2014 IEEE International Conference on Robotics & Automation (ICRA)
Hong Kong Convention and Exhibition Center
May 31 - June 7, 2014. Hong Kong, China

Learning Relational Object Categories Using Behavioral Exploration and Multimodal Perception

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

***Abstract*—This paper proposes a framework for learning human-provided category labels that describe individual objects, pairwise object relationships, as well as groups of objects. The framework was evaluated using an experiment in which the robot interactively explored 36 objects that varied by color, weight, and contents. The proposed method allowed the robot not only to learn categories describing individual objects, but also to learn categories describing pairs and groups of objects with high recognition accuracy. Furthermore, by grounding the category representations in its own sensorimotor repertoire, the robot was able to estimate how similar two categories are in terms of the behaviors and sensory modalities that are used to recognize them. Finally, this grounded measure of similarity enabled the robot to boost its recognition performance when learning a new category by relating it to a set of familiar categories.**

I. Introduction

The ability to learn and use object categories is an important aspect of human intelligence and has been extensively studied in psychology (see [1] for a review). Researchers have postulated that, with a few labeled examples, humans at various stages of development are able to identify common features that define category memberships as well as distinctive features that relate members and non-members of a target category [2], [3]. Other lines of research have highlighted the importance of active object exploration for learning object categories [4], [5]. Studies have also demonstrated that many object properties cannot always be detected by passive observation alone (see [6] and [7]).

Recently, several research groups have started to explore how robots can learn object category labels that can be generalized to novel objects [8], [9], [10], [11], [12]. Most studies have examined the problem exclusively in the visual domain or have used a relatively small number of objects and categories. Using vision alone, however, would preclude a robot from perceiving the tactile, auditory, and proprioceptive properties of the objects, and thus could severely limit the space of categories that may be learned. On the other hand, if only a small number of objects is used, then there is the potential to severely over-estimate the performance of the classification method (see [13] for a discussion).

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

II. Related Work

Supervised methods for object categorization attempt to establish a direct mapping between the robot's object representation and human-provided semantic category labels. A wide variety of computer vision methods have been developed that attempt to solve this problem using visual image features coupled with machine learning classifiers [14], [15], [16]. Several such methods have been developed for use by robots, almost all working exclusively in the visual domain [8], [17], [10], [18], [12], [19].

Other studies have also demonstrated the ability of robots to assign category labels to objects based on interaction with them [20], [11], [21], [22], [23], [24]. For example, [20] demonstrated that a robot can classify 9 different objects as either a rigid object, a paper object, or a plastic bottle using auditory and joint angle data obtained while the robot shook the objects. Also, [21] described a robot that learned to associate words describing an object (e.g., "cup") with object clusters discovered using an unsupervised method.

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

The second broad limitation of most existing approaches is that they only deal with semantic labels that can be expressed

978-1-4799-3684-7/14/$31.00 ©2014 IEEE

as *unary* relations, i.e., labels that apply to individual objects. Many semantic labels, however, cannot be expressed as unary relations. For example, the label "heavier than", can only be expressed as a *binary* relation. While existing methods can enable robots to classify individual objects, they do not yet allow robots to detect categories that describe object pairs or object groups.

To address these limitations, this paper proposes a relational approach to representing semantic category labels that describe objects, pairwise object relationships, and object groups. Unlike our previous work in object categorization [11], [22], object individuation [25], and object recognition [13], the proposed model can handle many types of object relations beyond simple unary object categories. In addition, the proposed model allows a robot to establish a measure of similarity between different object categories that is grounded in the robot's own sensorimotor repertoire.

III. Experimental Methodology

A. Robot

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

B. Objects and Categories

The robot explored 36 objects in this study. The objects were semi-transparent plastic jars with a height of 8.6 centimeters and a diameter of 9.4 centimeters. The objects varied according to their color, their weight, and their contents, as shown in Figure 1. Thus, each object was either red, green, or blue in color, heavy (337g), medium (250g), or light (177g) in weight, and had glass marbles, rice, beans, or screws inside of it. Every possible combination was included, resulting in a set of $3 \times 3 \times 4 = 36$ objects.

