots to assign category labels to objects based on interaction with them (Takamuku et al., 2007; Sinapov and Stoytchev, 2011; Araki et al., 2011; Sinapov et al., 2012; Yürüten et al., 2012; Chu et al., 2013). For example, Takamuku et al. (2007) demonstrated that a robot can classify 9 different objects as either a rigid object, a paper object, or a plastic bottle using auditory and joint angle data obtained while the robot shook the objects. Also, Araki et al. (2011) described a robot that learned to associate words describing an object (e.g., "cup") with object clusters discovered using an unsupervised method.

Despite all of these advances, current work on category recognition suffers from two broad

limitations. First, most object category recognition approaches are entirely vision-based and as such, they would be unable to detect object properties that cannot be extracted using vision alone. While some research has focused on using different sensory modalities coupled with actions, most studies to date use a small number of behaviors (typically just one) and a small number of sensory modalities.

The second broad limitation of most existing approaches is that they only deal with semantic labels that can be expressed as *unary* relations, i.e., labels that apply to individual objects. Many semantic labels, however, cannot be expressed with unary relations. For example, the label "heavier than", can only be expressed as a *binary* relation. Furthermore, in most learning tasks, the robot is only tasked with learning to detect the value of a given attribute (e.g., the color of an object). Such a robot would be able to classify a red ball as having the label "red," but would still be unable to detect that a *set* of objects vary by the attribute "color."

To address these limitations, this chapter proposes a relational approach to representing semantic category labels that describe objects, pairwise object relationships, and object groups. Unlike our previous work in object categorization (see Chapter 7 and Chapter 8), the proposed model can handle many types of object relations beyond simple unary object categories. In addition, the proposed model allows a robot to establish a measure of similarity between different object categories that is grounded in the robot's own sensorimotor repertoire.

9.3 Experimental Methodology

9.3.1 Robot

The experiments described in this chapter were conducted using an upper-torso humanoid robot. The robot had two 7-DOF Barrett Whole-Arm-Manipulators (WAMs) for arms, each equipped with a Barrett Hand as an end effector. During the experiments, only the right arm was used while the left arm was taken off the robot for maintenance. The robot captured proprioceptive, auditory and visual feedback using three types of sensors: 1) joint-torque sensors in the WAM that measure torques for all 7 joints at 500 Hz, 2) an Audio-Technica U853AW cardioid microphone mounted inside the head, and 3) a Microsoft Kinect sensor mounted at

a) The 36 objects used in this study

b) Color: red, green, and blue

c) Contents: glass, rice, beans, and screws

d) Weight: light, medium, and heavy

Figure 9.1 a) The 36 objects used in this study. b)-d) The three types of variations present within the set of objects explored by the robot: b) color, c) contents, and d) weight.

the robot's base.

9.3.2 Objects and Categories

The robot explored 36 objects in this study. The objects were semi-transparent plastic jars with a height of 8.6 centimeters and a diameter of 9.4 centimeters. The objects varied according to their color, their weight, and their contents, as shown in Figure 9.1. Thus, each object was either red, green, or blue in color, heavy (337g), medium (250g), or light (177g) in weight, and had glass marbles, rice, beans, or screws inside of it. Every possible combination was included, resulting in a set of $3 \times 3 \times 4 = 36$ objects.

In this work, the robot learned a diverse set of relational categories that can be applied on single objects, pairs of objects, and groups of objects:

- Categories on single objects: *red*, *green*, *blue*, *light*, *medium*, *heavy*, *glass*, *rice*, *beans*, *screws*.
- Categories on object pairs: *heavier*, *lighter*, *same weight*, *same color*, *same contents*.
- Categories on object groups: *vary by weight*, *vary by color*, *vary by contents*.

9.3.3 Exploratory Behaviors

The robot explored the 36 objects using 10 exploratory behaviors: grasp, lift, hold, shake, rattle, drop, tap, poke, push, and press. Figure 9.2 shows before and after images for each behavior. The behaviors were designed to mimic the exploratory behaviors used by infants (Gibson, 1988; Power, 2000) and were encoded as joint-space trajectories using the Barrett API.

In addition to these 10 interactive behaviors, the robot also performed the *look* behavior at the start of each object exploration trial. During the execution of each of the 10 exploratory behaviors, the robot captured auditory and proprioceptive data. During the execution of the *look* behavior, the robot used the Kinect sensor to take an RGBD image of the object, which was subsequently used to compute two types of visual features. The next sub-section describes the routines used to extract auditory, proprioceptive, and visual features.

9.3.4 Data Collection

The robot explored the objects in a series of trials. During each trial the robot recorded static images of the object on the table and then performed its full set of 10 exploratory behaviors in a sequence. Ten trials were performed on each object, resulting in a total of $36 \times 10 \times 10 = 3600$ behavioral interactions. To minimize any transient noise effects, after a single trial with an object, the object was not explored again until the robot had finished exploring all other objects.

Figure 9.2 Before and after images of the 10 exploratory behaviors that the robot used to learn about the objects.

9.3.5 Sensorimotor Feature Extraction

9.3.5.1 Visual Features Extraction

During the *look* behavior, the robot recorded static images of the object on the table for 1.0 second. These images were then used to extract two types of visual features. To do that, first, the object was segmented from the background using a pre-defined region of interest. Next, an $8 \times 8 \times 8$ color histogram was computed in RGB space based on the segmented object over the sequence of images. The color histogram served as the first type of visual features, $\mathbf{x}_{hist} \in \mathbb{R}^{512}$, that were used by the robot.

For the second type of visual features, for each image, the segmented region was divided into

$8 \times 8 = 64$ evenly spaced patches. The HSV values for the pixels in each patch were averaged together, resulting in a vector of size $8 \times 8 \times 3 = 192$. This was repeated for all images in the sequence and the values of these vectors were averaged, resulting in a single feature vector $\mathbf{x}_{patch} \in \mathbb{R}^{192}$.

9.3.5.2 Auditory Feature Extraction

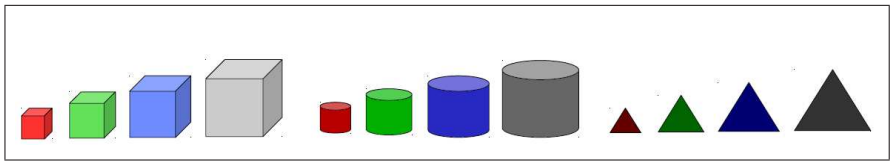
During the execution of the 10 interactive behaviors, the robot extracted features from the audio waveform recorded by the robot's microphones. For each waveform, first, the log-normalized Discrete Fourier Transform (DFT) was computed using 33 frequency bins. The resulting DFT matrix encoded the intensity for each frequency bin at each time step. The matrix was highly-dimensional and was therefore binned into a lower-dimensional 10×10 matrix. The value in each bin was set to the average of the values in the DFT matrix that fell into that bin. Thus, each sound was represented as a feature vector $\mathbf{x}_{audio} \in \mathbb{R}^{100}$.

9.3.5.3 Proprioceptive Feature Extraction

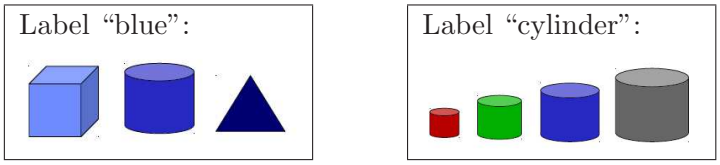
During the execution of an interactive behavior, the robot recorded joint-torque values for all 7 joints at 500 Hz, resulting in a $n \times 7$ matrix (where n is the number of time steps). To reduce dimensionality, the temporal axis was discretized into 10 equally spaced bins. This resulted in a lower dimensional feature vector $\mathbf{x}_{proprio} \in \mathbb{R}^{10 \times 7}$ which encoded proprioceptive features produced by the robot's interaction with the object.

9.3.6 Sensorimotor Contexts

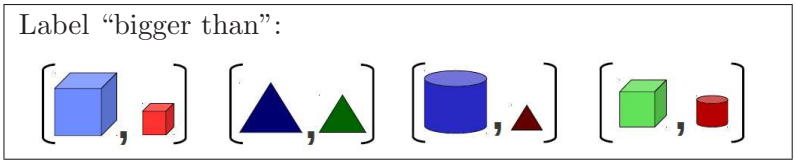
Each valid combination of a behavior and sensorimotor features is deemed a unique *sensorimotor context*. In this work, the robot used 22 sensorimotor contexts denoted by the set $\mathcal{C}$. For each context $c \in \mathcal{C}$, N_c denotes the dimensionality of the sensorimotor features detected that context (e.g., for the *shake-audio* context, $N_c = 100$, while for the *look-histogram* context, $N_c = 512$). Ten of those contexts correspond to proprioceptive features coupled with the 10 different exploratory behaviors. Similarly, another 10 of them correspond to the auditory features extracted from the detected sounds. Finally, two of the sensorimotor contexts correspond



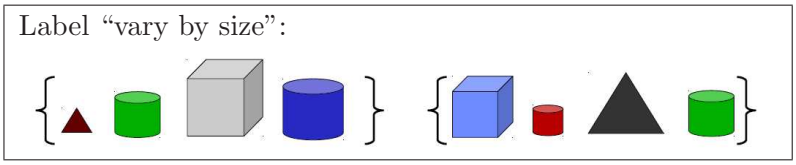
a) Objects



b) Example unary relations



c) Example binary relation



d) Example relation on sets

Figure 9.3 An illustration of how relations can be used to encode a variety of category labels. In this example, the set of objects consists of 4 triangles, 4 squares, and 4 circles, such that each set varies by color as well as by size. Unary relations can be used to represent categories such as “blue” or “cylinder.” Binary relations, on the other hand, can be used to represent the category label “bigger than.” Finally, unary relations whose ground is the power set of the set of objects can be used to encode labels such as “vary by size.”

to the two types of visual features extracted from the static images captured by the robot’s camera during the *look* behavior.

9.4 Theoretical Model

9.4.1 Representing Object Categories with Relations

In logic and set theory, a relation is typically defined as a property that assigns truth values to k-tuples of objects. When $k = 1$ the relation is called a *unary* relation. When $k = 2$ the relation is called a *binary* relation. Such relations are extremely common in mathematics (e.g., equality), as well as in everyday human language that is used to describe how two items relate

to each other (e.g., "heavier than" and "same color as"). Such relations may be reflexive (e.g., "similar to") or transitive (e.g., "heavier than"). Figure 9.3 illustrates how several different types of category labels can be represented using relations. In this example, the set of objects consists of 4 triangles, 4 squares, and 4 circles, such that the objects with identical shapes vary by color as well as by size. Unary relations can be used to represent categories such as "blue" or "cylinder." Binary relations, on the other hand, can be used to represent the category label "bigger than." Unary relations whose ground is the power set of the set of objects can be used to encode labels such as "vary by size."

More formally, let $\mathcal{O}$ be a set of objects. Let L be a k-ary relation over the sequence of domains $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_k$ such that each domain $\mathcal{D}_i \subseteq \mathcal{O}$ or $\mathcal{D}_i \subseteq \mathcal{P}(\mathcal{O})$, where $\mathcal{P}$ denotes the power set. This sequence of domains determines the *ground* of the relation, $G(L) = \mathcal{D}_1 \times \mathcal{D}_2 \times \ldots \times \mathcal{D}_k$. In other words, the set $G(L)$ contains all possible tuples for which the relation may hold. The set $F(L) \subset G(L)$ denotes the *floor* of the relation L and contains only tuples for which the relation holds.

Using this notation, a wide variety of categories can be modeled as relations. For example, the category "red" can be expressed as a unary relation L^{red} with ground $G(L^{red}) = \mathcal{O}$. The relation "heavier than" can be modeled as a 2-ary relation $L^{heavier}$ with ground $G(L^{heavier}) = \mathcal{O} \times \mathcal{O}$. This notation also allows the expression of semantic categories that describe *sets* of objects, rather than individual objects. For instance, the label "vary by color" can be modeled as a unary relation L^{color} with ground $G(L^{color}) = \mathcal{P}(\mathcal{O})$.

9.4.2 Learning Relational Object Categories

Let $\mathcal{L}$ be the set of relations that the robot must learn. For each relation $L \in \mathcal{L}$, the task of the robot is to learn a model that can classify a tuple $t \in G(L)$ as either positive (i.e., the relation holds for t) or negative (i.e., the relation does not hold for t). In other words, if L is a k-ary relation over the sequence of domains $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_k$, then the goal is to learn a model that can recognize whether the relation holds for a tuple of the form $t = (a_1, \ldots, a_k)$. Note that the value of k may be different for some other relation in $\mathcal{L}$.

In this work, the robot used a supervised machine learning method to learn a model for

each relation. Let $t_i \in G(L)$ be the i^{th} data point and let $y_i \in \{-1, +1\}$ be the class label, such that $y_i = +1$ if and only if the relation holds true for t_i (i.e., $t_i \in F(L)$) and -1 otherwise. Let $\mathcal{X}_{t_i}$ be a set of sensorimotor observations with all objects referenced by the tuple t_i. Thus, given a data set of the form $(t_i, \mathcal{X}_{t_i}, y_i)_{i=1}^{N}$, a classifier can be trained to recognize the class label of a novel data point t_{test} given sensorimotor observations $\mathcal{X}_{t_{test}}$. The main challenge consists of constructing an appropriate feature representation for a data point t_i that is suitable for learning.

Next, we describe an approach to computing relational features that are based on the robot's own sensorimotor interaction with the objects in a given tuple.

9.4.2.1 Relations On Single Objects

When $k = 1$ and the domain $\mathcal{D}_1 = \mathcal{O}$, the problem is reduced to the standard binary classification problem in which a single item (in this case, an object) is classified as either a positive example (i.e., class label of $+1$) or a negative example (i.e., class label of -1). To solve this problem, for each relation L and each sensorimotor context $c \in \mathcal{C}$, the robot trained a function M_L^c such that given a sensorimotor observation $\mathbf{x}_a^c \in \mathbb{R}^{N_c}$, obtained by interacting with object o_a, the model $M_L^c(\mathbf{x})$ computes a probabilistic estimate for whether or not the relation L holds for the tuple $t = (o_a)$. In other words, each model M_L^c can be used to compute the estimate $\hat{Pr}(t \in F(L)|\mathbf{x}_a^c)$.

To classify a novel object, let $\mathcal{X}_{test}$ denote a set of sensorimotor observations with a single object $o_{test} \in \mathcal{O}$ and let the tuple $t = (o_{test})$. The robot can then estimate the probability that $t \in F(L)$ (i.e., the relation holds for o_{test}) by:

$$\hat{Pr}(t \in F(L)|\mathcal{X}_{test}) = \alpha \sum_{\mathbf{x}_a^c \in \mathcal{X}_{test}} w_c \times \hat{Pr}(t \in F(L)|\mathbf{x}_a^c),$$

where each w_c is a weight corresponding to the estimated reliability of each context-specific model M_L^c and α is a normalization factor to ensure that the probabilities sum up to 1.0. In our experiments, the models M_L^c were C4.5 decision trees as implemented in the WEKA library (Witten and Frank, 2005) and probabilistic estimates were obtained using the class label distributions at the leaves.

