Object Category Recognition by a Humanoid Robot Using Behavior-Grounded Relational Learning

Jivko Sinapov and Alexander Stoytchev
Developmental Robotics Laboratory
Iowa State University
{jsinapov, alexs}@iastate.edu

***Abstract*—The ability to form and recognize object categories is fundamental to human intelligence. This paper proposes a behavior-grounded relational classification model that allows a robot to recognize the categories of household objects. In the proposed approach, the robot initially explores the objects by applying five exploratory behaviors (lift, shake, drop, crush and push) on them while recording the proprioceptive and auditory sensory feedback produced by each interaction. The sensorimotor data is used to estimate multiple measures of similarity between the objects, each corresponding to a specific 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 a novel object based on the object's similarity to the set of familiar objects. The framework was evaluated on an upper-torso humanoid robot with two large sets of household objects. The results show that the robot's model is able to recognize complex object categories (e.g., metal objects, empty bottles, etc.) significantly better than chance.**

I. 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 (e.g., [1], [2], [3], [4]). 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 [5]. 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 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 [6] for a literature survey).

To address these limitations, this paper 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.

II. Related Work

The importance of forming and recognizing object categories has lead to several important lines of research in the robotics community. Most object classification and categorization systems in use by robots today rely almost exclusively on visual and/or 3D laser scan data [1], [2], [3], [4]. While such systems have many useful applications, they suffer from several important limitations. For example, they are of little use when the object is not in a direct line of sight. In addition, they cannot distinguish between objects that look identical, but differ in other properties (e.g., material type or weight). The human visual system is also subject to these same limitations, which is why humans need other sensory modalities to better capture knowledge about objects [5], [7], [8].

To address these shortcomings, several projects have focused on recognizing and representing objects using proprio-

ceptive, auditory, and/or tactile feedback [9], [10], [11], [12], [13]. Others have focused on coupling visual-based object representations with exploratory behaviors [14], [15]. Natale *et al.* [9] have demonstrated that a robot can recognize objects with the help of a self-organizing map using proprioceptive data extracted from the robot's hand as it grasped an object. In other related work, Nakamura *et al.* [11] described a robot that uses proprioceptive, visual and auditory information when interacting with objects in order to infer the outputs of one modality from another (e.g., whether an object would make noise when picked up after only looking at it). Other research has demonstrated that robots can successfully recognize objects using only auditory feedback [16], [17], [10].

Similarly, our own previous research has demonstrated frameworks for interactive behavior-grounded object recognition [13], [18] as well as unsupervised object categorization [19], [20] using combinations of auditory, proprioceptive and visual sensory modalities. In another study, we demonstrated that a robot can solve the odd-one-out task (i.e., find the object that does not belong in a given set) using an unsupervised model operating on an object similarity graph [21]. This paper builds on such graph-based object representations by introducing a relational learning model that uses multiple object similarity relations. In contrast to our previous work, the new model is fully supervised and extracts relational features estimated from multiple sensorimotor contexts in order to solve a wide variety of category recognition tasks.

III. Experimental Setup

The experimental setup and the data set used to evaluate our model were previously published and presented by Bergquist *et al.* [18] for the different task of object recognition using proprioceptive feedback. Both are briefly summarized here.

A. Robot

The robot used in our experiments 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.

B. Exploratory Behaviors

The robot applied five exploratory behaviors on the objects: *lift*, *shake*, *drop*, *crush*, and *push* (see Fig. 1). 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.

C. Sensory Feedback Feature Extraction

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) [22]. The feature extraction routines that were used are identical to the ones described in [18] (for proprioception) and [13] (for audio), and are only briefly summarized here.

