Interactive Object Recognition Using 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 the 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, coupled with exploratory behaviors, can be used successfully for object recognition. Furthermore, the robot was able to integrate feedback from the two modalities, to achieve even better recognition accuracy. Finally, the results show that the robot can boost its recognition rate even further by applying multiple different exploratory behaviors on the object.**

Fig. 1. The robot used in this study, shown here holding one of the 50 household objects used in the experiments.

I. Introduction

HUMAN beings have the remarkable ability to represent object knowledge using multiple modalities, including vision, touch, and proprioception [7]. Research in psychology has shown that multiple modalities are required to capture many object properties such as weight, roughness, and stiffness. [21]. In contrast, most object recognition systems used in robotics today use almost exclusively computer vision techniques and thus rely on a single modality [30], [39], [34], [31]. With a clear view of the target object, such systems can achieve high accuracy, but 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 (e.g., grasping an object inside of a bag). The human visual system is also subject to these same limitations - not surprisingly, humans need other sensory modalities to capture knowledge about objects [21], [35], [11].

To address the inherent limitations of the visual sensory modality, this paper proposes a novel behavior-grounded method for interactive recognition of household objects using proprioceptive and auditory feedback. While vision-based approaches typically use passive observation, our framework uses active interaction to recognize the objects. 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, which is used to convert the high-dimensional input from each modality into a discrete sequence of most-highly activated states in the map. This feature representation reduces the dimensionality of the sensory feedback, which allows the use of standard machine learning methods designed to handle sequential data. Using these extracted features, the proposed method enables the robot to learn behavior-grounded object recognition models, each of which is coupled with a specific behavior and sensory modality.

The framework was tested with an upper-torso humanoid robot (see Fig. 1), which interacted with 50 different household objects, one of the largest number of objects used in robotics experiments. The robot recognized the objects by extracting features from its proprioceptive and auditory sensory streams, while applying five different exploratory behaviors on the objects: *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 sensory modalities used in this paper. The results show that both auditory and proprioceptive feedback, coupled with specific behaviors, contain information indicative of the object being manipulated. In addition, the robot was able to integrate feedback from multiple modalities and multiple behaviors performed on each test object, which resulted in recognition accuracy of over 98%. Further analysis of these results gives a strong indication that equipping robots with a diverse set of exploratory behaviors is necessary in order to scale up interactive recognition methods to a large number of objects.

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.* [35] described a study 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 [11]) 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 the sound's source (e.g., bouncing ball, car engine, footsteps, etc.), rather than the properties of the sound itself (e.g., pitch, tone, etc.) [8]. Thus, the human auditory system plays a crucial role in both 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 multiple experimental studies. For example, Giordano *et al.* [9] demonstrated 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.* [10] showed 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, 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 [30], [39], [31] and/or 3D laser scan data [34]. 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.* [25] in which a robot was able to recognize seven objects with the help of a Self-Organizing map using proprioceptive data extracted from the robot's hand as it grasped an object.

Proprioceptive data has also been used to estimate an object's mass and moment of inertia [16], [17]. Methods for estimating the dynamics of a robot's body (see [2], [12], [24], [14]) could also be applied to estimate the mass of an object or some other properties. 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 that 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.* [23] 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 of another (e.g., whether an object would make noise when picked up after only looking at it). Metta *et al.* [22] 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.* [15]. Their 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 conducted by Richmond *et al.* [33] [32], who described a platform for measuring contact sounds between a robot's end-effector and objects made 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.

Torres-Jara *et al.* [40] 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 was matched to one that was already in the training set, which resulted in a prediction for the object's type. This allowed the robot to correctly recognize four different objects.

More recently, Sinapov *et al.* [38] have shown that object recognition using auditory feedback can be scaled up to a larger number of objects - 36 - and extended to 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., exploratory behavior) using only the detected sound.

Following this line of research, this paper describes a method for interactive object recognition using a combination of proprioceptive and auditory feedback. While most published experiments with robots typically use less than 10 objects, our method was 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 in acoustic [38], [36] and proprioceptive [4] object recognition. This paper uses the same data set as in [4], but also uses the auditory feedback, which was previously ignored. This study also improves the object recognition model developed in [38] by allowing the robot to use a weighted combination rule when combining feedback from multiple sensory modalities and multiple behavior-grounded recognition models.

.

III. Experimental Setup

A. Robot

The robot used in this study was 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

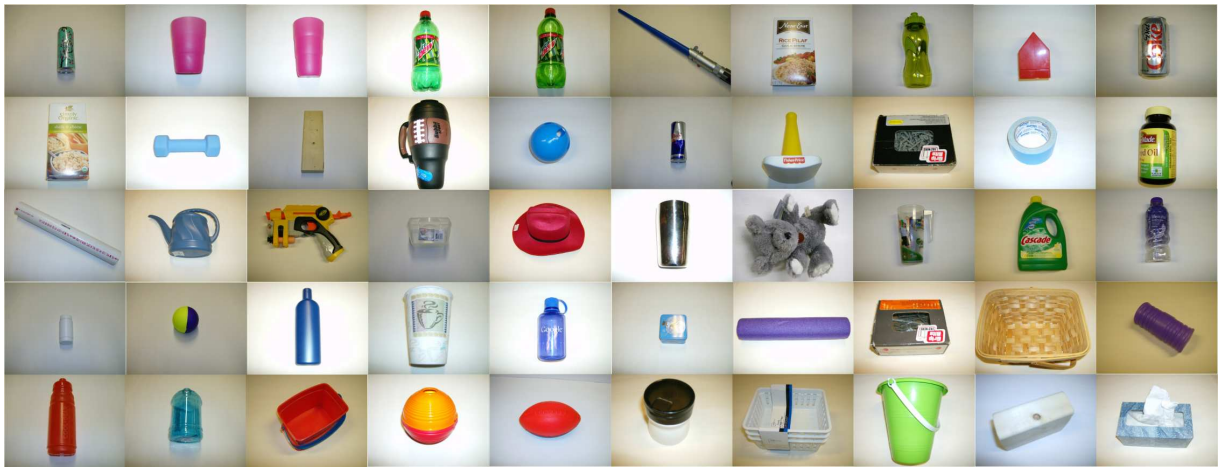


Fig. 2. The 50 household objects used in this study (not shown to scale). The object set includes cups, toys, balls, bottles, and containers. The objects are made of various materials, including plastic, metal, wood and paper.

was 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.