9.4.2.2 Relations on Object Pairs

Let L be a binary relation over the set of objects, i.e., $k = 2$ and the two domains are $\mathcal{D}_1 = \mathcal{O}$ and $\mathcal{D}_2 = \mathcal{O}$. As before, given a tuple $t = (o_a, o_b)$, where $o_a, o_b \in \mathcal{O}$, the task is to learn a model that can compute $\hat{Pr}(t \in F(L))$. To construct features that are suitable for learning, let $\mathbf{x}_a^c \in \mathbb{R}^{N_c}$ and $\mathbf{x}_b^c \in \mathbb{R}^{N_c}$ be two sensorimotor observations with objects o_a and o_b detected in the same context c. Three types of features are extracted by comparing the two features vectors:

- *Absolute Distance Features:* Let $\mathbf{f}_{absolute}^c$ be a feature vector such that each entry $\mathbf{f}[i] = |\mathbf{x}_a^c[i] - \mathbf{x}_b^c[i]|$. In other words, the vector $\mathbf{f}_{absolute}^c \in \mathbb{R}^{N_c}$ has the same length as the original sensorimotor observations and represents the absolute difference between those two observations.

- *Signed Distance Features:* Similarly, let $\mathbf{f}_{signed}^c \in \mathbb{R}^{N_c}$ be a feature vector such that each entry $\mathbf{f}[i] = \mathbf{x}_a^c[i] - \mathbf{x}_b^c[i]$.

- *Global Distance Features:* Finally, a third set of features were constructed to represent the global distance between the feature vectors $\mathbf{x}_a^c$ and $\mathbf{x}_b^c$:

  1. L2 distance:

$$d(\mathbf{x}_a^c, \mathbf{x}_b^c) = \sqrt{\sum_{i=1}^{N_c} (\mathbf{x}_a^c[i] - \mathbf{x}_b^c[i])^2}.$$

  2. Angle-based distance:

$$d(\mathbf{x}_a^c, \mathbf{x}_b^c) = \frac{\sum_{i=1}^{N_c} \mathbf{x}_a^c[i]\mathbf{x}_b^c[i]}{\sqrt{\sum_{i=1}^{N_c} (\mathbf{x}_a^c[i])^2 \sum_{i=1}^{N_c} (\mathbf{x}_b^c[i])^2}}.$$

  3. Canberra distance:

$$d(\mathbf{x}_a^c, \mathbf{x}_b^c) = \sum_{i=1}^{N_c} \frac{|\mathbf{x}_a^c[i] - \mathbf{x}_b^c[i]|}{|\mathbf{x}_a^c[i]| + |\mathbf{x}_b^c[i]|}.$$

4. Chi-square distance:

$$d(\mathbf{x}_a^c, \mathbf{x}_b^c) = \sum_{i=1}^{N_c} \frac{(\mathbf{x}_a^c[i] - \mathbf{x}_b^c[i])^2}{\mathbf{x}_a^c[i] + \mathbf{x}_b^c[i]}.$$

5. Modified Sum Squared Error-based distance:

$$d(\mathbf{x}_a^c, \mathbf{x}_b^c) = \frac{\sum_{i=1}^{N_c} (\mathbf{x}_a^c[i] - \mathbf{x}_b^c[i])^2}{\sum_{i=1}^{N_c} (\mathbf{x}_a^c[i])^2 \sum_{i=1}^{N_c} (\mathbf{x}_b^c[i])^2}.$$

Thus, given $\mathbf{x}_a^c$ and $\mathbf{x}_b^c$, a feature vector $\mathbf{f}_{global}^c \in \mathbb{R}^5$ was computed by calculating the five different distance measures between the input vectors.

The three types of features were subsequently appended in a single feature vector $\mathbf{f}_{a,b}^c = [\mathbf{f}_{absolute}^c, \mathbf{f}_{signed}^c, \mathbf{f}_{global}^c] \in \mathbb{R}^{2\times N_c+5}$. Given this feature representation and a set of training data, for each sensorimotor context c and for each binary relation L a model M_L^c was trained to output the estimated probability that an object pair $t = (o_a, o_b)$ is a member of the relation L, i.e.,

$$M_L^c(\mathbf{f}_{a,b}^c) \rightarrow \hat{Pr}(t \in F(L)|\mathbf{x}_a^c, \mathbf{x}_b^c).$$

Given the sets $\mathcal{X}_a^c$ and $\mathcal{X}_b^c$ that contain multiple sensorimotor observations with two novel objects o_a and o_b in context c, the robot computes the estimate for $\hat{Pr}(t \in F(L)|\mathcal{X}_a^c, \mathcal{X}_b^c)$ (i.e., the probability that the object pair belongs to the category L) according to:

$$\frac{1}{|\mathcal{X}_a^c| \times |\mathcal{X}_b^c|} \sum_{\mathbf{x}_a^c \in \mathcal{X}_a^c} \sum_{\mathbf{x}_b^c \in \mathcal{X}_b^c} M_L^c(\mathbf{f}_{a,b}^c).$$

Finally, using information from all sensorimotor contexts, an estimate for $\hat{Pr}((o_a, o_b) \in F(L))$ can be obtained by:

$$\alpha \sum_{c \in \mathcal{C}} w_c \times \hat{Pr}(t \in F(L)|\mathcal{X}_a^c, \mathcal{X}_b^c),$$

where α is a normalization factor and w_c is a weight associated with context c that corresponds to the estimated classification performance of the model M_L^c.

9.4.2.3 Relations on Object Groups

Semantic categories that describe groups of objects can be represented by relations with arity $k = 1$ and with domain $D_1 = \mathcal{P}(\mathcal{O})$, i.e., the power set of objects. Let L be the target relation and let $\mathcal{G} \subset \mathcal{O}$ be a group of objects. To construct a fixed length feature representation for the object group, pairwise object features are computed as described in the previous subsection and their expected values are estimated from all possible object pairs in the group, i.e.,

$$\mathbf{f}^c_{\mathcal{G}} = \mathbf{E}[\mathbf{f}^c_{a,b} | o_a \in \mathcal{G}, o_b \in \mathcal{G}].$$

More specifically, each element in the feature vector $\mathbf{f}^c_{\mathcal{G}}$ is estimated by

$$\mathbf{f}^c_{\mathcal{G}}[i] = \frac{1}{M} \sum_{o_a, o_b \in \mathcal{G}} \mathbf{f}^c_{a,b}[i],$$

where $M = |\mathcal{G}| \times (|\mathcal{G}| - 1)/2$, i.e., the number of edges in a fully connected graph when we consider the objects in $\mathcal{G}$ as vertices. Given this feature representation, for each sensorimotor context c, the robot trained a model $M^c_L(\mathbf{f}^c_{\mathcal{G}})$ that can estimate whether the semantic label L can be applied on the group of objects $\mathcal{G}$.

As before, the outputs of all context-specific models were combined using a weighted combination rule in which each model is weighted by its estimated reliability. In other words,

$$\hat{Pr}(\mathcal{G} \in F(L)) = \alpha \sum_{c \in \mathcal{C}} w_c \times \hat{Pr}(\mathcal{G} \in F(L) | \mathcal{X}^c_{\mathcal{G}}),$$

where $\mathcal{X}^c_{\mathcal{G}}$ is the set of sensorimotor observations in context c with all objects in $\mathcal{G}$.

The next subsection describes the incremental algorithm that was used to learn the full set of relations $\mathcal{L}$.

9.4.3 Incremental Learning of Relational Object Categories

In the proposed model, the robot learns target relations by incrementally exploring objects one at a time. After exploring an object, the robot is provided with labels that describe this object, labels that describe object pairs that contain this object, as well as labels that describe object groups containing this object. Let $\mathcal{O}_{known}$ be the currently known set of objects and

let $\mathcal{O}_{train}$ be the full set of training objects. At the start of the training process $\mathcal{O}_{known} = \{\}$. Each iteration consists of adding a new object to the set of known objects and can be described by the following steps:

1. Interaction Step: Randomly select an object o_{next} from the set $\mathcal{O}_{train}$. Let $\mathcal{X}_{next}$ be the set of sensorimotor observations produced after the robot performs its full set of exploratory behaviors on that object.

2. Learning Step: Candidate training points are randomly generated that describe the object o_{next} as well as pairs and groups of objects that contain it. Let $t_{single} = (o_{next})$, i.e., t_{single} is a tuple representing a single object. Let the set $\{t^1_{pair}, t^2_{pair}, \ldots, t^p_{pair}\}$ be a set of binary tuples of the form $t^i_{pair} = (o_{next}, o_i)$ where $o_i \in \mathcal{O}_{known}$. Finally, let the set $\{t^1_{group}, \ldots, t^q_{group}\}$ be a set of tuples where each $t^i_{group} \in \mathcal{P}(\mathcal{O}_{known} \cup \{o_{next}\})$ and $o_{next} \in t^i_{group}$. In our experiments $p = 5$ and $q = 6$, while the size of each group $|t^i_{group}| = 3$. Let $U = \{t_{single}, t^1_{pair}, \ldots, t^p_{pair}, t^1_{group}, \ldots, t^q_{group}\}$ denote the full set of candidate tuples generated with object o_{next}.

At each iteration, for each label $L \in \mathcal{L}$, let D_L be the full set of positive and negative example tuples associated with label L obtained up until exploring object o_{next}. Let $\mathcal{M}_L$ be the set of context-specific recognition models associated with label L. The candidate training points in the set U are then used to update the robot's relational category recognition models as shown in Algorithm 9.1. Here, the set U_L denotes a labeled dataset of tuples added in the current update step, where each tuple t is labelled as positive if $t \in F(L)$. After the labeled datasets are constructed (lines 4-12), the models for each label L are re-trained.

In addition, for each label L and each sensorimotor context c, the robot keeps track of the confusion matrix produced when evaluating the model M^c_L on new data. Thus, before re-training the classifiers, they are first evaluated on the novel data (line 14). Once the confusion matrix for a given model M^c_L is updated, the *kappa* statistic (described in the following section) is computed and used as the weight w_c, i.e., the measure of reliability that is used when combining multiple contexts.

3. Performance Evaluation Step: At the end of each iteration, the robot's model is evaluated using a hold out set of objects, $\mathcal{O}_{test}$. To do that, tuples are generated that describe individual

Algorithm 9.1 update-models($U, \{D_L\}_{L\in\mathcal{L}}, \{\mathcal{M}_L\}_{L\in\mathcal{L}}$)

for $L \in \mathcal{L}$ **do**
    Let $U_L = \{\}$.
end for
for $t_i \in U$ **do**
    for $L \in \mathcal{L}$ **do**
        if $t_i \in F(L)$ **then**
            Add $(t_i, +1)$ to dataset U_L.
        else if $t_i \in G(L)$ **then**
            Add $(t_i, -1)$ to dataset U_L.
        end if
    end for
end for
for $L \in \mathcal{L}$ **do**
    evaluate($\mathcal{M}_L$, U_L)
    $D_L = D_L \cup U_L$.
    train($\mathcal{M}_L$, D_L)
end for
return $[\{D_L\}_{L\in\mathcal{L}}, \{\mathcal{M}_L\}_{L\in\mathcal{L}}]$

objects, object pairs and object groups constructed using the set $\mathcal{O}_{test}$. More precisely, the test set contained $|\mathcal{O}|$ tuples describing individual objects, $|\mathcal{O}| \times |\mathcal{O}|$ tuples describing pairs of objects and $\binom{|\mathcal{O}|}{3}$ tuples describing groups of objects.

9.5 Results

9.5.1 Relational Category Recognition Rate

The first experiment was designed to evaluate the model's performance as more and more objects were incrementally added to the robot's training set. To do that, the proposed model was evaluated using 200 runs. For each run, the full set of 36 objects was randomly split into two sets $\mathcal{O}_{train}$ and $\mathcal{O}_{test}$ such that there were 24 objects in $\mathcal{O}_{train}$ and 12 objects in $\mathcal{O}_{test}$. For each relational category $L \in \mathcal{L}$, Cohen's kappa coefficient (Cohen, 1960) was chosen as the performance metric:

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)},$$

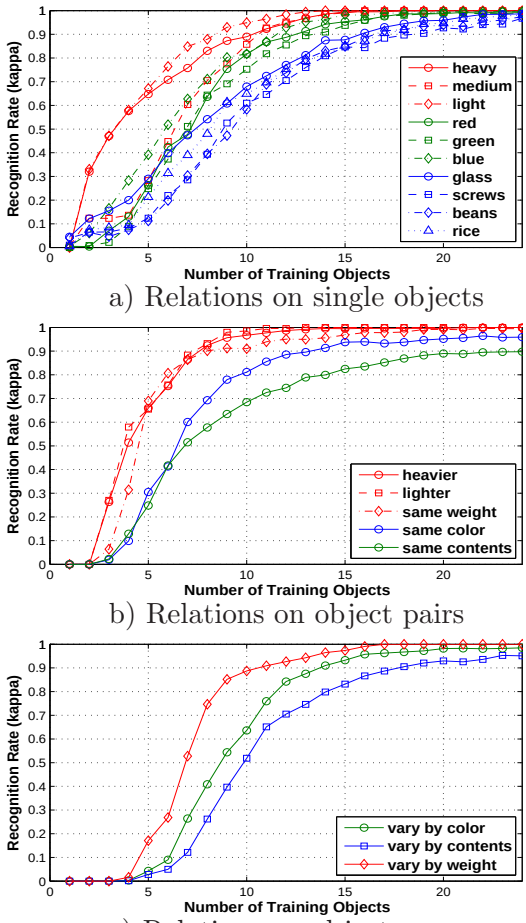


Figure 9.4 Relational category recognition performance as the number of objects explored by the robot is increased from 1 to 24. The figure shows the recognition rates for categories on single objects (a), pairs of objects (b), and groups of objects (c).

where $Pr(a)$ is the probability of correct classification by the model while $Pr(e)$ is the probability of correct classification by chance. This was necessary as reporting accuracy alone could be misleading, e.g., a model that always predicts -1 as the class label is bound to achieve high accuracy.

The results of this experiment are shown in Figure 9.4. The figure shows the recognition rates for categories on single objects (top), pairs of objects (middle), and groups of objects (bottom). Plots related to weight are colored in red, those related to color are colored in green, and finally, the plots related to the objects' contents are colored in blue.

As the robot explores more objects and obtains more training examples, the recognition rates for most relational categories reach a *kappa* of 1.0. Some categories are easier to learn than

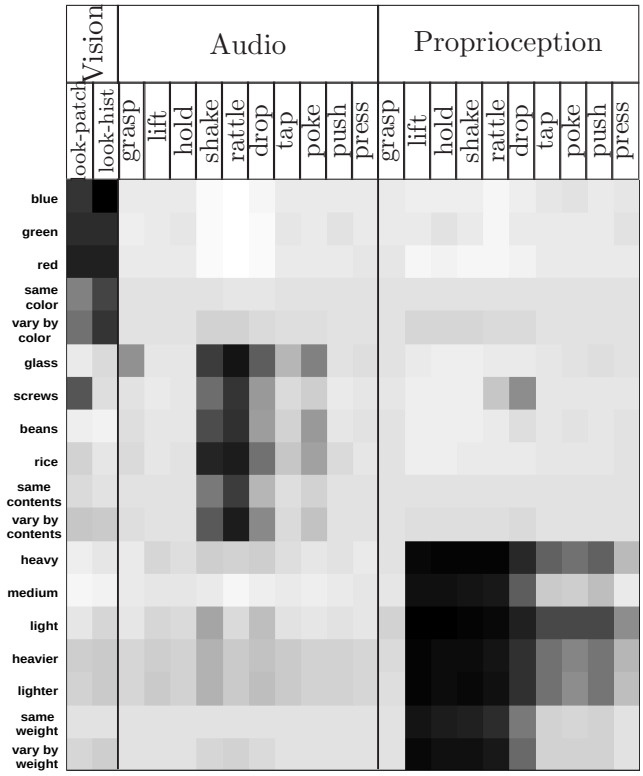


Figure 9.5 Estimated reliability weights associated with each sensorimotor context for each category. Each square corresponds to a recognition model M_L^c and is associated with a specific category and sensorimotor context. The shade of each square shows the estimated *kappa* statistic of the model, where white indicates *kappa* of 0.0 while black indicates 1.0.

others. In particular, concepts related to weight are learned much quicker than the rest. One potential explanation is that nearly all sensorimotor contexts produce proprioceptive feedback that is influenced by the weight of the object, while for concepts related to the object's color and contents, there are only a few contexts that produce the relevant information.

