Behavior-Grounded Object Identification, Grouping and Ordering by a Humanoid Robot

Connor Schenck, Jivko Sinapov, Kerrick Staley
Developmental Robotics Laboratory
Iowa State University
{cschenck, jsinapov, kerrick}@iastate.edu

***Abstract*—From an early stage in development, infants show a profound drive to explore the objects around them. Research in psychology has shown that in doing so, they solve a vast array of problems, including the formation and establishment of object representations, recognition of objects based on the stimuli they produce, object grouping and ordering, as well as learning words that describe objects and their properties. This project proposes a behavior-grounded framework for object perception that will enable a robot to solve some of these very same problems. Our robot interacted with 100 different objects by performing 10 different behaviors on them (e.g., grasp, shake, lift, etc.), while using several sensory modalities, including vision, proprioception and audio. Our robot was tasked with recognizing objects, grouping objects together, recognizing category labels of novel objects and ordering objects based on criteria such as heights and weight. Our results show that robots need to perceive objects interactively and in multiple sensorimotor contexts in order to scale up object perception skills to a large number of objects.**

I. Introduction

Our ability to explore physical objects is unparalleled in the natural world. From an early age, human beings spend much of their time manipulating objects while simultaneously observing the resulting stimuli (e.g., visual movement, auditory events, etc.). A long line of research in psychology has revealed that humans (as well as animals) acquire information about objects through the use of a number of manipulation behaviors, commonly referred to as *exploratory procedures* [25] or *exploratory behaviors* [11], [36]. For example, scratching an object can inform us of its roughness, while lifting it can inform us of 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.

Other research in psychology has established that the sensory feedback produced by objects can be crucial for solving several key tasks:

Fig. 1. The humanoid robot used in our experiments, along with the 100 objects that it explored.

1) *object identification*, i.e., the ability to individuate objects, recognize the object identity of a given object stimulus, and recognize when a stimulus is produced by a novel object [21], [19].
2) *object sorting*, i.e., the ability to spontaneously group items into sets, or orders, without being given a specific criteria [52], [35].
3) *category and relational learning*, i.e., the ability to assign category membership to novel objects as well as infer how two objects should be ordered, based on a criteria specified by a series of example objects with known labels and/or orderings [3].

The goal of this project was the development of a multi-modal behavior-grounded framework for object perception that would enable a robot to solve these problems in an experimental setting. To achieve this aim, the robot in our framework (show in Figure 1) actively performed exploratory behaviors (e.g., grasping, lifting, shaking, dropping, pushing and tapping) when learning

about objects as opposed to just passively observing them. While most robots perceive objects using vision alone, the robot in our framework also used the auditory, and proprioceptive and sensory modalities, which are necessary to capture many object properties [8], [29].

The rest of the paper is organized as follows: Section II gives an overview of the related work in psychology and robotics. Section III describes our experimental setup, including the robot, its exploratory behaviors, its sensors and the objects used in our experiments. Section IV describes the feature extraction methodology used to extract sensory feedback features from the robot's sensory streams. Section V describes the theoretical model used by the robot to identify, categorize, and order objects. Section VI details our experiment results, followed by a discussion and future work.

II. Related Work

A. Psychology and Cognitive Science

The ability of humans to individuate objects and recognize their identities has been extensively studied in psychology. The problem of object identification is typically defined as that of inferring how many objects the environment contains (also referred to as individuation) as well as recognizing when the same object is encountered twice (sometimes referred to as identification as well as recognition) [19]. Studies in developmental psychology have shown that this process is fundamental to establishing an internal object representation that can handle the large number of objects that humans encounter in their day to day lives [57], [21].

For this reason, how infants establish an object representation and subsequently use it to recognize the identities of objects is a question of significant interest to developmental psychology. For example, a study in infants showed that even at the age of 12-months, humans are able to individuate objects using both shape and color information [57]. The study also found that while both object features were used for the task of figuring out how many objects exist, 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 and adults often make different judgments based on the differences in the objects' features [59], indicating that at such an early age, the biological circuits that allow the problem to be solved are still developing.

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 they observed, or match an object stimulus to one of the estimated object identities. For example, one such study with human adults showed that as the number of objects observed increases, the likelihood that a novel object will be classified as a previously observed object goes down [19].

A closely related area of developmental psychology studies how infants group objects. An important finding is that certain experimental settings can elicit spontaneous sorting and grouping behaviors by infants [33], [52]. Starkley [52] 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 [35], [50]. For example, Spinozzi *et al.* [50] found that human-encultured Bonobos and Chimpanzees are capable of spontaneously partitioning a set of objects into two categories. The authors also report that chimpanzees' predominant means of partitioning a set of objects is by manipulating objects from one object class only. This procedure is consistent with the behavior of 3 year old infants [50]. 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 fundamental for language acquisition – for example, Nelson argued that children form primitive conceptual categories which are later used when binding the meaning of a word [33]. Similarly, based on a large volume of experimental research, Bloom 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 [6]. An example of what this may look like is provided by Kemp *et al.* [18] 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." [18, p. 216]*

Not surprisingly, a large volume of research has focused on revealing how humans learn the names of categories [3]. In this framework, the participants are typically presented with several examples from each object category and subsequently asked to categorize a

novel item. Researches postulate that humans use two different strategies (sometimes in combination) to learn categories from examples - the first involves finding the common features of members of an individual category, while the second consists of identifying the distinctive features among the non-members of that category [16], [15]. Experiments have shown that adults can learn categories even when presented only with pairs of objects of different categories [16]. Children between the ages of 6-9 years old, however, could only learn the same categories when provided with object pairs in which the two objects are of the same category class, indicating that the two strategies for solving the task have different developmental trajectories [16].

In addition to learning discrete categories, researchers have also examined how adult and infant humans learn real-valued comparative relations such as "A is bigger than B" [48], [10]. 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. Hence, the robot in this work will be tested in a similar fashion – after initially interacting with the objects, computational models will be evaluated using both discrete categorization as well as real-valued ordering tasks.

B. Robotics

Traditionally, most object recognition systems used by robots have relied heavily on computer vision techniques [37], [51], [38] and/or 3D laser scan data [41]. But studies in psychology indicate that not only is there a link between neural activations and different sensory inputs for the same object in the brain [2], but that often multiple senses are necessary to correctly recognize an object. In a study by Sapp *et al.*, toddlers were presented with sponges painted as rocks and only by grasping the sponges could they realize that they were being deceived [42]. Other studies involving proprioception or audition have also shown that not only is it possible to use sensory modalities other than vision to recognize objects and their properties, but in some cases it is necessary [17], [9], [12], [13].

Recently, there have been multiple studies in robotics that have focused on object recognition using sensory modalities other than vision or 3D laser scan data. A study by Natale *et al.* [32] showed that proprioceptive information obtained by grasping an object can be used to successfully recognize objects. Other studies have estimated physical parameters of objects from proprioceptive data [23], [24], which can be used to recognize objects. A study by Bergquist *et al.* [5] showed that a robot can use proprioceptive information alone to recognize an object from a large set of objects. A study by Sinapov *et al.* [47] showed a similar result using auditory information alone. Other studies have confirmed that audition can be used for object recognition [40], [39] as well as for determining properties of objects [22]. Another study by Metta *et al.* showed that integrating proprioception and vision can bootstrap a robot's ability to manipulate objects. All of these studies strongly imply that sensory modalities other than vision (e.g. audition, proprioception) are useful for object recognition in addition to vision. The robot in our experiments takes advantage of this by combing multiple sensory modalities when solving object perception tasks.

One of the major drawbacks of virtually all of the methods cited above is that during the training stage, the robot has to be told which object it is exploring at any given trial. In other words, the training trials must be grouped by object identity. In order to relax that assumption, a robot must be able to autonomously figure out how many objects it has interacted with as well as organize its sensorimotor data according to object ID (i.e., solve the object individuation problem). There has been relatively little work in robotics in that area - a study by Modayil and Kuipers [30] showed how a robot could use data gathered from a laser range finder to build an ontology of objects. Another study by Southey and Little [49] used a stereo camera to detect depth features in the robot's environment, which were combined based off 3D movement patterns to create representations of each object in the environment.

In addition to object recognition, there has been much work in robotics studying how robots can form object categories in an unsupervised manner. Some of them have focused on how robots can estimate similarity between objects and use that similarity to develop meaningful object categories [34], [32], [31], [55], [47], [54]. In [32] a Self-Organizing Map was used to illustrate the haptic similarities between objects, while [47] showed that a robot can use auditory data generated from performing multiple behaviors on an object to estimate similarities. Griffith *et al.* [14] showed that a robot can form categories of "container" and "non-container" by observing the movement of an object dropped in the vicinity of another object. Sinapov and Stoytchev [46] showed that a robot can use these object similarities to detect which object in a set of objects is the odd one out. While all of these studies showed how a robot can group objects in an unsupervised manner, they all

suffer from one main drawback: They all require the type of sorting to be specified in advance - for example, in [14], the robot's categorization model used the X-means algorithm, which can find clusters in data, but not orders or hierarchies. In [45], on the other hand, the categorization algorithm assumed that the objects can be organized in a hierarchy, as opposed to some other structure.