The robot's head was equipped with an Audio-Technica U853AW cardioid hanging microphone. The microphone's output was first routed 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.

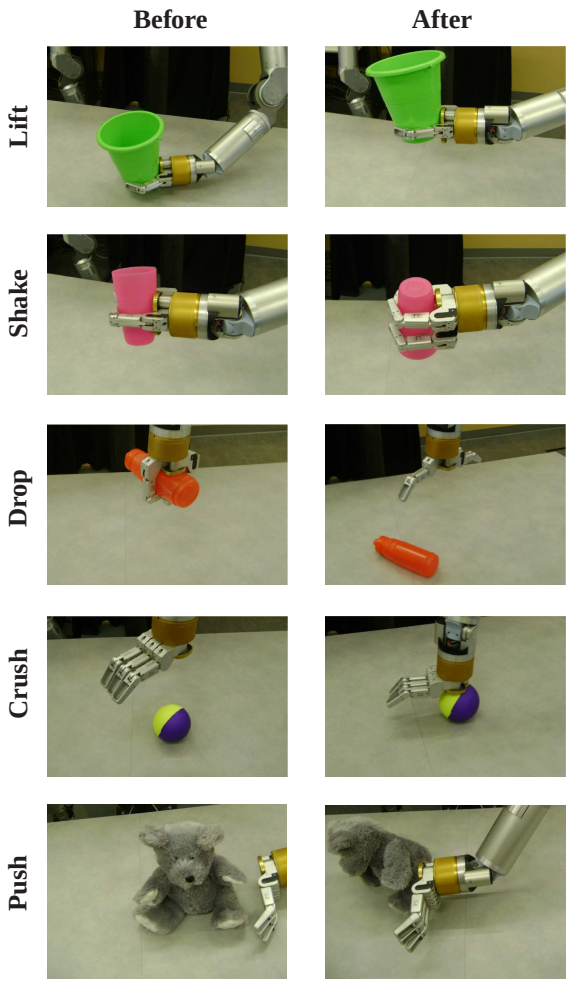


Fig. 3. *Before* and *after* snapshots of the five behaviors used by the robot.

B. Objects

The robot interacted with a set of objects, $\mathcal{O}$, consisting of 50 common household objects, including cups, bottles, and toys (see Fig. 2). The objects were 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.

C. Behaviors

The set of behaviors, $\mathcal{B}$, consisted of five exploratory behaviors that the robot performed 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 (position and orientation). Due to human error, however, there was still some 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 was to noise filter the raw joint torque values of the left arm, which were recorded during each interaction. As can be seen in Fig. 4, the raw values were somewhat noisy, containing many spike readings. To handle this noise, the raw data was filtered using a filter of width 10, which checked for data points that lie more

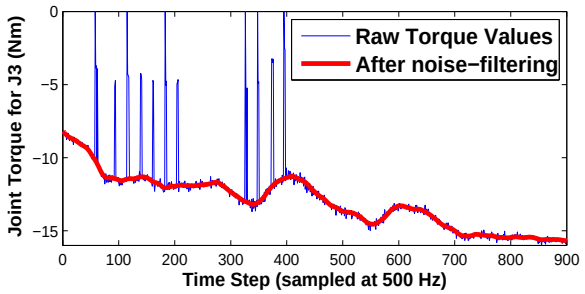


Fig. 4. Joint torque values for J_3 as the robot lifts the dumbbell object. The thinner line shows the raw joint torques recorded using the robot's low-level API. The thicker line shows the filtered joint torques. See the text for filter details.

than 3 standard deviations away from the window median. Any such values were thrown out and replaced with the window median. The time series was then smoothed using a moving-average filter of size 10. The solid line in Fig. 4 shows the resulting smoothed torque values after the noise-filtering procedure was performed.

The proprioceptive feedback, P_i, from the i^{th} interaction was represented as a sequence of states in a Self-Organizing Map (SOM) [13], one of several ways to quantize data vectors. This representation was 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 set of joint torque records $\mathcal{T} = \{T_i\}_{i=1}^K$, collected over K interactions with different objects, a set of individual joint torque vectors was sampled at random and used as an input training data set for the SOM. In other words, the SOM was trained with seven-dimensional input vectors, $t_j^i \in \mathbb{R}^7$, where each data point denoted a particular record of joint torque values (for all 7 joints). To avoid overfitting and to speed up the training process, only $1/5$ of the available input data points were used for training. The Growing Hierarchical SOM toolbox was used to train a 6 by 6 SOM (i.e., 36 total nodes) using the default parameters[1] for a non-growing 2-D single layer map [5]. Figure 5 gives an overview of the training procedure while Figure 6 shows how a torque record, T_i, can be mapped to a discrete sequence of states in the SOM.

