A Behavior–Grounded Approach to Forming Object Categories: Separating Containers from Non-Containers

Shane Griffith, Jivko Sinapov, Vladimir Sukhoy, and Alexander Stoytchev, *Member, IEEE*

***Abstract*—This paper introduces a framework that allows a robot to form a single behavior–grounded object categorization after it uses multiple exploratory behaviors to interact with objects and multiple sensory modalities to detect the outcomes that each behavior produces. Our robot observed acoustic and visual outcomes from 6 different exploratory behaviors performed on 20 objects (containers and non-containers). Its task was to learn 12 different object categorizations (one for each behavior–modality combination), and then to unify these categorizations into a single one. In the end, the object categorization acquired by the robot matched closely the object labels provided by a human. In addition, the robot acquired a visual model of containers and non-containers based on its unified categorization, which it used to label correctly 29 out of 30 novel objects.**

***Index Terms*—Artificial intelligence, intelligent robots, learning systems, robots, object categorization, developmental robotics**

I. INTRODUCTION

OBJECT categorization is a fundamental skill that emerges early in the course of human infant development [3]. From the moment infants begin to manipulate objects, they can identify differences between them in terms of the sensations that the objects produce [4]. As infants gain more control over their bodies, they begin to grasp, mouth, scratch, and bang objects in order to learn about them [5]. These exploratory behaviors and the sensations that they produce lay the foundations for forming many different object categories [6].

Each object category that infants learn in this way is associated with a set of functional and perceptual properties [7]. For example, containers have the functional property that a block placed inside of a container will start to move when the container is moved. Containers also have the perceptual property that they look concave. Different object categories are represented by different collections of properties. Over time, infants' category representations become more diverse [8].

In contrast, the majority of object categorization systems in artificial intelligence and robotics are almost entirely image–based. Given a clear view of the object these disembodied classifiers can accurately categorize objects using visual appearance alone [9]. Because they do not use the robot's body, however, the functional properties of objects cannot be learned by these systems [10]. Additional information sources are required for learning object categories that capture something about the functional properties of objects.

The problem of learning object categories becomes more complex when multiple information sources are available. For example, consider a robot that has microphones and cameras, which record information streams while the robot interacts with objects. Objects that were traditionally categorized only by their visual appearance can now also be categorized by the sounds that they produce or by their movements as the robot performs different behaviors on them. To make things even more complicated, each behavior–modality combination results in a different object categorization. It is not straightforward to figure out which of these categorizations are more meaningful or if it is possible to combine them into a single categorization.

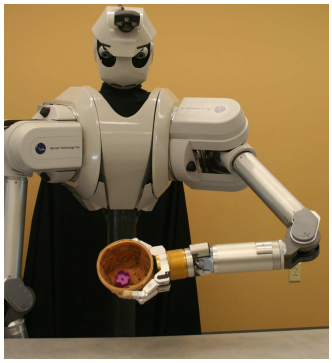

Fig. 1. The upper-torso humanoid robot, shown here shaking one of the objects used in the experiments. The small plastic block inside the object produces auditory and visual events, which the robot can detect and use to categorize the object as a container.

Research in developmental robotics has shown that robots can form meaningful behavior–grounded object categories using a single exploratory behavior and a single sensory modality [1][2][11]. Because these categories are grounded in the robot's own behavior, the robot can test, verify, and correct that knowledge autonomously without human intervention [12][13]. More work is needed, however, to show how a robot with an extensive behavioral and perceptual repertoire can reconcile the different object categorizations that result from each behavior–modality combination.

This paper introduces a computational framework that allows a robot to form a single behavior–grounded object categorization after it uses multiple exploratory behaviors to interact with objects and multiple sensory modalities to detect the outcomes that each behavior produces. Our robot (see Fig. 1) observed acoustic and visual outcomes from 6 different exploratory behaviors performed on 20 objects (containers and non-containers). Its task was to learn 12 different object categorizations (one for each behavior–modality combination), and then to unify these categorizations into a single one. In the end, the robot divided the objects into object categories that a human would call containers and non-containers. Furthermore, the robot was able to learn a visual model of the two categories and use this model to categorize novel objects.

Manuscript received December 22, 2010. This paper combines methodology from our earlier work that appeared in ICDL 2009 [1] and ICRA 2010 [2], and uses the ICRA 2010 dataset. The first author was supported by a National Science Foundation Graduate Research Fellowship.

All four authors are with the Developmental Robotics Laboratory, Department of Electrical and Computer Engineering, Iowa State University, Ames, IA 50011 USA (e-mail: shaneg@iastate.edu).

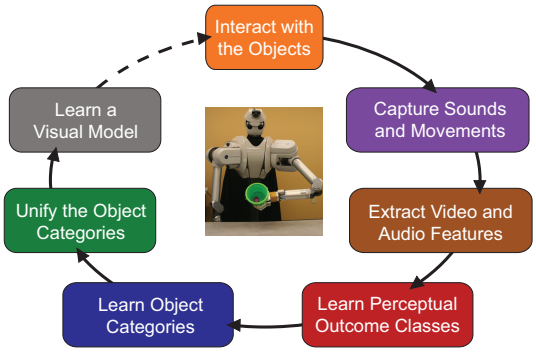

Fig. 2. The framework used by the robot to learn object categories. First, the robot interacts with the objects and observes the outcomes that are produced. The extracted auditory and visual features are used to learn perceptual outcome classes. These are used to form object categories, one for each behavior–modality combination. The categories are unified using consensus clustering into a single category. Finally, a visual model is trained that can recognize the categories of novel objects. The dotted line is to show that the visual model could be used to guide and refine future interactions with objects.

Fig. 2 shows a high-level overview of the framework described in this paper. First, the robot interacts with the objects and observes the sounds and the movement patterns that the objects produce. Perceptual features are learned from the raw sensory data, and feature extraction is performed. Next, the robot captures the different functional properties of the objects by clustering the extracted features into outcome classes. Object categories are learned by clustering the objects based on how often the different functional properties occur with each object. The object categorization step also includes a unification process, which unifies the object categories produced from multiple behaviors and sensory modalities. A visual model that can predict the categories of novel objects is learned in the last step. The visual classifier is trained using the object category labels produced by the unified clustering procedure. The visual classifier could also help guide and refine the robot's future interactions with objects.

II. RELATED WORK

A. Developmental Psychology

The object categorization framework described in this paper was motivated by work in developmental psychology, which attempts to explain how infants perform categorization tasks. Psychologists have found that if infants are presented with a set of objects, in which several of the objects have a common functional property, then the infants will categorize the objects based on this property [14]. In this context, infants categorize the objects by the sounds that they make or by their visual movement patterns, and not by static perceptual properties like object shape or color [14][15].

Infants may categorize objects in this way because they learn from the events that capture their attention [16]. For example, an object that makes noises will automatically draw their attention [17]. Events that violate their expectations (e.g., an unexpected movement pattern) also capture their attention [18]. Spelke argues that from birth infants can predict the movement patterns of objects and form expectations about their trajectories [19]. Infants know that there is no action without contact, that two objects cannot merge into one, that one object cannot split into multiple objects, etc. [19].

Typically, the expectations of infants seem to agree with the laws of real-world physics, but there are some exceptions. Needham *et al.* [20] found that when 7.5-month-old infants see a key-ring with keys, they perceive two distinct objects and thus predict that the key-ring and the keys will *move separately* when the key-ring is moved. More experienced 8.5-month-old infants, however, expect that the key-ring and the keys will move together because they have seen and heard the two 'distinct' objects move together many times [20]. A similar shift in expectations has been observed while studying infants' knowledge of containers: infants come to expect that an object inside a container will move with the container when the container is moved [21][22][23].

Together, these findings suggest that there is a gradual process of object category learning, in which object category representations are progressively grounded in different actions and their outcomes. Indeed, it is believed that infants first represent "what actions can be done on objects of certain kinds" [24], before they incorporate the object's visual shape into their representations. This may imply that behaviors and their outcomes form the bases of infants' initial concepts. Baillargeon has shown that only after infants have formed an "initial concept" do they begin to incorporate variables into their representations that serve to refine their predictions [25]. Passively observable object properties such as shape are learned gradually over time if they consistently appear with members of a category [22][26][27][28][29][30].

The fact that infants gradually improve their object category representations as they gain more experience supports Cohen's hypothesis that there is an information processing mechanism underlying object categorization [31]. One information processing mechanism known to be used by humans of all ages is the detection of the frequency of occurrence of a stimulus [32]. Humans implicitly extract the frequency information for a variety of naturally occurring phenomena [32]. It is reasonable to assume that infants may also use frequency information to separate objects into categories.

Neuroscientists have suggested that multimodal representations of categories are formed in high–level convergence zones in the hierarchical organization of the brain [8][33][34]. At these convergence zones, fragments of data from multiple modalities are bound together if they occur coincidentally or sequentially in space and time [33][34]. The representation of an abstract object category involves a multiregional activation of the brain, which reaches many different sensory areas [33][34]. The resulting multimodal representation encodes how objects in the category sound, move, look, feel, etc.

This paper introduces an object categorization framework for robots that was inspired by the studies with infants mentioned above. We believe that object category learning by robots may be more generalizable (i.e., less tailored toward a specific categorization problem) if it is modeled after the object categorization abilities of infants. One of the first abstract categories that infants learn is that of containers [35], which is why we evaluated the framework on a container/non-container categorization task.

B. Object Categorization in Robotics

Relatively few papers have addressed the task of interactive object categorization by a robot. Pfeifer and Scheier [36] were among the first to tackle this problem. They programmed a mobile robot with an ability to learn how to move differently–sized objects for the purpose of cleaning its environment. The robot learned that it could carry small objects and push medium–sized objects. It ignored the large objects that it could not push or carry, which allowed it to learn faster. Thus, the robot implicitly categorized objects by their *movability*.

Metta and Fitzpatrick [11] showed how a humanoid robot could simplify the problem of object segmentation by pushing objects. When the robot made contact with an object, the object was easily segmented from the background, which allowed the robot to construct a model for it. The robot used this procedure to interact with 4 different objects and to implicitly categorize them by their *rollability*.

Ugur *et al.* [37] showed how a mobile robot could learn about the traversability of objects in a simulated environment. The robot attempted to traverse an area that had randomly dispersed spheres, cylinders, and cubes. It learned which objects could be pushed aside (spheres and cylinders in lying orientations), and which could not (cubes and cylinders in upright orientations).

Learning the similarity between objects is a problem closely related to object categorization. Sinapov and Stoytchev [38] showed how a humanoid robot could describe different tools using a hierarchical taxonomy of outcomes. The robot constructed outcome taxonomies for 6 different stick-shaped tools based on its interactions with them and used the outcome taxonomies to measure the similarity between the tools. Montesano *et al.* [39] introduced a system that a robot could use to learn relationships between its actions, the perceptual properties of objects, and the observed effects. The system was evaluated with data from interactions with differently–sized spheres and cubes.

Sinapov *et al.* [40] demonstrated that acoustic object recognition is feasible even with a large set of objects and when multiple behaviors are performed. The robot listened to the acoustic outcomes produced by 36 objects as it grasped, shook, dropped, pushed, and tapped them. Individually, some behaviors were more useful for acoustic object recognition than others. As the robot performed more behaviors on an object, however, the recognition accuracy approached 99%. In a follow up study [41], the robot categorized the material type of the objects and whether or not they had contents inside them. The object categorizations were grounded in the acoustic object recognition models used by the robot.