Supervised learning for object category classification has also been studied in robotics, though not as extensively as identification. A study by Lopes and Chauhan [28] had a robot use vision to extract features from an object. They then used a set of classifiers to classify each object into different categories specified by a human. A study by Sinapov and Stoytchev [44] showed how a robot can use proprioceptive and auditory feedback to classify objects into six human-labeled categories. Other studies have examined relations among objects. The study by Griffith *et al.* [14] examined the relationship between objects dropped in the vicinity of a container/non-container, and how the two objects moved when the robot interacted with them. The research here presents methods for categorizing objects into predefined categories and learning relations between objects as they relate to ordering objects (e.g. bigger than). To the authors' knowledge, there has been no previous research in robotics on ordering objects.

III. Experimental Platform

A. Robot and Sensors

The robot in our experiments was an upper torso humanoid robot, which has as its actuators two 7-DOF Barrett WAMs, each with an attached 3-finger Barrett Hand. The WAMs have built-in sensors that measure joint angles and torques for all 7 joints at 500 Hz; auditory feedback is captured by an Audio-Technica U853AW cardioid microphone mounted in the head, which samples 1 channel (mono) at the standard 16-bit/44.1 kHz resolution and rate. A digital accelerometer device [53], mounted on one of its fingertips, samples acceleration of the fingertip at 1600 Hz, allowing detection of minute vibrations due to rubbing between the robot's fingertip and the objects' surfaces. The robot's vision sensors include a Logitech webcam (right eye) and a ZCam, an RGBD camera from 3DV systems that records standard 640×480 RGB video in addition to 320×240 depth images accurate to within 1-2 cm.

Fig. 2. The hundred objects that our robot explored. The objects are grouped according to twenty object categories. From left to right and top to bottom: 1) wicker baskets, 2) weights (objects vary by weight only), 3) small stuffed animals, 4) big stuffed animals, 5) metal objects, 6) wooden blocks, 7) pasta boxes, 8) metal tin containers, 9) PVC pipes, 10) cups, 11) pop cans, 12) plastic bottles, 13) canned food, 14) medicine pill bottles, 15) coffee containers with different types of contents, 16) green cones, 17) pink noodles, 18) egg coloring cups (vary only by color) 19) easter eggs (vary by material, and 20) balls.

B. Objects

For this project, the robot explored 100 different household objects. To our knowledge, this is the largest number of objects explored by a humanoid robot over the course of a single experiment. The 100 different objects consists of 20 object categories, with 5 objects per category. The objects within each category vary along certain dimensions while remaining constant along others. For example, the *PVC pipes* category includes 5 pipe cross sections which vary by width (and consequently, weight) but are constant in shape, color and material type. The object set was designed in this manner in order to test models for object recognition as well as object

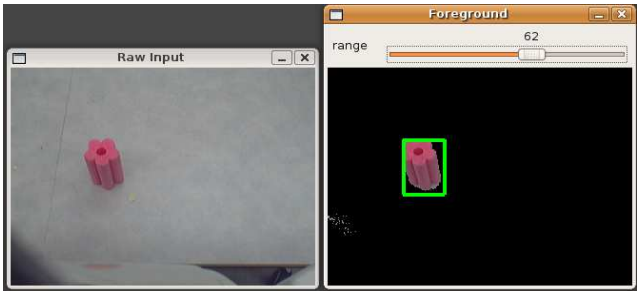


Fig. 3. Illustration of object detection in the robot's visual field of view.

categorization. Figure 2 shows all 100 objects grouped in the 20 object categories used in our experiments.

C. Behaviors

The robot's set of behavior consisted of 10 actions: *look*, *grasp*, *lift*, *hold*, *shake*, *drop*, *tap*, *push*, *poke*, and *crush*. The *look* behavior consisted of simply taking an RGB snapshot of the object on the table. All other behaviors, with the exception of *grasp* and *tap* were encoded as recorded trajectories, i.e., they were executed using pre-defined joint position coordinates.

The *grasp* and *tap* behaviors, on the other hand, were performed by the robot according to the detected visual location of the object. Visual object detection was performed using the following steps:

1) A background visual model of the table was created by taking a snapshot of the empty table before any objects are placed on it.
2) When an object is in place and the robot needs to determine its position, the robot moved its hand out of its field of view, calculated the deviation of each pixel observed from the value predicted by the background model, and then used a threshold to classify them as either "background" or "foreground".
3) The largest connected component in the "foreground" was detected and a bounding box was fit on it as shown in Figure 3.

The robot was trained to grasp and tap objects at various table positions using a simple learning by demonstration framework. During the training stage, the robot detected the location of objects on the table, after which the human-programmer physically moved the robot's arm to the appropriate joint-coordinates given the location of the object. Both models were trained on 12 demonstrations. After training, the robot used the 3-nearest neighbor algorithm which outputs interpolated joint-positions for a given object location by finding the three neighbors in the train set with object locations most similar to the one being observed. The object's location in the visual field of view was encoded by the pixel coordinates of the lower left corner of the bounding box around the detected object.

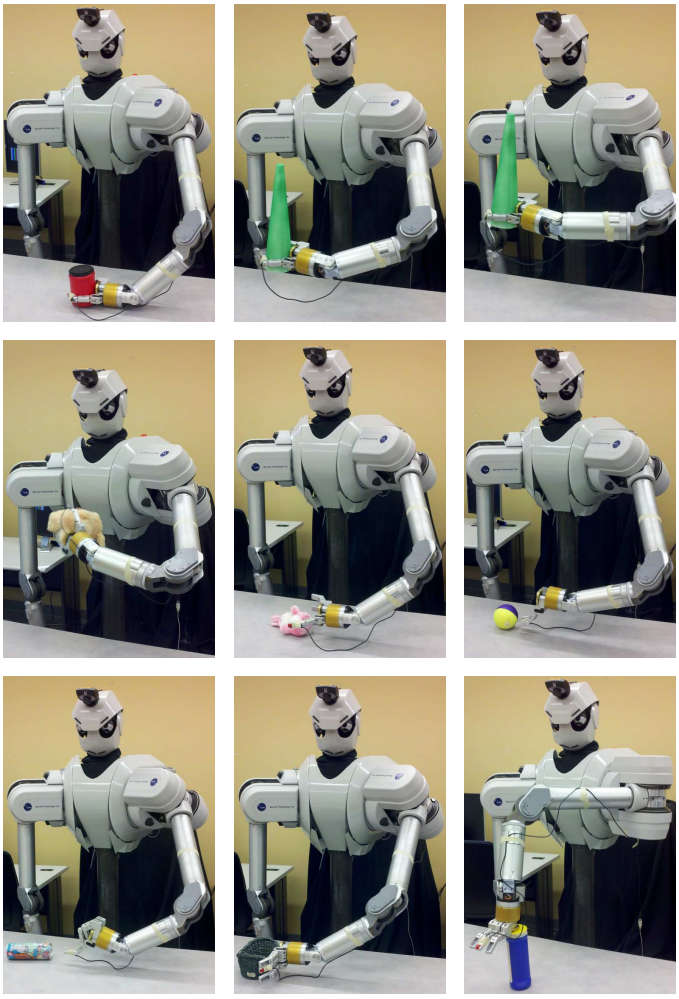
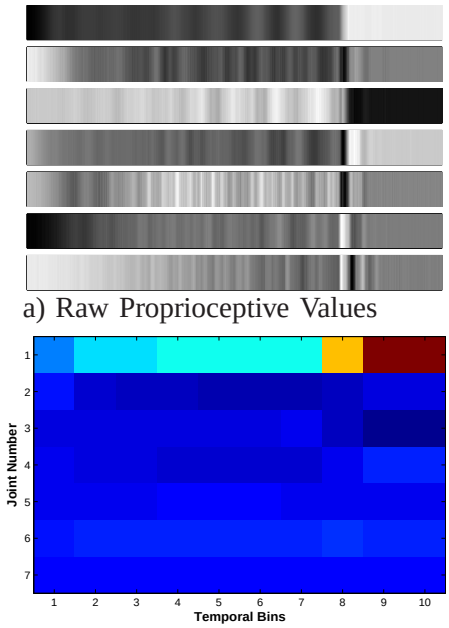


Fig. 4. The exploratory behaviors (excluding the "look" behavior) that the robot performs on objects. From top to bottom and left to right: 1) grasp, 2) lift, 3) hold, 4) shake, 5) drop, 6) tap, 7) poke, 8) push, 9) crush.

Figure 4 shows images of the 9 interactive behaviors (i.e., all except *look*) that the robot was programmed with.