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

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

C. Exploratory Behaviors

The robot explored the 36 objects using 10 exploratory behaviors: grasp, lift, hold, shake, rattle, drop, tap, poke, push, and press. Figure 2 shows before and after images

a) The 36 objects used in this study

b) Color: red, green, and blue

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

d) Weight: light, medium, and heavy

Fig. 1: a) The 36 objects used in this study. b)-d) The three types of variations present within the set of objects explored by the robot: b) color, c) contents, and d) weight.

for each behavior. The behaviors were designed to mimic the exploratory behaviors used by infants ([4], [5]) and were encoded as joint-space trajectories using the Barrett API.

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

D. Data Collection

The robot explored the objects in a series of trials. During each trial the robot recorded static images of the object on the table and then performed its full set of 10 exploratory behaviors in a sequence. Ten trials were performed on each object, resulting in a total of $36 \times 10 \times 10 = 3600$ behavioral

Fig. 2: Before and after images of the 10 exploratory behaviors that the robot used to learn about the objects.

interactions. To minimize any transient noise effects, after a single trial with an object, the object was not explored again until the robot had finished exploring all other objects.

E. Sensorimotor Feature Extraction

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

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

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

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

F. Sensorimotor Contexts

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

IV. Theoretical Model

A. Representing Object Categories with Relations

In logic and set theory, a relation is defined as a property that assigns truth values to k-tuples of objects. When $k = 1$ the relation is a *unary* relation. When $k = 2$ the relation is a *binary* relation. Such relations are common in mathematics (e.g., equality), as well as in everyday human language that describes how two items relate to each other (e.g., "heavier than" and "same color as"). Relations may be reflexive (e.g., "similar to") or transitive (e.g., "heavier than").

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

Using this notation, a wide variety of categories can be modeled as relations. For example, the category "red" can be expressed as a unary relation L^{red} with ground $G(L^{red}) = \mathcal{O}$.

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

B. Learning Relational Object Categories

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

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

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

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

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

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

where w_c corresponds to the estimated reliability of model M_L^c and α is a normalization factor to ensure that the probabilities sum up to 1.0. In our experiments, the models M_L^c were C4.5 decision trees as implemented in the WEKA library [26] and probabilistic estimates were obtained using the class label distributions at the leaves.

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

- *Absolute Distance Features:* Let $\mathbf{f}_{absolute}^c$ be a feature vector such that each entry $\mathbf{f}[i] = |\mathbf{x}_a^c[i] - \mathbf{x}_b^c[i]|$. In other words, the vector $\mathbf{f}_{absolute}^c \in \mathbb{R}^{N_c}$ has the same length as the original sensorimotor observations and represents the absolute difference between those two observations.
- *Signed Distance Features:* Similarly, let $\mathbf{f}_{signed}^c \in \mathbb{R}^{N_c}$ be a feature vector such that each entry $\mathbf{f}[i] = \mathbf{x}_a^c[i] - \mathbf{x}_b^c[i]$.
- *Global Distance Features:* Finally, a third set of features were constructed to represent the global distance between the feature vectors $\mathbf{x}_a^c$ and $\mathbf{x}_b^c$:

  1. L2 distance:
  $$d(\mathbf{x}_a^c, \mathbf{x}_b^c) = \sqrt{\sum_{i=1}^{N_c} (\mathbf{x}_a^c[i] - \mathbf{x}_b^c[i])^2}.$$
  2. Angle-based distance:
  $$d(\mathbf{x}_a^c, \mathbf{x}_b^c) = \frac{\sum_{i=1}^{N_c} \mathbf{x}_a^c[i]\mathbf{x}_b^c[i]}{\sqrt{\sum_{i=1}^{N_c} (\mathbf{x}_a^c[i])^2 \sum_{i=1}^{N_c} (\mathbf{x}_b^c[i])^2}}.$$
  3. Canberra distance:
  $$d(\mathbf{x}_a^c, \mathbf{x}_b^c) = \sum_{i=1}^{N_c} \frac{|\mathbf{x}_a^c[i] - \mathbf{x}_b^c[i]|}{|\mathbf{x}_a^c[i]| + |\mathbf{x}_b^c[i]|}.$$
  4. Chi-square distance:
  $$d(\mathbf{x}_a^c, \mathbf{x}_b^c) = \sum_{i=1}^{N_c} \frac{(\mathbf{x}_a^c[i] - \mathbf{x}_b^c[i])^2}{\mathbf{x}_a^c[i] + \mathbf{x}_b^c[i]}.$$
  5. Modified Sum Squared Error-based distance:
  $$d(\mathbf{x}_a^c, \mathbf{x}_b^c) = \frac{\sum_{i=1}^{N_c} (\mathbf{x}_a^c[i] - \mathbf{x}_b^c[i])^2}{\sum_{i=1}^{N_c} (\mathbf{x}_a^c[i])^2 \sum_{i=1}^{N_c} (\mathbf{x}_b^c[i])^2}.$$

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

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

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

Given the sets $\mathcal{X}_a^c$ and $\mathcal{X}_b^c$ that contain sensorimotor observations with objects o_a and o_b in context c, the robot computes

the estimate for $\hat{Pr}(t \in F(L)|\mathcal{X}_a^c, \mathcal{X}_b^c)$ (i.e., the probability that the object pair belongs to the category L) according to:

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

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

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

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

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

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

More specifically, each element of $\mathbf{f}_{\mathcal{G}}^c$ is estimated by

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

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

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

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

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

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

C. Incremental Learning of Relational Object Categories

In the proposed model, the robot learns target relations by incrementally exploring objects one at a time. After exploring an object, the robot is provided with labels that describe this object, labels that describe object pairs that contain this object, as well as labels that describe object groups containing this object. Let $\mathcal{O}_{known}$ be the currently known set of objects and let $\mathcal{O}_{train}$ be the full set of training objects. At the start of the training process $\mathcal{O}_{known} = \{\}$. Each iteration consists of adding a new object to the set of known objects and can be described by the following steps:

Algorithm 1 update-models$(U, \{D_L\}_{L \in \mathcal{L}}, \{\mathcal{M}_L\}_{L \in \mathcal{L}})$

1: **for** $L \in \mathcal{L}$ **do**
2: Let $U_L = \{\}$.
3: **end for**
4: **for** $t_i \in U$ **do**
5: **for** $L \in \mathcal{L}$ **do**
6: **if** $t_i \in F(L)$ **then**
7: Add $(t_i, +1)$ to dataset U_L.
8: **else if** $t_i \in G(L)$ **then**
9: Add $(t_i, -1)$ to dataset U_L.
10: **end if**
11: **end for**
12: **end for**
13: **for** $L \in \mathcal{L}$ **do**
14: *evaluate*$(\mathcal{M}_L\ ,\ U_L)$
15: $D_L = D_L \cup U_L$.
16: *train*$(\mathcal{M}_L\ ,\ D_L)$
17: **end for**
18: **return** $[\{D_L\}_{L \in \mathcal{L}}, \{\mathcal{M}_L\}_{L \in \mathcal{L}}]$

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

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

At each iteration, for each label $L \in \mathcal{L}$, let D_L be the full set of positive and negative example tuples associated with label L obtained up until exploring object o_{next}. Let $\mathcal{M}_L$ be the set of context-specific recognition models associated with label L. The candidate training points in the set U are then used to update the robot's relational category recognition models as shown in Algorithm 1. Here, the set U_L denotes a labeled dataset of tuples added in the current update step, where each tuple t is labelled as positive if $t \in F(L)$. After the labeled datasets are constructed (lines 4-12), the models for each label L are re-trained.[1]

In addition, for each label L and each sensorimotor context

[1]Each execution of the algorithm requires training $|\mathcal{L}|$ classifiers. The number of training samples for each classifier is bound by $|\mathcal{O}_{known}|+1$ for unary relations, $(|\mathcal{O}_{known}|+1) \times p$ for binary relations, and $(|\mathcal{O}_{known}|+1) \times q$ for relations on object groups. Since p and q are constants, the runtime complexity of the algorithm is $O(|\mathcal{L}| \cdot |\mathcal{O}_{known}| \cdot d^2)$ where d is an upper-bound for the dimensionality of the feature space used to train the C4.5 decision trees (C4.5 is linear in terms of training samples and quadratic with respect to input dimensionality).