After training the SOM, each torque record $T_i = [t_1^i, t_2^i, \ldots, t_{l^i}^i]$ was mapped to a sequence of SOM nodes, by mapping each vector t_j^i to a node in the map. A mapping function was defined, $Map(t_j^i) \rightarrow p_j^i$, where $t_j^i \in \mathbb{R}^7$ is the input torque vector and p_j^i is the node in the SOM with the highest activation value given the current input t_j^i. Thus, each torque record T_i was represented as a sequence, $P_i = p_1^i p_2^i \ldots p_{l^i}^i$, where $p_k^i \in \Gamma_p$, Γ_p was the set of nodes of the proprioceptive SOM, and l^i was the number of samples in the torque record T_i, as shown in Fig. 6. Thus, each P_i was represented as a discrete sequence over a finite alphabet. This representation reduced the dimensionality of the proprioceptive feedback, thus affording the use of standard machine learning algorithms designed to work on sequential data.

[1]Planar SOM with Euclidean distance metric, learning rate $\lambda = 0.7$, and 5 training cycles. The size of the SOM (6 by 6) was heuristically chosen based on prior work [38] and was not tuned to maximize performance. Parameters governing the growth of the map did not affect the results because the training option for a non-growing map was used.

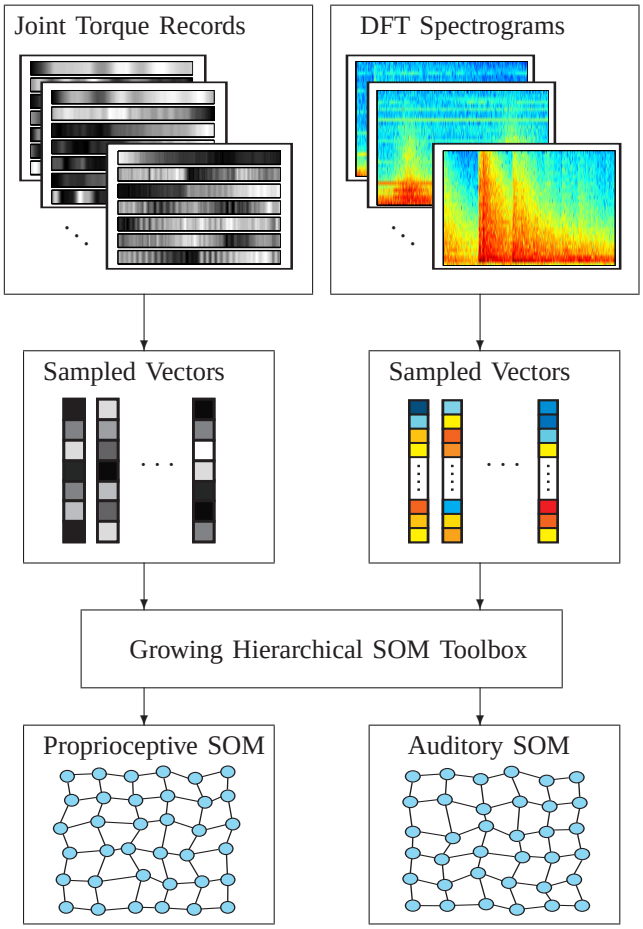


Fig. 5. Illustration of the procedure used to train the proprioceptive and auditory Self-Organizing Maps. **Proprioception** (left column): Given a set of joint torques recorded at 500 Hz during multiple interactions with different objects, a set of vectors is sampled at random and used as a data set 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 (the most highly activated state). **Audio** (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 data set for training the auditory SOM. The trained SOM can then map any particular DFT column vector from a novel spectrogram to the SOM node with the highest activation value.

B. Auditory Feature Extraction

Similarly, the auditory feedback from each interaction, A_i, was also represented as a sequence of states in another Self-Organizing Map (SOM) (see Figure 7). To do this, features from each sound were 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 was used to compute the DFT [19]. 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).

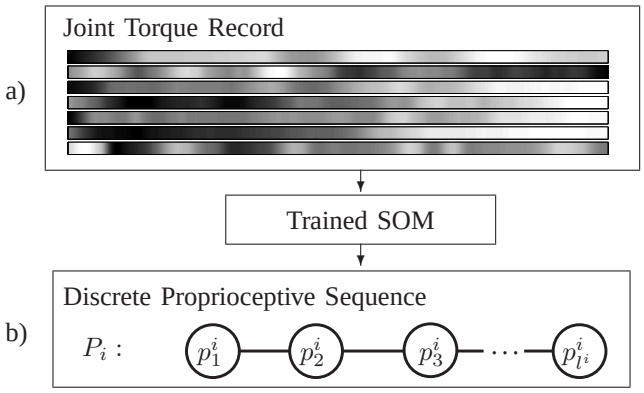


Fig. 6. Processing the proprioception data stream: a) The noise-filtered torque data for all 7 joints recorded while the robot lifts the dumbbell object. The horizontal axis denotes time while the color in each band indicates the torque values for each particular joint (white indicates low values while 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.

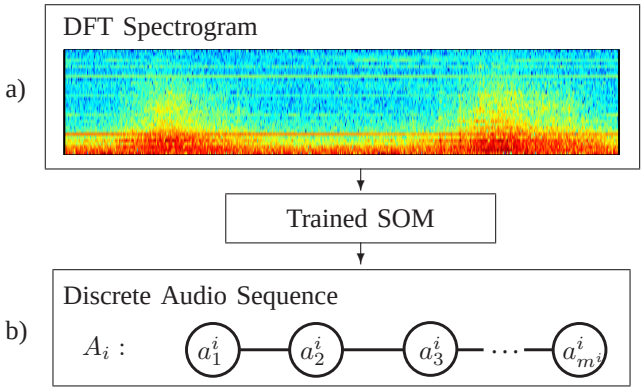


Fig. 7. Processing the auditory data stream: a) The Discrete Fourier Transform (DFT) spectrogram of the detected sound during one execution of the *shake* behavior on the mac&cheese box. The horizontal axis denotes time, while the vertical dimension denotes the 33 frequency bins. Orange-yellow color 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 was mapped to a node in the SOM. Thus, the length of the sequence A_i is equal to the number of column vectors of the input spectrogram. Each sequence token $a^i_j \in \Gamma_a$, where Γ_a is the set of SOM nodes in the auditory SOM.

