2019 Joint IEEE 9th International Conference on Development
and Learning and Epigenetic Robotics (ICDL-EpiRob)
Oslo, Norway, August 19-22, 2019

Sensorimotor Cross-Behavior Knowledge Transfer for Grounded Category Recognition

Gyan Tatiya, Ramtin Hosseini, Michael C. Hughes, and Jivko Sinapov
Department of Computer Science
Tufts University
{Gyan.Tatiya, Ramtin.Hosseini, Michael.Hughes, Jivko.Sinapov}@tufts.edu

***Abstract*— 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 that cannot always be detected using visual input alone. However, such learned representations are specific to each individual robot and cannot be directly transferred to another robot with different actions, sensors, and morphology. To address this challenge, we propose a framework for knowledge transfer across different behaviors and modalities that enables a source robot to transfer knowledge about objects to a target robot that has never interacted with them. The intuition behind our approach is that if two robots interact with a shared set of objects, the produced sensory data can be used to learn a mapping between the two robots' 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 a target robot to perform category recognition on a set of novel objects and categories without the need to physically interact with the objects to learn the categorization model.**

I. Introduction

From an early stage in development, humans and many other species use exploratory behaviors (e.g., shaking, lifting, pushing) to learn about objects [1]. Such behaviors produce not only visual but also auditory and haptic feedback [2], which is fundamental to grounding the meaning of many nouns and adjectives that cannot be represented using vision alone [3]. For example, to perceive whether an object is full or empty, a human may lift it; to perceive whether it is soft or hard, a human may press it [4]. In a sense, the behavior acts as the question which is subsequently answered by the sensory signal produced during its execution.

Recent advances in robotics have shown that robots too can use such exploratory actions for a variety of tasks, including object recognition [5], category acquisition [6], and language grounding [7]. Despite the significant advancement in interactive and multisensory object perception for robots [8], one challenge is that multisensory representations such as haptic, proprioceptive, auditory, and tactile perceptions cannot be easily transferred from one robot to another, as different robots may have different behaviors, bodies, and sensors. Since each robot has a unique morphology and sensor suite, each individual robot needs to learn its task-specific multisensory models of objects from scratch and cannot use models learned by a robot with different embodiment. Even in the case of two physically identical robots, it is not always possible to transfer multisensory object models as the robots' behaviors may be different.

To address these existing limitations, this paper proposes using an encoder-decoder neural network to project sensorimotor features that the source robot has observed when interacting with an object to a semantically similar feature space that the target robot would observe when it interacts with the same object. For example, if the source robot and the target robot had observations of what the same objects feel like when grasped and shook, the pair of datasets would be used to learn a shared latent space which in turn can be used to generate observations of new objects using the source robot's observations to teach the target robot. This generated feature space can be used to train a task-specific recognition model allowing the target robot to identify objects of novel classes that it has not previously interacted with. The benefit of this approach is that the target robot would not have to learn the recognition task from scratch, but instead could use the generated features obtained from the source robot.

The proposed method is evaluated on a dataset in which a humanoid robot explored a set of 100 objects, corresponding to 20 categories using 9 exploratory behaviors while recording haptic and auditory data. The results show that certain combinations of the sensory modality and the behavior performed by the source and the target robot to learn the encoder-decoder network can generate features that achieve recognition accuracy almost as good as if the target robot learned by actually interacting with the objects.

II. Related Work

A. Object Exploration in Cognitive Science

Cognitive neuroscience shows that it is important for humans to interact with objects in order to learn their tactile, haptic, proprioceptive and auditory properties [1], [4], [9]. Studies show that infants start learning how objects feel, sound, and move at an early stage and this ability becomes more goal-driven as we grow older [10]. Research has also shown that humans are able to integrate multiple sensory modalities to recognize objects and each modality contributes to the final decision [11], [12]. Inspired by these findings, we propose a method of knowledge transfer from the source robot to the target robot to facilitate the learning process

978-1-5386-8128-2/19/$31.00 ©2019 IEEE

of the target robot, as collecting multiple sensory data by interacting with objects is an expensive process.

