Interactive Object Recognition Using Only Proprioceptive and Auditory Feedback

Jivko Sinapov, Taylor Bergquist, Connor Schenck, Ugonna Ohiri, Shane Griffith and Alexander Stoytchev
Developmental Robotics Laboratory
Iowa State University
{jsinapov, knexer, cschenck, ucohiri, shaneg, alexs}@iastate.edu

***Abstract*—This paper proposes a method for interactive recognition of household objects using proprioceptive and auditory feedback. In our experiments, the robot observed the changes in its proprioceptive and auditory sensory streams while performing five exploratory behaviors (lift, shake, drop, crush, and push) on 50 common household objects (e.g., bottles, cups, balls, toys, etc.). The robot was tasked with recognizing objects it was manipulating by feeling them and listening to the sounds that they make without using any visual information. The results show that both proprioception and audio can be used for object recognition. Furthermore, the robot was able to intelligently integrate both modalities, achieving even better recognition accuracy than using either one alone.**

I. Introduction

HUMAN beings have the remarkable ability to represent object knowledge using multiple modalities (e.g., visual, tactile, proprioceptive, etc.) [1]. Research in psychology has shown that multiple modalities are required to capture many object properties such as weight, roughness, stiffness, etc. [2]. In contrast, most object recognition systems used in robotics rely heavily on computer vision techniques [3], [4]. With a clear view of the target object, such systems can achieve high accuracy, but still suffer from several limitations. For example, using vision alone, a robot cannot distinguish between a heavy object and a light object that otherwise look the same. Furthermore, such a system would be of little use if the object is outside the robot's field of view. The human visual system is also subject to these same limitations - not surprisingly, humans need other sensory modalities to capture object knowledge [2], [5], [6].

This paper proposes a method for interactive recognition of household objects using proprioceptive and auditory feedback. More specifically, proprioceptive feedback is extracted from the joint-torque values of the robot over the course of an interaction, while auditory feedback is extracted from the Discrete Fourier Transform of the sound detected during the interaction. The robot learns a model for each sensory modality using a Self-Organizing Map (SOM), which allows the robot to turn the high-dimensional input from each modality into a discrete sequence of most-highly activated states in the SOM. This representation reduces the dimensionality of the sensory feedback and allows the robot to use standard machine learning methods designed to handle sequences as input.

In this work, the robot interacted with 50 objects using five exploratory behaviors (*lift*, *shake*, *drop*, *crush*, and *push*). The robot was evaluated on the task of object recognition given the feedback from either one or both of the modalities investigated in this paper. The results show that both auditory and proprioceptive feedback contain information indicative of the object in the interaction. In addition, the robot was able to intelligently integrate predictions from multiple modalities and multiple behaviors performed on each test object, which resulted in recognition accuracy of over 98%.

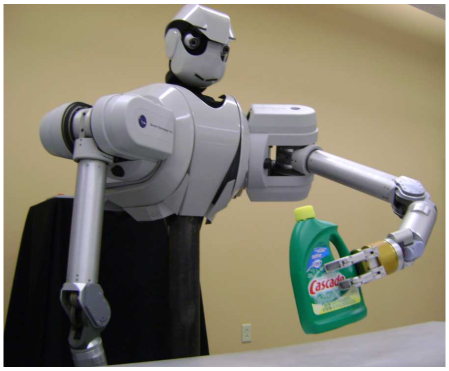

Fig. 1. The robot used in this study, shown holding one of 50 objects that the robot has experience with.

II. Related Work

A. Psychology and Cognitive Science

The work presented in this paper is directly inspired by research in psychology and cognitive science which highlights the importance of sensory modalities other than vision for object recognition tasks. For example, Sapp *et al.* describe a study [5] in which toddlers were presented with a sponge that was deceptively painted as a rock. As expected, the toddlers believed that the object was a rock until the moment they interacted with it (by touching it or picking it up). This and several other studies (see [6]) illustrate that proprioceptive information about objects can be very useful when vision alone is insufficient.

Natural sound is also an important source of information - it allows us to perceive events, and to recognize objects and their properties even when a direct line of sight is not available. The ecological approach to perception provides the insight that *listening* consists of perceiving the properties of a sound's source (e.g., a bouncing ball, a car engine, footsteps, etc.), rather than the properties of a sound itself (e.g., pitch, tone, etc.) [7]. Thus, our auditory system plays a crucial role in understanding and representing object knowledge. Our hypothesis is that this association can be learned by coupling behaviors performed on objects with the sounds produced during these interactions.

These insights have been confirmed by many experimental studies. For example, Giordano *et al.* [8] conducted a study which shows that humans can accurately recognize an object's material (e.g., wood, glass, steel or plexiglass) when listening to the sounds generated when the object is struck. Sound also allows us to perceive many physical properties of objects. Grassi *et al.* [9] show that human subjects were able to provide reasonably good estimates for the size of a ball dropped on plates by simply hearing the impact sound. Motivated by these and other examples, the work in this paper investigates a method that allows a robot to use sound as a source of information about objects in a similar manner.

B. Robotics

Traditionally, most object recognition systems used by robots have relied heavily on computer vision techniques [3], [4] and/or 3D laser scan data [10]. There has been relatively little previous work dealing exclusively with proprioceptive and auditory object recognition. One of the few examples is the work by Natale *et al.* [11] in which a robot was able to recognize seven objects using proprioceptive data extracted from the robot's hand as it grasped an object using a Self-Organizing Map.