9.5.2 Estimating Category Similarity

So far, the results show that the robot could learn a wide variety of relational categories in an incremental setting. An important question is whether or not the robot's model can relate those categories in a meaningful way. One way in which the different categories can be related is by considering the weights associated with each sensorimotor context for each category. Figure 9.5 shows the estimated reliability weights for each sensorimotor context and each category, averaged over all 200 simulated runs. Here, each square corresponds to a recognition model

M_L^c. The shade corresponds to the model's estimated *kappa* statistic, where 1.0 is black and 0.0 is white. The figure shows that there is great diversity in terms of which sensorimotor contexts are useful for which categories. Furthermore, it also shows that there is a greater number of sensorimotor contexts relevant to weight-related categories, which may explain why those categories are learned quicker than categories related to the object's color and contents.

Figure 9.6 shows a 2D ISOMAP (Tenenbaum et al., 2000) projection in which two categories are close if the same contexts are useful for recognizing them. The projection was computed by associating a weights vector $\mathbf{w}_L$ of length $|\mathcal{C}|$ with each category such that each element of the vector was equal to the *kappa* reliability measure of the corresponding sensorimotor context. The vectors were used to compute a $|\mathcal{L}| \times |\mathcal{L}|$ distance matrix by computing the Euclidean distance for each pair. The matrix was then used as input to the ISOMAP algorithm (Tenenbaum et al., 2000).

The visualization of the context weights and the 2D projection show that the learned relational object categories can be broadly classified into three types: *visual*, *auditory*, and *proprioceptive*. As expected, categories referring to the color of objects could only be recognized using the two types of visual features detected when performing the *look* behavior. Categories relating to the types of contents, on the other hand, were best perceived using the auditory sensory modality in conjunction with the *shake* and *rattle* behaviors. Finally, the categories related to the objects' weight could be perceived using a wide variety of behaviors, including *lift*, *hold*, and *shake*, coupled with the proprioceptive feedback detected using the robot's joint torque sensors.

One possible use of this representation is to improve performance when learning a new category by providing the robot with prior information about how the new category relates to ones that are already learned. For example, if the robot has already learned the relations *red*, *green*, and *blue*, it may be possible to improve its performance when learning the category *same color* if some prior information links the new category with the three familiar categories.

To test this, the robot's model was first trained on $\mathcal{L}_{known}$ categories and was then further trained on the remaining relational category L_{test} using the same procedure. Given a set of similar categories $\mathcal{L}_{similar} \subset \mathcal{L}_{known}$, a set of context weights $\mathbf{w}_{L_{test}}$ was computed such that

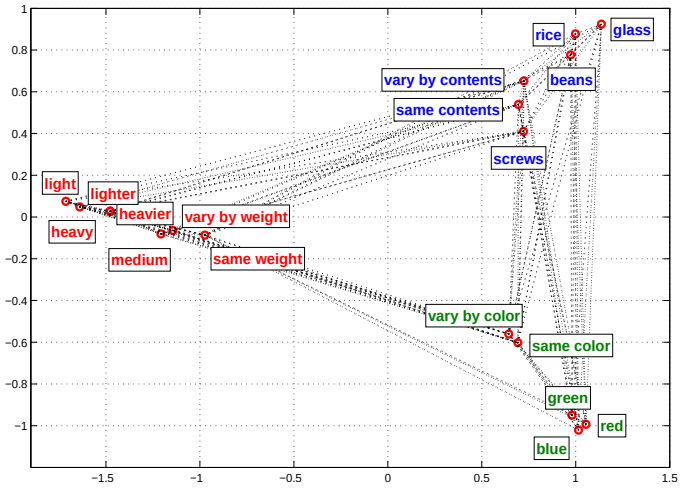


Figure 9.6 An ISOMAP projection (see Tenenbaum et al. (2000)) showing the similarity of the learned categories. Closeness in the projection indicates that the two categories can be recognized well using the same sensorimotor contexts.

$w^c_{L_{test}} = \mathbf{E}[w^c_L | L \in \mathcal{L}_{similar}]$. This process was repeated such that each relation in $\mathcal{L}$ was used once as L_{test}. Figure 9.7 shows the results of this test, where training was halted after exploring 5 training objects. The figure shows that by relating a new category to ones that are known, a robot can substantially improve its performance at test time, even if trained on a much smaller set of objects. This result is especially important because using exploration to estimate which behaviors and sensory modalities are useful for a given category may become more difficult as the set of categories grows larger and larger.

9.6 Conclusion and Future Work

While robot categorization abilities have been constantly improving, the state of the art methods still cannot account for categories that describe relations between objects. To address this need, this chapter proposed a novel framework that enables a robot not only to assign labels to individual objects, but also to detect relational categories that describe how objects relate to each other. The robot learned to recognize individual object properties, such as their color, weight, and contents. Furthermore, the robot learned to classify pairs of objects according to several labels such as “same color”, “heavier than”, etc. Finally, the robot also learned to recognize whether a group of objects varies by any of the three object properties.

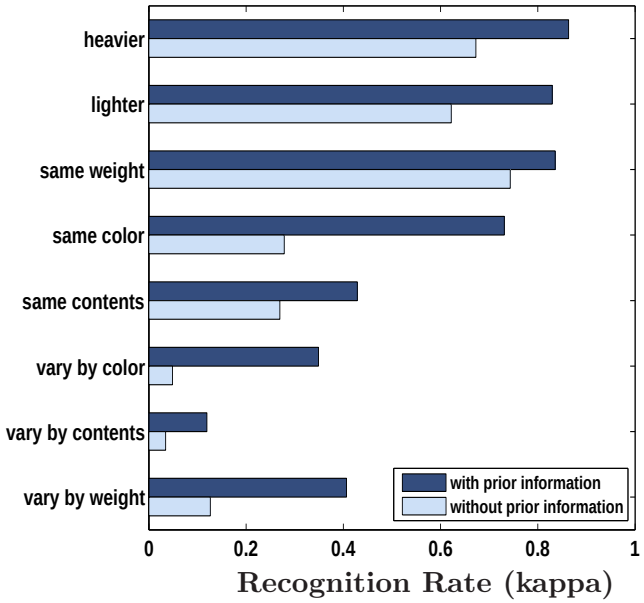


Figure 9.7 Visualization of the recognition improvement obtained when using prior information that relates a novel category to categories that are already learned. For this test, only 5 training objects were used and the results were averaged over 50 different runs. This figure shows that prior information that links the target category to familiar categories can be used to substantially improve the recognition rate.

In addition to achieving high recognition rates for all three types of categories, the robot was also able to establish a measure of similarity between the different relational categories that it learned. More specifically, two categories were deemed similar if they could be recognized using the same behaviors and sensory modalities and dissimilar otherwise. Our results showed that this type of representation is especially useful when the robot is tasked with learning a new relational category that is similar to already known categories.

Scaling up to an even larger number of categories and objects remains a challenge and is a direct line for future work. One possible avenue for tackling the problem is to further investigate how a robot can bootstrap learning of new categories using categories that are already known. For example, linking sensorimotor contexts associated with a known category to a novel category can be used not only to reduce the number of training objects as was shown here, but it could also be useful for reducing object exploration time during learning. Finally, it is also necessary to further expand the space of relational categories that can be handled by the model so that a robot can learn other relational categories (e.g., the label "ordered by height") that cannot be modeled as relations over object pairs or groups of objects.

CHAPTER 10. GROUNDED OBJECT INDIVIDUATION BY A HUMANOID ROBOT*

10.1 Introduction

Humans learn to individuate objects by first learning to detect whether two perceptual stimuli were produced by the same object or by two different objects (Krojgaard, 2004). This ability allows humans to infer how many unique objects they have observed and to establish an object representation that can be used to map individual experiences with an object to a unique object identifier (Kemp et al., 2009). Studies in developmental psychology have shown that this skill is fundamental to establishing an internal object representation that can handle the large number of objects that humans encounter in their daily lives (Krojgaard, 2004; Tremoulet et al., 2000).

In contrast, most methods used by robots to recognize objects start with a fixed object representation in which the robot's training data is labeled with one of a finite number of object identities (see Torres-Jara et al. (2005); Sinapov et al. (2009); Natale et al. (2004); Rasolzadeh et al. (2010); Bergquist et al. (2009); Rusu et al. (2008); Sinapov et al. (2011a); Marton et al. (2012) for a representative sample of such approaches). These methods implicitly make the assumption that the object individuation task has already been solved. In other words, training the robot's object recognition models requires that the training observations are labeled with the correct object identity. Providing labeled data, however, becomes increasingly more difficult as the number of objects increases. Furthermore, an autonomous robot operating in human environments is bound to encounter new objects that were not in its training dataset. Therefore,

*This chapter is based on the following paper: Sinapov, J. and Stoytchev, A., "Grounded Object Individuation by a Humanoid Robot", *In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, Karlsruhe, Germany, May 6-10, 2013.

Figure 10.1 The humanoid robot used in our experiments, along with the 100 objects that it explored.

in addition to recognizing objects, robots must also be able to individuate novel objects.

To address these challenges, this chapter describes a behavior-grounded approach to object individuation that enables a robot to estimate how many objects it has interacted with, and group its sensorimotor experience with objects according to the estimated object identities. The method was tested using a large-scale experiment in which the robot interacted with 100 different objects using 10 different exploratory behaviors. The results demonstrate that by using a small amount of prior training, the model can successfully individuate novel objects that were not present in the robot's training set.

10.2 Related Work

10.2.1 Psychology

When psychologists study how humans individuate and identify objects they typically use an experimental design in which the participant is presented with a sequence of objects and at the end is asked to infer how many unique objects were encountered (Kemp et al., 2009). In this setting, the subject cannot observe multiple objects at the same time, and thus must rely on the objects' perceptual features when solving the task. The results of the experiment conducted by Kemp et al. (2009) show that prior experience with objects with known object identities is necessary in order to solve the object individuation task on a novel set of objects.

Therefore, it is not surprising that humans use a variety of cues, other than object features, when individuating objects (Kemp et al., 2009; Krojgaard, 2004). For example, spatial cues can be used to individuate objects since observing two objects next to each other indicates that the two objects are not the same (Xu and Chun, 2009). Humans also use temporal cues, e.g., they assume that an object would remain the same object over the course of contiguous manipulation or observation (Becchio and Bertone, 2003). Most importantly, such spatial and temporal cues can inform the observer that the featural differences between the objects are not due to noisy observations, but due to the two objects being different (Kemp et al., 2009; Xu and Chun, 2009).

Inspired by these results from psychology, this chapter describes a learning approach to object individuation in which the robot was initially trained to detect whether two sensorimotor experiences are produced by the same object or by two different objects. Subsequently, the trained model was used to partition the robot's sensorimotor experience with novel objects in order to individuate them. The results of our experiments suggest that, just as for humans, prior information, in the form of a training set with known object identities, is necessary for solving this problem.

10.2.2 Robotics

Object individuation has received relatively little attention in robotics. In contrast, a wide variety of methods have been developed that allow robots to recognize previously observed objects. The majority of these methods use 2D and 3D visual features (see Quigley et al. (2007); Srinivasa et al. (2009); Rasolzadeh et al. (2010); Rusu et al. (2008); Marton et al. (2012)). Other vision-based approaches have also been proposed for finding image regions from multiple views that contain the same object Kang et al. (2012). In addition, experiments have demonstrated that robots can also recognize objects and their categories using proprioceptive (Natale et al., 2004; Bergquist et al., 2009), auditory (Torres-Jara et al., 2005; Sinapov et al., 2009), tactile (Bhattacharjee et al., 2012; Fishel and Loeb, 2012) and multi-modal (Sinapov and Stoytchev, 2011; Sinapov et al., 2011a, 2012) sensory feedback. The main limitation of these systems is that the object recognition models can only be trained on fixed datasets containing

labeled data for all objects that the robot may encounter. In other words, while such systems can recognize previously observed objects, they cannot individuate novel objects that they encounter after training time.

It is worth noting that this limitation does not only plague object recognition methods, but also affects a variety of other robotic systems. For example, to learn the affordances of a tool, the methods described by Stoytchev (2005) and Sinapov and Stoytchev (2008) assume that the robot's sensorimotor data is cleanly partitioned according to the identity of each tool. Similarly, when categorizing objects as either containers or non-containers, the robot described in Griffith et al. (2012) started with the implicit assumption that it already knows the identities of all objects that it has to interact with. These and many other examples show that today's robots typically start with fixed object representations, and thus lack the ability to individuate objects that they may encounter in the future.

10.3 Experimental Methodology

10.3.1 Robot

The upper-torso humanoid robot used in our experiments (shown in Figure 10.1) has two 7-DOF Barrett Whole Arm Manipulators (WAMs), each equipped with the 3-finger Barrett Hand. The robot's head was equipped with an Audio-Technica U853AW cardioid microphone that was used to capture auditory feedback. Proprioceptive feedback was captured by the built-in sensors in each WAM, which measure joint-torques at 500 Hz. Finally, visual feedback was detected using the robot's right eye, a 640 by 480 resolution Logitech webcam.

10.3.2 Objects

To test the proposed model, the robot explored 100 different household objects, which are shown in front of the robot in Figure 10.1. Some of the objects are visually identical, but they differ in other properties – for example, the five red containers were filled with different contents that produced different sounds when the objects were shaken. The five blue containers, on the other hand, contained varying amounts of rice, and thus they differed only in weight. To the

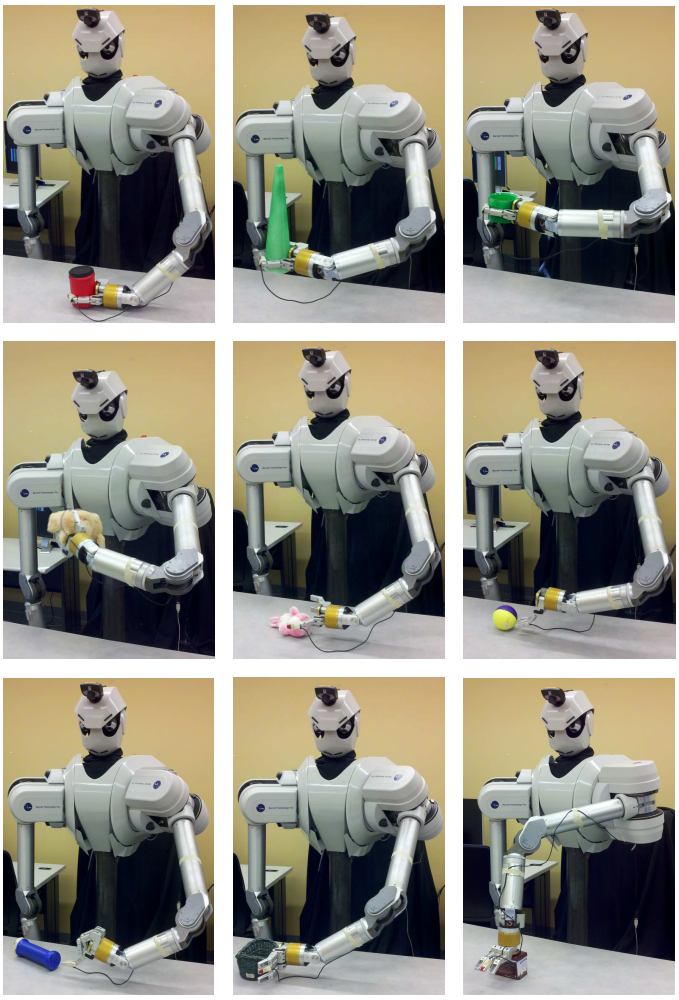


Figure 10.2 The exploratory behaviors that the robot performed on all objects. From top to bottom and from left to right: *grasp, lift, hold, shake, drop, tap, poke, push,* and *press*. In addition to the 9 behaviors pictured above, the robot also performed the *look* behavior, which consisted of taking an RGB snapshot of the object on the table.

best of our knowledge, this dataset contains the largest number of objects ever explored by a robot over the course of a single experiment.