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

Once the auditory SOM was trained, a column vector from any particular spectrogram could be efficiently mapped to a unique state in the SOM that has the highest activation value given the input vector. Thus, each sound was 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 was the set of nodes in the auditory SOM, and m^i was the number of column vectors in the spectrogram (see 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 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 $B_i \in \mathcal{B}$ was the executed behavior, $O_i \in \mathcal{O}$ was the object in the current interaction, $P_i = p^i_1 p^i_2 \ldots p^i_{l^i}$ was the proprioceptive sequence of most highly activated states in the proprioceptive SOM, and $A_i = a^i_1 a^i_2 \ldots a^i_{m^i}$ was the auditory sequence of most highly activated states in the auditory SOM. The recorded data set and the source code for this paper are available on-line at http://www.ece.iastate.edu/~alexs/lab/datasets/.

D. Object Recognition from a Single Modality

Given a proprioceptive or an auditory feedback sequence, P_i or A_i, detected as the robot performed behavior B_i on the test object, the task of the robot was to estimate the correct object label O_i for the object in the interaction. The robot solved this problem by learning recognition models as follows. For each behavior $B \in \mathcal{B}$, the robot learned recognition models $\mathcal{M}^B_p$ and $\mathcal{M}^B_a$, which could 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 performed the *lift* behavior, the proprioceptive recognition model $\mathcal{M}^{lift}_p$ could estimate the probability $Pr^{lift}_p(O_i = o|P_i)$ for each object $o \in \mathcal{O}$. Similarly, the auditory recognition model could estimate the probability of the object class $Pr^{lift}_a(O_i = o|A_i)$ given the auditory feedback sequence A_i. In both cases, the test object was assigned the label with the highest estimated probability.

In practice, the models $\mathcal{M}^B_p$ and $\mathcal{M}^B_a$ can be implemented by any machine learning method that can handle discrete sequences over a finite alphabet (i.e., strings) as an input. In this paper, these models were implemented by the k-Nearest Neighbors algorithm, a distance-based method, which does not build an explicit model of the training data [1], [3]. 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, was computed by counting the class labels of the k neighbors. For instance, if two of the three neighbors had an object class label *plastic ball* then $Pr(O_i = \textit{plastic ball}) = \frac{2}{3}$. Similarly, if the class label of the remaining neighbor was *plastic cup*, then $Pr(O_i = \textit{plastic cup}) = \frac{1}{3}$. The value for k was chosen heuristically, such that it is both large enough to allow probabilistic interpretation of the model's output, and also small enough relative to the number of trials per object that were used to train each of the robot's behavior-grounded recognition models (e.g., 9 trials when performing 10-fold cross-validation).

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 was represented as a sequence over a finite alphabet, the Needleman-Wunsch

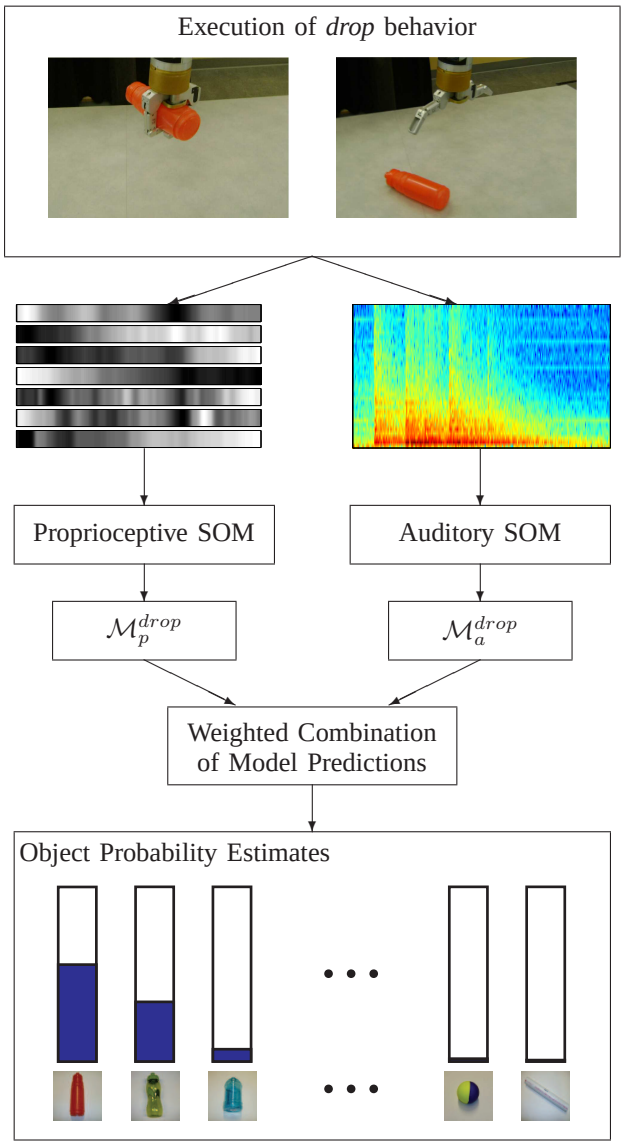


Fig. 8. Illustration of the procedure used to combine predictions from the proprioceptive and auditory object recognition models. In this trial, the robot dropped the test object and recorded the joint torque data and the Discrete Fourier Transform of the audio signal. They were subsequently discretized using the trained Self-Organizing Maps (one per modality). The resulting proprioceptive and auditory sequences were fed as input to the object recognition models $\mathcal{M}_p^{drop}$ and $\mathcal{M}_a^{drop}$, whose outputs were combined using weights corresponding the estimated performance of each model. The final output was a probability estimate for each object label (the object pictures are used for visualization only).

global alignment algorithm [26], [27] was used to 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. Section V.A describes the object recognition performance of the models $\mathcal{M}_p^B$ and $\mathcal{M}_a^B$ for all behaviors $B \in \mathcal{B}$.