D. Data Collection

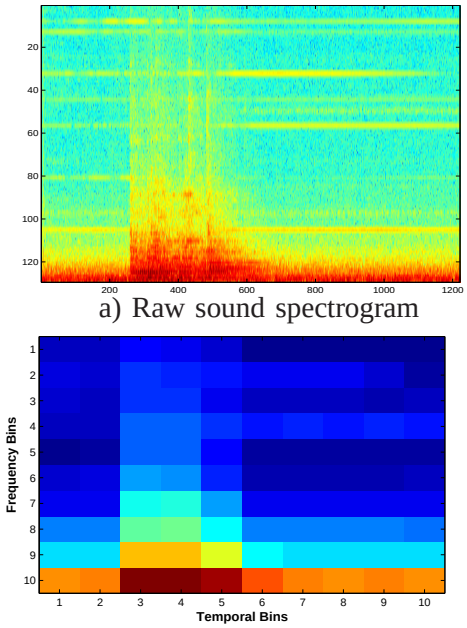
In our experiments, the robot interacted with the 100 objects over the course of a series of exploration trials. During each trial, an object was placed on the table, after which the robot performed a series of behaviors on the object. This was repeated until the robot had performed



a) Raw Proprioceptive Values

b) Proprioceptive Features

Fig. 5. a) The raw torque values for all seven joints as the robot performed the *crush* behavior on the smallest *green cone* object in our dataset.



a) Raw sound spectrogram

b) Spectro-temporal features

Fig. 6. a) The raw spectrogram of the sound detected as the robot performed the *tap* behavior on the *coke can* object. b) The resulting 10×10 spectro-temporal features.

its full set of 10 exploratory behaviors on each object for a total of five times, resulting in $10 \times 100 = 5000$ behavior executions.

Over the course of each behavior execution, the robot recorded sensory feedback from its microphones, joint-torque sensors, vibrotactile sensor, RGB webacm (right eye) and the RGB-D ZCam. The next section describes the feature extraction routines that were used to compute features from several of the recorded sensory input streams.

IV. Feature Extraction

A. Proprioceptive Feature Extraction

During each interaction, the robot recorded the torque applied to each of its 7 joints at 500hz, resulting in a joint torque record for each interaction. The joint torque record is a series of column vectors through time where each $x_{i,t}$ is the amount of torque being applied to joint i at time t.

To extract features, an n-bin average was used. To compute this, first each joint torque record was split into n bins of size $b_{size} = \frac{T}{n}$ based off temporal relation of each column (e.g. the first b_{size} columns in the first bin and so on) where T is the temporal length of the joint torque record. For each bin b_k, the average of each joint torque was computed, resulting in one column vector $\bar{\mathbf{x}}_\mathbf{k}$ for each b_k where $\bar{x}_{i,k}$ is the average torque for joint i in bin b_k.

For most experiments in this paper, the number of temporal bins, n, was set to 10. For the experiments involving object ordering, n was set to 1.

B. Auditory Feature Extraction

Auditory features were extracted using the log-normalized Discrete Fourier Transform (DFT) which was computed for each sound, using $2^7 + 1 = 129$ frequency bins. The SPHINX4 natural language processing library package was used to compute the DFT for each sound [27]. The DFT encodes the detected intensity for all 129 frequency bins over time. The DFT is highly-dimensional and thus cannot be used directly as an input to most machine learning algorithms. Therefore, given the DFT matrix for each sound, a 2-D histogram is computed by discretizing time into k_t bins and frequencies into k_f bins. The value for each bin in the histogram is set to the average of the values in the DFT matrix that fall into it. In all experiments conducted, k_t was set to 10 and k_f was set to 5. Hence, each sound is represented by feature vector, S, where $S \in \mathbb{R}^{5\times 10}$. Figure 6 shows an example of how the DFT of a sound is transformed into a 2-D histogram across time and frequency.

C. Visual Feature Extraction

The robot's learning model extracted visual object features from the RGB color images taken during the execution of the *look* behavior. Three types of features were computed: 1) color: the distribution of colors in the object's image, 2) aspect: the width and height of a bounding box centered on the object, and 3) size: the area of the object in the robot's image. The image of the object and the surrounding table area was taken by the robot's left eye webcam. The set of pixels representing the object was computed immediately afterwards to allow the robot to grasp the object by classifying. Pixels not representing the object were blackened in the saved copy of the image. The pixels were classified as "object" or "non-object" using a learned visual model of the background (see Section III.C).

When the images were loaded for feature computation, pixels with values very close to black were completely blackened, because almost all pixels having values in this region were compression artifacts and were originally black. Afterward, the cvFindContours() algorithm from the OpenCV library was used to locate contiguous regions of pixels. Finally, all pixels not within the bounding rectangle of the largest contiguous region were blackened, which eliminated regions of the environment that were labeled "object" due to chance variations in their appearance during the experiment. After the object was segmented in the image, visual features were computed as follows:

Color: For each trial with each object, 4 color histograms were computed, each of which separated the RGB color space into one of 4^3, 8^3, 12^3, or 16^3 bins. This was done by dividing the $[0, 256)$ range along each color axis into equal-length segments and classifying each pixel according which segment each of its channels fell onto. Mathematically, for each

$$n \in \{4, 8, 12, 16\}$$

we assigned a triplet

$$(bin_r, bin_g, bin_b) \in \mathbb{N}^3$$

to each pixel with coordinates

$$(r, g, b) \in \mathbb{N}^3 \mid 0 \leq r, g, b < 256$$

such that

$$\frac{bin_r}{n} \cdot 256 \leq r < \frac{bin_r + 1}{n} \cdot 256$$

and etc. for g, b. A feature vector in $\mathbb{N}^{n^3}$ was produced for each of the 4 histograms which gave the numbers of pixels falling into the bins of its corresponding histogram. In the following experiments described in this paper, the histograms produced when using 4 bins for each color channel were used.

Aspect: A feature vector in $\mathbb{N}^2$ was produced giving the width and height (in pixel units) of the bounding rectangle of the object.

Size: A feature vector in $\mathbb{N}^1$ was produced giving the total number of pixels comprising the object.

D. Hand Proprioception Feature Extraction

During the execution of the *grasp* behavior, the resulting finger joint angles were recorded. Thus, the *grasp* behavior was the only one that produced hand proprioception features. Each recorded feature vector was 3-dimensional, where each value indicates the end position for one of the three corresponding fingers of the Barett Hand. The end position of each finger was always in the range of 0 (fully open) to 20000 (fully closed).

E. Summary

In our experiments, the robot extracted proprioceptive, auditory, visual and hand features from each interaction. The visual features (color histogram, aspect ratio and visual size) were extracted from the RGB image taken by the robot at the start of each exploration trial. The auditory and proprioceptive features were extracted from the feedback detected over the course of each manipulation behavior (i.e., all behaviors except *look*). The hand proprioptive features were extracted only for the *grasp* behavior. Note that vibrotactile and Z-Cam RGBD data were also recorded for each behavior, but are not used in the experiments described in this report (extracting features from those two sensory streams will be done in future work). The next section describes the theoretical model which uses detected object features for the problems of recognizing, categorizing and ordering objects.

V. Theoretical Model

A. Notation

Let $\mathcal{B}$ be the set of exploratory behaviors and let $\mathcal{S}$ be the set of sensory modalities available to the robot. Let $\mathcal{C}$ be a set of behavior-modality contexts such that each context $c_j \in \mathcal{C}$ refers to a unique combination of a behavior and a sensory modality (e.g., *drop-audio*). Note that it is not necessary for every combination to be present in the set $\mathcal{C}$, since in our case certain behaviors do not produce sensations in certain modalities.

During each object exploration trial, the robot is presented with an object $o \in \mathcal{O}$, the set of all objects, and subsequently applies its set of exploratory behaviors on the object. Hence, when executing behavior $b \in \mathcal{B}$, the robot observes a set of sensory signals $\mathcal{X}_b = \{x_1 \dots x_{m_b}\}$ where each x_j represents the sensory feedback observed from some known sensory modality in $\mathcal{S}$. Note that the number of sensory feedback signals detected when performing some specific behavior, $|\mathcal{X}_b| = m_b$, may be less than the number of sensory modalities, $|\mathcal{S}|$, since certain behaviors do not produce sensations in certain modalities (e.g., looking at an object does not produce tactile sensations).

After all behaviors are applied on the test object, the i^{th} exploration trial may be summarized by the collection of observed sensory feedback signals, $T_i = \{X_b\}_{b \in \mathcal{B}}$. In practice, the signals x_j may be encoded as numerical vectors, real-valued time series, or discrete sequences. For this project, several different representations will be used, including sequences

B. Object Recognition

For this problem, the robot is tasked with recognizing the identity of the object (one out of the 100) being explored, given some sensory feedback x_i^c detected in sensorimotor context c. To solve this task, an object recognition model M_c is trained for each context $c \in \mathcal{C}$, such that given input x_i^c, the robot outputs $M_c(x_i^c) \rightarrow \hat{o}$, such that $\hat{o}$ is the estimated object identity of the object present in the interaction. In other words, for each viable combination of behavior and sensory modality, the robot learns a recognition model specifically adapted for data from that behavior-modality combination. Given sensory feedback features x_i^c, the model M_c outputs a probabilistic object identity estimate $Pr_c(o|x_i^c)$ for each object $o \in \mathcal{O}$.