Nakamura *et al.* [42] introduced an unsupervised approach to multimodal object categorization, in which objects were categorized by the similarity of their perceptual features. A robot interacted with 40 different objects, which included 8 different categories of children's toys. The robot squeezed objects to observe hardness, viewed objects from different angles to obtain visual appearance features, and shook objects to capture acoustic properties. Results showed that when all three modalities were used the robot's object categorization closely resembled human–provided ones. Further results showed that visual appearance information could be used to infer the hardness of a novel object, but not its acoustic properties [42].

Griffith *et al.* [1] introduced a framework for interactive object categorization by a robot. A humanoid robot dropped a block above an object and observed co-movement patterns between the two as it pushed the object. The robot categorized 5 containers and 5 non-containers using the frequency with which different co-movement patterns occurred with each object. The behavior-grounded categorization allowed the robot to learn a perceptual model of containers, which it used to infer the functional properties of novel objects.

In a follow-up study, Sahai *et al.* [43] used this object categorization framework for a robotic writing task. A humanoid robot scribbled with 12 different objects on 12 different surfaces. The robot categorized objects by the frequency with which each object left a mark on a surface. Also, it categorized surfaces by the frequency with which each surface preserved the traces left by each object. The categorizations separated the objects and the surfaces that provided the most utility in robotic writing tasks from those that provided the least.

This paper employs some of the same methodology that Sinapov *et al.* [40] have used for object recognition tasks. We were first motivated to conduct acoustic object categorization experiments in [2]. In that study, a robot categorized containers and non-containers based on the sounds they produced while interacting with them. The robot observed acoustic outcomes as it dropped a block above an object, grasped the object, moved the object, shook the object, flipped the object, and dropped the object. The results showed that some of the robot's behaviors produced sounds that were useful for categorizing the objects; other behaviors were not as useful.

This paper extends our previous work [1][2][43] by using multiple sensory modalities (audio and vision) to learn object categories. It uses the dataset from [2], which also contained visual data that is analyzed here for the first time. The new framework described here also unifies the robot's categorizations from multiple behaviors and modalities into a single one. The robot also formed a visual model of containers and non-containers based on its unified object categorization, which it used to infer the category of novel objects. It should be noted that in this paper, the identity of each object is assumed to be known. In other words, the acoustic data and the visual data corresponding to actions on a specific object is labeled with the object's ID. What is unlabeled is the category (container versus non-container).

III. EXPERIMENTAL SETUP

A. Robot

All experiments were performed using the upper-torso humanoid robot shown in Fig. 1. The robot's arms are two 7-DOF Whole Arm Manipulators (WAMs) manufactured by Barrett Technology. They are mounted in a configuration similar to that of human arms. Two Barrett Hands are used as end effectors. The WAMs are controlled in real time at 500 Hz over a CAN bus interface.

Fig. 3. The robot's vision system: a) the 3D camera (a ZCam [44]); b) a color image of the red non-container captured by the camera when mounted on the robot; c) the depth image corresponding to b).

The robot's auditory system consists of two Audio-Technica U853AW Hanging Microphones, which are mounted in the robot's head. The microphones' output is fed through two ART Tube MP Studio Microphone pre-amplifiers. The signal from the amplifiers is fed to a Lexicon Alpha bus-powered interface, which transmits the signal to a Linux PC over USB. For the experiments in this paper, sound was captured from a single microphone using the Java Sound API at 44.1 KHz over a 16-bit mono channel.

The robot's visual system consists of a single 3D camera—a ZCam manufactured by 3DV Systems [44]. The ZCam captures 320x240 depth images and 640x480 color images. The resolution of the depth images is accurate to $\pm$1-2 cm. The depth images are calculated by first pulsing infrared light in two frequencies and then detecting and processing the reflected pulses of light. Figure 3 shows a close up of the ZCam and its field of view when mounted on the robot.

B. Objects

The robot interacted with a small plastic block and 10 different objects (shown in Fig. 4). Each of the 10 objects was a container in one orientation and a non-container when flipped over. Flipping the containers was an easy way for the robot to learn about non-containers while preserving the dimensions of the objects in the two categories.

The objects were selected to have a variety of shapes, sizes, and materials. Objects were tall, short, rectangular and round. They were made of plastic, metal, wicker, and foam. A few objects that were initially selected could not be used because they were too large to be grasped. Also, the aluminum fingers of the Barrett Hand did not create a firm grip with some objects, which was important for a large-scale experimental study like this one. Therefore, rubber fingers were stretched over each of the robot's three fingers to achieve more reliable grasps.

C. Robot Behaviors

The robot performed six behaviors during each trial: 1) *drop block*, 2) *grasp* object, 3) *move* object, 4) *shake* object, 5) *flip* object, and 6) *drop object*. Before the start of each trial a person placed the block and the object at specific locations. The robot grasped the block and positioned its hand in the area above the object before executing the six behaviors. Figure 5 shows the sequence of interactions for two separate trials (one with a container and one with a non-container). The individual behaviors are described below.

Fig. 4. The objects used in the experiments. (**Containers**) The first two rows show the 10 container objects: wicker basket, metal trash can, potpourri basket, flower pot, bed riser, purple bucket, styrofoam bucket, car trash can, green bucket, and red bucket. (**Non-containers**) The second two rows show the same 10 objects as before but flipped upside down, which makes them non-containers for this particular robot with this particular set of behaviors.

1) Drop Block Behavior: The height from which the robot dropped the block over the object was the same for all trials/objects. The drop positions were randomly selected from a 2D Gaussian distribution centered above the object in a plane parallel to the table. The standard deviation of this distribution was empirically set to be proportional to the width (in pixels) of each object. Inverse kinematics was used to move the robot's hand to the drop position. Thus, the small block fell inside the container during approximately 70% of all trials with containers. During the other 30% of the trials with containers (and during trials with non-containers) the block fell on the table. In some cases the block rolled off the table (approximately 5% of all trials). In these cases, the block was left off the table for the duration of the trial.

Dropping the block produced a lot of noise and large visual movements. During trials when the block fell into a container, however, the block moved less and made less noise.

2) Grasp Behavior: The robot grasped the object after dropping the block above it. Grasping the object produced little noise and only slightly moved the object. Thus, we expected that the categorizations resulting from this behavior would be somewhat less meaningful.

The robot failed to grasp the object in some cases (63 out of 2000 trials), which occurred when one of the robot's three fingers did not properly close. In these situations, a person monitoring the experiments recorded the error. All of the problematic trials were repeated after the initial round of experiments were completed.

3) Move Behavior: After grasping the object with its left hand, the robot moved the object toward the right side of its body. Moving the object produced little noise, as the object made little contact with the table and the block laid still either on the table or inside a container. So, we expected that the robot would only form meaningful categories based on its visual observations from this behavior.

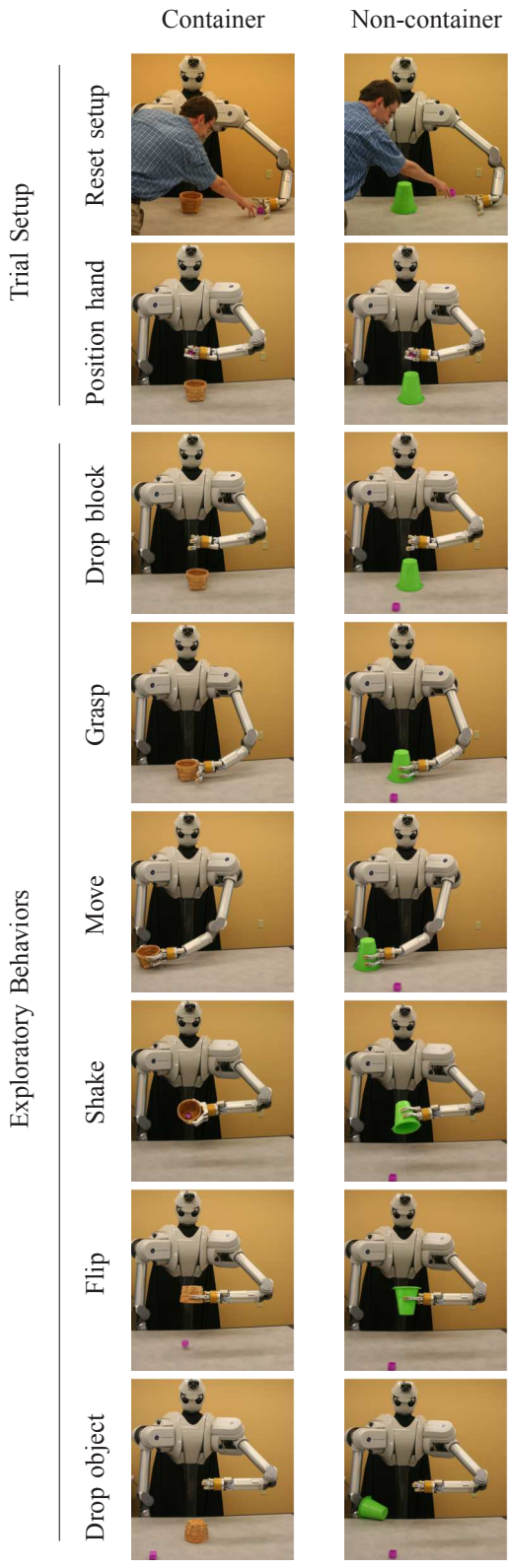

Fig. 5. Snapshots from two separate trials with a container and a non-container object. Before each trial a human experimenter reset the setup by placing the block and the object at marked locations. After grasping the block and positioning its arm at a random location above the object the robot performed the six exploratory behaviors one after another.

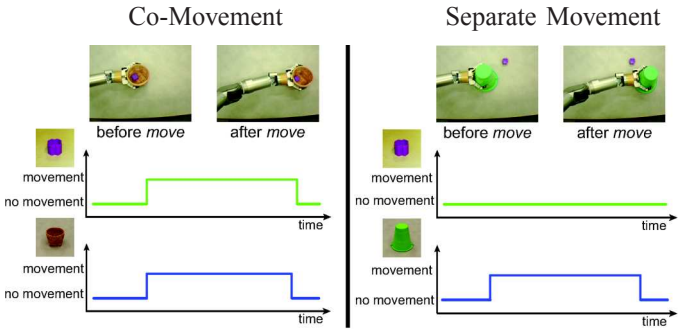

Fig. 6. Transformation of the video data into movement sequences for two different executions of the *move* behavior. (**Left**) Co-movement was observed during trials in which the block moved when the object moved. Here, the block was inside a container and moved with it when the robot performed the *move* behavior. (**Right**) Separate movement outcomes occurred when the block fell to the side of a container or during trials with non-containers.

4) Shake Behavior: The robot shook the object after moving it. Shaking took place well above the table to avoid banging the object into the table. Shaking the object caused a lot of movement and produced a lot of noise when the block was inside a container. During trials with non-containers, however, the behavior produced little noise and rarely caused co-movement between the block and the object. So, we expected that meaningful categorizations would be produced for this behavior.

5) Flip Behavior: The robot flipped the object over after shaking it. Flipping the object produced sounds only during trials in which the block was inside a container. During these trials, the block fell out of the container and crashed into the table. Thus, we expected that the robot would capture differences between containers and non-containers using this behavior.