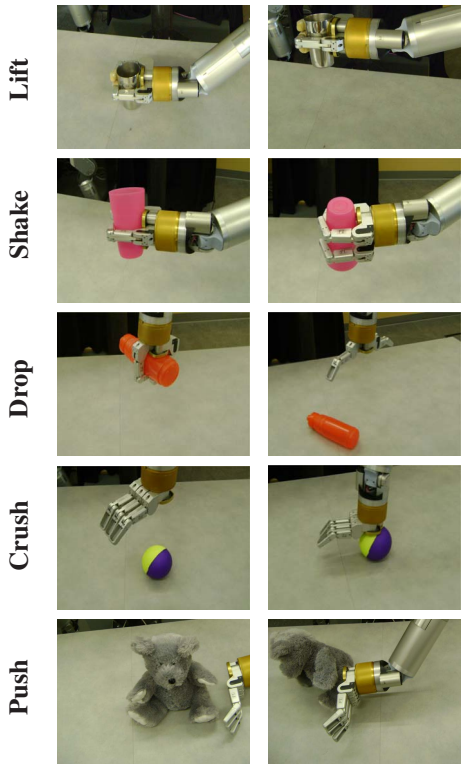

Fig. 1. *Before* and *after* snapshots of the five behaviors used by the robot.

The raw joint torque data was converted into a discrete sequence as follows. Given a specific joint-torque configuration of the robot's left arm (i.e., a vector in $\mathbb{R}^7$) detected over the course of an interaction, the data point was fed as input to the proprioceptive SOM and the index of the most highly activated state in the map was appended as the next token in the proprioceptive sequence for that behavioral interaction. The SOM was trained with sample joint-torque configurations experienced by the robot while executing all five behaviors.

The Discrete Fourier Transform (DFT) of each recorded sound was processed in a similar way: each column vector of the DFT was given as input to the auditory SOM and the index of the most highly activated state was added as the next token in the auditory feedback sequence. The auditory SOM was trained with a set of column vectors extracted from the recorded DFTs. For both sensory modalities, 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 [23].

After each behavioral execution, the robot recorded two sequences, $X_{prop} = p_1 p_2 \dots p_k$ and $X_{audio} = a_1 a_2 \dots a_l$, as shown in Fig.2. The theoretical model proposed in this paper requires a metric that can establish the similarity between two sequences from the same sensory modality. In our experiments, the global alignment similarity function was used, which is a common choice for comparing discrete sequences (the same

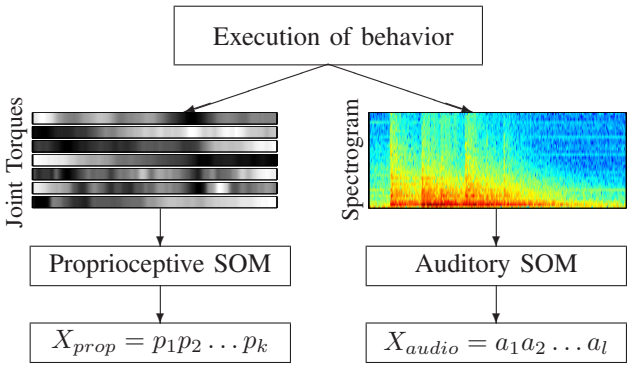

Fig. 2. Turning high-dimensional proprioceptive and auditory input into low dimensional discrete sequences. While performing a behavior on an object, the robot records the joint-torque data and the Discrete Fourier Transform of the audio signal. Each 7-dimensional joint-torque vector is given as input to the proprioceptive SOM and the index of the most highly activated SOM state is added as a token in the resulting sequence, X_{prop}. Similarly, each DFT column vector from the recorded spectrogram is mapped to a state in the auditory SOM, resulting in the sequence X_{audio}.

Fig. 3. The 6 object categories. An object may belong to multiple categories, e.g., the three pop cans also belong to the set of metal objects.

metric was also used in [18] and [13]).

D. Objects

The set of objects, $\mathcal{O}$, consisted of 25 common household objects, including cups, bottles, and toys (see Fig. 3). The objects were labeled according to six natural object categories, which were used in the evaluation of the robot's performance on the task of object category recognition. 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 V.E., 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 [13].