The models M_c are trained on data points $[x_i^c, o_i]$ for which the true object identity, o_i, is known. In the experiments presented in this report, the recognition models for each sensorimotor context were implemented by the k-Nearest Neighbor (kNN) classifier, a memory-based algorithm, which does not build an explicit model of the data [1], [4]. Instead, given a test data point, k-NN finds the k closest neighbors in its training set and outputs a prediction, which is a smoothed average over those neighbors. In this study, the parameter k was set to 3. Class label probabilities for each object $o \in \mathcal{O}$ were computed by counting the labels of the k neighbors. For example, if two of the three neighbors had object identity *A* then $Pr(o_i = A) = \frac{2}{3}$. Similarly, if the class label of the remaining neighbor was *B*, then $Pr(o_i = B) = \frac{1}{3}$. The k-NN implementation included in the WEKA machine learning library [60] was used to obtain the results.

After executing its full set of behaviors $\mathcal{B}$ on the test object, the robot combined the outputs of each individual context-specific model M_c in order to get a more accurate estimate for the identity of the object. Let $\mathcal{X}_\rangle = [x_i^1, x_i^2, \dots, x^{|\mathcal{C}|}{}_i]$ be the resulting set of sensory inputs detected in all sensorimotor contexts $\mathcal{C}$. The robot can get a combined probabilistic estimate for the identity of the object, $Pr(o|\mathcal{X}_\rangle)$ by summing up the outputs of the individual models:

$$Pr(o|\mathcal{X}_\rangle) = \sum_{c \in \mathcal{C}} Pr_c(o|x_i^c)$$

The robot's recognition model is evaluated using 5-fold cross validation: during each round of cross-validation, data from 4 of the exploratory trials with each object is used for training the models, while the data from the remaining trial is used for testing whether the recognition model is correct. This is repeated five time, such that each trials is used once in the test set and four times in the training set. The model's performance is reported in terms of percent object recognition accuracy (% Accuracy), defined as:

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

C. Object Grouping

In a typical categorization experiment in psychology, the participant is presented with a set of objects and then either asked to group them or allowed to freely explore them to see if spontaneous sorting behavior occurs. Hence, in this task the robot's categorization model is given sensorimotor experience with objects from two object categories and outputs an object grouping consisting of two groups of objects. For example, if presented with the object categories *cones* and *pop cans*, we epxect that the robot's categorization model will group the items into two groups, each corresponding to one of the two categories.

More specifically, the robot's categorization model takes as input a set of 50 exploration trials $\mathcal{T}_{input} = [T_1, T_2, \dots, T_{50}]$ in which the robot explored a set of 10 objects, $\mathcal{O}_{input}$, (with known object identity) from two different categories (unknown to the robot). The model is tasked to output two object sets $\mathcal{O}_a$ and $\mathcal{O}_b$, representing the estimated object categories, such that $\mathcal{O}_a \cup \mathcal{O}_b \equiv \mathcal{O}_{input}$ and $\mathcal{O}_a \cap \mathcal{O}_b \equiv \emptyset$.

The robot produces the categorization using the following steps. First, the robot's object recognition models are evaluated on the set of input trials $\mathcal{T}_{input}$ by performing 5-fold cross-validation as described in the previous subsection. The result of this procedure is a confusion matrix $\mathbf{A} \in \mathbb{R}^{|\mathcal{O}_{input}| \times |\mathcal{O}_{input}|}$ such that each entry A_{ij} specifies how often object o_i was recognized as object o_j. Next, an object similarity matrix $\mathbf{W} \in \mathbb{R}^{|\mathcal{O}_{input}| \times |\mathcal{O}_{input}|}$ is computed such that each entry $W_{ij} = \frac{A_{ij}+A_{ji}}{2}$. Finally, the resulting symmetric object similarity matrix $\mathbf{W}$ is given as input to a partitioning algorithm, which splits the set of objects into two groups such that the similarity (as encoded in $\mathbf{W}$) between objects in the same set is maximized while the similarity between objects in two different sets is minimized.

In our experiments, the robot used the *Spectral Clustering* partitioning algorithm, which falls into the family of *graph-based* or *similarity-based* clustering algorithms [58]. Given a similarity matrix, $\mathbf{W}$, the algorithm partitions the set of object into disjoint clusters by exploiting the eigenstructure of the matrix $\mathbf{W}$. Because finding an optimal graph partitioning is NP-complete, Shi and Malik [43] proposed an approximation that optimizes the *normalized cut* objective function. The algorithm, can be summarized with the following steps:

1) Let $\mathbf{W}_{n \times n}$ be the symmetric matrix containing the similarity score for each pair of objects.
2) Let $\mathbf{D}_{n \times n}$ be the degree matrix of $\mathbf{W}$, i.e., a diagonal matrix such that $\mathbf{D}_{ii} = \sum_j W_{ij}$.
3) Solve the eigenvalue system $(\mathbf{D} - \mathbf{W})x = \lambda \mathbf{D}x$ for the eigenvector corresponding to the second smallest eigenvalue and use it to bi-partition the graph.

The resulting partition encodes how the robot would group the object based on its experience with them. The robot's categorizations are evaluated in terms of whether the discovered partitioning matches the true categories assigned to the objects.

D. Object Category Recognition

The third task consists of training the robot to recognize the category labels of objects given a certain amount of objects with known labels. For example, if the robot interacts with a large set of objects, and if the user specifies that two of those objects are called "cups", then the robot's model should be able to infer what other objects are cups as well.

To solve this problem, for each sensorimotor context c, the robot learns a category recognition model trained on input datapoints with known category labels. As in the case of object recognition, the robot uses the k-NN classifier model this task. Similarly, the robot is also evaluated on how well it performs as model outputs from different sensorimotor contexts are combined.

The robot's category recognition model is evaluated by performing object-based cross-validation as follows: during each round of cross validation, the full set of trials in split into a test and train set, such that the train set contains trials with 4 out of the 5 objects for each category, while the test set contains the trial for the remainder object of each category. This is repeated 5 times, such that each object serves once as a test object with unknown category label and four times as a training object with a known label. The performance of the category recognition model is reported in terms of % accuracy. In addition, for each of the 20 categories, the f-Measure is reported. The f-Measure is the harmonic mean between the precision and recall for a given category label and is computed by:

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

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

E. Object Ordering

For this problem, the robot is tasked with correctly sorting objects by some external criteria E given some sensory feedback x_j^c and x_i^c detected in sensorimotor context c. To solve this task, an object comparison model M_c is trained for each context $c \in \mathcal{C}$, such that given x_j^c and x_i^c, the robot outputs $M_c(x_j^c, x_i^c) \rightarrow \{>, <\}$ where $<$ indicates that the robot estimates that o_j is less than o_i and $>$ indicates that it is greater than. In other words, the robot learns a model for each viable context that outputs the comparison of two objects.

The models M_c are trained on data points $[x_j^c, x_i^c, \{>, <\}]$ for which the true outcome of the comparison for o_j and o_i is known. In the experiments presented in this report, the comparison models used for each sensorimotor context were implemented by the Support Vector Machine (SVM) algorithm, as implemented in the WEKA machine learning library [60].

To evaluate the robot, first the objects were split into 20 sets of 5. For each external criteria E, sets were pruned if they did not have at least a $K\%$ difference between each pair of objects in the set, leaving the set $\mathcal{O}_E$ as the set of all objects that vary based on criteria E. For this experiment K was set to 10.

Each object o_j was interacted with m_b times, generating m_b feature vectors in the context c. To train the model, object-based cross-validation was used. For each object o_j, all feature vectors x_j generated by o_j were removed from the set. Then the model was trained on every pair of feature vectors x_i and x_k such that o_i and o_k are in the same set of 5 objects and $o_i \neq o_k$ given the true comparison between the two objects. Then every feature vector for object o_j, x_j, was paired with every other feature vector x_i such that o_j and o_i are in the same set of 5 objects and $o_j \neq o_i$.

The accuracy for each context c is reported as

$$A_c = \frac{\sum_{o_j \in \mathcal{O}_E} r_j}{\sum_{o_j \in \mathcal{O}_E} t_j}$$

where r_j is the number of correct comparisons when evaluating object o_j and t_j is the total number of comparisons when evaluating o_j.

To evaluate multiple contexts together, a weighted voting approach was used. For each $\mathcal{C}' \subseteq \mathcal{C}$ (for each combination of contexts), to calculate the accuracy $A_{\mathcal{C}'}$, each model M_c, where $c \in \mathcal{C}'$, voted on the outcome of every pair of feature vectors x_j and x_i such that o_j and o_i are in the same set of 5 objects and $o_j \neq o_i$. The estimated comparison then for x_j and x_i is

$$\hat{v}_{i,j} = \sum_{c \in \mathcal{C}'} A_c \hat{v}^c_{i,j}$$

where $\hat{v}^c_{i,j}$ is the vote of context c for x_j and x_i. The accuracy of $\mathcal{C}'$ is calculated as

$$A_{\mathcal{C}'} = \frac{\sum_{x_j, x_i \in \mathcal{O}_E} [\hat{v}_{i,j} = v_{i,j}]}{T_E}$$

where $[\hat{v}_{i,j} = v_{i,j}]$ is 1 iff the estimated comparison for x_j and x_i is equal to the actual comparison and T_E is the total number of comparisons for objects in O_E.