6) Drop Object Behavior: The robot dropped the object after flipping it. Dropping the object always produced sounds and sometimes caused movement patterns between the block and the object. The acoustic outcomes and the visual movement patterns, however, were seldom sufficient to discriminate containers from non-containers. So, we expected that the robot might capture differences in size or material properties using this behavior, but not functional differences.

IV. METHODOLOGY

A. Data Collection

Multiple audio and video sequences were collected by the robot while it was performing the six exploratory behaviors, $\mathcal{B}$ = [*drop block*, *grasp*, *move*, *shake*, *flip*, *drop object*]. The six behaviors were organized into trials and always performed one after another (see Fig. 5). For each of the 20 objects, the robot performed 100 trials, for a total of $20 \times 100 = 2000$ trials. Because each trial consisted of 6 behaviors, the robot performed $6 \times 2000 = 12000$ behavioral interactions.

Another way to describe this dataset is to say that each behavior (e.g., *shake*) was performed 100 times on each of the 20 objects. Thus, each of the six behaviors was performed 2000 times. During every interaction the robot recorded the tuple (B, O, A, V), where $B \in \mathcal{B}$ was one of the six behaviors

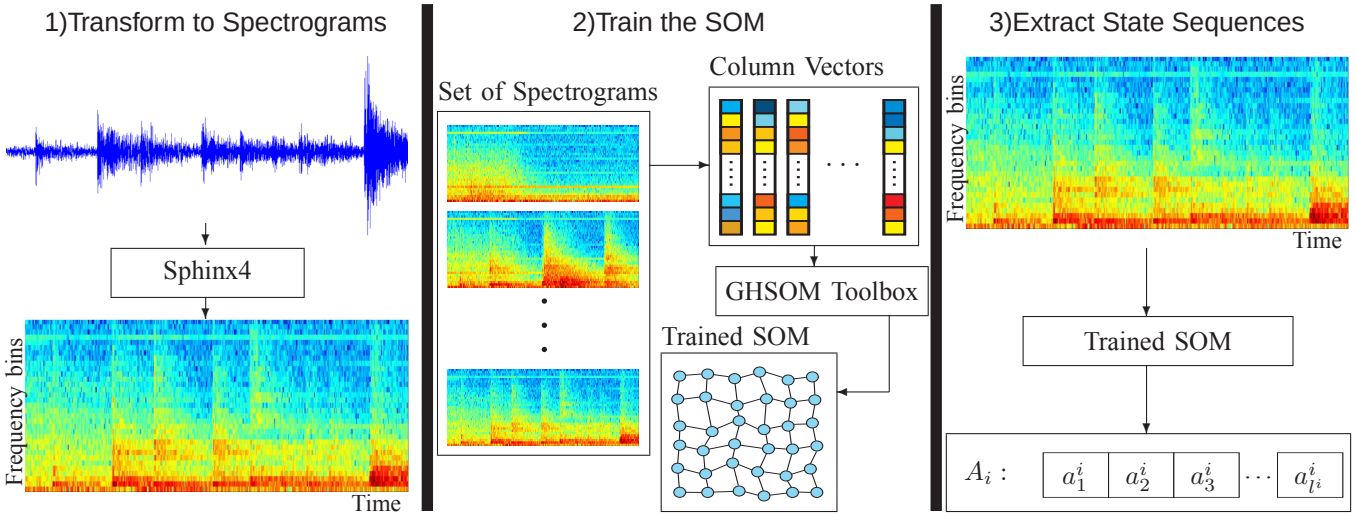

Fig. 7. The feature extraction process for acoustic observations: 1) The raw sound wave produced by each behavior is transformed to a spectrogram. Each spectrogram has 33 bins (represented as column vectors), which capture the intensity of the audio signal for different frequencies at a given time slice. Red color indicates high intensity while blue color indicates low intensity. 2) An SOM is trained using randomly selected column vectors from the spectrograms for a given behavior. 3) The column vectors of each spectrogram are mapped to a discrete state sequence using the states of the SOM. Each column vector is mapped to the most highly activated SOM node when the column vector is used as an input to the SOM. See the text for more details.

performed on object $O \in \mathcal{O}$, A was the recorded audio sequence, and V was the recorded video sequence. Audio data was sampled at 44.1 KHz over a 16-bit mono channel and stored as wave files. Visual data was captured from the robot's 3-D camera as a sequence of 640x480 color images and 320x240 depth images recorded at roughly 20 fps. The six behaviors lasted between 1 and 4 seconds each. *Drop object* and *grasp* took 1 second to complete; *drop block* and *flip* 2 seconds; *move* 3 seconds; and *shake* 4 seconds.

The order in which the robot interacted with the objects was chosen to minimize the effect of changing background noise. In a dataset of this magnitude, transient ambient noise can negatively impact the results (e.g., noise from the air conditioning system or computer fans). Therefore, the robot performed one trial with each of the twenty objects shown in Fig. 4 before moving on to the second trial with the first object, and so on.

B. Movement Detection

The robot processed the frames from the ZCam to track the positions of the block and the object and to detect their movements. During each trial, the object was tracked using the center of mass of the largest blob with the corresponding color. The same was done for the block, which had a different color from the object. Movement was detected when the $[x, y]$ position of the block or the $[x, y]$ position of the object changed by more than a threshold, δ, over a short temporal window $[t', t'']$. The threshold, δ, was empirically set to 2.5 pixels per two consecutive frames. A box filter with a width of 3 was used to filter out noise in the movement detection data. The movement detection data for the block and the object from one behavioral interaction was used to create a movement sequence (see section. IV.D). Figure 6 shows the sequence of detected movements of the block and the object for two different executions of the *move* behavior.

C. Auditory Feature Extraction

Auditory features were extracted automatically by representing the sounds produced by each behavioral interaction as a sequence of nodes in a Self-Organizing Map (SOM). The feature extraction process is the same as in our previous work [40]. The three stage process includes: 1) a Discrete Fourier Transform which takes a 44.1 KHz audio sample, A^i, and converts it to a 33 bin spectrogram, $P_i = [p_1^i, \ldots, p_l^i]$, where $p_j^i \in \mathbb{R}^{33}$ (the DFT window length was 26.6 ms, computed every 10 ms); 2) a 2D SOM that is trained with the spectrograms corresponding to one of the robot's six exploratory behaviors; and 3) a mapping, $\mathcal{M}(p_j^i) \rightarrow a_j^i$, of each spectrogram column vector, p_j^i, to the most highly activated state, a_j^i, in the SOM when p_j^i is presented as an input to the SOM (see Fig. 7). The mapping process results in a state sequence $A_i = a_1^i a_2^i \ldots a_{l^i}^i$, where each a_j^i stands for one of the SOM nodes. For each behavioral interaction, the corresponding SOM was trained using only 5% of the available column vectors (see Fig. 7), which were randomly selected from the spectrograms captured during this behavior.

The robot performed this procedure six times, once for every behavior. It acquired a set of state sequences, $\{A_i\}_{i=1}^{2000}$, for each of its six behaviors. This feature extraction method was chosen because it does not require a human to select the acoustic features that the robot will have to use. The algorithm identified and computed features in an unsupervised way. See [40] for further details.

D. Visual Feature Extraction

The robot extracted visual features using a procedure similar to that used for extracting auditory features (see Fig. 8). That is, visual features were extracted automatically by representing the movement sequences of the block and the object produced by each behavioral interaction as a sequence of nodes in a Self-Organizing Map (SOM). The three stage process includes: 1) a

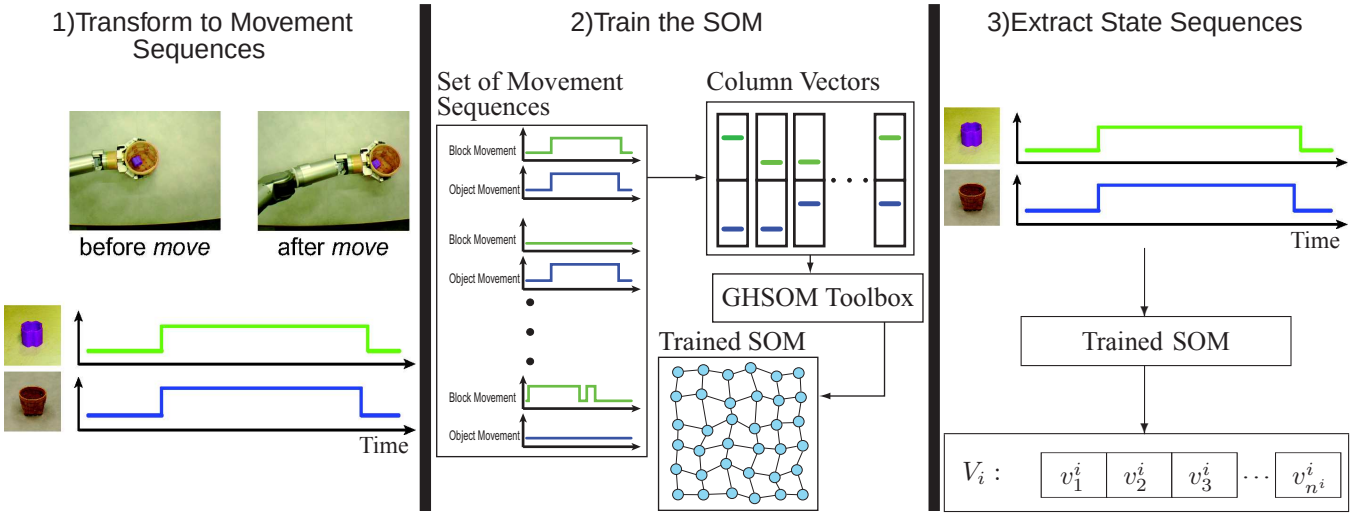

Fig. 8. The feature extraction process for visual observations: 1) The video data recorded during each execution of a given behavior is transformed into a movement sequence. The co-movement sequence pictured here was obtained after the robot performed the *move* behavior with one of the containers. 2) An SOM is trained using randomly selected column vectors from the set of all movement sequences for a given behavior. 3) Each movement sequence is mapped to a discrete state sequence of SOM states. To do this, each column vector of the movement sequence is mapped to the most highly activated SOM node when the column vector is used as an input to the SOM. See the text for more details.

movement detection step that takes a recorded video sequence, V^i, captured at 20 frames per second, and converts it into a movement sequence, $M_i = [m_1^i, \ldots, m_l^i]$, where $m_j^i \in \mathbb{R}^2$; 2) a 2D SOM that is trained with the movement sequence corresponding to one of the robot's six exploratory behaviors; and 3) a mapping, $\mathcal{M}(m_j^i) \rightarrow v_j^i$, of each co-movement column vector, m_j^i, to the most highly activated state, v_j^i, in the SOM when m_j^i is presented as an input to the SOM (see Fig. 8). The mapping process results in a state sequence $V_i = v_1^i v_2^i \ldots v_{n^i}^i$, where each v_j^i stands for one of the SOM nodes.

Again, the robot performed this procedure six times, once for every behavior. It acquired a set of state sequences, $\{V_i\}_{i=1}^{2000}$, for each of its six behaviors. The parameters used for training each visual SOM were the same parameters used for training the acoustic SOMs. The only difference between the two feature extraction procedures was the size of the column vectors. The column vectors used to represent spectrograms had 33 rows; the column vectors used to represent co-movement sequences had 2 rows.