10.3.3 Exploratory Behaviors

The robot was equipped with 10 different behaviors that it applied on all objects: *look*, *grasp*, *lift*, *hold*, *shake*, *drop*, *tap*, *poke*, *push*, and *press*. The *look* behavior consisted of taking an RGB snapshot of the object while the other nine behaviors (see Figure 10.2) were encoded as joint-space trajectories that were executed using Barrett's default PID controller. The robot performed its set of 10 exploratory behaviors on each of the 100 objects 5 different

times. This resulted in a total of 5000 behavioral interactions, which were organized into 500 exploratory trials, where each trial corresponds to the 10 different behaviors performed in a sequence on a single object. During the execution of each behavior, the robot recorded auditory, proprioceptive, and visual feedback, which were used to extract different features as described below.

10.3.4 Sensorimotor Feature Extraction

10.3.4.1 Color

For each exploratory trial, the robot extracted an $8 \times 8 \times 8$ color histogram in RGB space with uniformly spaced bins from the RGB image of the object recorded during the *look* behavior. For each image, background subtraction was used to segment the object from the background.

10.3.4.2 SURF

The Speeded-Up Robust Features (SURF) described by Bay et al. (2008) were computed for all images captured by the robot's camera. Figure 10.3.a shows an example image captured by the robot's camera along with the detected SURF interest points. The X-means (Pelleg and Moore, 2000) algorithm was used to quantize the detected SURF feature descriptors using 0.5% of all detected feature descriptors. This resulted in a dictionary containing 200 visual "words." Using the learned quantization, for each of the 5000 behavioral interactions, a 200-dimensional feature vector was computed encoding a histogram of the SURF descriptors detected over the course of executing the behavior.

10.3.4.3 Optical Flow

During the execution of each behavior (except *look*), the stream of images captured by the camera was used to compute dense optical flow using the algorithm and MATLAB implementation proposed by Sun et al. (2010a). For each pixel in a given image in the sequence, the algorithm computed a real-valued vector (u, v) encoding the direction of motion (i.e., the vector's angle) as well as the magnitude of the motion (i.e., the vector's norm). Figure 10.3.b shows the detected optical flow for a single frame captured during the execution of the *poke*

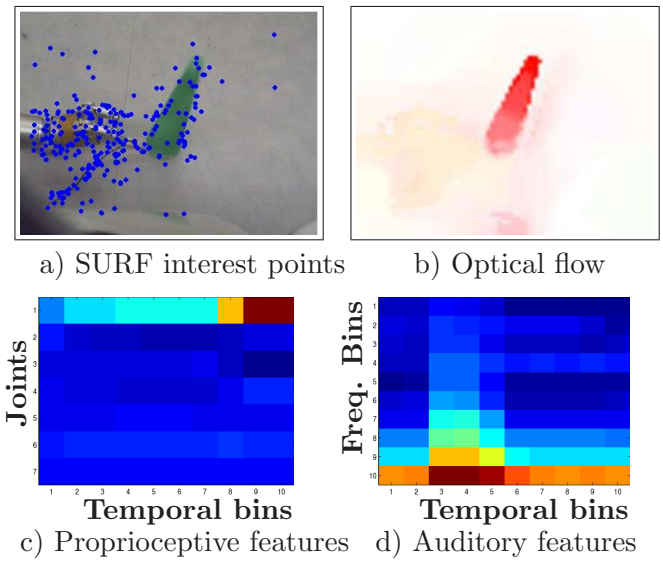
behavior on one of the green cones (the hue encodes the angle of the optical flow vector, while the intensity corresponds to the vector's norm). To reduce the dimensionality of the optical flow feedback, *weighted angular histogram* features were extracted from the sequence of optical flow images by binning the angles into 10 equally spaced bins. In other words, the norms of the optical flow vectors with angles ranging from 0 to $2\pi/10$ were added to bin number 1, while those in the range of $2\pi/10$ to $2 \times 2\pi/10$ were added to bin number 2, and so forth.

10.3.4.4 Proprioception

Proprioceptive features were extracted from the recorded joint torques for all 7 joints of the robot's left arm for all behaviors except *look*. The torques were recorded at 500Hz. To reduce the dimensionality of the signal, the series of torque values for each joint were discretized into 10 temporal bins (i.e., each bin encoded the average torque that was measured over its corresponding time window). This resulted in lower-dimensional data points $\mathbf{x} \in \mathbb{R}^{10\times7}$, which were subsequently used to represent the robot's proprioceptive experience with the objects. Figure 10.3.c shows an example 10×7 feature vector, visualized as a matrix in which the rows correspond to the 7 joints and the columns correspond to the 10 temporal bins. In addition to the joint-torque proprioceptive features, at the end of the *grasp* behavior, the final joint position for each of the three fingers was recorded and used as an additional source of proprioceptive feedback.

10.3.4.5 Audio

After the execution of each of the 9 interactive behaviors, the log-normalized Discrete Fourier Transform (DFT) was computed for the recorded waveform. The DFT was computed with the SPHINX4 natural language processing library package (Lee et al., 1990) using $2^7 + 1 = 129$ frequency bins. To reduce dimensionality, the DFT was further discretized using 10 temporal bins and 10 frequency bins, where the value for each bin was set to the average of the values in the DFT matrix that fell into it. Figure 10.3.d shows one discretized DFT that was calculated after performing the *drop* behavior.



a) SURF interest points b) Optical flow

c) Proprioceptive features d) Auditory features

Figure 10.3 Visualization of some of the sensorimotor features used by the robot. a) Sample SURF interest points computed from a single image; b) Sample dense optical flow computed while executing the *poke* behavior; c) Sample proprioceptive features detected while executing the *press* behavior; d) Sample audio features computed from the DFT for the *drop* behavior.

In summary, during each exploratory trial, the robot performed 10 exploratory behaviors on one of the 100 objects. Five of these trials were recorded for each object. During the execution of each behavior, the robot extracted features from several sensory modalities, where each viable combination of behavior and sensory modality (e.g., *drop-audio* or *look-color*) determined a unique sensorimotor context. The auditory, proprioceptive, and optical flow features were extracted while performing all 9 interactive behaviors. SURF features were extracted for all 10 behaviors. Color features were extracted from the static images captured during the *look* behavior while hand-proprioceptive features were extracted during the execution of the *grasp* behavior. Thus, the total number of sensorimotor contexts available to the robot was $9 \times 3 + 10 + 1 + 1 = 39$.

10.4 Theoretical Model

10.4.1 Notation and Problem Formulation

Let $\mathcal{S}$ be the set of sensorimotor contexts available to the robot, where each context refers to a specific combination of a behavior and a sensory modality. Also, let $\mathcal{T}$ be the full set of 500

exploratory trials with all objects. During each trial, the robot applies its set of exploratory behaviors on some object $o \in \mathcal{O}$. The i^{th} exploration trial can be represented with the collection of observed sensory feedback signals, $T_i = \{x_i^s\}_{s \in \mathcal{S}}$, where each feature $x_i^s \in \mathbb{R}^{d_s}$.

The object individuation task can be formulated as follows. Let $\mathcal{T}_{test} = \{T_i\}_{i=1}^{n}$ be a test set containing n interaction trials in which the robot explored a test set of objects, $\mathcal{O}_{test} \subset \mathcal{O}$. The individuation task is to separate the set of trials $\mathcal{T}_{test}$ into groups, such that each group contains only the trials with one of the objects in $\mathcal{O}_{test}$.

In other words, the object individuation task is a special case of clustering in which each data point corresponds to a sensorimotor observation with a physical object. In contrast to fully unsupervised clustering methods, the approach described here uses prior information in the form of a set of training trials for which the object identities are known. Let $\mathcal{O}_{train} \subset \mathcal{O}$ be the objects in the robot's training set such that $\mathcal{O}_{train} \cap \mathcal{O}_{test} = \emptyset$. The set $\mathcal{T}_{train} = \{T_i, o_i\}_{i=1}^{n_{train}}$ contains the exploratory trials with the training objects, where each trial T_i is labeled with the corresponding object identity $o_i \in \mathcal{O}_{train}$.

The method for object individuation described here consists of the following three stages:

1. *Distance Estimation Stage:* During this step, the robot estimates pair-wise distances for each pair of trials in $\mathcal{T}$, and for each sensorimotor context s.

2. *Learning Stage:* The data in $\mathcal{T}_{train}$ is used to learn a model that can classify a pair of trials as either "same", i.e., belonging to the same object, or "different", i.e., belonging to two different objects.

3. *Individuation Stage:* The learned model is applied on each pair of trials in the set $\mathcal{T}_{test}$, and in conjunction with a graph-based clustering algorithm is used to produce the labels of the final object individuation.

The next three subsections provide a detailed description for each of these three stages.

10.4.2 Distance Estimation Stage

In the first stage, the task is to estimate the perceptual dis-similarity for each pair of trials in the set $\mathcal{T}$. Given a sensorimotor context $s \in \mathcal{S}$, let $x_i^s \in \mathbb{R}^{d_s}$ and $x_j^s \in \mathbb{R}^{d_s}$ be the feature

vectors detected in that context for trials T_i and T_j. In this work, the dis-similarity between trials T_i and T_j in context s was estimated by computing the Euclidean distance between the feature vectors x_i^s and x_j^s. Thus, for each context $s \in \mathcal{S}$, the robot estimated a pair-wise trial distance matrix, $\mathbf{W}^s \in \mathbb{R}^{|\mathcal{T}|\times|\mathcal{T}|}$, such that each entry $W_{ij}^s \in \mathbb{R}$ encoded the perceptual dis-similarity between trials T_i and T_j in that context. Finally, for each matrix, the values of all elements were linearly rescaled to lie in the range from 0.0 to 1.0.

10.4.3 Learning Stage

A fundamental pre-requisite for object individuation is the ability to detect whether two perceptual stimuli were produced by the same object or by two different objects (Krojgaard, 2004). In the method proposed here, this is accomplished by learning a model that can classify a pair of trials as either "same" or "different", where the label depends on whether the same object was present in both trials or not. To learn such a model, two types of features were extracted for each pair of trials:

- *Perceptual dis-similarity features*: given a pair of trials T_i and T_j, a feature vector $\mathbf{f}^{ij} \in \mathbb{R}^{|\mathcal{S}|}$ was computed where each element $f_s^{ij} = W_{ij}^s$ for $s = 1$ to $|\mathcal{S}|$. In other words, $\mathbf{f}^{ij}$ encodes the perceptual distances between trials T_i and T_j in all available sensorimotor contexts.

- *Dis-similarity histogram features*: given a pair of trials T_i and T_j, and the computed feature vector $\mathbf{f}^{ij}$, the values in $\mathbf{f}^{ij}$ were used to construct a histogram that encodes the distribution of dis-similarities for the two trials. The histogram was constructed using 10 equally spaced bins, resulting in a 10-dimensional feature vector $\mathbf{h}^{ij}$.

During the learning stage, the two types of features were computed for all pairs of trials T_i and T_j from the set $\mathcal{T}_{train}$. This resulted in two datasets, $\mathcal{D}_{dist} = \{\mathbf{f}^{ij}, y_{ij}\}$ and $\mathcal{D}_{hist} = \{\mathbf{h}^{ij}, y_{ij}\}$, where each $y_{ij} = +1$ if trials T_i and T_j were performed with the same object and -1 otherwise. The first dataset, $\mathcal{D}_{dist}$, contained the raw perceptual distance features for each pair of trials, while the second, $\mathcal{D}_{hist}$, was based on features that encode the distribution of the raw perceptual distances.

The datasets were subsequently used to train two machine learning classifiers, $\mathcal{M}_{dist}$ and $\mathcal{M}_{hist}$ on the task of detecting whether two trials were performed on the same object. Thus, given a trial pair (T_i, T_j), the model $\mathcal{M}_{dist}$ produced an estimate for $\Pr_{dist}(\text{“same”} \mid \mathbf{f}^{ij})$, i.e., the probability that the two trials contained the same object. Similarly, given the same trial pair, the model $\mathcal{M}_{hist}$ produced the same estimate based on the histogram features for the trial pair, i.e., $\Pr_{hist}(\text{“same”} \mid \mathbf{h}^{ij})$. In the experiments described in this chapter, each of the two models was implemented using the WEKA (Witten and Frank, 2005) implementation of the AdaBoost (Freund and Schapire, 1996) algorithm with C4.5 decision tree (Quinlan, 1993) as a base classifier.

10.4.4 Individuation Stage

Given a test set of trials $\mathcal{T}_{test}$, the outputs of the classifiers $\mathcal{M}_{dist}$ and $\mathcal{M}_{hist}$, computed for each pair of trials in $\mathcal{T}_{test}$, were used to individuate the objects as described below. Let $\mathbf{A} \in \mathbb{R}^{|\mathcal{T}_{test}| \times |\mathcal{T}_{test}|}$ be the resulting individuation matrix where each entry was computed as:

$$A_{ij} = \frac{\Pr_{dist}(\text{“same”} \mid \mathbf{f}^{ij}) + \Pr_{hist}(\text{“same”} \mid \mathbf{h}^{ij})}{2}.$$

In other words, each entry A_{ij} corresponds to the estimated probability that trials T_i and T_j were performed with the same object. This probability was computed using a uniform combination of the outputs of the two classifiers.

To construct an object individuation using the matrix $\mathbf{A}$, the robot used the *spectral clustering* algorithm, which is one of several *graph-based* or *similarity-based* clustering algorithms (von Luxburg, 2007). Given an affinity matrix, i.e., $\mathbf{A}$, the algorithm partitions the set of trials into disjoint clusters by exploiting the eigenstructure of the matrix $\mathbf{A}$. To solve the problem efficiently, Shi and Malik (2000) proposed an algorithm that optimizes the *normalized cut* objective function. Given an input individuation matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, the algorithm can be summarized with the following steps:

1. Let $\mathbf{D} \in \mathbb{R}^{n \times n}$ be the degree matrix of $\mathbf{A}$, i.e., a diagonal matrix such that $\mathbf{D}_{ii} = \sum_j A_{ij}$.
2. Solve the eigenvalue system $(\mathbf{D} - \mathbf{A})x = \lambda \mathbf{D}x$ for the eigenvector corresponding to the second smallest eigenvalue and use it to bipartition the graph.

3. If necessary, recursively bipartition each subgraph that was obtained in Step 2.

This procedure recursively bipartitions the graph induced by the matrix $\mathbf{A}$ until the spectral clustering algorithm fails to find a bipartition with a high score according to the normalized cut objective function or until it fails to find a solution to the eigenvalue system. The code for the spectral clustering algorithm (Steps 1 and 2) used in our experiments is listed on the WEKA machine learning repository website (Dragone, 2006).