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

3. Performance Evaluation Step: At the end of each iteration, the robot's model is evaluated using a hold out set of objects, $\mathcal{O}_{test}$. To do that, tuples are generated that describe individual objects, object pairs and object groups constructed using the set $\mathcal{O}_{test}$. More precisely, the test set contained $|\mathcal{O}|$ tuples describing individual objects, $|\mathcal{O}|\times|\mathcal{O}|$ tuples describing pairs of objects and $\binom{|\mathcal{O}|}{3}$ tuples describing groups of objects.

V. Results

A. Relational Category Recognition Rate

The first experiment was designed to evaluate the model's performance as more and more objects were incrementally added to the robot's training set. To do that, the proposed model was evaluated using 200 runs. For each run, the full set of 36 objects was randomly split into two sets $\mathcal{O}_{train}$ and $\mathcal{O}_{test}$ such that there were 24 objects in $\mathcal{O}_{train}$ and 12 objects in $\mathcal{O}_{test}$. For each relational category $L \in \mathcal{L}$, Cohen's kappa coefficient [27] was chosen as the performance metric:

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

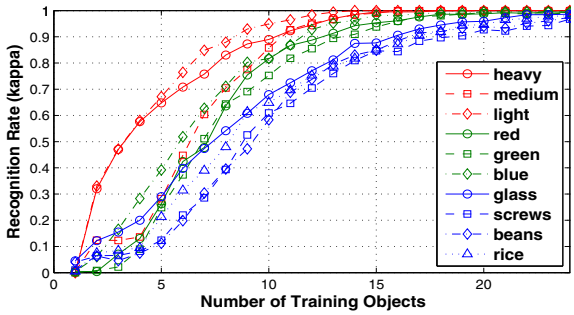
where $Pr(a)$ is the probability of correct classification by the model while $Pr(e)$ is the probability of correct classification by chance. This was necessary as reporting accuracy alone could be misleading, e.g., a model that always predicts -1 as the class label is bound to achieve high accuracy.

The results of this experiment are shown in Figure 3. The figure shows the recognition rates for categories on single objects (top), pairs of objects (middle), and groups of objects (bottom). Plots related to weight are colored in red, those related to color are colored in green, and finally, the plots related to the objects' contents are colored in blue.

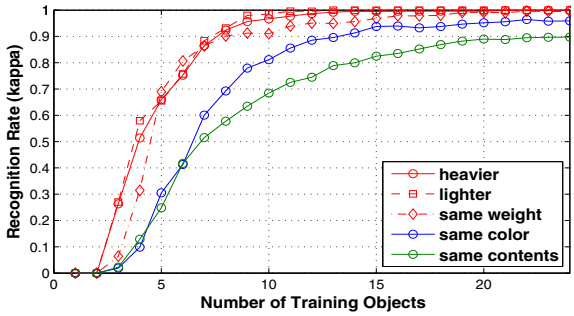
As the robot explores more objects and obtains more training examples, the recognition rates for most relational categories reach a *kappa* of 1.0. Some categories are easier to learn than others. In particular, concepts related to weight are learned much quicker than the rest. One potential explanation is that nearly all sensorimotor contexts produce proprioceptive feedback that is influenced by the weight of the object, while for concepts related to the object's color and contents, there are only a few contexts that produce the relevant information.