B. Multisensory Object Perception in Robotics

While most of the object recognition methods in robotics use visual sensing, several research studies have considered multiple sensory modalities coupled with exploratory actions [8]. A number of approaches and feature extraction techniques have been proposed for recognizing objects and their properties using auditory [13], [14], haptic [15], and tactile feedback [16], [17]. Besides recognizing objects, non-visual sensory modalities have also proven useful for learning object categories [18]–[21], object relations [22], and more generally, grounding language that humans use to describe objects [23]. Despite all of these advances, current work in this area is limited by the fact that each new robot is required to learn object models from scratch as different robots have different embodiment and sensors, resulting in excessive time required for individual robots to carry out the necessary object exploration, prohibiting rapid learning. In our work, we propose a method that would enable multisensory object knowledge learned by one robot to be transferred to another, thus reducing time spent on object exploration.

C. Encoder-Decoder Networks

Encoder-decoder networks consist of two feed-forward neural networks: an *encoder* and a *decoder* [24], [25]. The encoder transforms an input feature vector (the sensory input from the source robot) into a fixed-length code vector. The decoder takes a code vector as input and produces a target feature vector as output (e.g. the sensory information for the target robot). Often, encoder-decoder architectures are used for dimensionality reduction by forcing the intermediate code vector to be a much smaller size than either input or output. When input and output vectors are identical, they are referred to as autoencoder networks [26]. When inputs and outputs differ, the more general term "encoder-decoder" applies. Encoder-decoder approaches have enjoyed success in applications such as translating sentences written in two different languages [27] or learning multi-scale features for image representation tasks [28]. We propose using encoder-decoder networks to predict sensorimotor features produced by an interaction with an object by one robot (the target robot) given such features produced by another robot (the source robot). Such an ability enables the target robot to use sensorimotor experience from the source robot and drastically reduce the amount of interaction and data collection needed for learning multisensory recognition models.

III. Learning Methodology

A. Notation and Problem Formulation

For the source robot, let $\mathcal{B}_s$ be the set of exploratory behaviors (e.g. *push*, *drop*), let $\mathcal{M}_s$ be the set of sensory modalities (e.g. *audio*, *haptic*), and let $\mathcal{C}_s$ be the set of sensorimotor contexts such that each context $c_s \in \mathcal{C}_s$ refers to a combination of a behavior $b_s \in \mathcal{B}_s$ and a sensory modality $m_s \in \mathcal{M}_s$ (e.g., each context c_s could be *push-audio*, *drop-haptic*, etc.). Similarly, for the target robot, let $\mathcal{B}_t$ be the set of exploratory behaviors, let $\mathcal{M}_t$ be the set of sensory modalities, and let $\mathcal{C}_t$ be the set of sensorimotor contexts.

For each exploration trial, the source robot and the target robot perform exploratory behaviors $b_s \in \mathcal{B}_s$ and $b_t \in \mathcal{B}_t$, respectively, on a specific object and record a sensory signal for each modality in $\mathcal{M}_s$ and $\mathcal{M}_t$, respectively. Thus, during the i^{th} exploration trial, the source robot observed features $x_i^{c_s} \in \mathbb{R}^{D_{c_s}}$ and the target robot observed features $x_i^{c_t} \in \mathbb{R}^{D_{c_t}}$. Here, D_{c_s} and D_{c_t} are the dimensions of the features observed by the source robot and the target robot, respectively, under contexts c_s and c_t.

We divide our total set of possible object categories $\mathcal{Y}$ into two mutually exclusive subsets: $\mathcal{Y}_{\text{shared}}$ and $\mathcal{Y}_{\text{source-only}}$. Categories in $\mathcal{Y}_{\text{shared}}$ are *shared*; both source and target robots have access to multiple examples from these categories during the exploration or training phase. Categories in $\mathcal{Y}_{\text{source-only}}$ are only experienced by the source robot during the training phase. The goal of our work is to effectively train the *target robot* to recognize an object at test time from one of the categories in $\mathcal{Y}_{\text{source-only}}$, even though it has never experienced any object from these categories before.

