Sensorimotor Cross-Perception Knowledge Transfer for Grounded Category Recognition

Gyan Tatiya and Jivko Sinapov
Department of Computer Science
Tufts University
{Gyan.Tatiya}|{Jivko.Sinapov}@tufts.edu

I. Introduction

Humans use exploratory behaviors coupled with multi-modal perception to learn about the objects around them. Research in robotics has shown that robots too can use such behaviors (e.g., grasping, pushing, shaking) to infer object properties by using multiple perceptual modalities such as vision, haptic, tactile and vibration [1]. However, in some situations, it is possible that one of the sensors fails and some of the modules in the system require its input. To address this challenge, we propose a framework for knowledge transfer from the set of available sensors to recover signal of the failed sensor. The intuition behind our approach is that if the robot interacts with a set of objects while all the sensors are working, the produced sensory data can be used to learn a mapping between different feature spaces. We evaluate the framework on a category recognition task using a dataset containing 9 robot behaviors performed multiple times on a set of 100 objects. The results show that the proposed framework can enable to recover the signal of failed sensor that can be used to perform the object category recognition.

II. Learning Methodology

A. Notation and Problem Formulation

Let $\mathcal{B}$ be the set of exploratory behaviors (e.g. *lift*), let $\mathcal{M}$ be the set of sensory modalities (e.g. *haptic*), and let $\mathcal{C}$ be the set of sensorimotor contexts such that each context $c \in \mathcal{C}$ refers to a combination of a behavior $b \in \mathcal{B}$ and a sensory modality $m \in \mathcal{M}$ (e.g. *lift-haptic*). For each exploration trial, the robot performs exploratory behaviors $b \in \mathcal{B}$ on a specific object and record a sensory signal for each modality in $\mathcal{M}$. Thus, during the i^{th} exploration trial, the robot observed features $x_i^c \in \mathbb{R}^{D_c}$. Here, D_c is the dimensions of the features observed by the robot under contexts c.

We divide our total set of possible object categories $\mathcal{Y}$ into two mutually exclusive subsets: $\mathcal{Y}_{\text{sensor-work}}$ and $\mathcal{Y}_{\text{sensor-fail}}$. All the sensors work while the robot interacts with objects in the categories in $\mathcal{Y}_{\text{sensor-work}}$, and these categories are used during the training phase. Categories in $\mathcal{Y}_{\text{sensor-fail}}$ are only experienced by the robot when one of its sensors stops working. The goal of our work is to recover the signal of the failed sensor using the available sensors and train the robot to recognize an object at test time from one of the categories in $\mathcal{Y}_{\text{sensor-fail}}$ using the recovered sensory signal.

B. Knowledge Transfer Model

We propose using encoder-decoder neural network to transfer knowledge from the available sensors to recover the failed sensor signal. First, the encoder network transforms the observed feature vector of the robot x_i^c, to a lower-dimensional, fixed-size code vector $z_i \in \mathbb{R}^{D_z}$ of size D_z. We denote this non-linear mapping by an encoder function f: $z_i = f_\theta(x_i^c)$, which takes network parameter weights θ. Next, a decoder network maps an input code vector z_i to create a vector of "recovered" target feature vector $\hat{x}_i^c$. We denote this non-linear mapping by a decoder function g: $\hat{x_i}^c = g_\phi(z_i)$, which takes network parameter weights ϕ.

Training the encoder-decoder requires observing features from the robot across a set of N total objects using two different contexts c_1, c_2. Given a dataset of feature pairs $\{x_i^{c_1}, x_i^{c_2}\}_{i=1}^N$, we wish to find parameters (θ, ϕ) that minimize the error between the real features $x_i^{c_2}$ observed by the robot and the model's "recovered" target features $\hat{x}_i^{c_2}$ obtained by applying the encoder-decoder to the corresponding source features $x_i^{c_1}$.

Given a pre-trained encoder-decoder, the robot can recover signal of the failed sensor by using available sensors, and classify objects from $\mathcal{Y}_{\text{sensor-fail}}$ categories. For each object j from a category y_j that only the robot has seen in context c_1, a "recovered" training set was created of failed sensor's feature, category label pairs: $\{g_\phi(f_\theta(x_j^{c_1})), y_j\}$, then a standard support vector machine (SVM) classifier was trained from this dataset. Then, when deployed in an environment with novel objects without category labels, the robot can measure observed features x^{c_2} and feed these features into its pretrained SVM classifier to predict which category within the set $\mathcal{Y}_{\text{sensor-fail}}$ it has observed. We assume that at test time, only categories from $\mathcal{Y}_{\text{sensor-fail}}$ are possible.

III. Evaluation and Results

A. Evaluation

We assume that all the sensors works when the robot interacts with 15 categories, but then a sensor fails while the robot interacts with the rest 5 categories. The objects of the 15 categories are used to train the encoder-decoder network that projects the sensory signal of the available sensors to the failed sensor. Subsequently, the trained encoder-decoder network is used to generate "recovered" sensory signals for the other 5 object categories in $\mathcal{Y}_{\text{source-fail}}$.

We consider two possible category recognition approaches: our proposed transfer-learning pipeline using the projected data from the available sensors (i.e., how well it would do if it transfered knowledge from the available sensors), and a non-transfer ideal baseline using ground truth features produced by the robot (i.e., the best the robot could do if all of its sensors were working). In both cases, real features observed by the robot are used as input to the classifier at test time. We used 5-fold object-based cross-validation, where the training set consisted of 4 objects from each of the 5 categories the robot interacted with a failed sensor and the test set consisted of the remaining objects.