Proprioceptive data has also been used to estimate an object's mass and moment of inertia [12], [13]. Other methods for estimating the dynamics of a robot's body (see [14], [15], [16], [17]), can also be applicable when estimating a grasped object's mass, etc. In contrast, the research presented in this paper explores how a general sequential representation for high-dimensional sensory data, coupled with standard machine learning algorithms, can be used by the robot to learn to recognize the objects it manipulates. Thus, the method described here is not specific to proprioception, but can instead be applied to two (and possibly more) different modalities.

In other related work, Nakamura *et al.* [18] describe a robot that uses proprioception along with visual and auditory information when interacting with objects. The robot used one modality to infer the outputs another (e.g. whether an object would make noise when picked up after only looking at it). Metta *et al.* [19] show that integrating proprioception with vision can bootstrap a robot's ability to manipulate objects.

Similarly, there has been some work on the use of auditory information for recognizing objects and their properties. One of the first studies in this area was conducted by Krotkov *et al.* [20] in which the robot was able to identify the material type (aluminum, brass, glass, wood, or plastic) of several objects by probing them with its end effector. Auditory-based material recognition has also been the topic of research by Richmond *et al.* [21] [22], which described a platform for measuring contact sounds between a robot's end-effector and objects of different materials. The robot was able to acquire acoustic models for four objects of different materials, by repeatedly striking the objects at different positions.

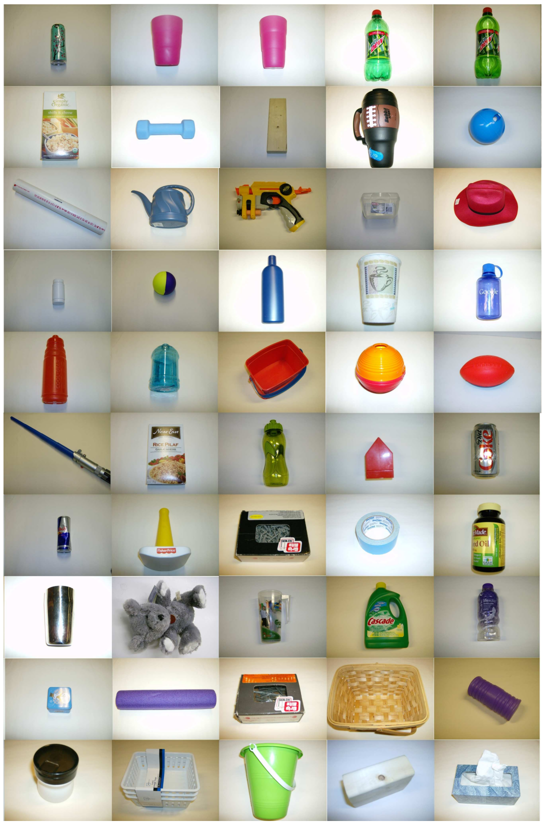

Fig. 2. The 50 objects used in this study (not shown to scale).

Torres-Jara *et al.* [23] demonstrated a robot that can perform acoustic-based object recognition using the sounds generated when tapping on the objects with its end effector. When tapping on a novel object, the spectrogram of the detected sound is matched to one that is already in the training set which results in a prediction for the object's type. This allowed the robot to correctly recognize four different objects.

More recently, Sinapov *et al.* [24] have shown that object recognition using auditory feedback can be scaled up to a much larger number of objects - 36 - and multiple robot behaviors (e.g., grasp, shake, tap, drop, push). The robot was able to recognize with high accuracy both the type of object and the type of interaction (i.e., behavior) using only the detected sound.

Following, this paper describes a method for interactive object recognition using a combination of proprioceptive and auditory feedback. The method is evaluated using a large-scale experimental study with 50 household objects (one of the largest number of objects reported in the robotics literature). We build upon our previous work on acoustic object recognition [24], [25] by showing that the model utilized for the representation of acoustic feedback is also very effective for proprioceptive feedback. Furthermore, we improve the object recognition model developed in [24] by allowing the robot to combine predictions from multiple exploratory behaviors and multiple sensory modalities in an intelligent manner.