The output of this procedure is a partitioning of the n trials into k clusters, which can be represented as a set of k sets of trials, $\mathcal{C} = \{C_\ell | \ell = 1, \ldots, k\}$ or as a label vector $\omega \in \mathbb{N}^n$ where each entry $\omega_i \in \{1, \ldots, k\}$ encodes the partition label for trial T_i. The next section describes several measures that were used to evaluate the robot's object individuation model.

10.5 Evaluation

10.5.1 Performance Measures

The estimated partitioning $\hat{\mathcal{C}}$ and the corresponding label vector $\hat{\omega}$ were evaluated by comparing them to the ground truth individuation, represented by the partitioning $\mathcal{C}$ and the vector ω, using several different methods.

10.5.1.1 Normalized Mutual Information

Normalized Mutual Information (NMI) has been proposed as a measure to capture the similarity between two different clusterings over the same dataset (Strehl and Ghosh, 2003). Given two clusterings ω^a and ω^b defined over the same set of n trials, let k^a and k^b be the number of clusters in ω^a and ω^b respectively. Let n_h^a be the number of trials in cluster C_h according to ω^a, and let n_ℓ^b the number of trials in cluster C_ℓ according to ω^b. Also, let $n_{h,\ell}$ be the number of trials that are in cluster C_h according to ω^a, as well as in cluster C_ℓ according to ω^b. Using these definitions, the NMI estimate, ϕ^{NMI}, is defined as:

$$\phi^{NMI}(\omega^a, \omega^b) = \frac{\sum_{h=1}^{k^a}\sum_{\ell=1}^{k^b} n_{h,\ell}\, log\left(\frac{n * n_{h,\ell}}{n_h^a * n_\ell^b}\right)}{\sqrt{\left(\sum_{h=1}^{k^a} n_h^a\, log(\frac{n_h^a}{n})\right)\left(\sum_{\ell=1}^{k^b} n_\ell^b\, log(\frac{n_\ell^b}{n})\right)}}.$$

This pairwise measure of mutual information is always in the range of 0.0 to 1.0, where 1.0 indicates that the two partitionings are identical while 0.0 means that the two partitionings were computed over two disjoint datasets.

10.5.1.2 Mean Partition Entropy

The second performance measure was chosen to evaluate the purity of each resulting cluster in the individuation with respect to object identity. Given a partition $C_\ell \in \mathcal{C}$, let $\Pr_\ell(o)$ be the probability that a randomly sampled trial from C_ℓ was performed on object $o \in \mathcal{O}$. Given the distribution over all objects for a given partition C_ℓ, Shannon's entropy (Shannon, 1948) can be computed by:

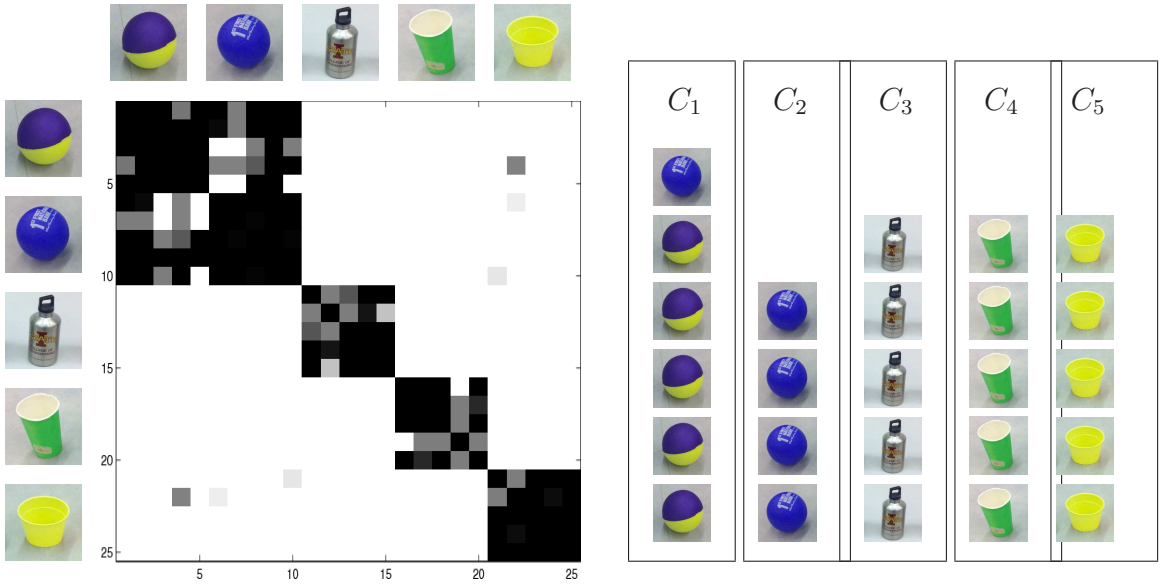
$$H_\ell = -\sum_{o \in \mathcal{O}} \Pr_\ell(o)\, log(\Pr_\ell(o)).$$

A value of 0.0 for a cluster C_ℓ would indicate that the cluster only contains trials with one object, while large values for H_ℓ would signify that the cluster contains trials with many different objects. Thus, given the full partitioning, $\mathcal{C}$, the Mean Partition Entropy (MPE) is defined as:

$$\text{MPE}(\mathcal{C}) = \frac{1}{|\mathcal{C}|}\sum_{C_\ell \in \mathcal{C}} H_\ell.$$

10.5.1.3 α-Individuation Rate

The last measure estimates the percentage of objects in the test set that were individuated correctly. An object o is considered individuated if there exists a partition C_ℓ in the set $\mathcal{C}$ that contains at least α trials with object o and no trials with any other objects. In this study, the robot performed 5 trials with each object, and therefore, the α-Individuation Rate was computed for $\alpha = 3$, 4, and 5.



a) Estimated Object Individuation Matrix b) Resulting Individuation

Figure 10.4 a) An example object individuation matrix **A**. The matrix encodes the estimated likelihood that a pair of trials in the test set were performed on the same object, where dark indicates high likelihood and white indicates low likelihood. In this example, the test set contained 25 trials with 5 different objects (5 trials per object). For better visualization, the entries of the matrix are sorted by object identity. b) The resulting object individuation. Each partition corresponds to a set of trials that, according to the trained model, were performed with the same object.

10.5.2 Baseline Comparison

The method for object individuation was also compared against an unsupervised approach in which the test set of trials is partitioned using only the pairwise distance matrices $\mathbf{W}^s$. To do so, a trial affinity matrix $\mathbf{U}$ was constructed such that each entry $U_{ij} = (1/|\mathcal{S}|) \sum_{s \in \mathcal{S}} (1.0 - W^s_{ij})$. In other words, each entry U_{ij} corresponds to the average perceptual similarity for the two trials computed across all sensorimotor contexts, with values close to 1.0 meaning highly similar and values close to 0.0 meaning highly dis-similar. The matrix $\mathbf{U}$ was then used as input to the partitioning algorithm described in Section 10.4.4 to produce a final object individuation.

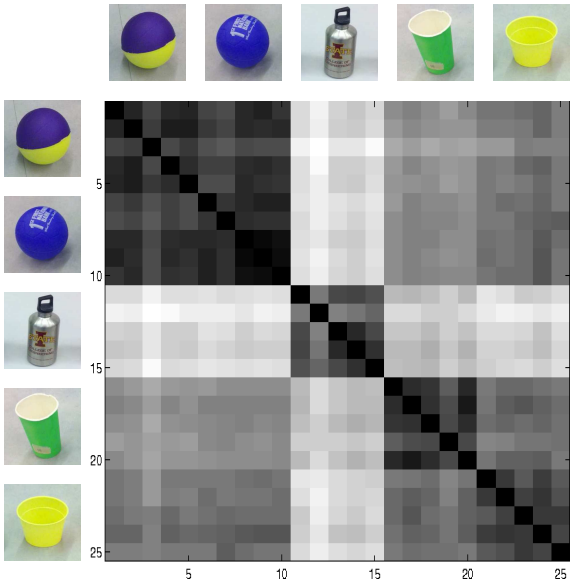


Figure 10.5 An example perceptual similarity matrix, $\mathbf{U}$, for 25 exploratory trials computed using the 39 raw context-specific distance matrices $\mathbf{W}^s$.

10.6 Results

10.6.1 Example

Figure 10.4.a shows a sample trial individuation matrix, $\mathbf{A}$, which was computed using a test set of 25 trials with 5 different objects (5 trials per object). Each entry in the matrix encodes the estimated probability that a pair of trials was performed with the same object, where dark indicates high likelihood and white indicates low likelihood. The individuation model used to fill in the entries of the matrix was trained on a separate set of 25 trials with another set of 5 objects.

For visualization purposes, the entries of the matrix are sorted by object identity. Because the matrix is sorted, the block pattern along the diagonal clearly shows that the learned model was able to detect which pairs of trials were performed with the same object far better than chance. For comparison, Figure 10.5 shows the perceptual similarity matrix, $\mathbf{U}$, for the same 25 exploratory trials, computed from the 39 raw context-specific distance matrices $\mathbf{W}^s$ using the unsupervised baseline approach. It is easy to see that the matrix $\mathbf{U}$ has more non-zero entries than the matrix $\mathbf{A}$ for pairs of trials that do not belong to the same object.

The estimated object individuation matrix **A** was used as an input to the partitioning algorithm to produce the final individuation shown in Figure 10.4.b. Each of the 5 partitions in the individuation corresponds to a set of trials that, according to the model, were performed with the same object. In this example, the model made one mistake as it incorrectly grouped one of the trials performed with the blue ball with the set of trials performed with the purple-yellow ball. The Normalized Mutual Information (NMI) between the output individuation and the ground truth individuation was 0.935. The α-Individuation Rate for $\alpha = 3$ and $\alpha = 4$ was 80.0% since in both cases there was one object (the first ball) that could not be individuated on its own. For $\alpha = 5$, the rate was 60.0% since only 3 of the objects were perfectly individuated (i.e., with all 5 trials in the same partition). To compare, when the perceptual similarity matrix **U** (see Figure 10.5) was used to partition the test trials the results were noticeably worse. The individuation had a substantially lower NMI of 0.809 and the α-individuation rate was only 20.0% for $\alpha = 3, 4$ and 5 (i.e., one partition contained 5 trials with a single object, while all others were mixed).

10.6.2 Baseline Comparison

The proposed individuation model was compared against the baseline unsupervised approach for partitioning the trials in the test set. During each test, the two approaches were evaluated using a randomly sampled set of 20 training objects and another randomly sampled set of 20 test objects, such that the two sets were disjoint. To compare against a chance model, the same experiment was performed with the added step of randomly shuffling the entries in the individuation matrix **A** before clustering it (i.e., multiple randomly chosen pairs of values in the matrix were swapped before using the matrix to compute the partitioning). Table 10.1 shows the results of these evaluations, averaged over 100 tests. Both the learned and the unsupervised models performed much better than chance. Furthermore, the superior performance of the learned model clearly shows that prior information, in the form of exploratory trials with known object identities, can substantially improve the robot's performance when individuating novel objects.

Table 10.1 Comparison between the learned individuation model, the baseline unsupervised model, and the chance model

	Normalized Mutual Information	Mean Partition Entropy	α-Individuation Rate (%)		
			$\alpha = 3$	$\alpha = 4$	$\alpha = 5$
Learned	0.964	0.056	87.1	74.5	71.5
Unsupervised	0.878	0.416	32.2	32.2	31.9
Random	0.506	1.373	0.0	0.0	0.0

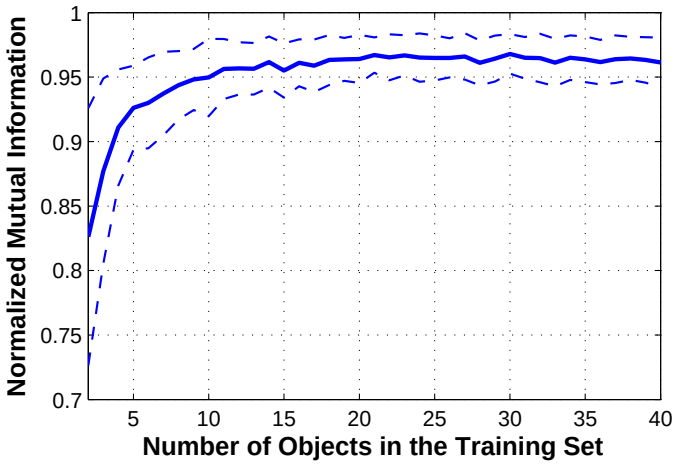


Figure 10.6 Performance of the robot's object individuation model, measured by the Normalized Mutual Information criterion, as a function of the number of objects used to train it. The dashed lines show the standard deviation, which was computed over 100 tests.

10.6.3 Performance vs. Number of Training Objects

The performance of the object individuation model was also evaluated as a function of the number of training objects, m, which was varied from 2 to 40. For each value, 100 tests were performed, such that during each test the model was evaluated using a randomly sampled set of m training objects and another randomly sampled set of 20 test objects.

The results of these tests, shown in Figure 10.6, indicate that the model's performance converges once there are at least 20 objects in the training set. Overall, even with a small number of training objects, the robot's model is able to successfully individuate novel objects substantially better than chance.

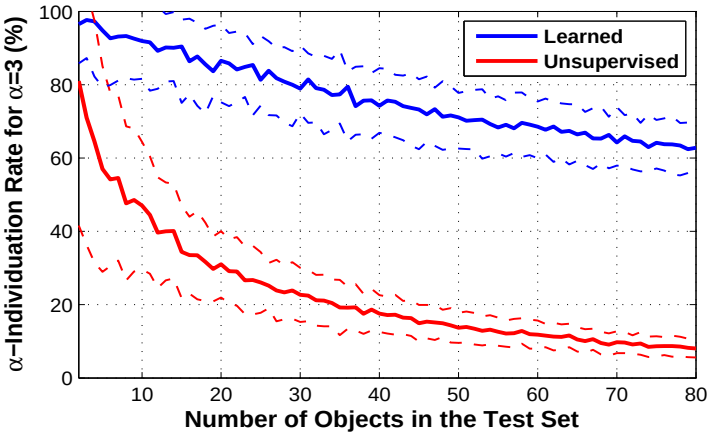


Figure 10.7 Performance of the learned individuation model and the baseline unsupervised model as a function of the number of objects in the test set.

10.6.4 Performance vs. Number of Test Objects

The last experiment explored the relationship between the number of objects in the test set and the performance of the robot's object individuation model. Studies in psychology have shown that there are inherent limits on the number of objects that humans can individuate at a time (Xu and Chun, 2009; Feigenson and Carey, 2005). To find out if the same is true for our robot, the number of objects in the test set was varied from 2 to 80, while the number of training objects was kept constant at 20.

Figure 10.7 shows the results of this experiment, where performance was measured using the Normalized Mutual Information measure. The results show that, just as with humans, the individuation task becomes more difficult as the number of test objects is increased. A possible explanation for this is that as the test set becomes larger, there are more pairs of perceptually similar objects that complicate the task. Nevertheless, even with a test set of 80 objects, the learned model was still able to successfully individuate over 60.0% of the objects. The unsupervised model, on the other hand, was able to individuate only 10% of the novel objects.