We used two metrics to evaluate the performance of the robot on the object category classification. The first was accuracy[1] (%). The whole evaluation process is repeated 10 times to get a mean accuracy and a standard deviation. The second metric was accuracy delta (%), which measures the drop in classification accuracy as a result of using projected features for training as opposed to the ground-truth features. We define this loss as $A\Delta = A_{truth} - A_{projected}$, where A_{truth} and $A_{projected}$ are the accuracies obtained when using real and projected features, respectively. Smaller accuracy delta indicates that it is easy to project available sensory features in the failed sensor's feature space, and the robot can use these projected features to learn a classifier that can achieve comparable performance as if the sensor works.

B. Results

1) Dataset Description: We used the dataset described in [1], in which a robot explores 100 objects belonging to 20 categories using 9 behaviors: *Crush*, *Grasp*, *Hold*, *Lift*, *Drop*, *Poke*, *Push*, *Shake* and *Tap*. During each behavior the robot recorded 4 modalities: visual (SURF), haptic, vibrotactile and audio, and we used its features as described in [1].

2) Accuracy Results of Category Recognition: Since there are 4 modalities, if a sensor fails there are 3 possible one-to-one mappings each from an available sensor, so there are 12 (4 x 3) possible mappings. There are 9 behaviors, so there are 108 (12 x 9) one-to-one projections. For many-to-one mapping, we concatenated the features of all the available sensors to recover the features of the failed sensor. Thus, there are 4 possible many-to-one mappings to recover each modality, and there are 36 (4 x 9) many-to-one projections. Fig. 1 and Fig. 2 show the 5 one-to-one and 5 many-to-one projections with the least accuracy delta, respectively.

For one-to-one projections, recovering haptic features from vibrotactile was the easiest task indicating that knowing what an object's surface feels like when performing a behavior can inform how much forces would be felt when performing that behavior. In most cases, many-to-one projections does not perform significantly better than one-to-one projections. However, in some cases, adding more modalities improved performance. For example, the accuracy delta to recover *shake-haptic* from *shake-vibro* is 34.4% and from *shake-audio, SURF, vibro* is 7.6%.

[1]Chance accuracy for 5 categories is 20%

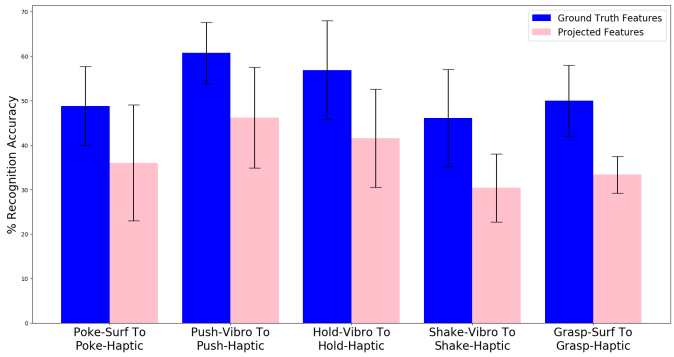

Fig. 1. Accuracies of one-to-one projections with the least Accuracy Delta.

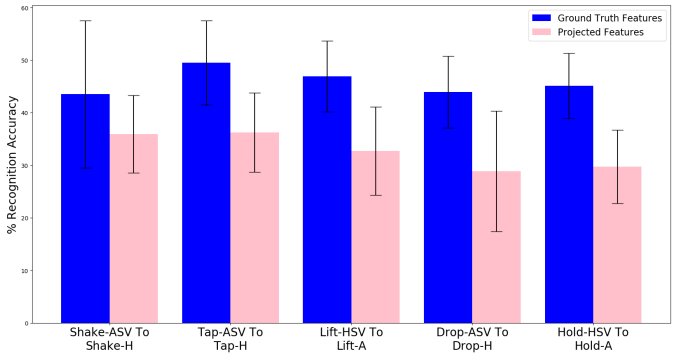

Fig. 2. Accuracies many-to-one projections with the least Accuracy Delta. Here A is Audio, H is Haptic, S is SURF, and V is vibrotactile

3) Accuracy Delta Results: Accuracy Delta results support the accuracy results that recovering haptic features from vibro is the easiest task as it is one of the lowest accuracy delta mappings, and adding audio, SURF and vibro features further improves the recovery of haptic features.

IV. Conclusion and Future Work

In many cases, it is possible that one of the sensors malfunctions, and a module in the system requires its input. We propose a framework for knowledge transfer that uses an encoder-decoder network to recover the features of the failed sensors from the available sensors. The recovered features were used to train a model for object category recognition. We discussed certain input sensors were able to recover features better than others. In future work, we would experiment with different encoder-decoder network architectures to recover more realistic features of the failed sensor. We would start by experimenting with Convolution Neural Networks (CNN), which is good in finding repeating local features in the data, and Recurrent neural network (RNN), which is good at learning temporal correlations from sequence inputs.

References

[1] J. Sinapov, C. Schenck, K. Staley, V. Sukhoy, and A. Stoytchev, "Grounding semantic categories in behavioral interactions: Experiments with 100 objects," *Robotics and Autonomous Systems*, vol. 62, no. 5, pp. 632–645, 2014.