VI. Results

A. Object Recognition

The first experiment evaluates the performance of the robot's recognition models for each possible sensorimotor context. Tables I and II shows the accuracy rates for each viable combination of a behavior and sensory modality.

TABLE I
Object Recognition from *Look* Behavior

Behavior	Color Histogram	Aspect Ratio	Visual Size
look	66.33 %	33.46 %	17.64 %

TABLE II
Object Recognition from a Single Behavior

Behavior	Audio	Proprioception (Arm)	Proprioception (Hand)
grasp	10.44 %	9.21 %	11.02 %
lift	17.44 %	37.07 %	–
hold	6.81 %	26.25 %	–
shake	30.26 %	47.90 %	–
drop	31.26 %	9.22 %	–
tap	31.86 %	14.23 %	–
push	39.4 %	43.0 %	–
poke	28.06 %	38.48 %	–
crush	28.25 %	64.12 %	–

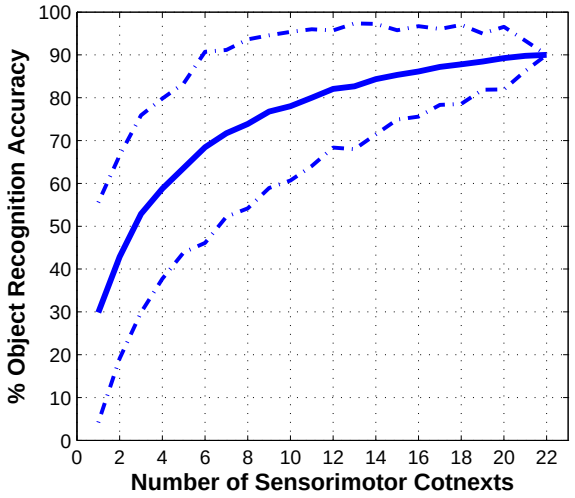


Fig. 7. Object recognition performance as the number of sensorimotor contexts (i.e., behavior-modality combinations) is varied from 1 to 22. At each level, the cross-validation is repeated 200 times with a random set of contexts selected. The solid line corresponds to the mean accuracy for the given number of contexts, while the dotted lines denote the standard deviation.

To compare, a model which randomly assigns object identity is expected to achieve 1.0% accuracy, since the number of object identities is 100. The results show that nearly every sensorimotor context contains information useful for object recognition. As expected, certain behaviors work better with certain modalities: for exampple, the proprioceptive features produced by the *lift* behavior are better for object recognition than the auditory features detected in that same context.

Following, the robot's performance at the object recognition task is also computed as a function of the mnumber of sensorimotor contexts available to the robot during the exploration trial (both for training and testing the recognition models). To do this, the number of

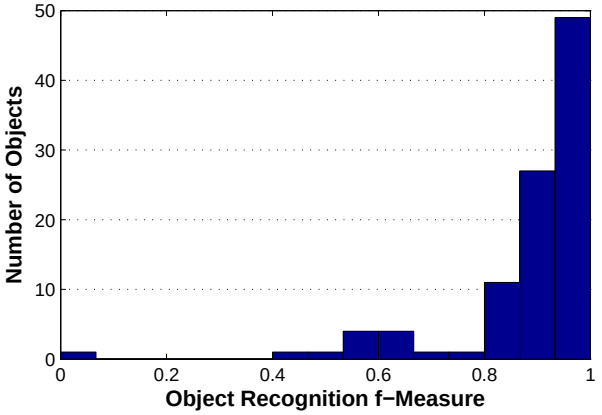


Fig. 8. A histogram of individual f-Measures per object identity. The histogram shows that most objects can be recognized almost perfectly (with 10 ten behaviors). One object (object 3 from the wooden blocks category) is almost impossible to recognize.

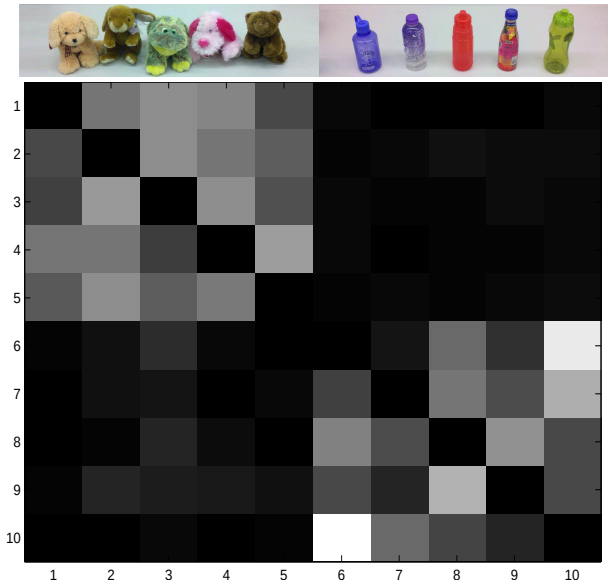


Fig. 9. Resulting object recognition confusion matrix after performing cross validation on the 10 objects in the *big stuffed animals* and *plastic bottles* categories. Each entry in the confusion matrix specifies how often object i was recognized as object j. In this example, the first five objects are from the *big stuffed animals* set while the last five objects are from the *plastic bottles* set. Values close to white indicate that a pair of objects are often confused. In this case, most errors happen within the category, i.e., stuffed animals are rarely recognized as a plastic bottle object.

contexts is varied from 1 to 22. At each level, the cross-validation is repeated 200 times with a random set of contexts selected. The results are then used to estimate the mean object recognition rate for a given number of contexts as well as its standard deviation.

Figure 7 shows the results of this experiment. The plot shows that as the robot experiences the objects with more behaviors and modalities, its object recognition rate improves substantially. With all 22 sensorimotor contexts, the robot's recognition rate hits 90.0%. This result shows that the diversity of the robot's behavioral repertoire is important (and necessary) in order to scale up object recognition methods to a large number of objects.

Finally, Figure 8 shows a histogram of individual f-Measures (as defined in the previous section) per object identity. Objects that are easy to recognize have high f-Measures while those that are difficult have low ones. The figure shows that most objects can be recognized almost perfectly when using all 10 ten behaviors. The right-most bar of the histogram corresponds to the 49 out of the 100 objects that are always correctly recognized with all 10 behaviors.

B. Object Grouping

The next set of experiments evaluates how the robot can group objects without knowing their true category labels. First, we look at whether the confusion matrix computed when performing cross-validation can be used for categorization. In the first test, the robot's model is presented with the set of trials performed on the 10 objects in the *big stuffed animals* and *plastic bottles* categories. Figure 9 shows the resulting confusion matrix after the robot has cross-validated its object recognition models trained to recognize these 10 specific objects. Each entry in the confusion matrix specifies how often

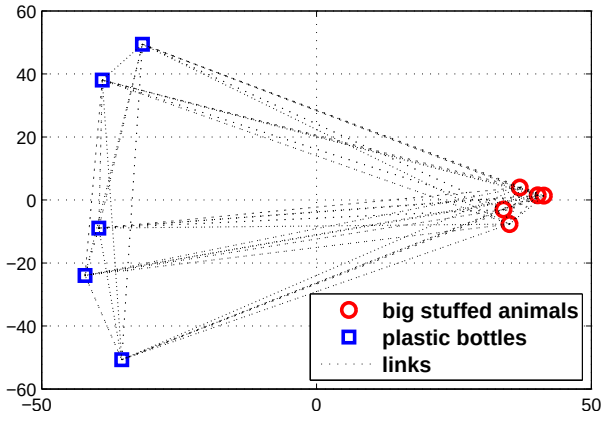


Fig. 10. An ISOMAP embedding of the similarity matrix $\mathbf{W}$ used by the robot's model to group the presented set of objects. In this example, the robot's model was presented with 10 objects, the five big stuffed animals and the five plastic bottles. The spectral clustering partitioning algorithm discovered two object clusters, each perfectly corresponding to one of the two human-provided object categories.