E. Learning Perceptual Outcome Classes

The acoustic outcome patterns produced by a given behavior can be clustered automatically to obtain auditory outcome classes. Similarly, the visual movement patterns produced by a given behavior can be clustered automatically to obtain visual outcome classes. In our case, the robot performed 6 behaviors and captured data from 2 modalities, so its task was to learn $6 \times 2 = 12$ separate sets of outcome classes. More formally, the robot learned k outcome classes from the set of SOM state sequences, $\{A_i\}_{i=1}^{2000}$ or $\{V_i\}_{i=1}^{2000}$, observed for one modality during the execution of one of the 6 behaviors. An unsupervised hierarchical clustering procedure based on the *spectral clustering* algorithm was used for this task (spectral clustering is a similarity-based clustering algorithm [45]). The procedure was performed 12 different times to obtain 6 different sets of acoustic outcome classes and 6 different sets of visual outcome classes. Figures 9 and 10 illustrate the process of learning acoustic outcome classes and visual outcome classes for one behavior, respectively.

The *spectral clustering* algorithm requires a similarity matrix as its input. The similarity between outcomes S_a and S_b, represented as sequences of SOM states produced by two different executions of the same behavior, was determined using the Needleman-Wunsch global alignment algorithm [46][47]. The algorithm[1] can estimate the similarity between any two sequences if the data is represented as a sequence over a finite alphabet. The general applicability of the algorithm has made it popular for other applications such as comparing biological sequences, text sequences, and more [47]. Computing the similarity of two sequences requires a substitution cost (i.e., a difference function) to be defined for any two tokens in the finite alphabet. Here the substitution cost is defined as the Euclidean distance between any two nodes in the SOM (each node in the 2D SOM has an x and a y coordinate).

The resulting similarity matrix, $\mathbf{W}$, was used as input to the unsupervised hierarchical clustering procedure, which partitions the input data points (i.e., either audio or video sequences) into disjoint clusters. The algorithm exploits the eigenstructure of the matrix to partition the data points. Finding the optimal graph partition is an NP-complete problem. Therefore, the Shi and Malik [49] approximation algorithm was used, which minimizes the *normalized cut* objective function. The following steps give a summary of the algorithm:

1) Let $\mathbf{W}_{n \times n}$ be the symmetric matrix containing the similarity score for each pair of outcome sequences.

[1]The Needleman-Wunsch algorithm maximizes the similarity between two sequences. An equivalent approach is to minimize the Levenshtein edit distance [48] between the sequences.

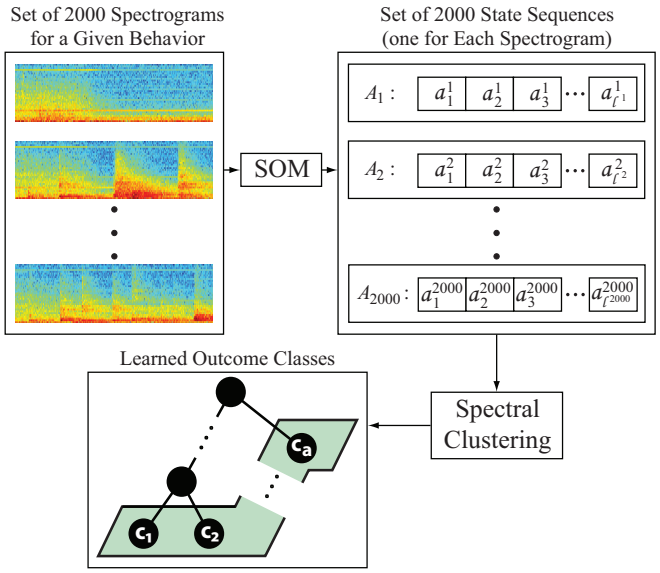

Fig. 9. Illustration of the process used to learn acoustic outcome classes. Each spectrogram is transformed into a state sequence using the trained SOM, which results in 2000 sequences, $\{A_i\}_{i=1}^{2000}$, for each behavior. The acoustic outcome classes are learned by recursively applying the spectral clustering algorithm on this set of sequences. The acoustic outcome classes, $C = \{c_1, ..., c_a\}$, are the leaf nodes of the tree created by the recursive algorithm.

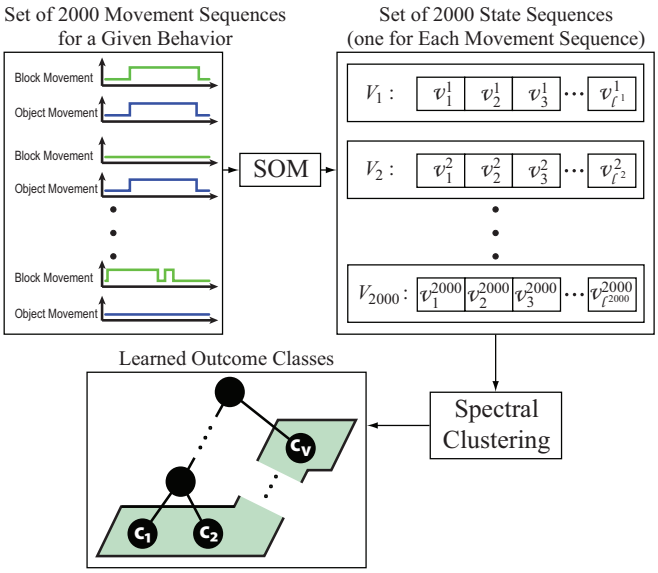

Fig. 10. Illustration of the process used to learn visual outcome classes. Each movement sequence is transformed into a state sequence using the trained SOM, which results in 2000 state sequences, $\{V_i\}_{i=1}^{2000}$, for each behavior. The set of sequences is recursively bi-partitioned using the spectral clustering algorithm in order to learn visual outcome classes, $C = \{c_1, ..., c_v\}$, which are the leaf nodes of the tree created by the recursive algorithm.

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.
4) Search for a threshold of the resulting eigenvector to create a bi-partition of the set of acoustic (or visual) outcomes that minimizes the normalized cut objective function. Accept this bi-partition if the resulting value of the objective function is smaller than a threshold α.
5) Recursively bi-partition subgraphs obtained in step 4 that have at least β audio or video sequences.

The output of this procedure is k outcome classes $C = \{c_1, ..., c_k\}$, which are represented as the leaf nodes in a tree structure (see Fig. 9 and Fig. 10). In our previous work [41], the value α used in step 4 was set to 0.995. The same value was used here as well. The value for β used in step 5 was empirically set to 40% of the size of the dataset that was initially passed to the spectral clustering algorithm.

V. Object Categorization

A. Learning Object Categories

The frequency with which some outcomes occur with different objects can be used to cluster the objects into categories. For example, when the robot drops a block over a container, it will hear the sound of the block bouncing inside the container more often than when it drops the block over a non-container, in which case the block falls on the table. Similarly, when the robot moves a container, it will see the block move with the container more often than when it moves a non-container, in which case the block does not move.

Given a set of **outcome classes** $C = \{c_1, \ldots, c_k\}$ extracted by the robot while interacting with objects $\mathcal{O} = \{O_1, \ldots, O_{20}\}$, the robot acquired an outcome occurrence vector $E_u = [e_1^u, \ldots, e_k^u]$ for each object O_u. The value of each e_j^u represents the number of times the outcome c_j occurred with object O_u, divided by the total number of interactions (100 interactions in this case). In other words, each outcome occurrence vector E_u encodes a probability distribution over the set of outcome classes, such that e_j^u specifies the probability of observing outcome class c_j with object O_u over the entire history of interactions.

The robot formed **object categories** by clustering the feature vectors $E_1, \ldots, E_{20}$ (one for each of the 20 objects shown in Fig. 4). The X-means unsupervised clustering algorithm was used for the procedure. X-means extends the standard K-means algorithm to automatically estimate the correct number of clusters, k, in the dataset [50]. Twelve different categorizations were constructed (one acoustic categorization and one visual categorization for each of the six exploratory behaviors).

B. Object Categorization Results

Figure 11 visualizes the twelve categorizations produced for each behavior–modality combination. The twelve categorizations are described in more detail below.

1) Acoustic Categorizations: Four of the six behaviors produced distinguishable acoustic signals that the robot could use to form object categories: *drop block*, *shake*, *flip*, and *drop object*. The (mostly silent) *grasp* and *move* behaviors produced acoustic signals that were very similar for all objects and the algorithm clustered all 20 objects into the same object class.

The *drop block* behavior produced three clusters that were almost homogeneous. The first cluster had only containers and the tall metal non-container (the only misclassified object). The second cluster had the rest of the non-containers. The

last cluster had the three soft material container baskets. The difference between the softness and hardness of the objects' materials was distinctive enough to create two container categories (cluster 1 and 3 in Fig. 11). For example, the two wicker baskets and the styrofoam bucket (in cluster 3) are made of soft materials, which muffled the block's sound. In contrast, when the block fell into one of the hard containers (in cluster 1) it bounced around longer and produced a louder sound.

The *shake* behavior produced results similar to the *drop block* behavior. In this case, however, there were only two clusters and the three soft-material container baskets were incorrectly classified as non-containers. These three objects produced very little sound when shaken, even if the block was inside them. Thus, they sounded similar to the non-containers, which seldom made noise during this interaction. The tall metal trash can was again misclassified.

The *flip* behavior was the most reliable way to discriminate between containers and non-containers in our experiments. It produced a perfect classification. Flipping the object over produced a distinct sound in the case of containers as the small block fell onto the table. In the case of non-containers, no sound was generated as the block was already on the table.

The *drop object* behavior resulted in clusters that were completely heterogeneous. The behavior did not produce different acoustic outcomes for containers and non-containers.

2) Visual Categorizations: All six behaviors produced visual movement patterns that the robot could use for object categorization (see Fig. 11). The *drop block* behavior was not a reliable way to categorize containers from non-containers. The categorization resulted in two noisy clusters, in which the robot incorrectly classified four containers and four non-containers. The categorization was similar to a random separation of the objects.

The *grasp* behavior resulted in a categorization with two clusters. Six containers were clustered together and the rest of the objects were classified to the other cluster. The behavior was more useful than expected because it generated a tiny amount of movement. However, in some trials the duration of movement was so short that it was filtered out. Four containers were misclassified due to this noisy data.

The *move* behavior produced a good categorization of containers and non-containers. Only three objects were misclassified: the metal non-container was incorrectly classified as a container; the tall metal trash can and the car trash can were incorrectly classified as non-containers. Each of the three objects has a unique shape, which may help to explain why the objects were misclassified. For example, the metal non-container sometimes functioned as a container since it had a 3/4" lip that could cause the block to come to rest on top of the object. Subsequently, during the *move* behavior the block frequently co-moved with the metal non-container.

The *shake* behavior produced results slightly better than the *move* behavior. Only two objects were misclassified in this case. Shaking the containers produced slight oscillations in the position of the block and the containers when the block was inside them, which allowed the robot to form a good categorization. The skinny car trash can was misclassified probably due to its width—it more readily occluded the block as it was shaken. The narrow shape also kept the block from falling inside the container as often as it fell inside the other containers.

The *flip* behavior produced a near–perfect classification of the objects. Flipping the object over produced a lot of block movement during trials when the block fell out of the containers. In all other trials, the block did not move. The skinny car trash can was again misclassified.