Figure 10.8 shows example object individuation and perceptual similarity matrices (**A** and **U**) for a test set of 400 exploratory trials with 80 objects (5 trials per object). As before, the entries in the matrices are sorted by object identity. Unlike the perceptual similarity matrix, the individuation matrix is sparse and has very few large values for pairs of trials that were

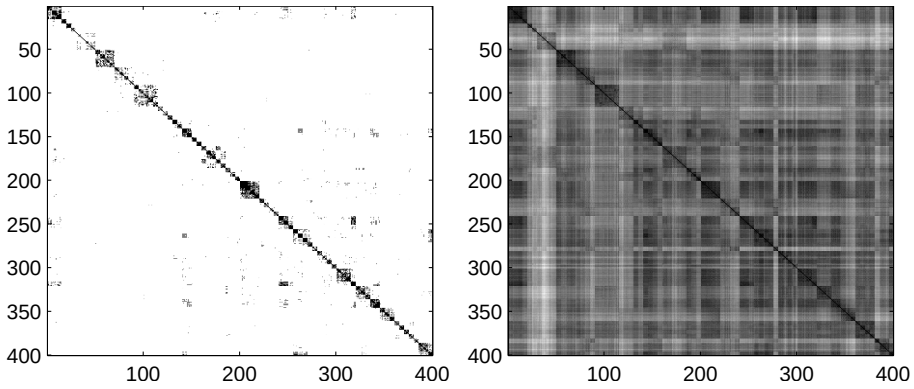


Figure 10.8 Example object individuation matrix, **A** (left), and perceptual similarity matrix, **U** (right), for a set 400 exploratory trials with 80 different objects (5 trials per object).

performed with two different objects. Furthermore, as shown in Figure 10.7, the performance of the unsupervised model drops at a much faster rate as the number of objects is increased, which showcases the need for prior training before attempting to individuate novel objects.

10.7 Conclusion and Future Work

While the problem of object recognition is well studied in robotics, the task of individuating novel objects that were not part of the robot's training set has received very little attention. To address this gap, this chapter proposed a method that allows a robot to successfully partition its sensorimotor experience with novel objects into clusters that correspond to the identities of the objects. The proposed method was tested with a large-scale dataset in which the robot explored 100 objects using a variety of exploratory behaviors and sensory modalities. Using prior information from exploratory trials for which the identities of the objects are known, the robot was able to achieve high performance on the task of object individuation as measured by several different performance measures.

A key result from this chapter is that unsupervised methods for partitioning of the robot's sensorimotor experience may not be sufficient for solving the object individuation problem. Instead, prior information, in the form of exploratory trials with known object identities, is needed in order to learn whether the observed perceptual differences between two sensorimotor

interactions are due to noise or due to the fact that the interactions were performed with two different objects. On average, the use of training data allowed the model to successfully individuate 87.1% of the objects in a test set of size 20, while, without it, the unsupervised model individuated only 32.2% of the 20 objects. Even with a larger test set of 80 objects, the learned model was able to individuate over 60% of the objects, while the model without prior training was able to individuate only 10% of the objects.

Another important result of this chapter is that, similar to studies with humans, performance was sensitive to the number of objects to be individuated. Therefore, one viable direction for future work is to explore ways of individuating a large number of objects by incrementally individuating smaller object subsets. Another direction for future work is to consider the effect of category and object labels on the individuation, since it has been shown that the presence of labels (i.e., words that describe the object) can improve the object individuation performance of human infants (Xu et al., 2005).

CHAPTER 11. CONCLUSION AND FUTURE WORK

This dissertation introduced a behavior-grounded object representation in which an object is represented by sensorimotor contingencies that span a diverse set of exploratory behaviors and sensory modalities. This object representation is inherently multi-modal and grounded in the robot's own sensorimotor experience. Results from several large-scale experiments with a humanoid robot showed that the proposed behavior-grounded framework can be used to solve a variety of problems that have typically been addressed only in the visual domain. Guided by findings in developmental psychology, this dissertation described an overarching framework that allows a robot to infer the identities of the objects that it interacts with, recognize them based on the sensory stimuli that they produce, group them according to multi-modal measures of perceptual similarity, and assign semantic category labels to individual objects as well as object relations.

The main contributions of this dissertation can be summarized as follows:

1. It develops a behavior-grounded framework that enables a robot to recognize objects by performing exploratory behaviors on them (Chapters 4 and 5).

2. It develops feature extraction methods that can be applied on a wide variety of sensory feedback coming from different sensory modalities (Chapters 4 and 8).

3. It demonstrates that sensorimotor interaction can be used to group objects according to their physical properties and human-provided labels using both unsupervised (Chapter 6) and supervised machine learning methods (Chapters 7 and 8).

4. It develops a novel framework that enables a robot not only to assign labels to individual objects, but also to detect relational categories that describe how objects relate to each other (Chapter 9).

5. It demonstrates a solution to the object individuation problem that enables a robot to infer the number of objects that it interacted with and group its sensorimotor data according to the estimated object identities (Chapter 10).

The next four sections describe in further detail how the studies described in this dissertation answered the research questions posed in Chapter 1.

11.1 Behavior-Grounded Object Recognition

Chapter 4 described the proposed framework for behavior-grounded object recognition. In a large-scale experiment, the robot explored 50 different objects using five different exploratory behaviors coupled with auditory and proprioceptive sensory stimuli. The feedback from the two sensory modalities, detected by the robot while interacting with an object, was represented as two sequences of the most highly activated nodes in two Self-Organizing Maps (one for each modality). Using global sequence comparison coupled with the k-Nearest Neighbors algorithm, the robot was able to recognize the explored object with accuracy substantially better than chance. The robot was also able to compute estimates for the reliability of each sensory modality and use them to improve its object recognition accuracy.

While object recognition has traditionally been viewed solely as a visual classification problem, this dissertation re-cast the problem as one that requires the use of exploratory behaviors coupled with different sensory modalities. Indeed, the results in Chapter 4 shows that even without the use of visual input, the robot's recognition accuracy reached 98.2% after applying all 5 exploratory behaviors on the test object. This gives a strong indication that traditional vision-based object recognition systems can be further improved by the additional use of auditory and proprioceptive feedback. This is particularly important for objects that may not be easily recognized using vision alone (e.g., a heavy and a light object that look identical). Thus, active interaction (as opposed to passive observation) is a necessary component for resolving perceptual ambiguities about objects.

In addition to the experiment described in Chapter 4, the study by Sinapov et al. (2011b) also demonstrated that by applying multiple different behaviors, a robot can also improve its

ability to discriminate between surface textures using vibrotactile feedback. To explain this improvement, Chapter 5 formulated a new metaphor, namely, *behaviors are classifiers*. In other words, the behavioral repertoire of the robot is an ensemble of classifiers, which can be boosted. The boosting effect generalizes not only to multiple exploratory behaviors, but also to multiple sensory modalities. Each new modality and each new behavior provides additional information that can be used to construct new classifiers.

Chapter 5 used two large datasets with 50 objects and 20 surfaces to generate the results, which clearly show that the metrics designed to measure the diversity of classifiers can be applied to measure the diversity of the behaviors in the robot's behavioral repertoire. In particular, the *disagreement measure* for two behavior-derived recognition models was found to be linearly related to the observed boost in recognition rate when both behaviors are applied. This is an important contribution as it establishes for the first time a link between empirical studies of exploratory behaviors in robotics and theoretical results on boosting in machine learning.

11.2 Grounding Object Categories in Behavioral Interactions

In addition to object recognition, this dissertation also explored novel methods that enable a robot to categorize objects using the sensory feedback produced during object exploration. Chapter 6 proposed an unsupervised method for solving the odd-one-out task, i.e., detecting which item does not belong in a given set. The experimental evaluation showed that the robot's choice for the odd object was consistent with human-defined object categories, with success rates varying from 45% to 100%, depending on the category. Certain behavior-modality combinations produced object similarity relations that were able to better capture the target category. These results show that sensorimotor interaction can capture many of the physical properties of objects that define an object category.

Chapter 7 showed that the same categories used in Chapter 6 can be explicitly learned by the robot using a graph-based learning approach. In contrast to traditional object classification methods that directly map visual object features to categories, the model presented here makes use of relational information that specifies how similar two objects are in a variety of

sensorimotor contexts. An important feature of our framework is its ability to simultaneously handle multiple robot behaviors, sensory modalities, and object attributes.

Chapter 8 described a method that scales to a much larger number of exploratory behaviors, sensory modalities, and objects than any previously published experiments in which robots have perceived objects by interacting with them. More specifically, in addition to doubling the number of objects, Chapter 8 also doubled the number of behaviors and more than tripled the number of sensorimotor contexts as compared to previous work.

The method was tested using a large-scale experiment in which the robot repeatedly interacted with 100 different objects from 20 object categories using 10 different behaviors (e.g., looking at the object, grasping it, shaking it, tapping it, etc.). The high recognition rates achieved by the robot (e.g., 97% using SVM) show that perceiving objects using a diverse set of behaviors and sensory modalities is crucial for scaling up object category recognition to a large number of objects and object categories. The model was also able to identify task-relevant sensorimotor contexts for a given categorization task, which allow a robot to learn what specific behaviors and sensory modalities are best for recognizing a specific category label in a novel object. Most importantly, by actively selecting which behavior to apply next, the model was able to reduce by half the exploration time required for classifying a new object. Finally, the robot's model was extended to detect if the test object does not belong to any of the known categories.

11.3 Beyond Simple Categories: Grounding Object Relations

Chapter 9 introduced a novel framework for object category learning that greatly increases the type of categories that can be learned by the robot as compared with our previous and related work. The proposed framework enables a robot not only to assign labels to individual objects, but also to detect relational categories that describe how objects relate to each other. The method was evaluated using a dataset in which the robot explored 36 different objects. Unlike the previous experiments, the objects in this dataset varied systematically according to their color, their weights, and their contents.

The robot learned to recognize individual object properties, such as their color, weight,

and contents. Furthermore, the robot learned to classify pairs of objects according to several labels such as "same color", "heavier than", etc. Finally, the robot also learned to recognize whether a group of objects varies by any of the three object properties. Thus, the proposed method could handle categories describing individual objects, categories describing pairwise object relationships, and categories that describe groups of objects.

In addition to achieving high recognition rates for all three types of categories, the robot was also able to establish a grounded measure of similarity between the different relational categories that it learned. More specifically, two categories were deemed similar if they could be recognized using the same behaviors and sensory modalities and dissimilar otherwise. Our results showed that this type of representation is especially useful when the robot is tasked with learning a new relational category that is similar to already known categories.

11.4 Behavior-Grounded Object Individuation

While the problem of object recognition is well studied in robotics, the task of individuating novel objects that were not part of the robot's training set has received very little attention. Because of this, most methods used by robots to recognize objects start with a fixed object representation in which the robot's training data is labeled with one of a finite number of object identities, i.e., they assume that the individuation problem has been solved. To address this gap, Chapter 10 proposed a method that allows a robot to successfully partition its sensorimotor experience with novel objects into clusters that correspond to the identities of the objects.

The proposed method was tested using the same large-scale dataset described in Chapter 8 in which the robot explored 100 objects using a variety of exploratory behaviors and sensory modalities. Inspired by research in developmental psychology, the robot learned an individuation model that was subsequently used to detect whether two distinct sensorimotor interactions were performed with the same object or with two different objects. Using prior information from exploratory trials for which the identities of the objects are known, the robot was able to achieve high performance on the task of object individuation as measured by several different performance measures.

A key result from Chapter 10 is that unsupervised methods for partitioning of the robot's

sensorimotor experience may not be sufficient for solving the object individuation problem. Instead, prior information, in the form of exploratory trials with known object identities, is needed in order to learn whether the observed perceptual differences between two sensorimotor interactions are due to noise or due to the fact that the interactions were performed with two different objects. On average, the use of training data allowed the model to successfully individuate 87.1% of the objects in a test set of size 20, while, without it, the unsupervised model individuated only 32.2% of the 20 objects. Even with a larger test set of 80 objects, the learned model was able to individuate over 60% of the objects, while the model without prior training was able to individuate only 10% of the objects.

11.5 Limitations

The research presented in this dissertation has several limitations. Some of them are due to the way in which the robot's behaviors are coded and represented in the experiments. For example, this research assumes that the robot can already interact with objects using a variety of exploratory behaviors. Furthermore, there is an assumption that all behaviors can be applied on all objects and that the objects are always placed on a tabletop, usually in a predefined location. Thus, this dissertation does not answer the question of where do these behaviors come from or how could a robot learn a new exploratory behavior. Both of these questions remain open and should be addressed in future work.

Another limitation of the methodology used in the studies described here is that each exploratory behavior was always applied on the same location of the object. In other words, the spatial relation between the robot's end effector and the object was assumed to be constant over multiple executions of the behavior on an object. As described in the next section, it would be highly desirable if the methodology can be extended to also take into account the spatial relation between the robot's hand and the object. Another general limitation of the methodology is that it requires the robot to exhaustively perform all behaviors on all objects, multiple different times. This was feasible to do in the lab with 100 objects, but it is probably impractical if the number of objects is scaled by one or more orders of magnitude, or if the robot is not under constant human supervision. While Chapter 8 provided a solution that

enables a robot to minimize behavioral exploration when classifying a novel object, the wider problem of minimizing exploration time remains open.

Finally, there are several limitations due to the sensory processing methods used by the robot to represent sensorimotor feedback that is produced over the course of an interaction. While the feature extraction routines used in Chapters 8, 9, and 10 were relatively simple, the resulting features required similar timing across multiple executions of the same behavior to be useful. While some of these routines could be adapted to other sensory modalities, many were specifically designed for a given sensory signal. Therefore, there is still a great need for general methods that can represent a wider set of modalities, including modalities that may not necessarily be known to the programmer in advance.

11.6 Future Work

A longstanding goal of the line of research presented in this dissertation is to enable robots to effectively use exploratory behaviors when learning about the objects in their environment. While the studies described here make small steps in that direction, there are still several open problems that should be addressed in future work.

Representing Space and Spatial Relations – As discussed in the previous section, there are several limitations due to the exploratory behaviors used in this study. The current methodology does not allow the robot to learn that a specific behavior should be applied at a specific object feature. Therefore, a direct line for future work is to extend the representation of the behaviors so that they are not only discrete entities, but also capture spatial relations and object geometry. Some early simulation results, which were not included in this dissertation, demonstrate that spatial frames of references can be part of that representation (Sinapov and Stoytchev, 2007, 2008). Enabling robots to perform a variety of actuating behaviors on a wider set of articulated objects also remains an open problem.

Alternative Methods for Sensory Processing – This dissertation proposed several methods for representing sensory feedback signals from a range of sensory modalities. Some of

those methods, such as the Self-Organizing Map representation described in Chapter 4, can be applied to a variety of sensory modalities (e.g., audio and proprioception), while others, such as the SURF and Optical Flow features described in Chapter 8 were specifically designed to represent visual feedback. Therefore, there is still a great need for methods that can represent a wider set of modalities, including modalities that may not necessarily be known to the programmer in advance.

Grounded Language Learning – Learning language by pairing words and sentences with percepts is a long standing problem in Artificial Intelligence. I believe the research described in this dissertation has many implications that can drive future work in embodied language acquisition. The experiments described here showed that a robot may acquire embodied representations of a large set of concepts (e.g., nouns and adjectives that describe objects, object pairs, and groups of objects). Furthermore, the methodology overcomes the two main limitations of grounded language learning systems, i.e., that they're often disembodied and purely vision-based. Nevertheless, the research stopped short of language learning as the target concepts were presented as discrete entities rather than structured sentences. This limitation may be overcome in future work by providing the robot with a narrated description of the objects and events that occur during a given sensorimotor interaction.

BIBLIOGRAPHY

W. Aha, D. Kibler, and M. Albert. Instance-based learning algorithm. *Machine Learning*, 6: 37–66, 1991.

T. Araki, T. Nakamura, T. Nagai, K. Funakoshi, M. Nakano, and N. Iwahashi. Autonomous acquisition of multimodal information for online object concept formation by robots. In *Proc. of the 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1540–1547, 2011.

R. Arkin. Motor schema based navigation for a mobile robot: An approach to programming by behavior. In *Proc. of the IEEE International Conference on Robotics and Automation*, pages 264–271, 1987.