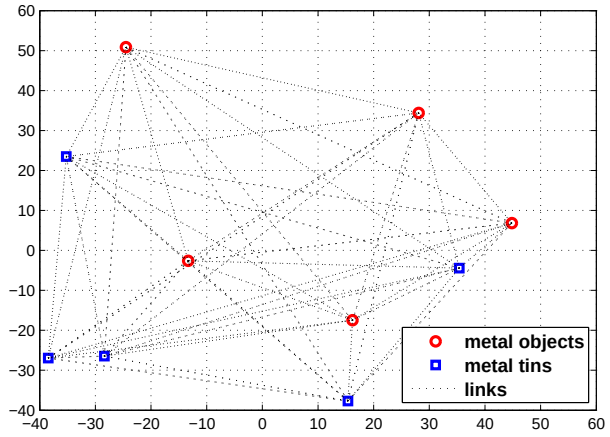


Fig. 11. An ISOMAP embedding of the similarity matrix $\mathbf{W}$ used by the robot's model to group the presented set of objects. In this example, the robot's model was presented with 10 objects, the five big stuffed animals and the five plastic bottles. The spectral clustering partitioning algorithm discovered two object clusters, each perfectly corresponding to one of the two human-provided object categories.

object i was recognized as object j. In this example, the first five objects are from the *big stuffed animals* set while the last five objects are from the *plastic bottles* set. Values close to white indicate that a pair of objects are often confused. In this case, most errors happen within the category, i.e., stuffed animals are rarely recognized as a plastic bottle object. This indicates that the robot may be able to use its object recognition models to perform cross-validation on a given set of objects, and subsequently use the resulting confusion matrix as a measure for similarity when grouping the given set of objects into two groups.

Following, the confusion matrix is converted into a symmetric similarity matrix and used as an input to a partitioning algorithm which groups the objects into two sets. In this example, the spectral clustering algorithm divided the set of objects into two groups, where each corresponded to one of the two object categories. Figure 10 shows an ISOMAP embedding [56] of the similarity matrix. The robot's model for grouping the objects produced two object sets, each perfectly corresponding to one of the two human-provided object categories.

Similar results were observed with other object category pairs. This results indicates that the robot's model for spontaneous object grouping produces object clusters which closely match human category names. Nevertheless, not all categories are perfectly separable by the robot's grouping model. Figure 11 shows an example in which the robot's model is presented with objects from two categories, *metal tin containers* and *metal objects*. The 2D embedding shows that the confusion matrix for this set of objects has many mistakes in which a tin objects is confused as one of the metal objects that is not a tin and thus the categories cannot be separated when looking at the confusion matrix. In this example, the robot's model produced two object groups, one containing 3 of the metal tins and 2 of the metal objects (non-tins) and the other containing the rest. Following, the next set of experiments examines how well the robot can explicitly learn to classify novel objects into one of the twenty categories.

TABLE III
OBJECT ORDERING ACCURACY BY HEIGHT USING SVM MODEL

Behavior	Audio	Proprioception
Grasp	69.1 %	69.8 %
Slow Lift	56.1 %	76.7 %
Hold	58.9 %	79.2 %
Shake	64.7 %	84.3 %
Drop	70.1 %	68.1 %
Tap	79.7 %	52.3 %
Crush	93.1 %	96.8 %
Poke	73.7 %	68.1 %
Push	69.1 %	85.4 %
Average	70.5 %	75.6 %

C. Category Recognition

The third sets of experiments evaluates the performance of the robot's category recognition models. In this setting, the model is trained with known labels for 4 out of 5 objects for each category and evaluated on the remaining set. As with object recognition, the evaluation is also performed when varying the number of sensorimotor contexts from 1 to 22. The reported performance measure is the f-Measure for each category type. An f-Measure of 1.0 indicates that the category was always recognized (see theoretical model section for further explanation).

Figure 12 shows the recognition rates for all 20 categories as the number of behavior-modality combinations used to train recognition models is varied from 1 to 22. As the robot is allowed to experience objects in more sensorimotor contexts, its ability to classify them into categories increases. With all 22 sensorimotor contexts, the robot can recognize the correct category of a novel object with 89.1% accuracy (a chance model is expected to achieve 5.0% accuracy as there are 20 categories).

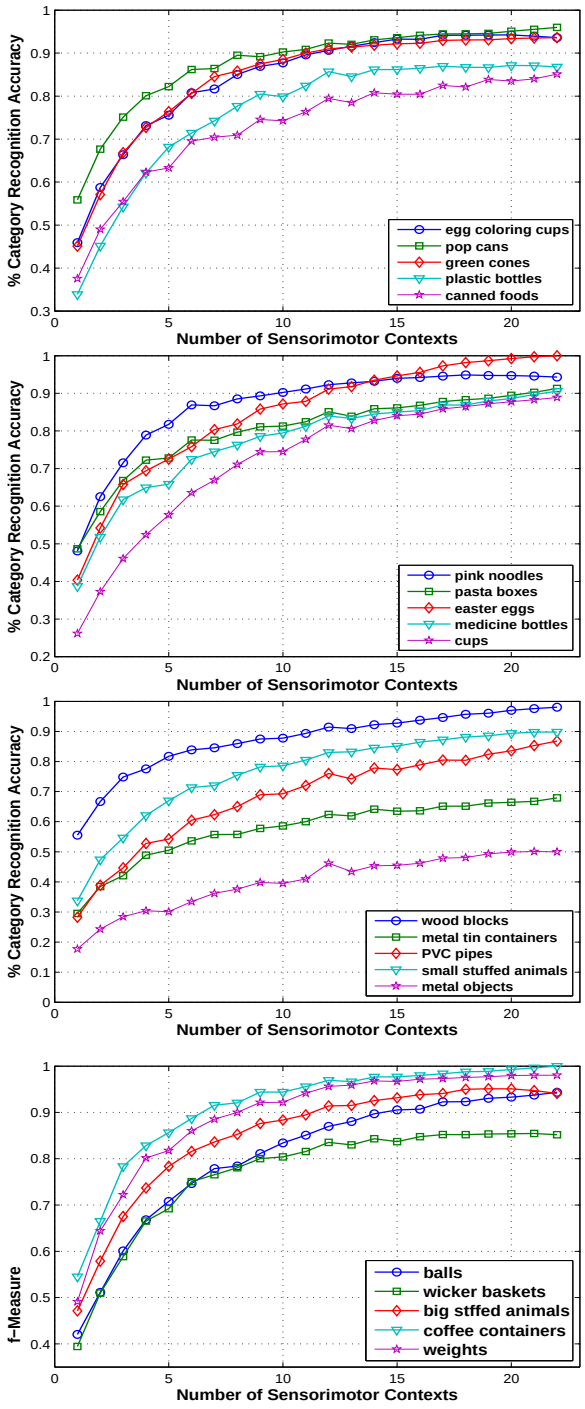


Fig. 12. Category Recognition

D. Object Ordering

The external criteria used to evaluate the robot were *height* and *weight*. Tables III and IV show the results for single contexts. The contexts used for ordering included audio and proprioception and every behavior excpet look. The object sets used for height evaluation are shown in

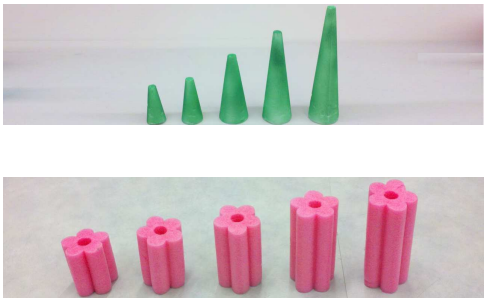


Fig. 13. Object sets for ordering by height

figure 13. The object sets used for weight evaluation are shown in figure 14.

The accuracy is reported as the number of comparisons between feature vectors for objects in the same category that were correctly predicted over the total number of comparisons. As a reference, chance accuracy is 50%. For *height*, there were 2 objects sets with 5 interactions with each object (and thus 5 feature vectors for each object), for a total of 500 comparisons. For *weight*, there were 7 sets with 5 interactions with each object (and thus 5 feature vectors for each object), for a total of 1750 comparisons. As with object recognition and category recognition, the evaluation is also performed when varying the number of sensorimotor contexts from 1 to 18. In this case though, every combination of contexts was used, rather than a random sample.

As expected, some behaviors, such as $\{proprioception, tap\}$ and $\{audio, slow\,lift\}$ for height, perform near chance. For height, the *crush* behavior for both audio and proprioception perform significantly better than chance. Also as expected, the contexts $\{proprioception, slow\,lift\}$, $\{proprioception, hold\}$, and $\{proprioception, shake\}$

TABLE IV
OBJECT ORDERING ACCURACY BY WEIGHT USING SVM MODEL

Behavior	Audio	Proprioception
Grasp	65.2 %	67.8 %
Slow Lift	77.9 %	97.3 %
Hold	64.2 %	96.5 %
Shake	81.1 %	96.9 %
Drop	65.9 %	83.2 %
Tap	67.3 %	72.2 %
Crush	63.1 %	63.5 %
Poke	75.4 %	71.0 %
Push	69.1 %	84.2 %
Average	69.9 %	81.4 %

Fig. 14. Object sets for ordering by weight

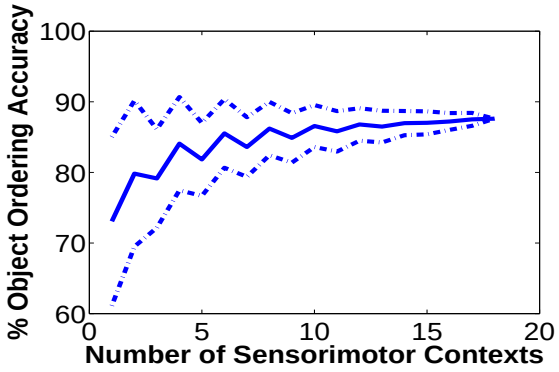


Fig. 15. Object ordering accuracy for height as the number of sensorimotor contexts (i.e., behavior-modality combinations) is varied from 1 to 18. At each level, every possible combinations of contexts is evaluated. The solid line corresponds to the mean accuracy for the given number of contexts, while the dotted lines denote the standard deviation.