E. Combining Multiple Modalities

Finally, we show how the robot can combine the outputs 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 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 could label the object with 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 same 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. Figure 8 illustrates the combination of auditory and proprioceptive feedback after performing an interaction with a test object.

It turns out that this method of integrating multiple modalities is similar to the way humans complete the same task [7]. For example, when asked to infer an object property given proprioceptive and visual feedback, humans use a weighted combination of the predictions of the two modalities, where the weights are proportional the estimated reliability of each modality [7]. The 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 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.

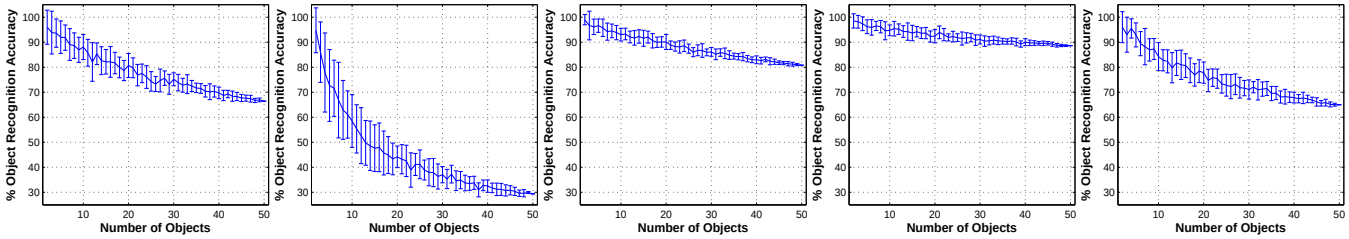
V. Results

A. Object Recognition with a Single Behavior

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 behavioral interaction with a test object. The performance of each model is reported in terms of the percentage of correct predictions (i.e., 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,



a) Lift b) Shake c) Drop d) Crush e) Push

Fig. 9. Recognition rates for the robot's behavior-grounded object recognition models (using both proprioceptive and auditory feedback) as a function of the number of objects, n, in the data set. For each value of n, 10-fold cross-validation was performed 20 different times, each with a different randomly selected object subset of size n. The solid lines indicate the resulting mean accuracy estimates while the error bars indicate the standard deviation of those estimates.

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 and behavior combination 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 auditory and proprioceptive recognition models perform substantially better than chance, with the auditory 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 proprioceptive model fares far better than the auditory 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. Proprioception, on the other hand, is most reliable when the object is crushed. The proprioceptive sequence implicitly captures the compliance and the height of the object through the initial contact force and the timing of the first contact with the object. As expected, proprioception is also useful when lifting the object, since it implicitly captures the object's weight.

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 results also show that combining the predictions from the two modalities improves the recognition accuracy for each of the five behaviors. This improvement is greatest for behaviors that yield reasonable performance for both modalities (e.g., *drop* and *crush*). However, even for behaviors where one of the modalities is far less reliable than the other (e.g., *lift*), there is still an 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. Scalability with a Single Behavior

The second experiment looks at how the object recognition performance varies as the robot interacts with more and more objects. Most studies in robotics typically use a small number of objects. Presumably, it may be possible to achieve a high recognition accuracy when dealing with a small set of objects, but low recognition accuracy when the number of objects is increased. To test this hypothesis, the number of objects, n, was varied from 2 to 50 and for each n smaller than 50, the model was evaluated on 20 different randomly chosen object subsets of size n. For each subset, the accuracy of each of the five behavior-grounded models was recorded and used to compute the expected accuracy (and standard deviation) for each value of n.

Figure 9 shows the mean accuracies and standard deviations for all five behavior-grounded recognition models as a function of the number of objects in the data set, when using the weighted combination of the proprioceptive-auditory model outputs. With a small number of objects, the robot is able to achieve a high recognition rate. As the robot interacts with more and more objects, however, the recognition rate drops since a larger set of objects inherently contains objects with similar physical properties. The same trend can also be seen in Figure 10, which shows the mean accuracy rates for the three modality conditions, averaged across all behaviors. Therefore, robots that learn about objects should ultimately be evaluated on large sets of objects in order to obtain more realistic and robust performance estimates.

C. Object Recognition with Multiple Behaviors

The next experiment evaluates whether the robot's object recognition performance on all 50 objects can be improved by applying multiple behaviors to the test object and combining

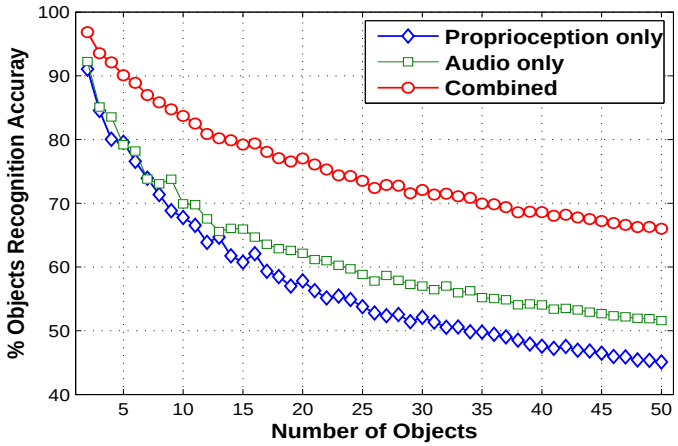


Fig. 10. Average recognition accuracies from a single behavioral interaction as the number of objects, n, is varied from 2 to 50. For each value of n, 10-fold cross-validation was performed 20 different times, each with a different randomly selected object subset of size n.