IV. 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.

A. 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 $c \in \mathcal{C}$.

B. 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.[1]

C. 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 3 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 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 4 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).[2]

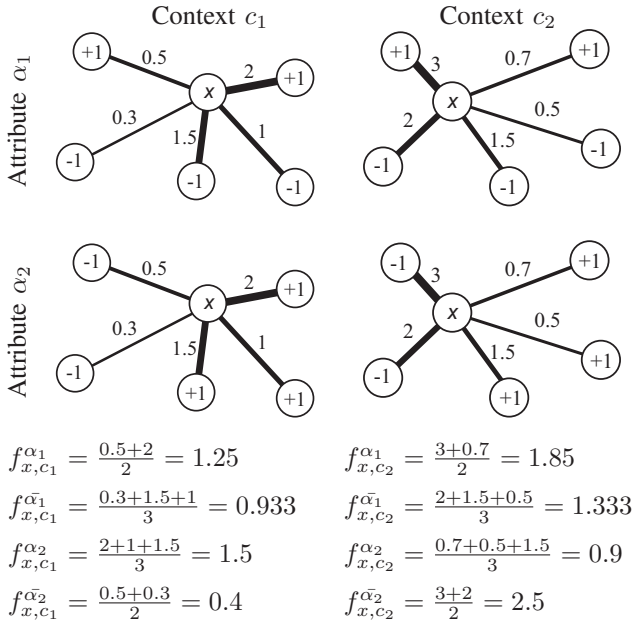

Fig. 4. A simple example of relational feature extraction. In this case, there are two contexts (c_1 and c_2) and two attributes (α_1 and α_2). There are five familiar objects with known labels (either -1 or $+1$) for both attributes and one unlabeled novel object (denoted with x). The edges correspond 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.

V. Results

A. 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

[1]Note that estimates of the variability, i.e., $\mathbf{Var}[sim(X_a, X_b)]$, may also be used as additional relations from which features can be extracted. In our experiments, this additional information was not used as it did not lead to an improvement in classification performance.

[2]The WEKA machine learning library, which provides implementations of k-NN, SVM, and C4.5, was used [24]. 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 I
INTERPRETING κ COEFFICIENT VALUES, AS PROPOSED IN [26]

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

TABLE II
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

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 [25], 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 probability 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 I shows how to interpret κ values as proposed in [26].

B. 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 II 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 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.

C. 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 5 shows the results of this experiment. There is a slight to moderate improvement in the classification perfor-

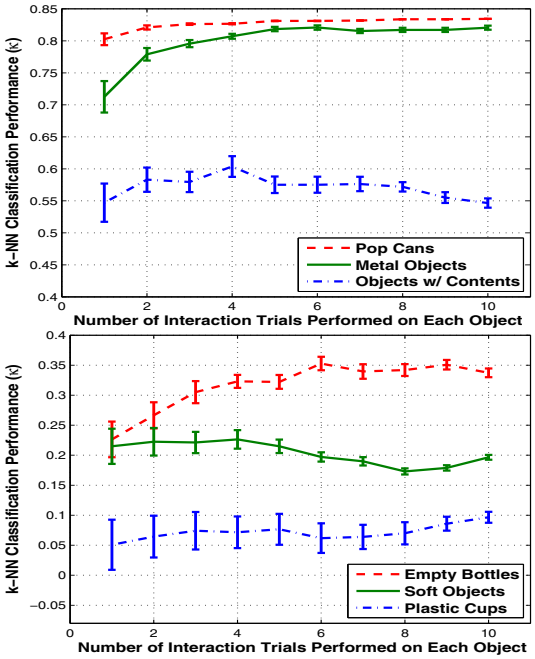

Fig. 5. Classification performance of the k-NN category recognition model as a function of the number of interaction trials used to estimate the object similarity matrices $\mathbf{W}^c$.