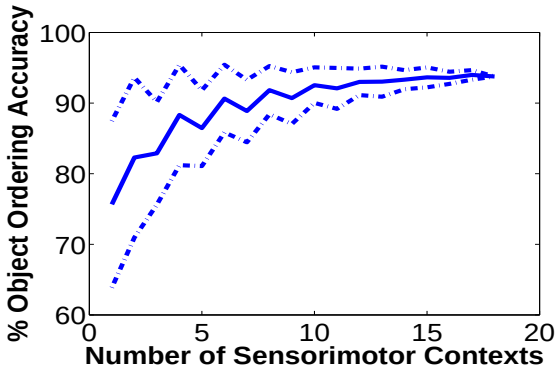


Fig. 16. Object ordering accuracy for weight as the number of sensorimotor contexts (i.e., behavior-modality combinations) is varied from 1 to 18. At each level, every possible combinations of contexts is evaluted. The solid line corresponds to the mean accuracy for the given number of contexts, while the dotted lines denote the standard deviation.

perform the best for weight, significantly above chance. Interestingly enough, the robot is able to get 81.1% for weight accuracy with the context $\{audio, shake\}$, suggesting that there is some relation between the noise an object makes when being shook and its weight.

Figures 15 and 16 show the accuracy when varying the number of sensorimotor contexts from 1 to 18 for height and weight respectively. For each number of sensorimotor contexts, every possible combination of that size was evaluated using weighted-voting (see theoretical model for details), and the mean and standard deviation is reported. The robot is able to acheive 87.6% accuracy for height and 93.8% accuracy for weight when using all combinations of contexts. This is lower than the maximum value for each (96.8% and 97.3% respectively) when using only the best single context. This suggests that if accuracy is known for each context *a priori*, then it is not beneficial to combine contexts; but if accuracy is not known *a priori*, then combing contexts will, on average, improve performance.

Figures 17 and 18 show the error rate plotted against the difference between the object pairs that were evaluated (where error rate is the number of incorrect predictions over the total number of predictions). The difference for height is reported in inches, and the difference for weight is reported in ounces. The figures show that there is a relation between the difference between objects

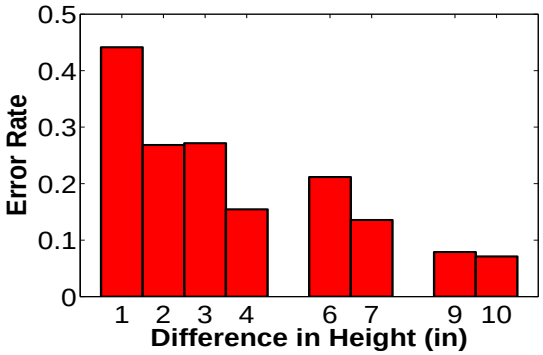


Fig. 17. The error rate (i.e. the number of incorrect predictions over the number of total predictions) for object ordering by height versus the actual difference in height between object pairs.

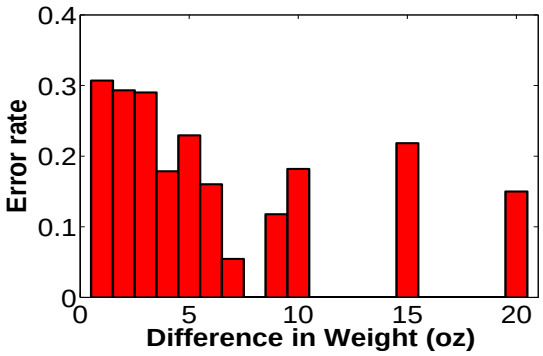


Fig. 18. The error rate (i.e. the number of incorrect predictions over the number of total predictions) for object ordering by weight versus the actual difference in weight between object pairs.

and the error rate: as the difference grows, the error rate declines. But for the weight that relation is not as clear as it is for height.

VII. Conclusion and Future Work

This project explored the development of a behavior-grounded framework for object perception by a humanoid robot. Several problems were addressed:

1) *object recognition*: the ability to recognize the object identity of the object that the robot interacts with, based on prior experience with that object.
2) *object grouping*: the ability to group objects into categories without specific human-provided guidelines.
3) *object category recognition*: the ability to learn object category labels, and classify novel objects according to the correct human-given category
4) *object ordering*: the ability to order novel objects according to various properties (e.g., height, and weight) based on pairs of objects for which the order relation is known.

The trained recognition model was able to estimate the identity of the present object by training a series of recognition models, each corresponding to a specific behavior-modality combination (i.e., sensorimotor context) that produces sensory feedback. This representation allowed the robot to estimate the object identity given sensory feedback detected with the object when performing any of the 10 behaviors that the robot was programmed with. In addition, the robot was able to significantly improve its recognition rate by combining the outputs of multiple models after performing a series of behaviors on the object and detecting the resulting sensory feedback features. These results make a strong case that robots should experience objects using a diverse set of behaviors and sensory modalities in order to scale up their recognition abilities to a large number of household objects.

Following, the robot's recognition models were used to estimate a measure for pair-wise object similarity, such that objects that are often confused with each other are considered similar, while objects that are never confused with each other are considered different. After the robot's model estimated the pair-wise object similarity, it used the resulting matrix to partition the object set into two groups using the spectral clustering graph-based algorithm. The results showed that the model's choice for object groups matched closes the human-provided category labels - for example, when the model was presented with 5 stuffed animals and 5 plastic bottles (without knowing the category of each object), it produced two object groups, such that each group was a perfect match to the human-provided category label. This result shows that robots can estimate object groups that match category nouns even without explicitly knowing that the objects fall into human-provided categories. In addition, the result also shows that if a robot can recognize objects, it can also categorize them based on how easy it is to distinguish between each pair of objects.

For the third task, a model was trained to explicitly estimate the category label of a *novel* object (i.e., one for which training data is not available) given training data with objects for which the category labels are provided by a human. The evaluation of the robot's category recognition model showed that, just as with object recognition, the number of behaviors and modal-

ities available to the robot can greatly influence the classification performance. The results imply that a robot may ground category nouns (e.g., *cup*, *container*, *ball*, etc.) in its own behavioral repertoire.

In the final task, a model was trained to predict the outcome of a pairwise comparison between two objects (i.e. greater than or less than). Given a *novel* object, it had to determine how it compared to objects it had previously interacted with. The evaluation of this model showed that a robot can in fact learn these object orderings and that certain contexts are best at comparing by certain properties, such as *slow lift*, *hold* and *shake* with *proprioception* for comparing by weight and *crush* for both *audio* and *proprioception* for comparing by height. Unlike the other models, though, this model has shown that if a robot knows *a priori* which sensorimotor contexts are best suited for comparing which properties, then combing modalities and behaviors does not improve accuracy. On the other hand, if the robot does not know the accuracies for individual contexts *a priori*, then combing them significantly improves accuracy. This suggests that at least for object ordering, there are certain sensorimotor contexts that specialize at perceiving certain properties about objects. A robot that wants to be able to order by a diverse set of properties, then, would find it beneficial to equip itself with multiple, diverse sensorimotor contexts for performing object interactions in.

There are many directions for future work. First, incorporating features extracted from the robot's vibrotactile sensor and the RGBD Z-Cam is a direct extension to this project that we plan to pursue. Using RGBD data gathered from the Z-Cam, more properties can be used for ordering such as color or volume. As well, adding in RGBD data would increase the diversity of the ordering predictors. Based on the results in this paper, we can safely predict that with an even richer experience with objects, a robot may be able to scale up object perception methods to an even larger object sets.

Finally, we also plan to pursue novel methods for unsupervised object grouping with the goal of enabling a robot to discover object concepts that may be relevant to many language learning tasks (e.g., learning objects' category nouns as well as the adjectives that describe them). The drawback of most existing algorithms is that they assume a specific form (e.g., a hierarchy, or a grouping) that describes how objects are related to each other. To avoid this pitfal, in future work, we plan to implement methods such as the one described in [20] to allow the robot to determine which structure type should be used to organize a particular set of objects – in other words, the structure used to sort the object is induced by the model, rather then specified by the programmer.