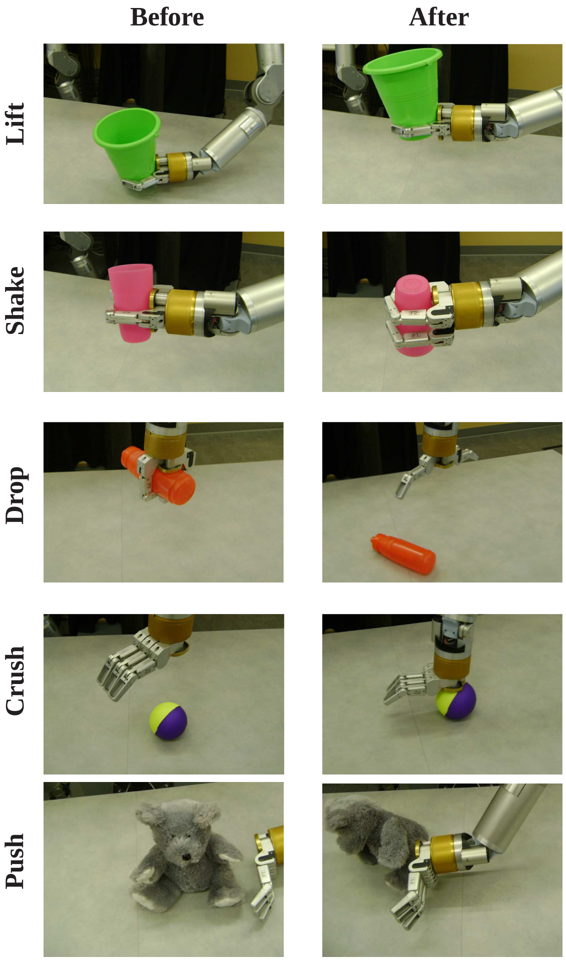

Fig. 3. *Before* and *after* snapshots of the five behaviors used by the robot.

III. Experimental Setup

A. Robot

The robot used in this study is an upper-torso humanoid robot, with two 7-DOF Barrett WAMs for arms and two 3-finger Barrett Hands as end effectors (see Fig.1). The robot is controlled in real time from a Linux PC at 500 Hz over a CAN bus interface. The raw torque data was captured and recorded at 500Hz using the robot's low-level API.

To capture auditory feedback, the robot's head was equipped with a U853AW Hanging Microphone. The microphone's output was first directed through an ART Tube MP Studio Microphone pre-amplifier, and subsequently processed through a Lexicon Alpha bus-powered audio interface which connects to the PC using USB. Sound input was recorded at 44.1 KHz using the Java Sound API over a 16-bit channel.

B. Objects

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

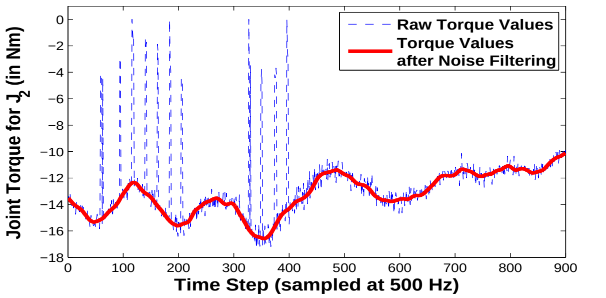

Fig. 4. Joint torque values for the shoulder joint (J_2) as the robot lifts the dumbbell object. The blue line shows the raw joint torques recorded using the robot's low-level API. The red line shows the filtered joint torques. See the text for filter details.

C. Behaviors

The set of behaviors, $\mathcal{B}$, consists of five exploratory behaviors that the robot performs on each object: *lift*, *shake*, *drop*, *crush*, and *push*. The behaviors were performed with the robot's left arm, and encoded with the Barrett WAM API. Fig. 3 shows *before* and *after* images for each of the five exploratory behaviors. The raw proprioceptive data (i.e., joint torques) and the raw audio were recorded for the duration of each behavior (start to end). Prior to the execution of each trial, each object was placed in roughly the same configuration (i.e., position and orientation). Due to human error, however, there was still variation of the grasp contact points, as well as the contact points with the object during the *push* and *crush* behaviors across multiple trials with the same object.

IV. Learning Methodology

A. Proprioceptive Feature Extraction

The first step in the feature extraction routine is to noise filter the raw joint torque values of the left arm recorded during each interaction. As can be seen in Fig. 4, the raw values are somewhat noisy and contain many spike readings. To handle this noise, the raw data is filtered using a filter of width 10 which checks for data points that lie more than 3 standard deviations away from the window median. Any such values are thrown out and replaced with the window median. The time series is then smoothed using a moving-average filter of size 10. The solid line in Figure 4 shows the resulting smoothed joint torque values after the noise-filtering procedure is performed.

The proprioceptive feedback, P_i, from the i^{th} interaction is represented as a sequence of states in a Self-Organizing Map (SOM) [26], one of several ways to quantize data vectors. This representation is obtained as follows: let $T_i = [t_1^i, t_2^i, \ldots, t_{l^i}^i]$ be the noise-filtered joint torque values for some interaction i, such that each $t_j^i \in \mathbb{R}^7$ denotes the torque values for all 7 joints of the left arm at time step j. Given a collection of joint torque records $\mathcal{T} = \{T_i\}_{i=1}^K$, a set of individual joint torque vectors is sampled and used as an input training dataset for the SOM. In other words, the SOM is trained with input datapoints $t_j^i \in \mathbb{R}^7$ where each data point denotes some particular recorded joint