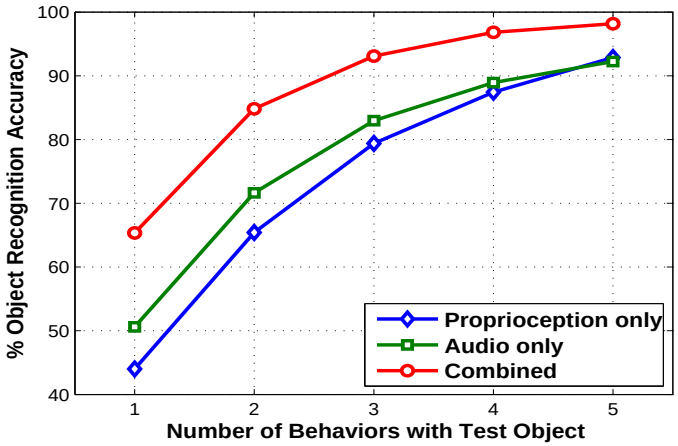


Fig. 11. Object recognition performance with all 50 objects using proprioceptive and auditory feedback with k-NN as the number of interactions with the test objects is varied from 1 (resulting in the average per behavior accuracies shown in Table I) to 5 (applying all five behaviors on the object, and combining the resulting predictions). Overall, when performing all five behaviors on the test object, the robot's object recognition accuracy is 98.2%.

the resulting predictions. For example, it should be easier to recognize the test object if the robot lifts, shakes and then drops the object, than if it applies 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 combinations 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 11 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, it reaches 98.2%. This shows that *interactive* object recognition can provide highly accurate classification for a large set of objects, as long as the robot is allowed to perform several behavioral interactions with the object and combine their resulting predictions in an efficient manner.

A subsequent question to answer is whether the same type of recognition improvement can be achieved by performing the same behavior multiple times on the test object (as opposed to applying multiple different behaviors). An evaluation experiment was conducted in which the data set was split into 5 folds (each containing 2 trials with all five behaviors performed on each object) and 5-fold cross validation was performed. In other words, during each of the five iterations, the model was trained on 4 of the folds, and tested on the remaining one. For each of the five behaviors, the test set now contains two instances of the same behavior applied on each of the 50 objects. The test set also contains 4 instances for each of the $\binom{5}{2} = 10$ unique combinations of different behaviors (e.g., lift-shake) per object. After all five rounds of cross-validation, the individual accuracies of the five behaviors were estimated from the recorded model outputs when compared to the actual object IDs. The accuracies for each combination of exploratory behaviors were also estimated and stored in a 5×5 matrix. The diagonal entries of this symmetric matrix contain the 5 accuracy estimates obtained when performing the same behavior twice, while the 10 lower-diagonal entries contain the accuracy obtained when combining feedback from each of the 10 unique pairs of behaviors. These estimates were used to compute the improvement in recognition accuracy for different combinations of behaviors as described below.

Let $acc(\mathcal{M}^i, \mathcal{M}^j)$ be the estimated recognition accuracy when combining the outputs of recognition models $\mathcal{M}^i$ and $\mathcal{M}^j$ associated with behaviors B_i and B_j, and let $acc(\mathcal{M}^i)$ and $acc(\mathcal{M}^j)$ be their individual accuracies estimated when performing a single behavior execution on the test object. Given two behaviors B_i and B_j (which may be the same if $i = j$), the recognition improvement (RI_{ij}) obtained when applying the two behaviors sequentially on the test object can be measured relative to the recognition accuracy of the individual behaviors, i.e.,

$$RI_{ij} = acc(\mathcal{M}^i, \mathcal{M}^j) - \frac{acc(\mathcal{M}^i) + acc(\mathcal{M}^j)}{2}$$

With this formulation we can test whether combining feedback from two different behaviors results in greater recognition boost than combining feedback from two executions of the same behavior. The results of this evaluation, shown in Figure 12, confirm that the recognition improvement is higher when two different exploratory behaviors are applied on the test object, as opposed to applying the same behavior twice. This result gives a strong indication that the diversity of the exploratory behaviors is more important than the number of times each behavior is executed when classifying an object.

D. The Boosting Effect of Exploratory Behaviors

The previous experiments showed that by performing multiple exploratory behaviors a robot can query multiple recognition models, each tied to a specific behavior and sensory modality, and thus dramatically improve its object recognition

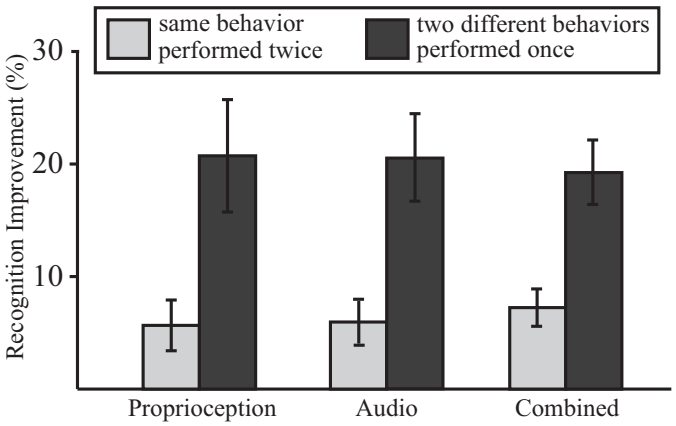


Fig. 12. Object recognition improvement obtained by combining model outputs after two executions of the *same* behavior as well as two executions of *different* behaviors, estimated using 5-fold cross-validation. In all cases, the recognition improvement is higher when combining feedback from two distinct exploratory behaviors. When applying the same behavior twice, the standard deviation of the recognition improvement was estimated from 5 samples, one for each behavior. When applying two different behaviors, the standard deviation was estimated from 10 samples, one for each unique pair of behaviors.