B. Knowledge Transfer Model

Our proposed encoder-decoder approach is designed to transfer knowledge from the source robot to the target robot. First, the encoder neural network transforms the observed feature vector of the source robot $x_i^{c_s}$, 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: $z_i = f_\theta(x_i^{c_s})$, which takes network parameter weights θ. Next, a decoder neural network maps an input code vector z_i to create a vector of "reconstructed" target feature vector $\hat{x}_i^{c_t}$. We denote this non-linear mapping by a decoder function g: $\hat{x_i}^{c_t} = g_\phi(z_i)$, which takes network parameter weights ϕ.

Training the encoder-decoder for a context pair c_s, c_t requires observing features from both source and target robot across a set of N total objects. Given a dataset of source-target feature pairs $\{x_i^{c_s}, x_i^{c_t}\}_{i=1}^N$, we wish to find parameters $(\theta,\ \phi)$ that minimize the error between the real features $x_i^{c_t}$ observed by the target robot and the model's "reconstructed" target features $\hat{x}_i^{c_t}$ obtained by applying the encoder-decoder to the corresponding source features $x_i^{c_s}$. We use root mean square error (RMSE) as the error to minimize:

$$\theta^\star, \phi^\star = \arg\min_{\theta,\phi} \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i^{c_t} - g_\phi(\underbrace{f_\theta(x_i^{c_s})}_{z_i}))^2} \quad (1)$$

$\hat{x}_i^{c_t}$

We emphasize that the objects used to train the encoder-decoder come from the set of shared categories $\mathcal{Y}_{\text{shared}}$.

C. Category Recognition Model using Transferred Features

Given a pre-trained encoder-decoder for a source context c_s (e.g. *push-audio* or *drop-haptic*), we can train the

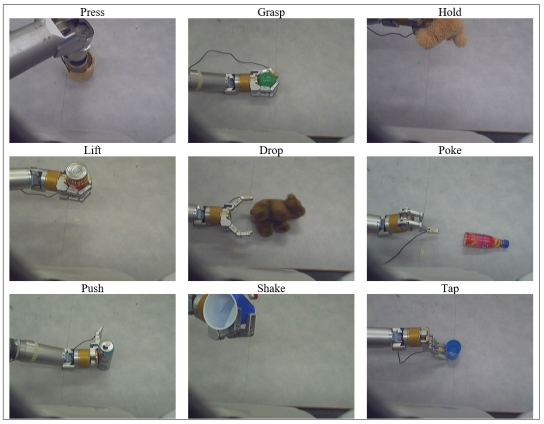

Fig. 1. The exploratory interactions that the robot performed on all objects. From top to bottom and from left to right: (1) *Press*, (2) *Grasp*, (3) *Hold*, (4) *Lift*, (5) *Drop*, (6) *Poke*, (7) *Push*, (8) *Shake* and (9) *Tap*.

target robot to classify objects from several categories it has never experienced before, as long as examples of these categories are seen by the source robot under context c_s. We denote this set of categories $\mathcal{Y}_{\text{source-only}}$. We assume the source robot has seen J total feature-label pairs from these categories: $\{x_j^{c_s}, y_j\}_{j=1}^{J}$, where $y_j \in \mathcal{Y}_{\text{source-only}}$. We can transfer this labeled dataset to the target robot by creating a "reconstructed" training set: $\{g_\phi(f_\theta(x_j^{c_s})), y_j\}_{j=1}^{J}$. This dataset can be used to train a standard multi-class classifier. Then, when the target robot is deployed in an environment with novel objects without category label, the target robot can measure observed features x^{c_t} and feed these features into its pretrained classifier to predict which category within the set $\mathcal{Y}_{\text{source-only}}$ it has observed. Throughout, we will assume that at test time, only categories from $\mathcal{Y}_{\text{source-only}}$ are possible for the target robot to encounter. However, it is straightforward to extend our approach the combined set of possible categories $\mathcal{Y}_{\text{source-only}}$ and $\mathcal{Y}_{\text{shared}}$ by combining a target robot's real and reconstructed training datasets.