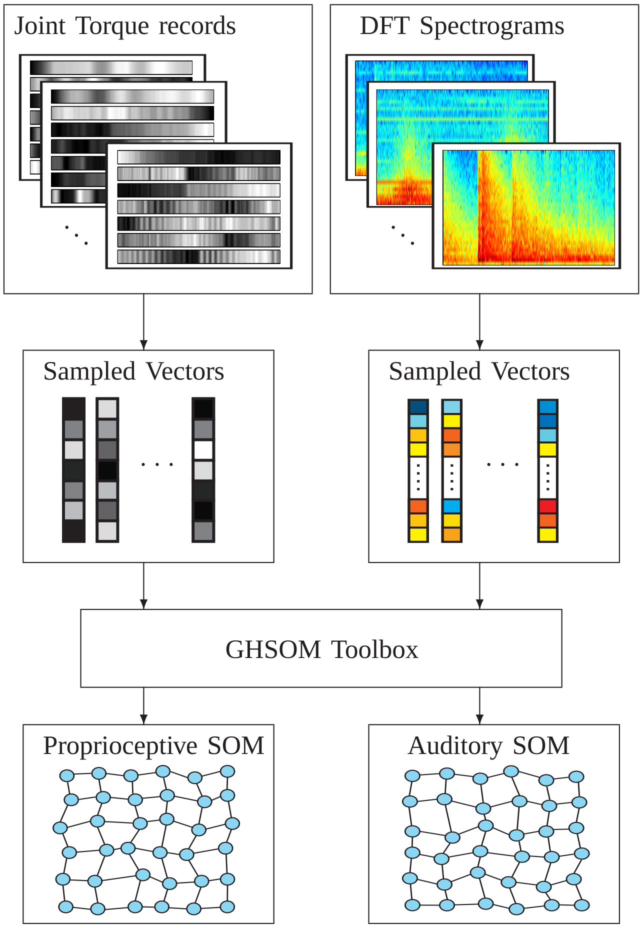

Fig. 5. Illustration of the procedure used to train the proprioceptive and auditory Self-Organizing Maps. **Proprioception** (left column): Given a set of joint torques recorded at 500 Hz during multiple interactions with different objects, a set of vectors is sampled at random and used as a dataset for training the SOM. Each of these vectors is in $\mathbb{R}^7$ and denotes the values of the 7 joint torques of the robot's left arm at a particular point in time. Once trained, the SOM can map any particular joint torque configuration to one of the SOM's states (i.e., the most highly activated state). **Audition** (right column): Given a set of Discrete Fourier Transform (DFT) spectrograms, a set of column vectors is extracted (each in $\mathbb{R}^{33}$) and used as a dataset for training the auditory SOM. The trained SOM can then map any particular DFT column from a novel spectrogram to the SOM node with the highest activation value given the input column vector.

torque values for all 7 joints. The Growing Hierarchical SOM toolbox was used to train a 6 by 6 SOM (i.e., 36 total nodes) using the default parameters for a non-growing 2-D single layer map [27]. Due to memory constraints, only $1/5$ of the available input data points $t^i_j \in \mathbb{R}^7$ were used for training. Figure 5 gives a visual overview of the training procedure while Figure 6 shows how each torque record T_i is mapped to a discrete sequence of states in the SOM.

After training the SOM, each torque record $T_i = [t^i_1, t^i_2, \ldots, t^i_{l^i}]$ is mapped to a sequence of SOM nodes, by mapping each vector t^i_j to a node in the map. A mapping function is defined, $Map(t^i_j) \rightarrow p^i_j$, where $t^i_j \in \mathbb{R}^7$ is the input torque vector and p^i_j is the node in the SOM with the highest activation value given the current input t^i_j. Hence, each torque record T_i is represented as a sequence, $P_i = p^i_1 p^i_2 \ldots p^i_{l^i}$, where $p^i_k \in \Gamma_p$, Γ_p is the set of nodes of the proprioceptive SOM, and l^i is the number of samples in the torque record T_i, as shown in Fig. 6. Thus, each P_i consists of a discrete sequence over a finite alphabet. This representation reduces the dimensionality of the proprioceptive feedback, thus affording the use of standard machine learning algorithms designed to work on sequential data.

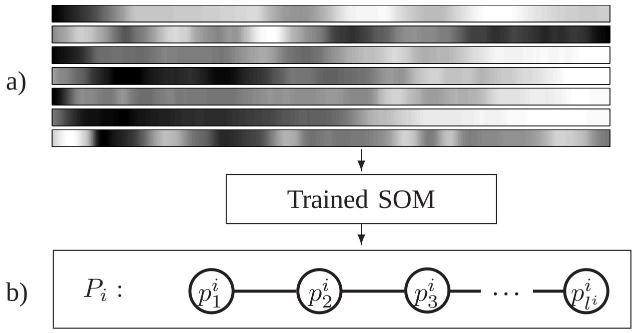

Fig. 6. Processing the proprioception data stream: a) The noise-filtered torque data for all 7 joints recorded while the robot lifts the dumbbell object. The horizontal axis denotes time while the color in each band indicates the torque values for each particular joint (white indicates low values while black indicates high values); b) The sequence of states in the SOM corresponding to the torques recorded during this interaction, obtained after each $\mathbb{R}^7$ column vector of torque data is mapped to a node in the SOM. The length of the sequence P_i is l^i, which is the same as the length of the horizontal time dimension of the torque data shown in a). Each sequence token $p^i_j \in \Gamma_p$, where Γ_p is the set of SOM nodes.

B. Auditory Feature Extraction

Similarly, the auditory feedback, A_i, from each interaction is also represented as a sequence of states in another Self-Organizing Map (SOM) (see Figure 7). To do this, features from each sound are first extracted using the log-normalized Discrete Fourier Transform (DFT), using $2^5 + 1 = 33$ frequency bins with a window of 26.6 milliseconds, computed every 10.0 milliseconds. The SPHINX4 natural language processing library (with default parameters) was used to compute the DFT [28]. Figure 7 shows the resulting spectrogram after applying the Fourier transform on a recorded sound. The spectrogram encodes the intensity level of each frequency bin (vertical axis) at each given point in time (horizontal axis).