rate. But what is the cause of this apparent boost in recognition rate? To answer this question, we turn to research in machine learning theory, which has shown that the rate of classification improvement of a classifier ensemble can be linked to pairwise measures of classifier diversity [6], [18]. For example, combining classifier predictions from two diverse or complementary classifiers generally results in higher recognition rates [18]. These machine learning results rely on the assumption that the data points in the original data set are identically and independently distributed (i.i.d.) and that each individual classifier is trained on a biased subset of the original data set. In our setting, each behavior-grounded recognition model is also trained on a biased subset of the data set, corresponding to the data produced by the specific behavior and sensory modality associated with that model. However, the i.i.d. assumption is clearly violated – for example, the sounds produced by dropping the objects come from a different distribution than the sounds produced when shaking the objects. The data points are also not independently distributed, since the initial grasp configuration can influence both the proprioceptive and the auditory feedback produced by the subsequent exploratory behaviors. Despite these differences, the next experiment examines whether the relationship between classifier diversity and recognition improvement still holds.

Given a behavior-grounded recognition model $\mathcal{M}^i$, let $\mathbf{y_i} = [y_{1,i}, \ldots, y_{K,i}]^T$ be a $K-$ dimensional binary vector, such that $y_{k,i} = 1$ if the model $\mathcal{M}^i$ correctly labels the object explored during trial k, and 0 otherwise. The pairwise diversity between two models $\mathcal{M}^i$ and $\mathcal{M}^j$ can be measured by comparing the corresponding vectors $\mathbf{y_i}$ and $\mathbf{y_j}$. One such metric is the disagreement measure [18], which is defined as:

$$DIS_{i,j} = \frac{N^{01} + N^{10}}{N^{11} + N^{10} + N^{01} + N^{00}}$$

where N^{pq} is the number of trials (out of K) for which $y_{k,i} = p$ and $y_{k,j} = q$, e.g., N^{10} is the number of trials in which

TABLE II
THE RELATIONSHIP BETWEEN A PAIR OF RECOGNITION MODELS $\mathcal{M}^i$ AND $\mathcal{M}^j$ CAN BE EXPRESSED USING A 2 X 2 TABLE, WHICH SHOWS HOW OFTEN THEIR PREDICTIONS COINCIDE (N^{11} AND N^{00}) AND HOW OFTEN THEY DISAGREE (N^{01} AND N^{10}).

	$\mathcal{M}^j$ correct	$\mathcal{M}^j$ wrong
$\mathcal{M}^i$ correct	N^{11}	N^{10}
$\mathcal{M}^i$ wrong	N^{01}	N^{00}

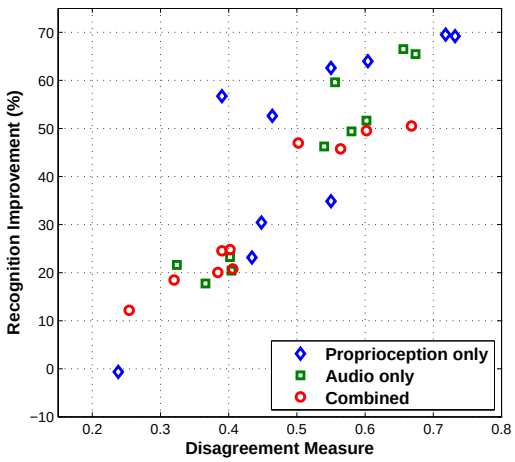


Fig. 13. The relationship between the pairwise disagreement measure and the recognition improvement for each of the 10 possible pairs of behaviors, under three different modality conditions (audio only, proprioception only, or combined), estimated using 10-fold cross-validation.

the object was correctly recognized by model $\mathcal{M}^i$, but misclassified by model $\mathcal{M}^j$ (see Table II). In other words, the disagreement measure is equal to the number of trials in which one model was correct and the other was wrong divided by the total number of trials. Calculating this measure is equivalent to computing the normalized Hamming distance between the vectors $\mathbf{y_i}$ and $\mathbf{y_j}$. The range of this measure is always in the interval of 0.0 to 1.0, with low values indicating that the two models tend to agree, regardless of whether they are right or wrong.

Figure 13 shows the relationship between the disagreement measure and the recognition improvement (as defined in the previous subsection) for each one of the 10 unique pairs of behaviors and for all three modality conditions: audio only, proprioception only, or combined. The plot shows that the amount of disagreement is linearly related to the expected improvement. Thus, the collection of the robot's behavior-grounded recognition models acts as an ensemble of classifiers.

This is indeed a surprising result, since in the machine learning literature it is assumed that the classifiers in the collection are trained on identically and independently distributed data points. Nevertheless, even when the i.i.d. assumption is violated, the concept of classifier diversity was found to be useful for explaining the improvement in recognition accuracy in the robot experiments. This finding establishes an important link between research in machine learning theory and studies in robotics that make use of multiple exploratory behaviors and sensory modalities (see [37] for further details). Overall, the results from this experiment highlight the importance of

enabling robots to perceive objects using multiple, as well as diverse, channels of information (e.g., different behaviors and different sensory modalities).

VI. Conclusions and Future Work

The main contribution of this paper is a framework for interactive object recognition, that can handle multiple exploratory behaviors and multiple sensory modalities. The proposed object recognition framework was evaluated using a large-scale experimental study, in which the robot manipulated 50 different objects using five exploratory behaviors (*lift*, *shake*, *drop*, *crush*, and *push*) and two sensory modalities (audio and proprioception). The feedback from the two sensory modalities, detected by the robot while interacting with an object, was represented as two sequences of the most highly activated nodes in two Self-Organizing Maps (one for each modality). Using global sequence comparison coupled with the k-Nearest Neighbors algorithm, the robot was able to recognize the explored object with accuracy substantially better than chance. The robot was also able to compute estimates for the reliability of each sensory modality and use them to improve its object recognition accuracy.