B. Estimating Category Similarity

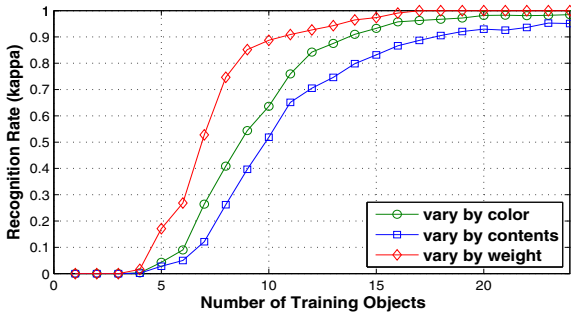
So far, the results show that the robot could learn a wide variety of relational categories in an incremental setting. An important question is whether or not the robot's model can relate those categories in a meaningful way. One way in which the different categories can be related is by considering the weights associated with each sensorimotor context for each



a) Relations on single objects



b) Relations on object pairs



c) Relations on object groups

Fig. 3: Recognition performance as the number of objects explored by the robot is increased from 1 to 24. The figure shows the recognition rates for categories on single objects (a), pairs of objects (b), and groups of objects (c).

category. Figure 4 shows the estimated reliability weights for each sensorimotor context and each category, averaged over all 200 simulated runs. Here, each square corresponds to a recognition model M_L^c. The shade corresponds to the model's estimated *kappa* statistic, where 1.0 is black and 0.0 is white. The figure shows that there is great diversity in terms of which sensorimotor contexts are useful for which categories. Furthermore, it also shows that there is a greater number of sensorimotor contexts relevant to weight-related categories, which may explain why those categories are learned quicker than categories related to the object's color and contents.

Figure 5 shows a 2D ISOMAP [28] projection in which two categories are close if the same contexts are useful for recognizing them. The projection was computed by associating a weights vector $\mathbf{w}_L$ of length $|\mathcal{C}|$ with each category such that each element of the vector was equal to the *kappa* reliability measure of the corresponding sensorimotor context.

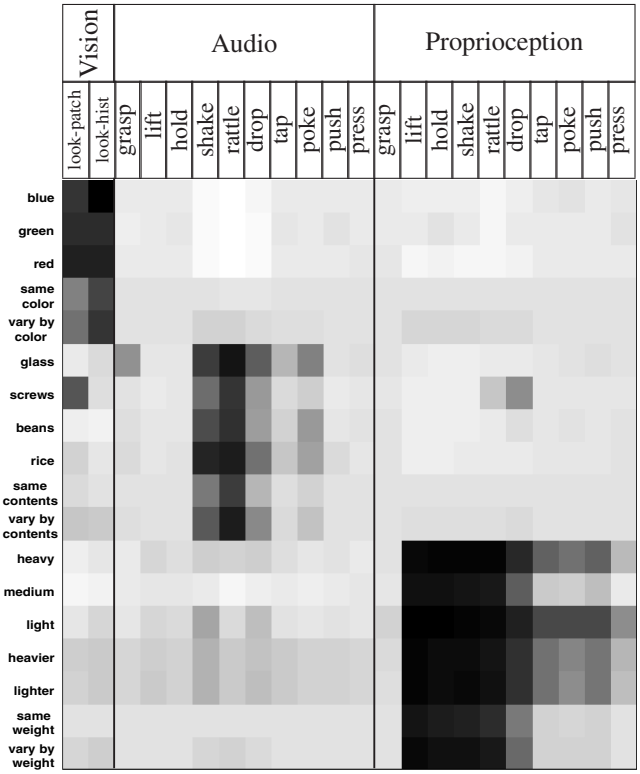


Fig. 4: Estimated reliability weights associated with each sensorimotor context for each category. Each square corresponds to a recognition model M_L^c and is associated with a specific category and sensorimotor context. The shade shows the estimated *kappa* statistic of the model, where white indicates *kappa* of 0.0 while black indicates 1.0.

The vectors were used to compute a $|\mathcal{L}| \times |\mathcal{L}|$ distance matrix by computing the Euclidean distance for each pair. The matrix was then used as input to the ISOMAP algorithm [28].