The block *appeared* to move, however, during several trials with the green non-container, which is why it was misclassified as a container. The block often came to rest at the perimeter of the visual field where the depth position fluctuated during these trials.

The *drop object* behavior was not a reliable way to categorize containers and non-containers. The behavior led to two arbitrary clusters. The red bucket and the purple bucket were classified together. The rest of the objects were placed in the other cluster.

C. Evaluating the Object Categorizations

To check whether the robot was able to extract meaningful object clusters we computed the category information gain. The information gain captures how well the object categories formed by the robot resemble the categories specified by a human. The information gain is high when the category labels assigned to the objects match human-provided category labels. It is low otherwise. In other words, if the information gain is high, then the robot has categorized the objects in a meaningful way (even though the robot does not know the human words corresponding to the categories).

Let $\lambda^{(f)} = [\mathcal{O}^1, \ldots, \mathcal{O}^{M_f}]$ define an object categorization over the set of objects $\mathcal{O}$, for a specific behavior–modality combination B_f, where $\mathcal{O}^i$ is the set of objects in the i^{th} cluster. Let p_c^i and p_{nc}^i be the estimated probabilities that an object drawn from the subset $\mathcal{O}^i$ will be a container or a non-container (as defined by human labels). Given a cluster of objects $\mathcal{O}^i$, the Shannon entropy of the cluster is defined as:

$$\mathcal{H}(\mathcal{O}^i) = -p_c^i log_2(p_c^i) - p_{nc}^i log_2(p_{nc}^i)$$

In other words, an object cluster containing mostly containers (or mostly non-containers) will have low entropy, while a cluster containing an equal number of containers and non-containers will have the maximum entropy. The information gain for the object categorization $\lambda^{(f)} = [\mathcal{O}^1, \ldots, \mathcal{O}^{M_f}]$, which was learned using behavior-modality combination B_f, is given by the following formula:

$$IG(\lambda^{(f)}) = \mathcal{H}(\mathcal{O}) - \sum_{i=1}^{M_f} \frac{|\mathcal{O}^i|}{|\mathcal{O}|} \mathcal{H}(\mathcal{O}^i)$$

To get a baseline information gain value for comparison, the information gain was computed for a random labeling. That is, the values for p_c^i and p_{nc}^i were estimated after randomly shuffling the labels of the objects in all clusters $\mathcal{O}^i$ (where $i = 1$ to M_f) while preserving the number of objects in each cluster. This procedure was repeated 100 times to estimate

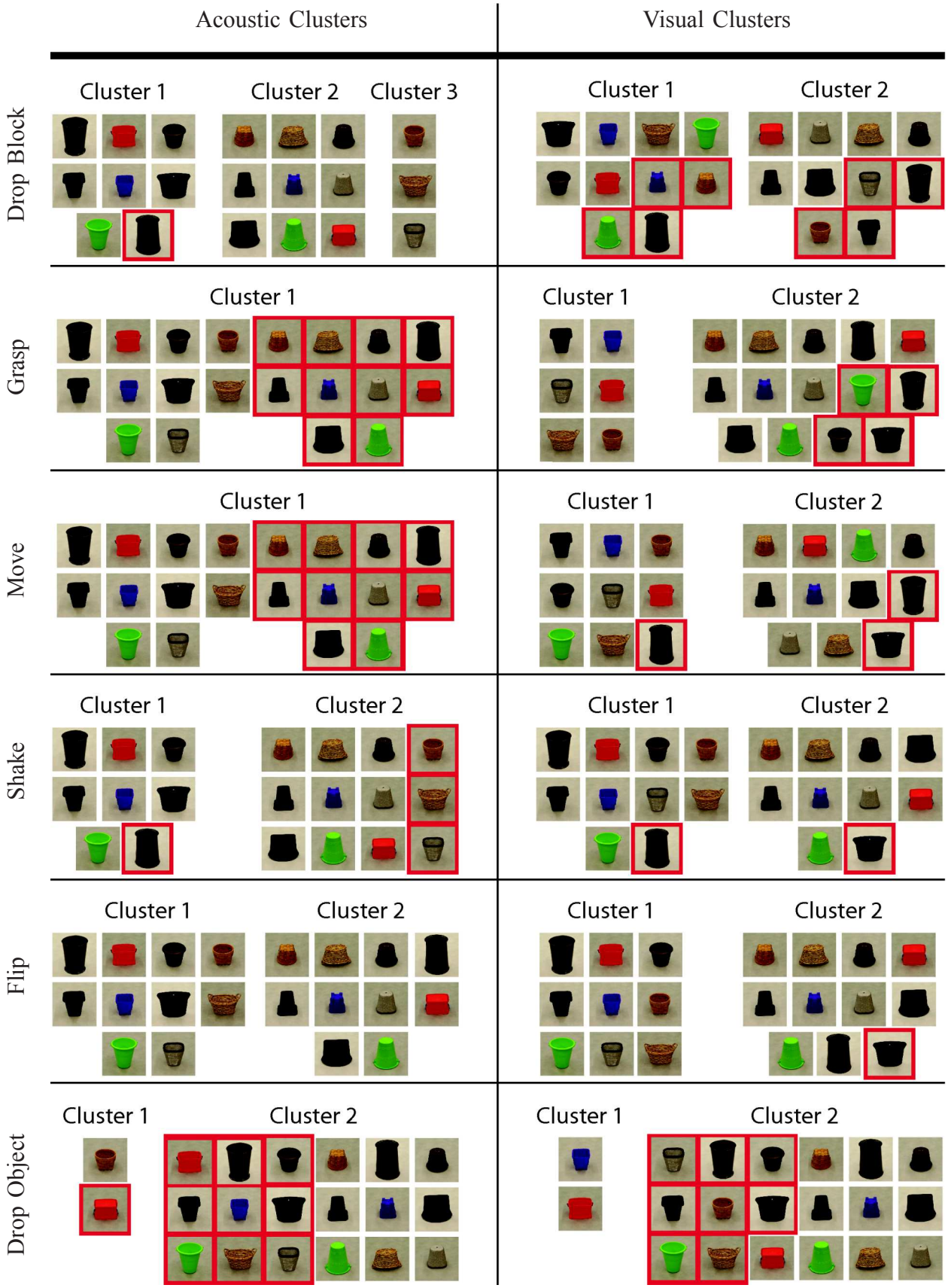

Fig. 11. Visualization of the object categories formed by the robot for the six exploratory behaviors and the two sensory modalities. Incorrect classifications are framed in red (based on category labels provided by a human and the majority class of the cluster). The quality of each categorization depends on the behavior that was performed and the sensory modality that was used for clustering. For example, the *flip* behavior produced both acoustic outcomes and visual movement patterns that resulted in a perfect and a near–perfect classification of containers and non-containers, respectively. Combinations of other behaviors and modalities produced clusters that were not always so pure.

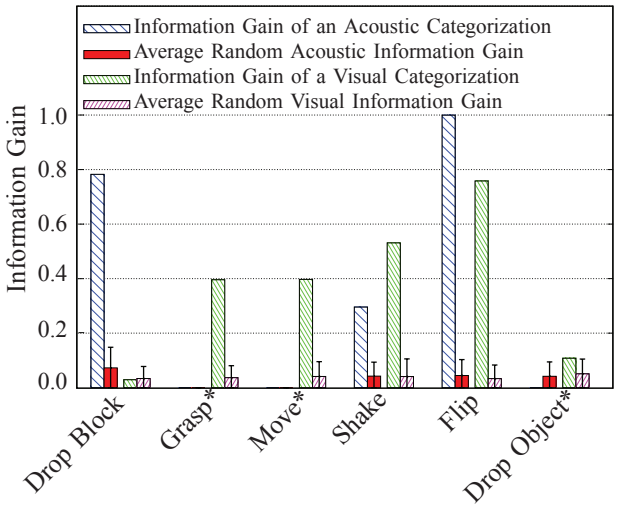

Fig. 12. Information gain of the object categories formed by the robot for each behavior–modality combination. For comparison, the information gain for a random classification is shown next to the object category information gain. The random information gain was computed by shuffling the labels 100 times and estimating the mean and the standard deviation. When computing the information gain, the correct object labels (container or non-container) were provided by a human. For some behaviors the acoustic information gain was zero, which is denoted with the * symbol.

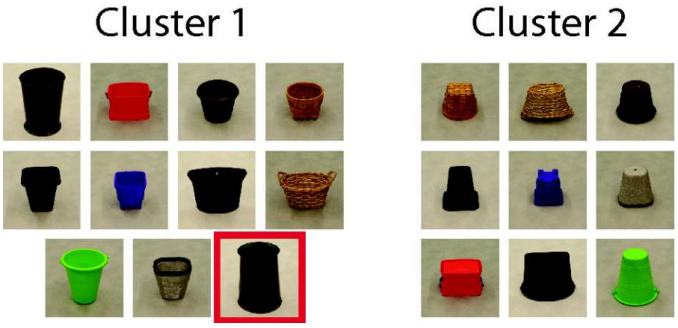

Fig. 13. Visualization of the unified object categorization produced by the consensus clustering algorithm, which searched for a consolidated clustering of the twelve input clusterings shown in Fig. 11. The unified categorization closely matches ground-truth labels provided by a human. Only one object was misclassified.

the mean and the standard deviation. Figure 12 shows the information gain for each categorization and compares it to the corresponding baseline average random information gain.

The figure shows that the categorization produced using acoustic outcomes from the *flip* behavior most closely matches the labels provided by an adult human. Next in order are: the categorization produced using the acoustic signals from the *drop block* behavior, the categorizations produced using the visual movement patterns from the *flip*, *shake*, *grasp* and *move* behaviors, and the categorization produced using the acoustic signals from the *shake* behavior. All of these categorizations have an information gain that is better than chance. The remaining categorizations have an insignificant information gain with respect to the human-provided labels, which shows that they are not suitable for capturing the functional properties of containers.

The fact that some clusterings formed by the robot were noisy was expected. Some behaviors are simply better at capturing certain object properties than others. With 20 objects of various shapes, sizes, and materials there are many ways the robot could have categorized the objects. However, no behavior completely separated objects by size or material. On the other hand, seven behavior–modality combinations captured the functional properties of the containers well (i.e., acoustic signals from the *drop block*, *shake*, and *flip* behaviors; and visual movement patterns from the *grasp*, *move*, *shake*, and *flip* behaviors). The next section shows how the different categorizations can be combined into a single one.

VI. Unified Object Categorization

A. Unification Algorithm

As Fig. 11 shows, by categorizing objects using multiple behaviors and multiple modalities, the robot can form many different categorizations of the objects. Some categorizations closely match the object labels provided by a human; others are noisy. Without a method to unify the different categorizations of the objects, however, an object categorization is *at most* meaningful with respect to the behavior and the modality that were used to produce it.