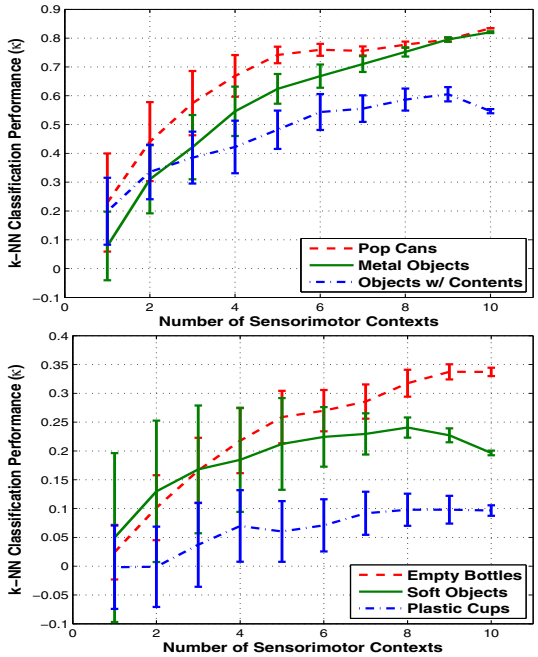

Fig. 6. Classification performance of the k-NN category recognition algorithm as a function of the number of sensorimotor contexts available to the relational recognition model.

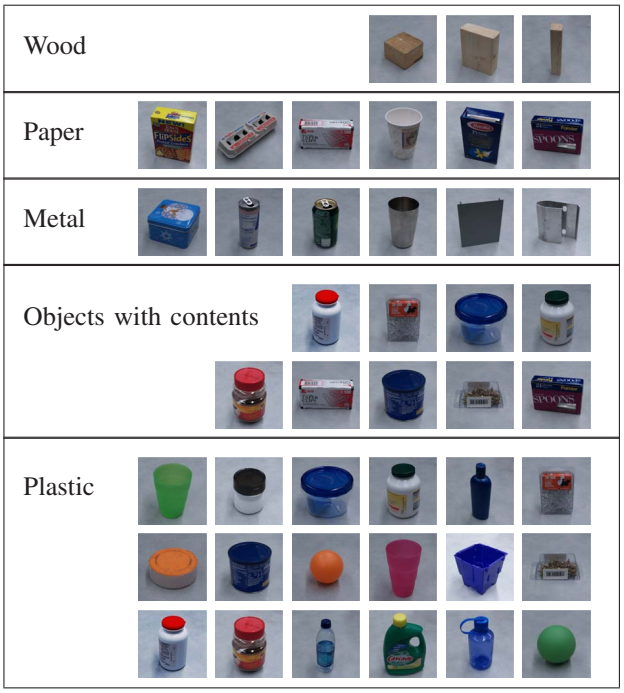

Fig. 7. The objects from the second data set and their corresponding categories, which were used to further validate the method presented in this paper. Some objects belong to multiple categories. Three of the objects in that data set do not belong to any of the five categories and are not shown here.

mance 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.

D. 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 II). 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 6 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 6 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, which has shown that exploratory behaviors act as classifiers that can be boosted [27]. Research in psychology has indeed shown that animals and humans use multiple exploratory behaviors and multiple sensory modalities to both learn and represent the properties of objects [28], [29].

E. 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 [13] and categorization [19]. In this experiment, the robot performed five exploratory behaviors (grasp, shake, drop, push and tap) on 36 household objects (see Fig. 7). The auditory data from each trial was recorded and converted into a discrete sequence as described earlier (as well as in [13]). 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 [19].

Table III 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

TABLE III
CLASSIFICATION PERFORMANCE (κ) ON THE DATA SET FROM [13]

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**

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.

VI. CONCLUSION AND FUTURE WORK

This paper proposed 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 [30]. 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.

REFERENCES

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

[2] 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*, vol. 56, no. 11, pp. 927 – 941, 2008.