As in the case with proprioceptive data, a 6 by 6 SOM is trained on extracted column vectors from the set of DFT spectrograms detected by the robot (see Figure 5). In other words, the SOM is trained with input data points in $\mathbb{R}^{33}$ which represent the intensity levels for each of the 33 spectrogram frequency bins at a given point in time.

Once the auditory SOM is trained, a column vector from any particular spectrogram can be efficiently mapped to a unique state in the SOM which has the highest activation value given the input vector. Thus, each sound is represented as a sequence, $A_i = a^i_1 a^i_2 \ldots a^i_{m^i}$, where each $a^i_k \in \Gamma_a$, Γ_a is the set of nodes on the auditory SOM, and m^i is the number of column vectors in the spectrogram, as shown in Fig. 7.

C. Data Collection

Let $\mathcal{B} = \{$*lift*, *shake*, *drop*, *crush*, *push*$\}$ be the set of exploratory behaviors available to the robot. For each of

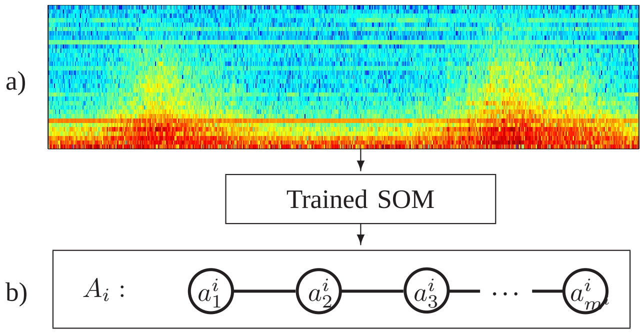

Fig. 7. Processing the auditory data stream: a) The discrete fourier transform (DFT) spectrogram of the detected sound after the robot performs the *shake* behavior on the box of mac&chees. The horizontal axis denotes time, while the vertical dimension denotes the 33 frequency bins. Orange-yellow color indicates high intensity, while blue-ish color denotes low intensity. b) The sequence of states in the SOM corresponding to the DFT recorded during this interaction, obtained after each $\mathbb{R}^{33}$ column vector of the DFT is mapped to a node in the SOM. Hence the length of the sequence A_i is equal to the number of column vectors of the input spectrogram. Each sequence token $a_j^i \in \Gamma_a$, where Γ_{audio} is the set of SOM nodes in the auditory SOM.

the five interactions, the robot performed ten trials with all 50 objects for a total of $5 \times 10 \times 50 = 2500$ recorded interactions. During the i^{th} trial, the robot recorded a data point of the form (B_i, O_i, P_i, A_i), where $B_i \in \mathcal{B}$ is the executed behavior, $O_i \in \mathcal{O}$ is the object in the current interaction, $P_i = p_1^i p_2^i \dots p_{li}^i$ is the proprioceptive sequence of most highly activated states in the proprioceptive SOM, and $A_i = a_1^i a_2^i \dots a_{m^i}^i$ is the auditory sequence of most highly activated states in the auditory SOM.

D. Object Recognition from a Single Modality

Given a proprioceptive or an auditory feedback sequence, P_i or A_i, detected as the robot performs behavior B_i on the test object, the task of the robot is to predict the correct object label O_i for the object in the interaction. The robot solves this problem by learning predictive models as follows. For each behavior $B \in \mathcal{B}$, the robot learns predictive models $\mathcal{M}_p^B$ and $\mathcal{M}_a^B$ which can estimate the correct object label O_i given the respective proprioceptive and auditory feedback sequences P_i and A_i. For example, given a proprioceptive sequence P_i detected as the robot performs the *lift* behavior, the proprioceptive predictive model $\mathcal{M}_p^{lift}$ estimates the probability $Pr_p^{lift}(O_i = o|P_i)$ for each object $o \in \mathcal{O}$. Similarly, the auditory predictive model estimates the probability of the object class $Pr_a^{lift}(O_i = o|A_i)$ given the auditory feedback sequence A_i. In both cases, the test object is assigned the label with the highest estimated probability.

In practice, the models $\mathcal{M}_p^B$ and $\mathcal{M}_a^B$ for each behavior $B \in \mathcal{B}$, can be realized by any machine learning method which can handle discrete sequences over a finite alphabet (i.e., strings) as an input. In this paper, these models are implemented by the k-Nearest Neighbors algorithm, a distance-based method which does not build an explicit model of the training data [29], [30]. Instead, given a test data point, it simply finds the k closest neighbors and outputs a prediction, which is a smoothed average over those neighbors. In this study, k was set to 3. An estimate for the probability of each object, given the sequences, is computed by counting the class labels of the k neighbors. For instance, if two of the three neighbors have object class label *plastic ball* then $Pr(O_i = \textit{plastic ball}) = \frac{2}{3}$. Similarly, if the class label of the remaining neighbor is *plastic cup*, then $Pr(O_i = \textit{plastic cup}) = \frac{1}{3}$.