Therefore, it is desirable to form one unified categorization from multiple categorizations of the objects. That is, given a set of object categorizations $\Lambda = \lambda^{(1)}, \ldots, \lambda^{(r)}$ and a desired number of object categories p, the robot forms a single, unified categorization $\hat{\lambda}$. The categorization $\hat{\lambda}$ defines p categories of objects and is determined to be representative of the input categorizations Λ using the objective function $\phi(\Lambda,\hat{\lambda})$. The function measures the total normalized mutual information between a set Λ containing r object categorizations and a single categorization $\hat{\lambda}$. More formally,

$$\phi(\Lambda, \hat{\lambda}) = \sum_{q=1}^{r} \phi^{NMI}(\hat{\lambda}, \lambda^{(q)})$$

where $\phi^{NMI}(\hat{\lambda}, \lambda^{(q)})$ is the normalized mutual information between categorizations $\hat{\lambda}$ and $\lambda^{(q)}$ (see [51]). Thus, the best unified categorization is defined as the clustering of the objects that has the highest possible total normalized mutual information with respect to the multiple input categorizations. Finding the best clustering, however, is intractable. Therefore, it is necessary to search for a clustering that is approximately the best. For this task, we used the hard consensus clustering algorithm [51]. The algorithm takes as input a set of object categorizations Λ and a value k, employs three functions that independently solve for a good approximation, and outputs the best unified clustering that it finds. The output of this procedure is a labeling $L_i \in \mathcal{L}$ for each object $O_i \in \mathcal{O}$.

In this case, the set of object categorizations Λ consisted of the twelve categorizations shown in Figure 11. The algorithm was run several times with p varying from 2 to 10. From these runs, the unified object categorization was chosen as the clustering that maximized the objective function. The result of unifying the twelve object categorizations is shown in Fig. 13. The figure shows that the hard consensus clustering algorithm was able to find a meaningful categorization even though only seven of the twelve behavior–modality combinations

produced a good clustering of the objects. Only one object was misclassified in the unified object categorization.

For completeness, a brief description of the three functions used by the hard consensus clustering algorithm is provided below. The algorithm runs these functions in parallel and picks one of the category results that maximizes the NMI. The three functions are: 1) The Cluster–based Similarity Partitioning Algorithm (CSPA) generates a similarity matrix for the objects. Each entry in this matrix represents the number of times that two objects appear in the same cluster. A similarity–based clustering algorithm is applied to this matrix to cluster the objects. 2) The HyperGraph Partitioning Algorithm (HGPA) constructs a hypergraph and partitions it into k disjoint components by cutting a minimal number of hyperedges. A hypergraph is a special graph in which an edge can connect to many vertices. In our case, the objects are the vertices of the hypergraph and the clusters of objects are the edges. 3) The Meta-CLustering Algorithm (MCLA) groups multiple similar clusters of objects until there are at most k disjoint clusters. For more details see [51].

B. Robustness of the Algorithm

To test the generalizability properties of the algorithm we ran three additional experiments that are briefly summarized below. In the first experiment, the consensus clustering algorithm was able to form a meaningful categorization when the object classes were skewed to have more containers than non-containers. The set of objects was skewed by using only 4 of the 10 non-containers. The robot categorized the objects using the same learning framework. This process was repeated 10 times with different sets of 4 randomly chosen non-containers. The algorithm misclassified 2 objects in 8 of these instances and 1 object in another instance. In the second experiment, the interaction data for one random container and one random non-container was removed for each behavior–modality combination. The robot categorized the objects using the same learning framework, and the process was repeated 10 times with different sets of removed data. All resulting unified categorizations matched the categorization shown in Fig. 13. Thus, the algorithm was able to form a meaningful categorization even when some of the interaction data was missing. The third experiment tested an alternative approach to forming a unified object categorization by directly concatenating and clustering the feature vectors used to produce the individual categorizations (see section V.A). The X-means algorithm was used to do the clustering, which produced three clusters with two misclassified objects. This result is inferior to the unified categorization shown in Fig. 13. Overall, the algorithm proved to be quite robust.

By combining the different categorizations into a single one, the robot effectively ruled out the nonsense categorizations that it acquired, allowing it to form two object categories that are close to what a human would call containers and non-containers. Furthermore, the unified categorization condensed a large amount of data into a single categorization, which described the functional properties of objects across the robot's whole sensorimotor repertoire. Having a single categorization also meant that a single perceptual model could be learned, and used to infer the object category of novel objects using only passive observation. The next section describes how the robot was able to form a perceptual model for the two object categories shown in Fig. 13.

VII. Categorizing Novel Objects

It is impractical for a robot to categorize all novel objects by first interacting with them for a long time. To reduce the exploration time, the robot can learn a perceptual model for each acquired object category in the unified object categorization (see Fig. 13) and use that model to estimate the category of a novel object. More specifically, let $\mathbf{f_i} \in \mathbb{R}^n$ be the visual feature vector for object O_i, and let $L_i \in \mathcal{L}$ be the category label of that object according to the learned unified categorization, where $\mathcal{L}$ is the set of object categories. Given training examples $(\mathbf{f_i}, L_i)_{i=1}^{i=N}$, the task of the robot is to learn a recognition model $\mathcal{M}$ that can estimate the correct category of a novel object O_{test} given the object's visual features $\mathbf{f_{test}}$. In other words, $\mathcal{M}(\mathbf{f_{test}}) \rightarrow L_{test}$, where $L_{test} \in \mathcal{L}$ is the estimated category of the novel object. The next subsection describes the feature extraction routine used to compute the visual features $\mathbf{f_i} \in \mathbb{R}^n$ for both familiar and novel objects.

A. Feature Extraction

To extract the visual features of objects, principal component analysis (PCA) was used to find compact representations for the unlabeled visual sensory stimuli. PCA transforms the input data into a new coordinate system, where each coordinate represents a different projection of the input data. The coordinates are ordered based on how well the projections explain the variance in the data. More formally, the input images $\mathbf{x_i} \in \mathbb{R}^m$ are transformed into a set of independent basis vectors $\mathbf{b_1}, \ldots, \mathbf{b_n} \in \mathbb{R}^m$ and a vector of weights $\mathbf{f_i} \in \mathbb{R}^n$ such that $\mathbf{x_i} - \bar{\mathbf{x}} \approx \sum_j \mathbf{b_j} f_i^j$, where $\bar{\mathbf{x}}$ is the mean of all input images. The weights $\mathbf{f_i} \in \mathbb{R}^n$ represent the compact features of the high-dimensional input image $\mathbf{x_i}$.

The algorithm was trained on 30x30 depth images, one for each of the 20 objects that the robot interacted with. The training images were extracted automatically from the larger 320x240 depth images captured by the ZCam. The objects were located using background subtraction and a boundary box was placed around them. The corresponding locations in the depth image were cropped and scaled to 30x30 pixels. The resulting images were used as input to the PCA algorithm. The first five basis vectors computed by the algorithm captured 90% of the variance in the data and are shown in Fig. 15. The figure shows that the first vector, which captures 43% of the variance, is a convex feature characteristic of non-containers. The second and the third vectors, which jointly capture 40% of the variance, represent a feature characteristic of containers. The next subsection describes the recognition algorithm that maps the visual features of an object to its estimated object category.

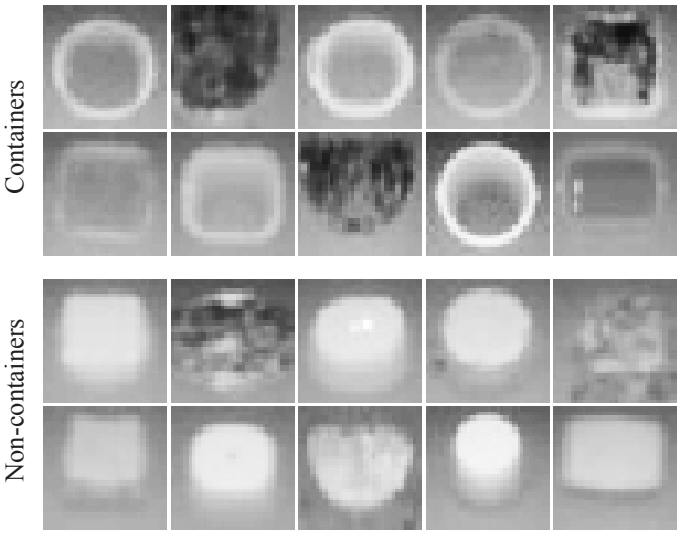

Fig. 14. The 20 depth images of the objects used as input to the PCA algorithm. Each image was generated by finding the object in the larger 320x240 depth image and scaling the region to 30x30 pixels.

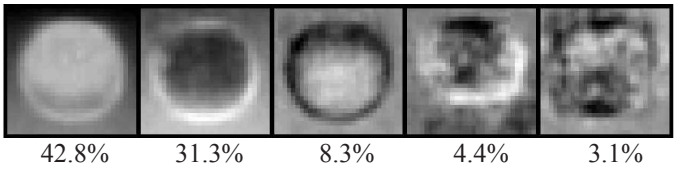

Fig. 15. A visualization of the first five principal components computed by the PCA algorithm using the images shown in Fig. 14 as input. The percentage of the variance explained by each component is listed below it. These five principal components, along with the category labels from Fig. 13, were later used to classify novel objects as 'containers' or 'non-containers.'

B. Recognition Algorithm

The object category recognition model $\mathcal{M}$ was trained on the visual features of the 20 objects that the robot interacted with. Let $\mathbf{f_i} \in \mathbb{R}^2$ represent the extracted visual features for the i^{th} object, and let $L_i \in \mathcal{L}$ be its category according to the unified categorization (see Fig. 13). Using this formulation, the robot acquired the set $(\mathbf{f_i}, L_i)_{i=1}^{20}$, which contains the 20 labeled training examples available to it.

The robot's recognition model, $\mathcal{M}$, was implemented as a k-Nearest Neighbors (k-NN) classifier with $k = 3$. K-NN is an instance-based learning algorithm that does not build an explicit model of the data, but simply stores all labeled data points and uses them when the model is queried to make a prediction. Given a novel object, O_{test}, the robot extracted its visual features $\mathbf{f_{test}}$, computed from a 30 x 30 depth image of the object, and the learned basis vectors. Subsequently, k-NN was used to find the k closest neighbors of $\mathbf{f_{test}}$ in the training set, using the Euclidean distance function. Finally, the novel object was labeled with the majority category of the k closest neighbors. For example, if 2 of the closest neighbors to $\mathbf{f_{test}}$ were containers, then the novel object was labeled as a container as well.

The classifier was tested on how well it could detect the object category of 30 novel objects by passively observing them. The set of novel objects included 15 containers, which were selected to have a variety of shapes, sizes, and material properties. The other 15 objects were non-containers, which were the same novel containers only flipped over (see Fig. 16). Using the extracted visual features and the k-Nearest Neighbor classifier, the robot assigned the correct object category to 29 of the 30 objects. This result implies that the robot not only has the ability to interactively distinguish between containers and non-containers, but also to learn a visual model that allows it to passively determine the functional category of novel objects.