F. Ashby and W. Maddox. Human category learning. *Psychology*, 56(1):149, 2005.

C. Atkeson, C. An, and J. Hollerbach. Estimation of inertial parameters of manipulator loads and links. *The International Journal of Robotics Research*, 5(3):101–119, 1986.

C. Atkeson, A. Moore, and S. Schaal. Locally weighted learning. *Artificial Intelligence Review*, 11(1-5):11–73, 1997.

H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-up robust features (SURF). *Computer Vision and Image Understanding*, 110(3):346–359, 2008.

C. Becchio and C. Bertone. Object temporal connotation. *Brain and cognition*, 52(2):192–196, 2003.

T. Bergquist, C. Schenck, U. Ohiri, G. S. Sinapov, J., and A. Stoytchev. Interactive object

recognition using proprioceptive feedback. In *Proceedings of the 2009 IROS Workshop: Semantic Perception for Robot Manipulation, St. Louis, MO*, 2009.

T. Bhattacharjee, J. Rehg, and C. Kemp. Haptic classification and recognition of objects using a tactile sensing forearm. In *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2012.

P. Bloom. *How children learn the meanings of words: Learning, development and conceptual change*. MIT Press, Cambridge, MA., 2000.

A. Booth and S. Waxman. Object names and object functions serve as cues to categories for infants. *Developmental Psychology*, 38(6):948, 2002.

M. Boutell, J. Luo, X. Shen, and C. Brown. Learning multi-label scene classification. *Pattern recognition*, 37(9):1757–1771, 2004.

R. Brooks. A robust layered control system for a mobile robot. *J. of Robotics and Automation*, 2(1):14–23, 1986.

M. Buckley, M. Booth, E. Rolls, and D. Gaffan. Selective perceptual impairments after perirhinal cortex ablation. *Journal of Neuroscience*, 21(24):9824, 2001.

C. J. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2:121–167, 1998.

M. Casasola and L. Cohen. Infant categorization of containment, support and tight-fit spatial relationships. *Developmental Science*, 5(2):247–264, 2002.

M. Casasola, L. Cohen, and E. Chiarello. Six-month-old infants' categorization of containment spatial relations. *Child development*, 74(3):679–693, 2003.

A. Chan and E. Pampalk. Growing hierarchical self organizing map (ghsom) toolbox: visualizations and enhancements. In *Proc. of the 9th Intl. Conf. on Neural Information Processing (NIPS)*, pages 2537–2541, 2002.

O. Chapelle, P. Haffner, and V. Vapnik. Support vector machines for histogram-based image classication. *IEEE Transactions of Neural Networks*, 10:1055–1064, 1999.

S. Chitta, J. Sturm, M. Piccoli, and W. Burgard. Tactile sensing for mobile manipulation. *IEEE Transactions on Robotics*, (99):1–11, 2011.

V. Chu, I. Mcmahon, L. Riano, C. G. Mcdonald, Q. He, J. M. Perez-tejada, M. Arrigo, N. Fitter, J. C. Nappo, T. Darrell, and K. Kuchenbecker. Using robotic exploratory procedures to learn the meaning of haptic adjectives. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2013.

J. Cohen. A Coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

J. Colombo, K. McCollam, J. Coldren, D. Mitchell, and S. Rash. Form categorization in 10-month-olds. *Journal of Experimental Child Psychology*, 49(2):173–188, 1990.

N. Dag, I. Atil, S. Kalkan, and E. Sahin. Learning affordances for categorizing objects and their properties. In *2010 IEEE International Conference on Pattern Recognition*, pages 3089–3092, 2010.

H. Dang and P. Allen. Robot learning of everyday object manipulations via human demonstration. In *Proc. of the 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1284–1289, 2010.

N. Davidson and S. Gelman. Inductions from novel categories: The role of language and conceptual structure. *Cognitive development*, 5(2):151–176, 1990.

T. Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting and randomization. *Machine Learning*, 40(2):139–157, 2000.

L. Dragone. Spectral clusterer for WEKA (Retrieved Jan. 2009). `www.luigidragone.com/software/spectral-clusterer-for-weka/`, 2006.

P. Eimas and P. Quinn. Studies on the formation of perceptually based basic-level categories in young infants. *Child development*, 65(3):903–917, 1994.

F. Endres, C. Plagemann, C. amd Stachniss, and W. Burgand. Unsupervised discovery of object classes from range data using latent Dirichlet allocation. In *Procceedings of Robotics: Science and Systems*, pages 113–120, 2009.

A. Erkan, O. Kroemer, R. Detry, Y. Altun, J. Piater, and J. Peters. Learning probabilistic discriminative models of grasp affordances under limited supervision. In *Proc. of the 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1586–1591, 2010.

M. Ernst and H. Bulthof. Merging the Senses into a Robust Percept. *Trends in Cognitive Science*, 8(4):162–169, 2004.

L. Feigenson and S. Carey. On the limits of infants' quantification of small object arrays. *Cognition*, 97(3):295–313, 2005.

R. Fergus, P. Perona, and A. Zisserman. A visual category filter for google images. *Computer Vision-ECCV*, pages 242–256, 2004.

J. Fishel and G. Loeb. Bayesian exploration for intelligent identification of textures. *Frontiers in Neurorobotics*, 6, 2012.

P. Fitzpatrick, G. Metta, L. Natale, S. Rao, and G. Sandini. Learning about objects through action - initial steps towards artificial cognition. In *Proceedings of the 2003 IEEE International Conference on Robotics and Automation*, 2003.

Y. Freund and R. E. Schapire. Experiments with a new boosting algorithm. In *Proc. of the Thirteenth International Conference on Machine Learning*, pages 148–156, San Francisco, 1996.

A. Fulkerson and S. Waxman. Words (but not tones) facilitate object categorization: Evidence from 6-and 12-month-olds. *Cognition*, 105(1):218–228, 2007.

T. Gao and D. Koller. Active classification based on value of classifier. In *Advances in Neural Information Processing Systems (NIPS 2011)*, 2011.

W. Gaver. What in the world do we hear? An ecological approach to auditory event perception. *Ecological Psych.*, 5:1–29, 1993.

D. Gentner and L. Namy. Analogical processes in language learning. *Current Directions in Psychological Science*, 15(6):297, 2006.

L. Getoor and C. P. Diehl. Link mining: a survey. *ACM SIGKDD Explorations Newsletter*, 7 (2):3–12, 2005.

E. J. Gibson. Exploratory behavior in the development of perceiving, acting, and the acquiring of knowledge. *Annual Review of Psychology*, 39:1–41, 1988.

A. Gijsberts, T. Tommasi, G. Metta, and B. Caputo. Object recognition using visuo-affordance maps. In *Proc. of the 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1572–1578, 2010.

B. Giordano and S. McAdams. Material identification of real impact sounds: Effects of size variation in steel, glass, wood, and plexiglass plates. *J. of the Acoustical Soc. of America*, 119(2):1171–81, 2006.

M. Grassi. Do we hear size or sound? Balls dropped on plates. *Perception and Psychophysics*, 67(2):274–284, 2005.

S. Griffith, J. Sinapov, M. Miller, and A. Stoytchev. Toward interactive learning of object categories by a robot: A case study with container and non-container objects. In *Proceedings of the 8th IEEE International Conference on Development and Learning (ICDL)*, pages 1–6. IEEE, 2009.

S. Griffith, J. Sinapov, V. Sukhoy, and A. Stoytchev. A behavior–grounded approach to forming object categories: Separating containers from non-containers. *IEEE Transactions on Autonomous Mental Development*, 4(1):54–69, 2012.

R. Hammer, G. Diesendruck, D. Weinshall, and S. Hochstein. The development of category learning strategies: What makes the difference? *Cognition*, 112(1):105–119, 2009.

R. Hammer, A. Brechmann, F. Ohl, D. Weinshall, and S. Hochstein. Differential category learning processes: The neural basis of comparison-based learning and induction. *NeuroImage*, 52(2):699–709, 2010.

T. Hastie and R. Tibshirani. Classification by pairwise couplling. *Advances in Neural Information Processing Systems*, 10, 1998.

P. Hauf and G. Aschersleben. Action–effect anticipation in infant action control. *Psychological Research*, 72(2):203–210, 2008.

P. Hauf, G. Aschersleben, and W. Prinz. Baby do-baby see! how action production influences action perception in infants. *Cognitive Development*, 22(1):16–32, 2007.

M. Heller. Haptic dominance in form perception: vision versus proprioception. *Perception*, 21 (5):655–660, 1992.

V. Hodge and J. Austin. A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22(2):85–126, 2004.

J. Hollerbach and C. Wampler. The calibration index and taxonomy for robot kinematic calibration methods. *The International Journal of Robotics Research*, 15(6):573–591, 1996.

J. Horst, L. Oakes, and K. Madole. What does it look like and what can it do? category structure influences how infants categorize. *Child Development*, 76(3):614–631, 2005.

K. Hosoda, Y. Tada, and M. Asada. Anthropomorphic robotic soft fingertip with randomly distributed receptors. *Robotics and Autonomous Systems*, 54(2):104 – 109, 2006.

D. Hume. *An enquiry concerning human understanding: a critical edition*, volume 3. Repreinted in 2000 by Oxford University Press, USA, 1776.

N. Jamali and C. Sammut. Material classification by tactile sensing using surface textures. In *Proc. of the 2010 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2336 –2341, May 2010.

S. Johnson, D. Amso, and J. Slemmer. Development of object concepts in infancy: Evidence for early learning in an eye-tracking paradigm. *Proceedings of the National Academy of Sciences*, 100(18):10568, 2003.

H. Kang, M. Hebert, A. Efros, and T. Kanade. Connecting missing links: Object discovery from sparse observations using 5 million product images. In *Proc. of the European Conference on Computer Vision*, 2012.

I. Kant. *Critique of pure reason*. Macmillan Education Limited, London, 1781.

D. Katz and O. Brock. Manipulating articulated objects with interactive perception. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 272–277, 2008.

C. Kemp, A. Edsinger, and E. Torres-Jara. Challenges for robot manipulation in human environments [grand challenges of robotics]. *IEEE Robotics & Automation Magazine*, 14(1): 20–29, 2007.

C. Kemp, A. Jern, and F. Xu. Object discovery and identification. *Advances in Neural Information Processing Systems*, 22, 2009.

C. Kemp, K. Chang, and L. Lombardi. Category and feature identification. *Acta psychologica*, 133(3):216–233, 2010. ISSN 0001-6918.

R. L. Klatzky, D. K. Pai, and E. P. Krotkov. Perception of material from contact sounds. *Presence: Teleoperators & Virtual Environments*, 9(4):339–410, 2000.

T. Kohonen. *Self-Organizing Maps*. Springer-Verlag, New York, 2001.

M. Krabbes and C. Döschner. Modelling of Robot Dynamics Bsed on Multi-Dimensional RBF-Like Neural Network. In *Proc. of Intl. Conf. on Information Intelligence and Systems (ICIIS)*, pages 180–187, 1999.

K. Kraebel and P. Gerhardstein. Three-month-old infants object recognition across changes in viewpoint using an operant learning procedure. *Infant Behavior and Development*, 29(1): 11–23, 2006.

P. Krojgaard. A review of object individuation in infancy. *British Journal of Developmental Psychology*, 22(2):159–183, 2004.

E. Krotkov. Perception of material properties by robotic probing: Preliminary investigations. In *In Proceedings of the International Joint Conference on Artificial Intelligence*, pages 88–94, August 1995.

E. Krotkov, R. Klatzky, and N. Zumel. Robotic perception of material: Experiments with shape-invariant acoustic measures of material type. In *Experimental Robotics IV*, volume 223 of *Lecture Notes in Control and Information Sciences*, pages 204–211. Springer Berlin, 1996.

D. Kubus and F. Wahl. Estimating Inertial Load Parameters Using Force/Torque and Acceleration Sensor Fussion. In *Robotic 2008, VDI-Beritche 2012 Munchen, Germany*, pages 29–32, 2008.

D. Kubus, T. Kroger, and F. Wahl. On-line rigid object recognition and pose estimation based on inertial parameters. In *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, pages 1402–1408, 2007.

L. Kuncheva and C. Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51(2):181–207, 2003.

K. Lai and D. Fox. 3D laser scan classification using web data and domain adaptation. In *Proc. of Robotics: Science and Systems (RSS)*, 2009.

K. Lai, L. Bo, X. Ren, and D. Fox. A large-scale hierarchical multi-view rgb-d object dataset. In *Proc. of the IEEE International Conference on Robotics & Automation (ICRA)*, 2011a.

K. Lai, L. Bo, X. Ren, and D. Fox. Sparse distance learning for object recognition combining RGB and depth information. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 4007–4013, 2011b.

L. Lam. Classifier combinations: implementations and theoretical issues. *Multiple Classifier Systems*, pages 77–86, 2000.

L. Lam and C. Suen. Optimal combinations of pattern classifiers. *Pattern Recognition Letters*, 16(9):945–954, 1995.

R. Landis and G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

S. Lederman. Chapter 4. the perception of texture by touch. In W. Schiff and E. Foulke, editors, *Tactual perception: A sourcebook*, pages 130–168. Cambridge Univ Press, 1982.

S. Lederman and R. Klatzky. Hand movements: A window into haptic object recognition. *Cognitive psychology*, 19(3):342–368, 1987.

S. Lederman and R. Klatzky. Haptic classification of common objects: knowledge-driven exploration. *Cognitive Psychology*, 22:421–459, 1990.

K. Lee, H. Hon, and R. Reddy. An overview of the SPHINX speech recognition system. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38(1):35–45, 1990.

A. Leonardis and S. Fidler. Learning hierarchical representations of object categories for robot vision. *Robotics Research*, pages 99–110, 2011.

J. Locke. *An essay concerning human understanding*. Collins, London, 1690. reprinted in 1847.

L. Lopes and A. Chauhan. Scaling up category learning for language acquisition in human-robot interaction. In *Proceedings of the Symposium on Language and Robots*, pages 83–92, 2007.

K. Lorenz. *Learning as Self-Organization*, chapter Innate bases of learning. Mahwah, NJ: Lawrence Erlbaum and Associates, Publishers, 1996.

Y. Luo, L. Kaufman, and R. Baillargeon. Young infants' reasoning about physical events involving inert and self-propelled objects. *Cognitive psychology*, 58(4):441–486, 2009.

A. R. Luria. *Cognitive development, its cultural and social foundations*. Harvard University Press, 1976.

D. Lynott and L. Connell. Modality Exclusivity Norms for 423 Object Properties. *Behavior Research Methods*, 41(2):558–564, 2009.

Z. Marton, R. Rusu, D. Jain, U. Klank, and M. Beetz. Probabilistic categorization of kitchen objects in table settings with a composite sensor. In *Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on*, pages 4777–4784. IEEE, 2009.

Z. C. Marton, F. Seidel, F. Balint-Benczedi, and M. Beetz. Ensembles of Strong Learners for Multi-cue Classification. *Pattern Recognition Letters (PRL), Special Issue on Scene Understandings and Behaviours Analysis*, 34:754–761, 2012.

C. Mash, M. Arterberry, and M. Bornstein. Mechanisms of visual object recognition in infancy: Five-month-olds generalize beyond the interpolation of familiar views. *Infancy*, 12(1):31–43, 2007.

M. Matarić. Designing emergent behaviors: From local interactions to collective intelligence. In *Proc of the International Conference on Simulation of Adaptive Behavior*, pages 432–441, 1992.

A. Meltzoff and M. Moore. Object representation, identity, and the paradox of early permanence: Steps toward a new framework. *Infant Behavior and Development*, 21(2):201–235, 1998.