The k-NN algorithm requires a distance measure, which can be used to compare the test data point to the training data points. Since each data point in this study is represented as a sequence over a finite alphabet, the Needleman-Wunsch global alignment algorithm [31], [32] was used, which can estimate the similarity between two sequences. While normally used for comparing biological or text sequences, the algorithm is applicable to other situations that require a distance measure between two strings. The algorithm requires a substitution cost to be defined over each pair of possible sequence tokens, e.g., the cost of substituting 'a' with 'b'. Since each token represents a node in a Self-Organizing Map, the cost for each pair of tokens was set to the Euclidean distance between their corresponding SOM nodes in the 2-D plane.

E. Combining Multiple Modalities

Finally, the robot needs to be able to combine the predictions from its proprioceptive and auditory recognition models in an efficient manner. Let $(B_i, O_{test}, P_i, A_i)_{i=1}^N$ be the recorded data after the robot has performed N behaviors on the object O_{test}. For example, this could be the sequential application of the *lift*, *shake*, and *drop* behaviors. For each of the models $\mathcal{M}_p^B$ and $\mathcal{M}_a^B$, let w_p^B and w_a^B, be the estimates for the models' object recognition performance (e.g., accuracy estimated by performing cross-validation on the training set). Given these estimates, and the input data $(B_i, O_{test}, P_i, A_i)_{i=1}^N$, the robot assigns the prediction to the object label o that maximizes:

$$\sum_i^N [w_p^{B_i} Pr_p^{B_i}(O_{test} = o|P_i) + w_a^{B_i} Pr_a^{B_i}(O_{test} = o|A_i)]$$

In other words, given one or more interactions with the object, the robot combines the predictions from different sensory modalities using estimates for the reliability of each channel of information. Note that the reliability weights for each modality are contingent on the behavior - e.g., auditory feedback may be very reliable when dropping the object, but much less reliable when the object is simply lifted.

It turns out that this means of integrating multiple modalities is similar to the way humans do the same task [1]. For example, when tasked to infer an object property given proprioceptive and visual feedback, humans use a weighted rule to combine the predictions of the two modalities, where the weights are proportional to each signal's estimated reliability [1]. Such a weighted combination of model predictions ensures that a sensory modality that is not useful in a given context will not dominate over other more reliable modalities or channels of information. For example, if it is expected that the auditory object recognition model will not achieve high accuracy when the robot performs the *lift* behavior (since the object will generate little if any sound), then in that context, the prediction from that model should be combined using a low reliability weight. The next section presents the results after evaluating

TABLE I
OBJECT RECOGNITION ACCURACY USING K-NN MODEL

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

the specific models $\mathcal{M}_p^B$ and $\mathcal{M}_a^B$ for each behavior $B \in \mathcal{B}$, as well as the weighted combination rule that was just presented.

V. RESULTS

A. Object Recognition from a Single Interaction

The first experiment evaluates the performance of the proprioceptive object recognition models $\mathcal{M}_p^B$ and auditory object recognition models $\mathcal{M}_a^B$ for each behavior $B \in \mathcal{B}$ using a single interaction with the test object. The performance of each model is reported in terms of the percentage of correct predictions (the accuracy) where:

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

The performance is estimated using 10-fold cross-validation: the set of data points $(B_i, O_i, P_i, A_i)_{i=1}^N$, where N = 2500, is split into ten folds corresponding to the ten trials performed with each object. During each of the ten iterations, nine of these folds are used for training the models and the remaining fold is used for evaluation.

The weighted combination of the auditory and proprioceptive models is also evaluated. During each round of the cross-validation procedure, the reliability weights w_p^B and w_a^B for each behavior $B \in \mathcal{B}$ are estimated by the robot by performing cross-validation on the training set only (i.e., the accuracy rate for each modality is estimated from the training set, without access to the test set).

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

Overall, the auditory stream is most informative when the object is dropped. The sound produced when the object hits the table implicitly captures many properties of the object: material type, size, and even shape (e.g., a ball will bounce when dropped, while a solid piece of wood will not). Proprioception on the other hand, is most reliable when the object is crushed, in which case the proprioceptive sequence captures the force detected when coming in contact with the object, as well as the timing that first contact. As expected, proprioception is also useful when lifting the object, since it implicitly captures the object's weight.

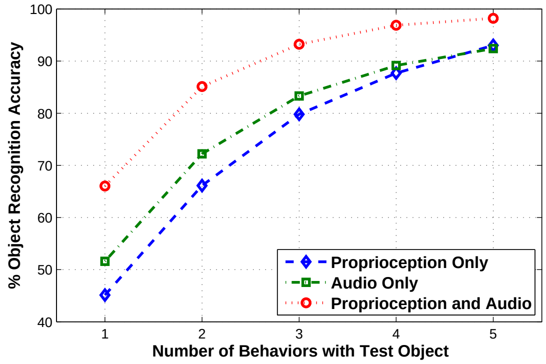

Fig. 8. Object recognition performance using proprioceptive and auditory feedback with k-Nearest Neighbor as the number of interactions with the test objects is varied from 1 (the default, used to generate Table I) to 5 (applying all five behaviors on the object). The results clearly show that the object recognition accuracy increases significantly after multiple interactions with the test object. Furthermore, by intelligently combining predictions from multiple modalities, the robot can recognize objects with higher accuracy than either modality alone. Overall, when performing all five behaviors on the test object, the robot's object recognition accuracy is 98.2%.