[3] S. Srinivasa *et al.*, "HERB: a home exploring robotic butler," *Autonomous Robots*, vol. 28, no. 1, pp. 5–20, 2010.

[4] F. Endres, C. Plagemann, C. amd Stachniss, and W. Burgand, "Unsupervised discovery of object classes from range data using latent Dirichlet allocation," in *Proc. RSS 2009, Seattle, WA, USA*, pp. 113–120.

[5] D. Lynott and L. Connell, "Modality Exclusivity Norms for 423 Object Properties," *Behavior Research Methods*, vol. 41, no. 2, pp. 558–564, 2009.

[6] L. Getoor and C. P. Diehl, "Link mining: a survey," *ACM SIGKDD Explorations Newsletter*, vol. 7, no. 2, pp. 3–12, 2005.

[7] F. Sapp, K. Lee, and D. Muir, "Three-year-olds' difficulty with the appearance-reality distinction," *Developmental Psychology*, vol. 36, no. 5, pp. 547–60, 2000.

[8] M. Heller, "Haptic dominance in form perception: vision versus proprioception," *Perception*, vol. 21, no. 5, pp. 655–660, 1992.

[9] L. Natale, G. Metta, and G. Sandini, "Learning haptic representations of objects," in *Proceedings of the International Conference on Intelligent Manipulation and Grasping*, 2004.

[10] E. Torres-Jara, L. Natale, and P. Fitzpatrick, "Tapping into touch," in *Proc. 5-th Intl. Workshop on Epigenetic Robotics*, 2005, pp. 79–86.

[11] 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*, 2007, pp. 2415–2420.

[12] S. Takamuku, K. Hosoda, and M. Asada, "Object Category Acquisition by Dynamic Touch," *Adv. Rob.*, vol. 22, no. 10, pp. 1143–1154, 2008.

[13] J. Sinapov, M. Weimer, and A. Stoytchev, "Interactive learning of the acoustic properties of household objects," in *Proc. of the IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2009, pp. 2518–2524.

[14] E. Ugur, M. Dogar, M. Cakmak, and E. Sahin, "Curiosity-driven learning of traversability affordance on a mobile robot," in *Proc. of the 6th IEEE Intl. Conf. on Development and Learning*, 2007, pp. 13–18.

[15] S. Hart, "An Intrinsic Reward for Affordance Exploration," in *Proc. of the 8th IEEE Intl. Conf. on Development and Learning*, 2009, pp. 1–6.

[16] E. Krotkov, R. Klatzky, and N. Zumel, "Robotic perception of material: Experiments with shape-invariant acoustic measures of material type," in *Experimental Robotics IV*, ser. Lecture Notes in Control and Information Sciences. Springer Berlin/Heidelberg, 1996, vol. 223, pp. 204–211.

[17] J. Richmond and D. Pai, "Active measurement of contact sounds," in *Proc. of the IEEE ICRA*, 2000, pp. 2146–2152.

[18] 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.

[19] 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.

[20] 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 *Proc. of the 8th IEEE Intl. Conf. on Development and Learning*, 2009.

[21] J. Sinapov and A. Stoytchev, "The Odd-One-Out Task: Toward an Intelligence Test for Robots," in *Proc. of the 9th IEEE Intl. Conf. on Development and Learning*, 2010, pp. 126–131.

[22] T. Kohonen, *Self-Organizing Maps*. Springer, 2001.

[23] 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*, 2002, pp. 2537–2541.

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

[25] J. Cohen, "A Coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[26] R. Landis and G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.

[27] J. Sinapov and A. Stoytchev, "The boosting effect of exploratory behaviors," in *Proceedings of the 24-th National Conferrence on Artificial Intelligence (AAAI)*, 2010, pp. 126–131.

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

[29] T. Power, *Play and Exploration in Children and Animals*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers, 2000.

[30] X. Zhu, Z. Ghahramani, and T. J. Mit, "Semi-supervised learning with graphs," Carnegie Mellon University, Tech. Rep., 2005.