G. Metta and P. Fitzpatrick. Early integration of vision and manipulation. *Adaptive Behavior*, 11(2):109–128, 2003.

J. Modayil and B. Kuipers. The initial development of object knowledge by a learning robot. *Robotics and Autonomous Systems*, 56(11):879–890, 2008.

L. Morency, C. Sidner, C. Lee, and T. Darrell. Contextual recognition of head gestures. In *Proceedings of the 7th international conference on Multimodal interfaces*, pages 18–24. ACM, 2005.

T. Nakamura, T. Nagai, and N. Iwahashi. Multimodal object categorization by a robot. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2415–2420, 2007.

T. Nanayakkara, K. Watanabe, and K. Izumi. Evolving Runge-Kutta-Gill RBF Networks to Estimate the Dynamics of a Multi-Link Manipulator. In *Proc. of Systems, Man, and Cybernetics*, pages 770–775, 1999.

L. Natale, G. Metta, and G. Sandini. Learning haptic representation of objects. In *Proceedings of the International Conference on Intelligent Manipulation and Grasping*, 2004.

G. Navarro. A guided tour to approximate string matching. *ACM Computing Surveys*, 33(1): 31–88, 2001.

S. Needleman and C. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, 48(3):443–453, 1970.

K. Nelson. Some Evidence for the Cognitive Primacy of Categorization and Its Functional Basis. *Merrill-Palmer Quarterly*, 19(1):21–39, 1973.

S. Nolfi and D. Marocco. Active perception: A sensorimotor account of object categorization. In *From Animals to Animats 7: Proc. of the Sixth International Conf. on Simulation of Adaptive Behavior*, 2002.

D. Norman. *The Design of Everyday Things.* Doubleday, 1988.

A. Opelt, A. Pinz, M. Fussenegger, and P. Auer. Generic object recognition with boosting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 416–431, 2006.

P. Oudeyer and F. Kaplan. What is intrinsic motivation? a typology of computational approaches. *Frontiers in Neurorobotics*, 1, 2007.

M. Paulus and P. Hauf. Infants' use of material properties to guide their actions with differently weighted objects. *Infant and Child Development*, 20(4):423–436, 2011.

D. Pelleg and A. W. Moore. X-means: Extending k-means with efficient estimation of the number of clusters. In *17th Int. Conf. on Machine Learning*, pages 727–734, 2000.

R. Peters, O. Jenkins, and R. Bodenheimer. Sensory-Motor Manifold Structure Induced by Task Outcome: Experiments with Robonaut. In *Proc. of IEEE Intl. Conf. on Humanoid Robots*, pages 484–489, 2006.

J. Piaget. *The origins of intelligence in the child*. Norton, New York, 1952.

K. Plunkett, J. Hu, and L. Cohen. Labels can override perceptual categories in early infancy. *Cognition*, 106(2):665–681, 2008.

J. Ponce. *Toward category-level object recognition*, volume 4170. Springer-Verlag New York Inc, 2006.

P. Potì. Logical structures of young chimpanzees' spontaneous object grouping. *International Journal of Primatology*, 18(1):33–59, 1997.

T. G. Power. *Play and Exploration in Children and Animals*. Lawrence Erlbaum Associates, Publishers, Mahwah, NJ, 2000.

M. Quigley, E. Berger, and A. Ng. STAIR: Hardware and software architecture. *Presented at AAAI 2007 Robotics Workshop*, 2007.

W. Quine. *Philosophy of logic*. Harvard University Press, 1986.

J. R. Quinlan. *C4.5: programs for machine learning*. 1993.

D. Rakison and G. Butterworth. Infants' attention to object structure in early categorization. *Developmental Psychology; Developmental Psychology*, 34(6):1310, 1998.

B. Rasolzadeh, M. Björkman, K. Huebner, and D. Kragic. An active vision system for detecting, fixating and manipulating objects in the real world. *International Journal of Robotics Research*, 29(2-3):133–154, 2010.

A. Rebguns, D. Ford, and I. Fasel. Infomax control for acoustic exploration of objects by a mobile robot. In *Lifelong Learning: Papers from the 2011 AAAI Workshop*, 2011.

J. Richmond. Automatic measurement and modelling of contact sounds. Master's thesis, University of British Columbia, 2000.

J. Richmond and D. Pai. Active measurement of contact sounds. In *Proc. of ICRA*, pages 2146–2152, 2000.

D. Roberson, J. Davidoff, and N. Braisby. Similarity and categorisation: Neuropsychological evidence for a dissociation in explicit categorisation tasks. *Cognition*, 71(1):1–42, 1999.

P. Rochat. Object manipulation and exploration in 2-to 5-month-old infants. *Developmental Psychology*, 25(6):871, 1989.

B. Rooks. The harmonious robot. *Industrial Robot: An International Journal*, 33(2):125–130, 2006.

H. A. Ruff. The development of perception and recognition of objects. *Child development*, pages 981–992, 1980.

H. A. Ruff. Infants' manipulative exploration of objects: Effects of age and object characteristics. *Developmental Psychology*, 20(1):9–20, 1984.

R. B. Rusu, Z. C. Marton, N. Blodow, M. Dolha, and M. Beetz. Towards 3D point cloud based object maps for household environments. *Robotics and Autonomous Systems*, 56(11):927 – 941, 2008.

H. Saal, J. Ting, and S. Vijayakumar. Active sequential learning with tactile feedback. In *Proc. 13th Int. Conf. on Artificial Intelligence and Statistics (AISTATS)*, 2010.

K. Saenko and T. Darrell. Object category recognition using probabilistic fusion of speech and image classifiers. In *Proceedings of the 4th international conference on Machine learning for multimodal interaction*, pages 36–47. Springer-Verlag, 2007.

F. Sapp, K. Lee, and D. Muir. Three-year-olds' difficulty with the appearance-reality distinction. *Developmental Psychology*, 36(5):547–60, 2000.

A. Saxena, J. Driemeyer, and A. Ng. Robotic grasping of novel objects using vision. *The International Journal of Robotics Research*, 27(2):157, 2008.

R. Schapire and Y. Singer. Boostexter: A boosting-based system for text categorization. *Machine learning*, 39(2):135–168, 2000.

C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27: 379–423, 1948.

J. Shi and J. Malik. Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

J. Sinapov and A. Stoytchev. Learning and generalization of behavior-grounded tool affordances. In *Proc. of the 7th IEEE Int. Conf. on Development and Learning (ICDL)*, 2007.

J. Sinapov and A. Stoytchev. Detecting the functional similarities between tools using a hierarchical representation of outcomes. In *Proc. of the 7th IEEE International Conference on Development and Learning (ICDL)*, pages 91–96, 2008.

J. Sinapov and A. Stoytchev. From acoustic object recognition to object categorization by a humanoid robot. In *Proc. of the RSS 2009 Workshop on Mobile Manipulation, Seattle, WA.*, 2009.

J. Sinapov and A. Stoytchev. The boosting effect of exploratory behaviors. In *Proc. National Conferrence on Artificial Intelligence (AAAI)*, pages 1613–1618, 2010a.

J. Sinapov and A. Stoytchev. The Odd One Out Task: Toward an Intelligence Test for Robots. In *Proceedings of the 8th IEEE International Conference on Development and Learning (ICDL)*, pages 126–131, 2010b.

J. Sinapov and A. Stoytchev. Object category recognition by a humanoid robot using behavior-grounded relational learning. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 184–190, 2011.

J. Sinapov, M. Weimer, and A. Stoytchev. Interactive learning of the acoustic properties of objects by a robot. In *Proceedings of the RSS Workshop on Robot Manipulation: Intelligence in Human Environments, Zurich, Switzerland*, 2008.

J. Sinapov, M. Weimer, and A. Stoytchev. Interactive learning of the acoustic properties of household objects. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 2518–2524, 2009.

J. Sinapov, T. Bergquist, C. Schenck, G. S. Ohiri, U., and A. Stoytchev. Interactive object recognition using proprioceptive and auditory feedback. *International Journal of Robotics Research*, 30(10):1250–1262, 2011a.

J. Sinapov, V. Sukhoy, R. Sahai, and A. Stoytchev. Vibrotactile recognition and categorization of surfaces by a humanoid robot. *IEEE Transactions on Robotics*, 27(3):488–497, 2011b.

J. Sinapov, C. Schenck, K. Staley, V. Sukhoy, and A. Stoytchev. Grounding semantic categories in behavioral interactions: Experiments with 100 objects. *Robotics and Autonomous Systems (in press)*, 2012.

D. Skalak. The sources of increased accuracy for two proposed boosting algorithms. In *Proc. American Association for Artificial Intelligence, AAAI-96, Integrating Multiple Learned Models Workshop*, 1996.

A. Sloman. Why symbol-grounding is both impossible and unnecessary. `www.cs.bham.ac.uk/research/projects/cogaff/misc/talks/models.pdf`, 2008. (Retrieved Oct. 2012).

A. Sloman and J. Chappell. The altricial-precocial spectrum for robots. In *International Joint Conference on Artificial Intelligence*, volume 19, pages 1187–1193, 2005.

L. Smith, N. Cooney, and C. McCord. What Is "High"? The Development of Reference Points for "High" and "Low". *Child Development*, 57(3):583–602, 1986.

M. Snowling, C. Hulme, A. Smith, and J. Thomas. The effects of phonetic similarity and list length on children's sound categorization performance. *Journal of Experimental Child Psychology*, 58(1):160, 1994.

K. Soska, K. Adolph, and S. Johnson. Systems in development: motor skill acquisition facilitates three-dimensional object completion. *Developmental psychology*, 46(1):129, 2010.

E. Spelke and K. Kinzler. Core knowledge. *Developmental science*, 10(1):89–96, 2007.

G. Spinozzi, F. Natale, J. Langer, and K. Brakke. Spontaneous class grouping behavior by bonobos (Pan paniscus) and common chimpanzees (P. troglodytes). *Animal Cognition*, 2(3): 157–170, 1999.

S. Srinivasa, C. Ferguson, D. Helfrich, D. Berenson, A. Collet, R. Diankov, G. Gallagher, G. Hollinger, J. Kuffner, and M. VandeWeghe. Herb: A Home Exploring Robotic Butler. *Autonomous Robots*, 28(1):5–20, 2009.

D. Stack and M. Tsonis. Infants haptic perception of texture in the presence and absence of visual cues. *British journal of developmental psychology*, 17(1):97–110, 1999.

D. Starkey. The origins of concept formation: Object sorting and object preference in early infancy. *Child Development*, 52(2):489–497, 1981.

R. Stephens and D. Navarro. One of These Greebles is Not Like the Others: Semi-Supervised Models for Similarity Structures. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, pages 1996–2001, 2008.

A. Stoytchev. Behavior-grounded representation of tool affordances. In *Proc. of IEEE Int. Conf. on Robotics and Automation (ICRA)*, pages 3071–3076, 2005.

A. Stoytchev. Some basic principles of developmental robotics. *IEEE Transactions on Autonomous Mental Development*, 1(2):122 –130, 2009.

A. Strehl and J. Ghosh. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *The Journal of Machine Learning Research*, 3:583–617, 2003.

M. Sugiyama and M. Kawanabe. *Machine Learning in Non-Stationary Environments*. MIT Press, Cambridge, MA, 2012.

V. Sukhoy, J. Sinapov, L. Wu, and A. Stoytchev. Learning to press doorbell buttons. In *Development and Learning (ICDL), 2010 IEEE 9th International Conference on*, pages 132–139, 2010.

D. Sun, S. Roth, and M. Black. Secrets of optical flow estimation and their principles. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2432 –2439, 2010a.

J. Sun, J. Moore, A. Bobick, and J. Rehg. Learning Visual Object Categories for Robot Affordance Prediction. *The International Journal of Robotics Research*, 29(2-3):174–197, 2010b.

R. Sutton. Verification, the Key to AI (Retrieved Oct. 2011). `http://www.cs.ualberta.ca/~sutton/IncIdeas/KeytoAI.html`, 2001.

S. Takamuku, K. Hosoda, and M. Asada. Shaking eases object category acquisition: Experiments with a robot arm. In *Proceedings of the Seventh International Conference on Epigenetic Robotics*, 2007.

S. Takamuku, K. Hosoda, and M. Asada. Object category acquisition by dynamic touch. *Advanced Robotics*, 22(10):1143–1154, 2008.

M. Tanaka, J. Levequem, H. Tagami, K. Kikuchi, and S. Chonan. The "haptic finger" a new device for monitoring skin condition. *Skin Research and Technology*, 9(1):131–136, 2003.

Y. Tanaka, M. Tanaka, and S. Chonan. Development of a sensor system for collecting tactile information. *Microsyst. Technol.*, 13:1005–1013, April 2007.

J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

E. Torres-Jara, L. Natale, and P. Fitzpatrick. Tapping into touch. In *Proc. 5-th Intl. Workshop on Epigenetic Robotics*, pages 79–86, 2005.

P. Tremoulet, A. Leslie, and D. Hall. Infant individuation and identification of objects. *Cognitive Development*, 15(4):499–522, 2000.

G. Tsoumakas and I. Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13, 2007.

G. Van de Walle, S. Carey, and M. Prevor. Bases for object individuation in infancy: Evidence from manual search. *Journal of Cognition and Development*, 1(3):249–280, 2000.

V. Vapnik. *Statistical Learning Theory*. Springer-Verlag, New York, 1998.

D. Vernon. Cognitive vision: The case for embodied perception. *Image and Vision Computing*, 26(1):127–140, 2008.

C. Von Hofsten, O. Kochukhova, and K. Rosander. Predictive tracking over occlusions by 4-month-old infants. *Developmental Science*, 10(5):625–640, 2007.

U. von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

T. Wilcox and R. Baillargeon. Object individuation in infancy: The use of featural information in reasoning about occlusion events. *Cognitive Psychology*, 37(2):97–155, 1998.

T. Wilcox, R. Woods, L. Tuggy, and R. Napoli. Shake, rattle, and one or two objects? young infants' use of auditory information to individuate objects. *Infancy*, 9(1):97–123, 2006.

T. Wilcox, R. Woods, C. Chapa, and S. McCurry. Multisensory exploration and object individuation in infancy. *Developmental psychology*, 43(2):479, 2007.

I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufman, San Francisco, 2nd edition, 2005.

W. Wohlkinger and M. Vincze. 3D object classification for mobile robots in home-environments using web-data. In *IEEE 19th International Workshop on Robotics in Alpe-Adria-Danube Region (RAAD)*, pages 247–252, 2010.

F. Xu, M. Cote, and A. Baker. Labeling guides object individuation in 12-month-old infants. *Psychological Science*, 16(5):372–377, 2005.

Y. Xu and M. Chun. Selecting and perceiving multiple visual objects. *Trends in cognitive sciences*, 13(4):167–174, 2009.

G. Yule. On the association of attributes in statistics. *Phil. Trans., A*, 194:257–319, 1900.

O. Yürüten, K. F. Uyanık, Y. Çalışkan, A. K. Bozcuoğlu, E. Şahin, and S. Kalkan. Learning adjectives and nouns from affordances on the icub humanoid robot. In *From Animals to Animats 12*, pages 330–340. Springer, 2012.

S. Zhou. Trace and determinant kernels between matrices. *Neural Information Processing Systems (NIPS)*, 2004.

Z. Zhou, D. Zhan, and Q. Yang. Semi-supervised learning with very few labeled training examples. In *Proceedings of the national conference on artificial intelligence*, volume 22, page 675, 2007.

X. Zhu, Z. Ghahramani, and T. J. Mit. Semi-supervised learning with graphs. Technical report, Carnegie Mellon University, 2005.