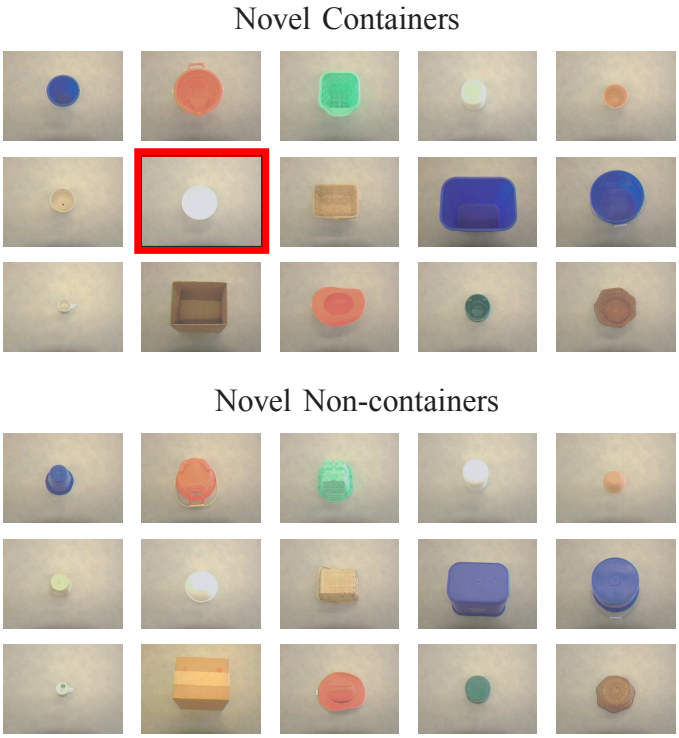

Fig. 16. The result of using a Nearest Neighbor classifier to label novel objects as 'containers' or 'non-containers'. The mixing bowl (outlined in red) was the only misclassified object. Visual features were extracted for each of the 30 novel objects and used in the classification procedure.

VIII. Evaluating the Effect of Experience on the Quality of Object Categorizations

Intuitively, the quality of an object categorization should depend on how much experience the robot has had with each object. As Fig. 12 shows, the robot formed seven meaningful categorizations after 100 interactions were performed with each object. The unification of all twelve categorizations also produced a meaningful categorization, in which only one object was misclassified. Even fewer interactions, however, may be required to reproduce these results.

To find out how much experience is necessary to form a good object categorization, the categorization quality was evaluated as the number of interactions, N, with each object was increased from 10 to 100. The learning framework described in sections IV and V was used to produce the categorizations for this evaluation (i.e., the same learning framework used to produce the results in Fig. 11, except that the trained SOM was

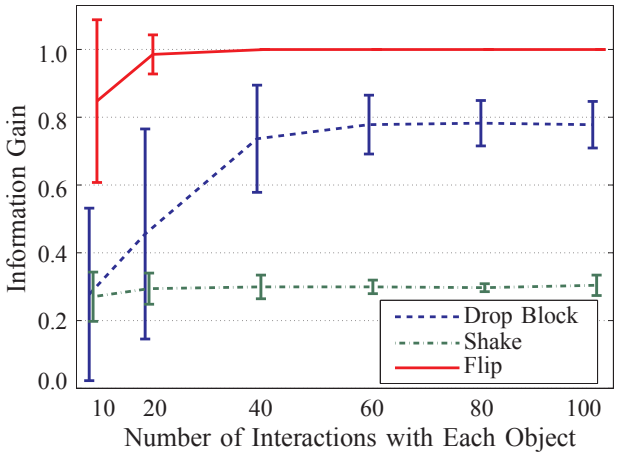

Fig. 17. Information gain for the acoustic categorizations formed by the *drop block*, *shake*, and *flip* behaviors as the number of interactions with each object is increased. This graph was computed by randomly sampling N interactions from the 100 interactions with each object and re-running the learning algorithms on the smaller dataset. This process was repeated 100 times for each value of N to estimate the mean and standard deviation. Human-provided category labels were used to compute the information gain.

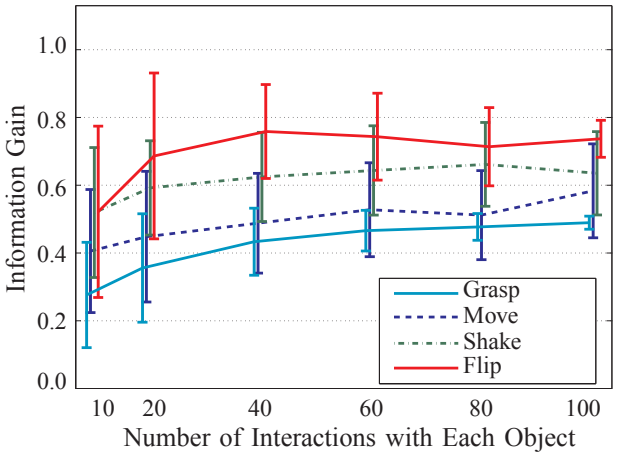

Fig. 18. Information gain for the visual categorizations formed by the *grasp*, *move*, *shake*, and *flip* behaviors as the number of interactions with each object is increased. This graph was computed using the same procedure as that described in Fig. 17.

reused to reduce computation time)[2]. The set of N trials used to compute a categorization was randomly sampled from the set of all 100 trials performed on each object. The quality of a categorization was determined by computing its information gain using human–provided labels. The process was repeated 100 times for each value of N to estimate the mean and the standard deviation. Although only seven out of twelve behavior–modality combinations originally led to a meaningful object categorization, the unification procedure was performed on all twelve combinations. The quality of the resulting unified clustering was also evaluated using its information gain.

The results are shown in two graphs to simplify their analysis. Figure 17 shows the quality of the acoustic categorizations for the *drop block*, *shake*, and *flip* behaviors. Figure 18 shows the quality of the visual categorizations for the *grasp*, *move*, *shake*, and *flip* behaviors. The other five behavior–modality combinations are not depicted in the graphs because their information gain remained near zero. Also not shown is the quality of the unified categorization, which remained fairly constant at 0.75 as the value for N increased from 10 to 100.

The mean information gain for all of the categorizations converged after about 40 interactions with each object were performed. The quality of the individual categorizations increased as the robot gained more experience. The mean information gain converged when the features used to represent the functional properties of each object stabilized. In contrast with the individual categorizations, the information gain of the unified categorization converged after only 10 interactions with each object. Thus, the unified categorization was meaningful even when the robot had an insufficient amount of data to fully characterize the functional properties of each object for the individual behavior–modality combinations.

The mean information gain for the individual categorizations converged to values that are comparable to those obtained using the original ordering of the dataset (see Fig. 12). Because order–dependent clustering algorithms were used in this framework (Spectral Clustering and X-means), some of the categorizations changed when the dataset was shuffled. For example, the mean information gain of the visual categorizations for the *grasp*, *move*, and *shake* behaviors improved slightly. The rest of the behavior–modality combinations reached information gain values similar to those in Fig. 12.

The effect that the order–dependent clustering algorithms had on the categorization performance is most clear when the number of interactions, N, with each object is 100. Instead of consistently identifying the same categorization of the objects for each behavior–modality combination, the framework identified a distribution of categorizations. The histograms in Fig. 19 show that certain behavior–modality combinations were affected more than others. For example, the *grasp–vision* behavior–modality combination was only slightly affected (the flower pot object was misclassified as a non-container 4 times; it was correctly classified as a container 96 times). The *move–vision* behavior–modality combination was significantly affected, however, as three objects oscillated between categories. This analysis was also performed for the unified categorization, where 62 out of 100 categorizations matched the results obtained using the original ordering of the dataset. In general, if the functional properties of an object placed it somewhere between containers and non-containers, then the resulting category label for this object tended to fluctuate. Fluctuations in the individual categorizations seldom reduced the quality of the unified categorization.

[2]It is important to point out that the trained SOM represents the robot's self–organized "feature extraction" mechanism. This does not have to be part of the robot's "object categorization" mechanism. In fact, evidence from developmental psychology suggests that the auditory features that infants learn are fixed by the time they learn words. When infants are around six months old they are sensitive to the sounds that are used in many different languages. By nine months, however, they have learned a fixed set of auditory features, which are specific to their native language [52]. For example, at this age, it would be more difficult for an infant raised in an English–speaking home to distinguish between some of the sounds that an infant raised in a Chinese–speaking home can distinguish without a problem.

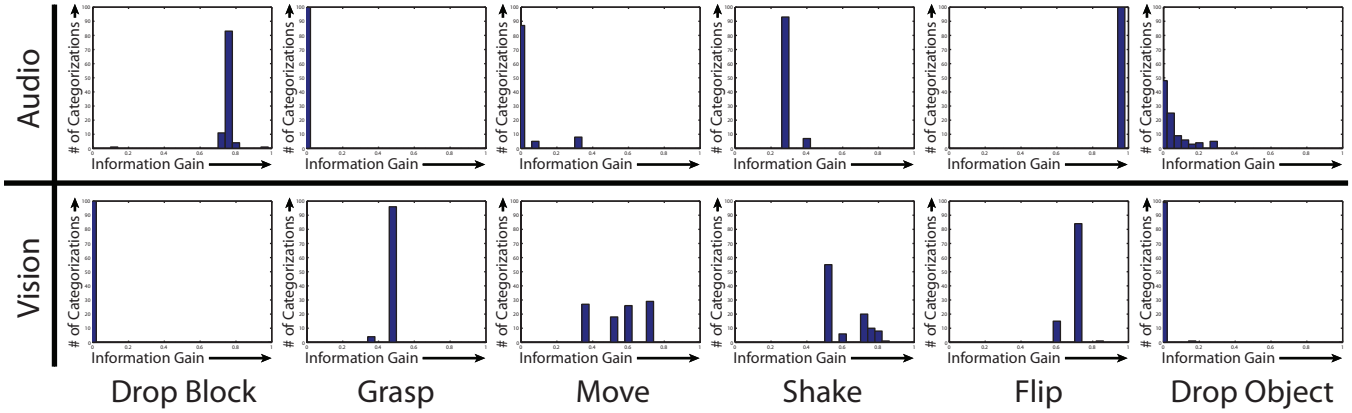

Fig. 19. The distribution of information gain values for different categorizations obtained with different behavior–modality combinations. Each histogram was generated by computing the information gain values for 100 different categorizations of the objects, which were obtained by running the framework 100 times on different orderings of the dataset.

IX. CONCLUSIONS AND FUTURE WORK

This paper described a computational framework for learning object categories, in which a robot explored objects using multiple behaviors and sensed the resulting outcomes using multiple sensory modalities. The framework was evaluated using an object categorization task with 20 containers and non-containers. The robot observed the acoustic signatures and the visual movement patterns of the objects as it performed six different exploratory behaviors. A separate object categorization was produced for each behavior–modality combination, which resulted in twelve different categorizations of the objects. These categorizations were then unified using consensus clustering into a single object categorization. It was shown that this behavior–grounded object categorization is meaningful when compared with human–provided object labels. It was also shown that this level of categorization performance was attainable after only 10 interactions were performed with each object for each behavior–modality combination. Finally, this paper also showed that a visual classifier can effectively categorize novel objects when it is trained using the category label for each object.

This is the first framework in which an object categorization formed by a robot was constructed by creating many different categorizations for a set of objects, which correspond to different behavior–modality combinations, and then unifying them into a single one. Our methodology is consistent with Leslie Cohen's definition that object categorization is about finding similarities among perceptually different objects [31]; whereas object recognition is about finding differences among perceptually similar objects [53]. The results showed that some of the perceptual differences (e.g., softness) between the objects were captured by the individual categorizations formed by the robot. However, by unifying many different individual categorizations, the robot ignored these perceptual differences and formed a categorization based only on the containment property, which was the most common thing between the objects.

In the end, the experience that the robot acquired in this large–scale experiment was condensed into a single object categorization. The robot had knowledge of the functional properties of each object in terms of the frequency with which different acoustic outcomes and different visual movement patterns occurred with it. The robot also knew the differences between the objects in terms of this frequency information, which served as the basis for categorizing the objects. Finally, the robot knew the visual appearance of containers and non-containers. Having the option to categorize an object by either its functional properties or its visual appearance is advantageous, and mirrors some of the characteristics of object categorization in humans [54].