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

B. Object Recognition using Multiple Interactions

The second experiment evaluates whether the robot's object recognition performance can be boosted by applying multiple behaviors to the test object and combining the resulting predictions. For example, it should be easier to recognize the test object if the robot lifts, shakes and then drops the object, rather than applying just a single behavior. In this experiment, the number of available interactions with the test object is varied from 1 (the default case, used to generate Table I) to 5 (i.e., performing all five behaviors). When estimating the performance for 2, 3 and 4 interactions with the object, all possible combination of behaviors are considered (e.g., for 2 interactions, there are 10 possible combinations), and the mean accuracy is reported. Model predictions from multiple interactions with the object are combined using the reliability weights estimated for each combination of behavior and modality, as described in the previous section.

Figure 8 shows the results of this experiment. Not surprisingly, the recognition accuracy improves dramatically as the robot interacts with the object using more and more behaviors - once all five behaviors are performed, the robot

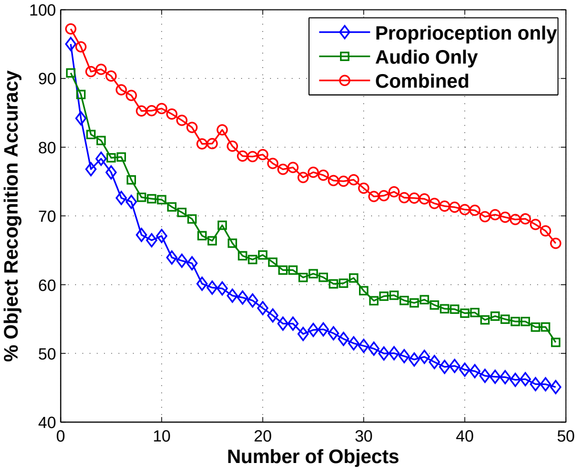

Fig. 9. Recognition accuracy from a single interaction as a function of the number of objects in the dataset.

achieves 98.2% recognition rate. Hence, *interactive* object recognition can provide highly accurate classification for a large set of objects, provided the robot is allowed to perform several interactions with the object and combine the resulting predictions in an intelligent manner.

C. Scalability

The last experiment looks at how the object recognition performance varies as the robot interacts with more and more objects. Most studies in robotics typically involve a small number of objects and presumably, it may be easier to achieve a higher recognition accuracy when dealing with a smaller set of objects. To test this assumption, the number of objects in the dataset was varied from 2 to 50 and for each value in that range, the algorithm was evaluated on a randomly selected subset of objects ten different times.

Figure 9 shows the mean accuracies for each of the three conditions (proprioception only, audio only, and combined) as a function of the number of objects the robot interacted with. As expected, with a small number of objects, the robot is able to achieve a high recognition rate even after a single interaction. A larger set of objects, on the other hand, inherently contains objects that have similar physical properties, thus increasing the difficulty of the recognition task. Therefore, robots which learn about objects should ultimately be evaluated on large object sets in order to get a more realistic estimate for their performance.

VI. Conclusions and Future Work

This paper described how a robot can couple proprioceptive and auditory feedback with exploratory behaviors in order to accurately recognize common household objects. Our robot interacted with 50 different objects by applying five different behaviors on them: *lift*, *shake*, *drop*, *crush*, and *push*. The proprioceptive and auditory feedback was represented as two sequences of the most highly activated nodes in two Self-Organizing Maps (one for each modality). Using the k-Nearest Neighbors algorithm, the robot was able to recognize the object in the interaction with accuracy substantially better than chance. The robot was also able to estimate the reliability of each modality in order to integrate them in a sound and intelligent manner. More importantly, after applying all 5 exploratory behaviors on the test object, the robot's recognition accuracy reached 98.2%, highlighting the importance of combining information extracted using multiple behaviors and multiple sensory modalities.

These results give a strong indication that traditional vision-based object recognition systems can be further improved by the additional use of auditory and proprioceptive feedback. This is particularly important for objects that may not be easily recognized using vision alone (e.g., a vision system cannot distinguish between a heavy and a light object that look identical). Thus, active interaction (as opposed to passive observation) is a necessary component in order to resolve such perceptual ambiguities about objects. Active object exploration is one of the hallmarks of human and animal intelligence (see [33], [34]) and therefore, robots may also be better suited for human environments if they can learn about objects by interacting with them.

There are several direct lines for future work. First, other methods for dimensionality reduction (e.g., vector quantization, or Spatio-Temporal Isomap, as used in [35]) can be applied in order to find meaningful patterns in the robot's proprioceptive sensory stream. Second, while the robot in our study was tasked to perform object recognition, it is also possible to use auditory and proprioceptive feedback to detect certain physical properties of the object (e.g., its material, whether it is hollow or solid, etc.). Some preliminary results indicate that after applying all 5 behaviors on a novel object, the robot can detect its material and other physical properties significantly better than chance [25]. Furthermore, the method of integrating information from proprioceptive and auditory feedback can be generalized to an arbitrary number of sensory modalities, allowing the robot to detect the reliability of each modality for each exploratory behavior. Integrating proprioceptive and tactile information from the robot's hand, as well as color and depth information from the robot's camera will allow the robot to further improve its ability to learn about common household objects. Robots that can interactively explore objects and make use of multiple sensory modalities will ultimately be better suited for working in human-inhabited environments.