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. Analysis of the results also showed that the learned behavior-grounded recognition models act as an ensemble of classifiers. Thus, by applying a set of diverse behaviors on an object, the robot can boost its recognition accuracy.

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 heavy and a light object that look identical). Thus, active interaction (as opposed to passive observation) is a necessary component for resolving perceptual ambiguities about objects. Active object exploration is one of the hallmarks of human and animal intelligence (see [29], [20]), which lends further credence to our approach to object recognition using exploratory behaviors.

There are several possible avenues for future work. First, other methods for dimensionality reduction (e.g., vector quantization, or Spatio-Temporal Isomap, as used in [28]) can be applied in order to find meaningful patterns in the robot's proprioceptive and auditory sensory streams. Second, while the robot in our study was tested on an object recognition task, it is also possible to use auditory and proprioceptive feedback to detect certain physical properties of the object (e.g., its material type, 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 type and other physical properties significantly better than chance [36]. Furthermore, the method for 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.

Acknowledgment

This work was funded in part by NSF Research Experience for Undergraduates (REU) Grant IIS-0851976.

References

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

[2] C. Atkeson, C.H. An, and J. Hollerbach. Estimation of inertial parameters of manipulator loads and links. *The International Journal of Robotics Research*, 5(3):101–119, 1986.

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

[4] T. Bergquist, C. Schenck, U. Ohiri, J. Sinapov, S. Griffith, and A. Stoytchev. Interactive object recognition using proprioceptive feedback. In *Proceedings of the 2009 IROS Workshop: Semantic Perception for Robot Manipulation, St. Louis, MO*, 2009.

[5] A. Chan and E. Pampalk. Growing hierarchical self organizing map (ghsom) toolbox: visualizations and enhancements. In *Proc. of the 9th Intl. Conf. on Neural Information Processing (NIPS)*, pages 2537–2541, 2002.

[6] T. Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting and randomization. *Machine Learning*, 40(2):139–157, 2000.

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

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

[9] B. Giordano and S. McAdams. Material identification of real impact sounds: Effects of size variation in steel, glass, wood, and plexiglass plates. *J. of the Acoustical Soc. of America*, 119(2):1171–81, 2006.

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

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

[12] J. Hollerbach and C. Wampler. The calibration index and taxonomy for robot kinematic calibration methods. *The International Journal of Robotics Research*, 15(6):573–591, 1996.

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

[14] M. Krabbes and C. Döschner. Modelling of Robot Dynamics Bsed on Multi-Dimensional RBF-Like Neural Network. In *Proc. of Intl. Conf. on Information Intelligence and Systems (ICIIS)*, pages 180–187, 1999.

[15] E. Krotkov, R. Klatzky, and N. Zumel. Robotic perception of material: Experiments with shape-invariant acoustic measures of material type. In *Experimental Robotics IV*, volume 223 of *Lecture Notes in Control and Information Sciences*, pages 204–211. Springer Berlin, 1996.

[16] D. Kubus, T. Kroger, and F.M. Wahl. On-line rigid object recognition and pose estimation based on inertial parameters. In *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, pages 1402–1408, 2007.

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

[18] L. Kuncheva and C. Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51(2):181–207, 2003.

[19] K. Lee, H. Hon, and R. Reddy. An overview of the SPHINX speech recognition system. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38(1):35–45, 1990.

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

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

[22] G. Metta and P. Fitzpatrick. Early integration of vision and manipulation. *Adaptive Behavior*, 11(2):109–128, 2003.

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

[24] T. Nanayakkara, K. Watanabe, and K. Izumi. Evolving Runge-Kutta-Gill RBF Networks to Estimate the Dynamics of a Multi-Link Manipulator. In *Proc. of Systems, Man, and Cybernetics*, pages 770–775, 1999.

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

[26] G. Navarro. A guided tour to approximate string matching. *ACM Computing Surveys*, 33(1):31–88, 2001.

[27] 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.*, 48(3):443–453, 1970.

[28] R. Peters, O. Jenkins, and R. Bodenheimer. Sensory-Motor Manifold Structure Induced by Task Outcome: Experiments with Robonaut. In *Proc. of IEEE Intl. Conf. on Humanoid Robots*, pages 484–489, 2006.

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

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

[31] B. Rasolzadeh, M. Björkman, K. Huebner, and D. Kragic. An active vision system for detecting, fixating and manipulating objects in the real world. *International Journal of Robotics Research*, 29(2-3):133–154, 2010.

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

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

[34] R. B. Rusu, Z. C. Marton, N. Blodow, M. Dolha, and M. Beetz. Towards 3d point cloud based object maps for household environments. *Robotics and Autonomous Systems*, 56(11):927 – 941, 2008.

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

[36] J. Sinapov and A. Stoytchev. From acoustic object recognition to object categorization by a humanoid robot. In *Proc. of the RSS 2009 Workshop on Mobile Manipulation, Seattle, WA.*, 2009.

[37] J. Sinapov and A. Stoytchev. The boosting effect of exploratory behaviors. In *Proc. National Conferrence on Artificial Intelligence (AAAI)*, pages 1613–1618, 2010.

[38] J. Sinapov, M. Weimer, and A. Stoytchev. Interactive learning of the acoustic properties of household objects. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 2518–2524, 2009.

[39] S. Srinivasa, C. Ferguson, D. Helfrich, D. Berenson, A. Collet, R. Diankov, G. Gallagher, G. Hollinger, J. Kuffner, and M. VandeWeghe. Herb: A Home Exploring Robotic Butler. *Autonomous Robots*, 28(1):5–20, 2009.

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