The framework presented here can be extended in several possible directions. One possible extension is to reduce the human input provided to the object categorization framework. For instance, the object IDs were provided by a human and used during the categorization procedure. It may be possible to use object recognition models to eliminate the dependency on human–provided object IDs. It is also desirable to let the robot learn its own exploratory behaviors. In the current framework, the behaviors were encoded by a human programmer. Presumably, it should be possible for a robot to learn these behaviors on its own.

Another possible extension of this work is to make the framework capable of categorizing many different types of objects. Intuitively, finding a meaningful categorization for a large number of object types would require the robot to have an increased amount of experience with each object. The robot could use more sensory modalities with each behavior, however, to reduce the amount of experience that is required. For example, tactile and proprioceptive sensory modalities could be added in order to capture more information during each interaction. Our previous work has shown that by adding more sensory modalities the robot could improve its object recognition abilities [55]. The same is probably true for object category recognition.

REFERENCES

[1] 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 (ICDL)*, Shanghai, China, June 2009.

[2] S. Griffith, J. Sinapov, V. Sukhoy, and A. Stoytchev, "How to separate containers from non-containers? A behavior-grounded approach to acoustic object categorization," in *Proc. of the IEEE Intl. Conf. on Robotics and Automation (ICRA)*, Anchorage, AK, May 3-8 2010, pp. 1852–1859.

[3] D. H. Rakison and L. M. Oakes, Eds., *Early Category and Concept Development: Making Sense of the Blooming, Buzzing Confusion*, 1st ed. Oxford University Press, USA, 2003.

[4] T. Power, *Play and Exploration in Children and Animals*. Mahwah, NJ: Laurence Erlbaum Associates, 2000.

[5] P. Rochat, "Object manipulation and exploration in 2- to 5-month-old infants," *Developmental Psychology*, vol. 25, no. 6, pp. 871–884, 1989.

[6] E. J. Gibson, "Exploratory behavior in the development of perceiving, acting, and the acquiring of knowledge," *Annual review of psychology*, vol. 39, no. 1, pp. 1–42, 1988.

[7] E. Rosch, *Cognition and Categorization*. Hillsdale, NJ: Erlbaum, 1978, ch. Principles of Categorization, pp. 189–206.

[8] L. W. Barsalou, W. K. Simmons, A. K. Barbey, and C. D. Wilson, "Grounding conceptual knowledge in modality-specific systems," *Trends in Cognitive Sciences*, vol. 7, no. 2, pp. 84–91, 2003.

[9] A. Pinz, "Object categorization," *Foundations and Trends in Computer Graphics and Vision*, vol. 1, no. 4, pp. 255–353, Dec 2005.

[10] L. B. Smith, *Functional Features in Language and Space*. Oxford University Press, 2005, ch. Shape: A Developmental Product, pp. 235–255.

[11] G. Metta and P. Fitzpatrick, "Early integration of vision and manipulation," *Adaptive Behavior*, vol. 11, no. 2, pp. 109–128, June 2003.

[12] R. Sutton, "Verification, the key to AI," on-line essay. [Online]. Available: http://www.cs.ualberta.ca/ sutton/IncIdeas/KeytoAI.html

[13] A. Stoytchev, "Some basic principles of developmental robotics," *IEEE Transactions on Autonomous Mental Development (TAMD)*, vol. 1, no. 2, pp. 122–130, 2009.

[14] J. S. Horst, L. M. Oakes, and K. L. Madole, "What does it look like and what can it do? Category structure influences how infants categorize," *Child Development*, vol. 76, no. 3, pp. 614–631, 2005.

[15] C. W. Robinson and V. M. Sloutsky, "Auditory dominance and its change in the course of development," *Child Development*, vol. 75, no. 5, pp. 1387–1401, 2004.

[16] R. Schmidt, *Attention and awareness in foreign language learning (Technical Report 9)*. Honolulu: University of Hawaii, Second Language Teaching and Curriculum Center, 1995, ch. Consciousness and foreign language learning: A tutorial on the role of attention and awareness in learning, pp. 1–63.

[17] G. Butterworth and M. Castillo, "Coordination of auditory and visual space in newborn human infants," *Perception*, vol. 5, no. 2, pp. 155–160, 1976.

[18] R. Baillargeon, "Infants' physical world," *Current Directions in Psychological Science*, vol. 13, no. 3, pp. 89–94, 2004.

[19] E. S. Spelke and K. D. Kinzler, "Core knowledge," *Developmental Science*, vol. 10, no. 1, pp. 89–96, 2007.

[20] A. Needham, J. Cantlon, and S. O. Holley, "Infants' use of category knowledge and object attributes when segregating objects at 8.5 months of age," *Cog. Psychology*, vol. 53, no. 4, pp. 345–360, 2006.

[21] S. Hespos and E. Spelke, "Precursors to spatial language: The case of containment," *The Categorization of Spatial Entities in Language and Cognition*, vol. 15, pp. 48–144, 2007.

[22] S. Hespos and R. Baillargeon, "Reasoning about containment events in very young infants," *Cognition*, vol. 78, no. 3, pp. 207–245, March 2001.

[23] A. M. Leslie and P. DasGupta, "Infants' understanding of a hidden mechanism: Invisible displacement," April 1991, paper presented at symposium on "Infants' reasoning about spatial relationships." SRCD Biennial Conference, Seattle.

[24] P. Rochat and T. Striano, "Primacy of action in early ontogeny," *Human Development*, vol. 41, pp. 112–115, 1998.

[25] R. Baillargeon, "How do infants learn about the physical world?" *Current Directions in Psychological Science*, vol. 3, no. 5, pp. 133–140, 1994.

[26] S. Wang and R. Baillargeon, "Can infants be "taught" to attend to a new physical variable in an event category? The case of height in covering events," *Cognitive Psychology*, vol. 56, pp. 284–326, 2008.

[27] S. J. Hespos and R. Baillargeon, "Decalage in infants' knowledge about occlusion and containment events: Converging evidence from action tasks," *Cognition*, vol. 99, pp. 207–245, 2006.

[28] R. Baillargeon and J. DeVos, "Object permanence in young infants: Further evidence," *Child Development*, vol. 62, pp. 1227–1246, 1991.

[29] A. Aguiar and R. Baillargeon, "Eight-and-a-half-month-old-infants' reasoning about containment events," *Child Development*, vol. 69, pp. 636–653, 1998.

[30] S. M. Sitskorn and A. W. Smitsman, "Infants' perception of dynamic relations between objects: Passing through or support?" *Developmental Psychology*, vol. 31, pp. 437–447, 1995.

[31] L. Cohen, "Unresolved issues in infant categorization," in *Early category and concept development*, D. Rakison and L. M. Oakes, Eds. New York: Oxford University Press, 2003, pp. 193–209.

[32] L. Hasher and R. T. Zacks, "Automatic processing of fundamental information: The case of frequency of occurrence," *American Psychologist*, vol. 39, pp. 1372–1388, 1984.

[33] W. Simmons and L. Barsalou, "The similarity-in-topography principle: reconciling theories of conceptual deficits," *Cognitive Neuropsychology*, vol. 20, no. 3, pp. 451–486, 2003.

[34] A. R. Demasio, "Time-locked multiregional retroactivation: a systems-level proposal for the neural substrates of recall and recognition," *Cognition*, vol. 33, no. 1, pp. 25–62, 1989.

[35] M. Casasola, L. B. Cohen, and E. Chiarello, "Six-month-old infants' categorization of containment spatial relations," *Child Development*, vol. 74, no. 3, pp. 679–693, May 2003.

[36] R. Pfeifer and C. Scheier, "Sensory-motor coordination: The metaphor and beyond," in *Robotics and Autonomous Systems*, vol. 20, 1997, pp. 157–178.

[37] E. Ugur, M. Dogar, M. Cakmak, and E. Sahin, "The learning and use of traversability affordance using range images on a mobile robot," in *Proc. of the IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2007, pp. 1721–1726.

[38] J. Sinapov and A. Stoytchev, "Detecting the functional similarities between tools using a hierarchical representation of outcomes," in *Proc. of the 7th IEEE Intl. Conf. on Development and Learning (ICDL)*, Monterey, CA, August 2008.

[39] L. Montesano, M. Lopes, A. Bernardino, and J. Santos-Victor, "Learning object affordances: From sensory-motor coordination to imitation," *IEEE Trans. on Robotics*, vol. 24, no. 1, pp. 15–26, 2008.

[40] J. Sinapov, M. Wiemer, and A. Stoytchev, "Interactive learning of the acoustic properties of household objects," in *Proc. of the IEEE Intl. Conf. on Robotics and Automation (ICRA)*, Kobe, Japan, May 2009, pp. 3937–3943.

[41] J. Sinapov and A. Stoytchev, "From acoustic object recognition to object categorization by a humanoid robot," in *Proc. of the Robotics Science and Systems (RSS) 2009 Workshop - Mobile Manipulation in Human Environments*, Seattle, WA, June 2009.

[42] T. Nakamura, T. Nagai, and N. Iwahashi, "Multimodal object categorization by a robot," in *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2007, pp. 2415–2420.

[43] R. Sahai, S. Griffith, and A. Stoytchev, "Interactive identification of writing instruments and writable surfaces by a robot," in *Proc. of the Robotics Science and Systems (RSS) 2009 Workshop - Mobile Manipulation in Human Environments*, Seattle, WA, June 2009.

[44] 3DV Systems. http://www.3dvsystems.com/technology/product.html.

[45] U. von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 395–416, 2007.

[46] S. Needleman and C. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *J. Mol. Biol.*, vol. 48, no. 3, pp. 443–453, 1970.

[47] G. Navarro, "A guided tour to approximate string matching," *ACM Computing Surveys*, vol. 33, no. 1, pp. 31–88, 2001.

[48] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Soviet Physics Doklady*, vol. 10, no. 8, pp. 707–710, 1966.

[49] J. Shi and J. Malik, "Normalized cuts and image segmentation," *Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[50] D. Pelleg and A. W. Moore, "X-means: Extending k-means with efficient estimation of the number of clusters," in *Proc. of the 7th Intl. Conf. on Machine Learning (ICML)*, 2000, pp. 727–734.

[51] A. Strehl and J. Ghosh, "Cluster ensembes - a knowledge reuse framework for combining multiple partitions," *Journal of Machine Learning Research*, vol. 3, pp. 583–617, 2002.

[52] P. Jusczyk, A. Cutler, and N. Redanz, "Preference for the predominant stress patterns of English words," *Child Development*, vol. 64, no. 3, pp. 675–687, 1993.

[53] J. J. DiCarlo and D. D. Cox, "Untangling invariant object recognition," *Trends in Cognitive Sciences*, vol. 11, no. 8, pp. 333–341, 2007.

[54] S. Lederman and R. Klatzky, "Hand movements: A window into haptic object recognition," *Cognitive Psychology*, vol. 19, pp. 342–368, 1987.

[55] J. Sinapov and A. Stoytchev, "The boosting effect of exploratory behaviors," in *Proc. of the 24th National Conference on Artificial Intelligence (AAAI)*, Atlanta, GA, July 2010, pp. 1613–1618.