References

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

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

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

[4] S. Srinivasa, C. Ferguson, D. Helfrich, D. Berenson, A. Collet, R. Diankov, G. Gallagher, G. Hollinger, J. Kuffner, and M. VandeWeghe, "Herb: A Home Exploring Robotic Butler," *Autonomous Robots - Special Issue on Autonomous Mobile Manipulation*, 2009.

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

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

[7] W. Gaver, “What in the world do we hear? An ecological approach to auditory event perception,” *Ecological Psychology*, vol. 5, pp. 1–29, 1993.

[8] B. Giordano and S. McAdams, “Material identification of real impact sounds: Effects of size variation in steel, glass, wood, and plexiglass plates,” *Journal of the Acoustical Society of America*, vol. 119, no. 2, pp. 1171–81, 2006.

[9] M. Grassi, “Do we hear size or sound? Balls dropped on plates,” *Perception and Psychophysics*, vol. 67, no. 2, pp. 274–284, 2005.

[10] R. B. Rusu, Z. C. Marton, N. Blodow, M. Dolha, and M. Beetz, “Towards 3d point cloud based object maps for household environments,” *Robotics and Autonomous Systems*, vol. 56, no. 11, pp. 927 – 941, 2008.

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

[12] D. Kubus, T. Kroger, and F. Wahl, “On-line rigid object recognition and pose estimation based on inertial parameters,” in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2007.

[13] D. Kubus and F. Wahl, “Estimating Inertial Load Parameters Using Force/Torque and Acceleration Sensor Fussion,” in *Robotic 2008, VDI-Beritche 2012 Munchen, Germany*, pp. 29–32.

[14] C. Atkeson, C. An, and J. Hollerbach, “Estimation of inertial parameters of manipulator loads and links,” *The International Journal of Robotics Research*, vol. 5, no. 3, pp. 101–119, 1986.

[15] J. Hollerbach and C. Wampler, “The calibration index and taxonomy for robot kinematic calibration methods,” *The International Journal of Robotics Research*, vol. 15, no. 6, pp. 573–591, 1996.

[16] T. Nanayakkara, K. Watanabe, and K. Izumi, “Evolving Runge-Kutta-Gill RBF Networks to Estimate the Dynamics of a Multi-Link Manipulator,” in *Systems, Man, and Cybernetics, IEEE SMC ’99 Conference Proceedings.*, 1999.

[17] M. Krabbes and C. Doschner, “Modelling of Robot Dynamics Bsed on Multi-Dimensional RBF-Like Neural Network,” in *Proceedings of 1999 International Conference on Information Intelligence and Systems (ICIIS)*, 1999.

[18] T. Nakamura, T. Nagai, and N. Iwahashi, “Multimodal object categorization by a robot,” in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2007, pp. 2415–2420.

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

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

[21] J. Richmond and D. Pai, “Active measurement of contact sounds,” in *Proc. of the IEEE Conference on Robotics and Automation*, 2000, pp. 2146–2152.

[22] J. L. Richmond, “Automatic measurement and modelling of contact sounds,” Master’s thesis, University of British Columbia, 2000.

[23] E. Torres-Jara, L. Natale, and P. Fitzpatrick, “Tapping into touch,” in *Proc. of the Fifth International Workshop on Epigenetic Robotics*, 2005, pp. 79–86.

[24] J. Sinapov, M. Weimer, and A. Stoytchev, “Interactive learning of the acoustic properties of household objects,” in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2009.

[25] J. Sinapov and A. Stoytchev, “From acoustic object recognition to object categorization by a humanoid robot,” in *Proceedings of the Workshop on Mobile Manipulation, part of 2009 Robotics Science and Systems conference, Seattle, WA.*, 2009.

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

[27] A. Chan and E. Pampalk, “Growing hierarchical self organizing map (ghsom) toolbox: visualizations and enhancements,” in *Proceedings of the 9th International Conference on Neural Information Processing (NIPS)*, 2002, pp. 2537–2541.

[28] K. E. Lee, H. W. Hon, and R. Reddy, “An overview of the SPHINX speech recognition system,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, no. 1, pp. 35–45, 1990.

[29] W. D. Aha, D. Kibler, and M. K. Albert, “Instance-based learning algorithm,” *Machine Learning*, vol. 6, pp. 37–66, 1991.

[30] C. G. Atkeson, A. W. Moore, and S. Schaal, “Locally weighted learning,” *Artificial Intelligence Review*, vol. 11, no. 1-5, pp. 11–73, 1997.

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

[32] 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.

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

[34] K. Lorenz, *Learning as Self-Organization*. Mahwah, NJ: Lawrence Erlbaum and Associates, Publishers, 1996, ch. Innate bases of learning.

[35] R. Peters, O. Jenkins, and R. Bodenheimer, “Sensory-Motor Manifold Structure Induced by Task Outcome: Experiments with Robonaut,” in *Proceedings of IEEE International Conference on Humanoid Robots*, 2006, pp. 484–489.

PLACE
PHOTO
HERE

Michael Shell Biography text here.

PLACE
PHOTO
HERE

Michael Shell Biography text here.

PLACE
PHOTO
HERE

Michael Shell Biography text here.

PLACE
PHOTO
HERE

Michael Shell Biography text here.

PLACE
PHOTO
HERE

Michael Shell Biography text here.

PLACE
PHOTO
HERE

Michael Shell Biography text here.