VIII. Appendix

A. Team

1) **Kerrick Staley** is a first-year student in Computer Engineering. He is interested, in general, in computer science, mathematics, and the physical sciences; he has specific interests in robotics, cryptography and data security, user interface design, and the practicalization of open source software. He programs primarily in C/C++ and Python. He enjoys reading Slashdot.org, and his Kirby skills in SSB64 will stomp most competitors. He has a website with further biographical details at kerrickstaley.com.
2) **Connor Schenck** is a senior in Computer Science. He has experience with C/C++, Java, and Matlab. He has used OpenCV, Weka, Java Swing, and MATLAB's Image Processing Toolkit. He has taken courses on Machine Learning, Artificial Intelligence, Algorithms, and Statistics. He is a coauthor for the paper *Interactive Object Recognition Using Proprioceptive Feedback* and *Interactive Object Recognition Using Proprioceptive and Auditory Feedback*. He has also worked on multiple projects in the Developmental Robotics Laboratory at Iowa State University.
3) **Jivko Sinapov** received the B.S. degree in Computer Science from the University of Rochester, NY in 2005. He is currently a PhD student in Computer Science and works at the Developmental Robotics Laboratory at Iowa State University, Ames. His research interests include developmental robotics, robotic perception, manipulation, and machine learning.

B. Software Packages

The following list of software libraries was used in for this project:

1) **The WEKA Java Machine Learning Library :** contains a number of implementations for popular machine learning algorithms for the tasks of classification, and unsupervised clustering [60].
2) **Structural Form Discovery MATLAB package:** implementation of the model proposed by Kemp

et al. [20] for the purposes of fitting structures to data.

3) **OpenCV:** C++ computer vision library, used when detecting the object on the table, as well as extracting visual object features.
4) **GHSOM package:** a Java library implementing the Growing-Hierarchical Self-Organizing Map algorithm [7] for dimensionality reduction. The package will be used to turn high-dimensional sensory feedback data into low dimensional discrete sequence.
5) **Sparse Coding MATLAB package:** a MATLAB library developed by Lee *et al.* [26], which will be used to extract features given depth images taken by the robot's ZCam.
6) **robocop:** C++ software, written by Vlad Sukhoy, which wraps the Barett WAM API and is used for recording the robot's sensorimotor data during object exploration trials.

C. Future Timeline

1) Submit paper to Humanoids Conference: May 26.

REFERENCES

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

[2] A. Amedi, K. Kriegstein, NM Atteveldt, MS Beauchamp, and MJ Naumer. Functional imaging of human crossmodal identification and object recognition. *Experimental Brain Research*, 166(3):559–571, 2005.

[3] F.G. Ashby and W.T. Maddox. Human category learning. *Psychology*, 56(1):149, 2005.

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

[5] T. Bergquist, C. Schenck, U. Ohiri, J. Sinapov, S. Griffith, 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.

[6] P. Bloom. How children learn the meanings of words: Learning, development and conceptual change, 2000.

[7] 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.

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

[9] W. Gaver. What in the world do we hear? An ecological approach to auditory event perception. *Ecological Psych.*, 5:1–29, 1993.

[10] D. Gentner and L.L. Namy. Analogical processes in language learning. *Current Directions in Psychological Science*, 15(6):297, 2006.

[11] E. J. Gibson. Exploratory behavior in the development of perceiving, acting, and the acquiring of knowledge. *Annual Review of Psychology*, 39:1–41, 1988.

[12] 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.

[13] M. Grassi. Do we hear size or sound? Balls dropped on plates. *Perception and Psychophysics*, 67(2):274–284, 2005.

[14] 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 *Development and Learning, 2009. ICDL 2009. IEEE 8th International Conference on*, pages 1–6. IEEE, 2009.

[15] 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.

[16] 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.

[17] M. Heller. Haptic dominance in form perception: vision versus proprioception. *Perception*, 21(5):655–660, 1992.

[18] C. Kemp, K.K. Chang, and L. Lombardi. Category and feature identification. *Acta psychologica*, 133(3):216–233, 2010.

[19] C. Kemp, A. Jern, and F. Xu. Object discovery and identification. 2009.

[20] C. Kemp and J.B. Tenenbaum. The discovery of structural form. *Proceedings of the National Academy of Sciences*, 105(31):10687, 2008.

[21] P. Krojgaard. A review of object individuation in infancy. *British Journal of Developmental Psychology*, 22(2):159–183, 2004.

[22] 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.

[23] D. Kubus, T. Kroger, and F.M. 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.

[24] D. Kubus and F.M. Wahl. Estimating Inertial Load Parameters Using Force/Torque and Acceleration Sensor Fussion. In *Robotic 2008, VDI-Beritche 2012 Munchen, Germany*, pages 29–32.

[25] S. Lederman and R. Klatzky. Haptic classification of common objects: knowledge-driven exploration. *Cognitive Psychology*, 22:421–459, 1990.

[26] H. Lee, A. Battle, R. Raina, and A.Y. Ng. Efficient sparse coding algorithms. *Advances in neural information processing systems*, 19:801, 2007.

[27] 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.

[28] L.S. 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. Citeseer, 2007.

[29] D. Lynott and L. Connell. Modality Exclusivity Norms for 423 Object Properties. *Behavior Research Methods*, 41(2):558–564, 2009.

[30] J. Modayil and B. Kuipers. Bootstrap learning for object discovery. In *Intelligent Robots and Systems, 2004.(IROS 2004). Proceedings. 2004 IEEE/RSJ International Conference on*, volume 1, pages 742–747. IEEE, 2005.

[31] 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.

[32] L. Natale, G. Metta, and G. Sandini. Learning haptic representation of objects. In *Proceedings of the International Conference on Intelligent Manipulation and Grasping*, 2004.

[33] K. Nelson. Some Evidence for the Cognitive Primacy of Categorization and Its Functional Basis. *Merrill-Palmer Quarterly*, 19(1):21–39, 1973.

[34] 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.

[35] P. Potì. Logical structures of young chimpanzees' spontaneous object grouping. *International Journal of Primatology*, 18(1):33–59, 1997.

[36] Thomas G. Power. *Play and Exploration in Children and Animals*. Lawrence Erlbaum Associates, Publishers, Mahwah, NJ, 2000.

[37] M. Quigley, E. Berger, and A.Y. Ng. STAIR: Hardware and software architecture. *Presented at AAAI 2007 Robotics Workshop*, 2007.

[38] 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.

[39] J. Richmond. Automatic measurement and modelling of contact sounds. Master's thesis, University of British Columbia, 2000.

[40] J. Richmond and D. Pai. Active measurement of contact sounds. In *Proc. of ICRA*, pages 2146–2152, 2000.

[41] 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.

[42] F. Sapp, K. Lee, and D. Muir. Three-year-olds' difficulty with the appearance-reality distinction. *Developmental Psychology*, 36(5):547–60, 2000.

[43] J. Shi and J. Malik. Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

[44] J. Sinapov and A. Stoytchev. Object Category Recognition by a Humanoid Robot Using Behavior-Grounded Relational Learning.

[45] 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.

[46] J. Sinapov and A. Stoytchev. The Odd One Out Task: Toward an Intelligence Test for Robots. 2010.

[47] 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.

[48] L.B. Smith, N.J. Cooney, and C. McCord. What Is" High"? The Development of Reference Points for" High" and" Low". *Child Development*, pages 583–602, 1986.

[49] T. Southey and J.J. Little. Object discovery through motion, appearance and shape. In *AAAI Workshop on Cognitive Robotics, Technical Report WS-06-03. AAAI Press*, 2006.

[50] G. Spinozzi, F. Natale, J. Langer, and K.E. Brakke. Spontaneous class grouping behavior by bonobos (Pan paniscus) and common chimpanzees (P. troglodytes). *Animal Cognition*, 2(3):157–170, 1999.

[51] 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.

[52] D. Starkey. The origins of concept formation: Object sorting and object preference in early infancy. *Child Development*, 52(2):489–497, 1981.

[53] V. Sukhoy, R. Sahai, J. Sinapov, and A. Stoytchev. Vibrotactile recognition of surface textures by a humanoid robot. In *Proceedings of the Humanoids 2009 Workshop "Tactile Sensing in Humanoids - Tactile Sensors and Beyond", Paris, France*, pages 57–60, Dec 7, 2009.

[54] J. Sun, J.L. Moore, A. Bobick, and J.M. Rehg. Learning Visual Object Categories for Robot Affordance Prediction. *The International Journal of Robotics Research*, 29(2-3):174, 2010.

[55] S. Takamuku, K. Hosoda, and M. Asada. Object category acquisition by dynamic touch. *Advanced Robotics*, 22(10):1143–1154, 2008.

[56] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[57] P.D. Tremoulet, A.M. Leslie, and D.G. Hall. Infant individuation and identification of objects. *Cognitive Development*, 15(4):499–522, 2000.

[58] U. von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

[59] T. Wilcox and R. Baillargeon. Object individuation in infancy: The use of featural information in reasoning about occlusion events. *Cognitive Psychology*, 37:97–155, 1998.

[60] I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufman, San Francisco, 2nd edition, 2005.