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

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

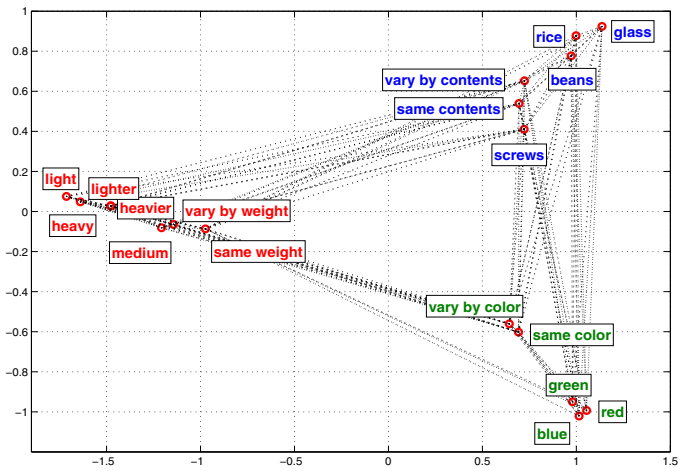


Fig. 5: An ISOMAP [28] projection showing the similarity of the learned categories. Closeness in the projection indicates that the two categories can be recognized well using the same sensorimotor contexts.

To test this, the robot's model was first trained on $\mathcal{L}_{known}$ categories and was then further trained on the remaining relational category L_{test} using the same procedure. Given a set of similar categories $\mathcal{L}_{similar} \subset \mathcal{L}_{known}$, a set of context weights $\mathbf{w}_{L_{test}}$ was computed such that $w^c_{L_{test}} = \mathbf{E}[w^c_L | L \in \mathcal{L}_{similar}]$. This process was repeated such that each relation in $\mathcal{L}$ was used once as L_{test}. Figure 6 shows the results of this test, where training was halted after exploring 5 training objects. The figure shows that by relating a new category to ones that are known, a robot can substantially improve its performance at test time, even if trained on a much smaller set of objects. This result is especially important because using exploration to estimate which behaviors and sensory modalities are useful for a given category may become more difficult as the set of categories grows larger and larger.

VI. Conclusion and Future Work

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

In addition to achieving high recognition rates for all three types of categories, the robot was also able to establish a measure of similarity between the different relational categories that it learned. More specifically, two categories were deemed similar if they could be recognized using the same behaviors and sensory modalities and dissimilar otherwise.

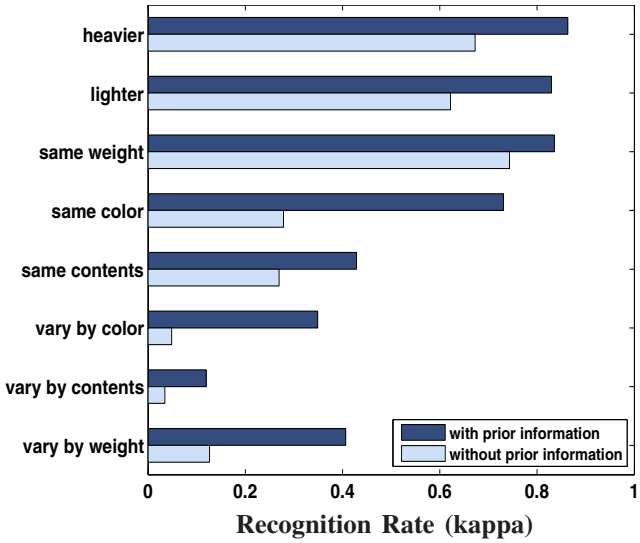


Fig. 6: Visualization of the recognition improvement obtained when using prior information that relates a novel category to categories that are already learned. For this test, only 5 training objects were used and the results were averaged over 50 different runs. This figure shows that prior information that links the target category to familiar categories can be used to substantially improve the recognition rate.

Our results showed that this type of representation is especially useful when the robot is tasked with learning a new relational category that is similar to already known categories.

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

REFERENCES

[1] F. Ashby and W. Maddox, "Human category learning," *Psychology*, vol. 56, no. 1, pp. 149–178, 2005.

[2] R. Hammer, G. Diesendruck, D. Weinshall, and S. Hochstein, "The development of category learning strategies: What makes the difference?" *Cognition*, vol. 112, no. 1, pp. 105–119, 2009.

[3] 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*, vol. 52, no. 2, pp. 699–709, 2010.

[4] 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.

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

[6] M. Ernst and H. Bulthof, "Merging the Senses into a Robust Percept," *Trends in Cognitive Science*, vol. 8, no. 4, pp. 162–169, 2004.

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

[8] L. Lopes and A. Chauhan, "Scaling up category learning for language acquisition in human-robot interaction," in *Proceedings of the Symposium on Language and Robots*, 2007, pp. 83–92.

[9] S. Griffith, J. Sinapov, V. Sukhoy, and A. Stoytchev, "A behavior–grounded approach to forming object categories: Separating containers from non-containers," *IEEE Transactions on Autonomous Mental Development*, vol. 4, no. 1, pp. 54–69, 2012.

[10] Z. Marton, R. Rusu, D. Jain, U. Klank, and M. Beetz, "Probabilistic categorization of kitchen objects in table settings with a composite sensor," in *Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2009, pp. 4777–4784.

[11] J. Sinapov and A. Stoytchev, "Object category recognition by a humanoid robot using behavior-grounded relational learning," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2011, pp. 184–190.

[12] A. Leonardis and S. Fidler, "Learning hierarchical representations of object categories for robot vision," *Robotics Research*, pp. 99–110, 2011.

[13] J. Sinapov, T. Bergquist, C. Schenck, U. Ohiri, S. Griffith, and A. Stoytchev, "Interactive object recognition using proprioceptive and auditory feedback," *International Journal of Robotics Research*, vol. 30, no. 10, pp. 1250–1262, 2011.

[14] R. Fergus, P. Perona, and A. Zisserman, "A visual category filter for Google images," *Computer Vision-ECCV*, pp. 242–256, 2004.

[15] J. Ponce, *Toward category-level object recognition*. Springer-Verlag New York Inc, 2006, vol. 4170.

[16] A. Opelt, A. Pinz, M. Fussenegger, and P. Auer, "Generic object recognition with boosting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 416–431, 2006.

[17] K. Lai and D. Fox, "3D laser scan classification using web data and domain adaptation," in *Proc. of Robotics: Science and Systems (RSS)*, 2009.

[18] W. Wohlkinger and M. Vincze, "3D object classification for mobile robots in home-environments using web-data," in *IEEE 19th International Workshop on Robotics in Alpe-Adria-Danube Region (RAAD)*, 2010, pp. 247–252.

[19] K. Lai, L. Bo, X. Ren, and D. Fox, "A large-scale hierarchical multi-view RGB-D object dataset," in *Proc. of the IEEE International Conference on Robotics & Automation (ICRA)*, 2011.

[20] S. Takamuku, K. Hosoda, and M. Asada, "Shaking eases object category acquisition: Experiments with a robot arm," in *Proceedings of the Seventh International Conference on Epigenetic Robotics*, 2007.

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

[22] J. Sinapov, C. Schenck, K. Staley, V. Sukhoy, and A. Stoytchev, "Grounding semantic categories in behavioral interactions: Experiments with 100 objects," *Robotics and Autonomous Systems*, 2012 (In Press).

[23] O. Yürüten, K. F. Uyanık, Y. Çalışkan, A. K. Bozcuoğlu, E. Şahin, and S. Kalkan, "Learning adjectives and nouns from affordances on the iCub humanoid robot," in *From Animals to Animats 12*. Springer, 2012, pp. 330–340.

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

[25] J. Sinapov and A. Stoytchev, "Grounded object individuation by a humanoid robot," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2013, pp. 4981–4988.

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

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

[28] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.