IV. Experiments and Results

A. Dataset Description

We used the dataset described in [18], in which an upper-torso humanoid robot used a 7-DOF arm to explore 100 different objects belonging to 20 different categories using 9 behaviors: *Crush*, *Grasp*, *Hold*, *Lift*, *Drop*, *Poke*, *Push*, *Shake* and *Tap* (shown in Fig. 1). During each behavior the robot recorded auditory and haptic feedback using two sensors: 1) an Audio-Technica U853AW cardioid microphone that captures audio sampled at 44.1 KHz, and 2) joint-torque sensors that capture torques from all 7 joints at 500 Hz. Each behavior was performed 5 times with each of the 100 objects, resulting in a total of 9 x 5 x 100 = 4,500 interactions.

We used the auditory and haptic features computed from raw sensory signals as described in [18]. For audio, the discrete Fourier transform was performed using 129 log-spaced frequency bins and a spectro-temporal histogram was computed by discretizing both time and frequencies into 10 equally spaced bins, resulting in a 100-dimensional feature vector. Haptic data was similarly discretized into 10 temporal bins, resulting in a 70-dimensional feature vector (the arm had 7 joints). Fig. 2 shows an example of *audio* and Fig. 3 shows an example of *haptic* features.

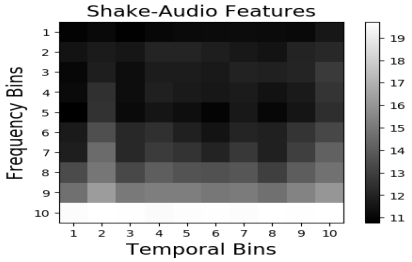

Fig. 2. Example *audio* features using *shake* behavior performed on an object from the *medicine bottles* category.

B. Knowledge Transfer Model Implementation

The encoder-decoder network[1] used consists of a multilayer perceptron (MLP) architecture of three hidden layers for both encoder and decoder, with 1000, 500, 250 hidden units and Exponential Linear Units (ELU) [29] as an activation function, and a 125-dimensional latent code vector as depicted in Fig. 3. The network parameters are initialized randomly and updated for 1000 training epochs using Adam optimization [30] with learning rate 10^{-4}, and was implemented using TensorFlow 1.12 [31].

C. Category Recognition Model Implementation

At test time, we performed classification of objects into categories from the set $\mathcal{Y}_{\text{source-only}}$ via a multi-class Support Vector Machine (SVM) [32]. Using the kernel trick, an SVM maps training examples to an (implicit) high-dimensional feature space where examples from different classes may be closer to linearly separable. We used the Radial Basis Function (RBF) kernel SVM implementation in the open-source scikit-learn package [33], with default hyperparameters. We also tested a k-nearest neighbors classifier (not shown) [34], which performed similarly to the SVM.

D. Evaluation

We assume that the source robot interacts with all 20 object categories, but the target robot interacts with only 15 randomly selected object categories. The objects of the 15 categories shared by both robots are used to train the encoder-decoder network that projects the sensory signal of the source robot to the target robot. Since the dataset we used has only one robot, we assume that the source and the target robots are physically identical, but they perform different behaviors on shared objects.[2] Subsequently, the trained encoder-decoder network is used to generate "reconstructed"

[1]Datasets and source code for study replication is available at: https://github.com/gtatiya/Knowledge-Transfer-in-Robots. The experiment pipeline is visually explained and complete results of SVM and K-NN are available on the GitHub page of the study.

[2]Note that the proposed transfer learning method makes no such assumption and is applicable in situations where the two robots are physically different and/or use different feature representations for a given modality.

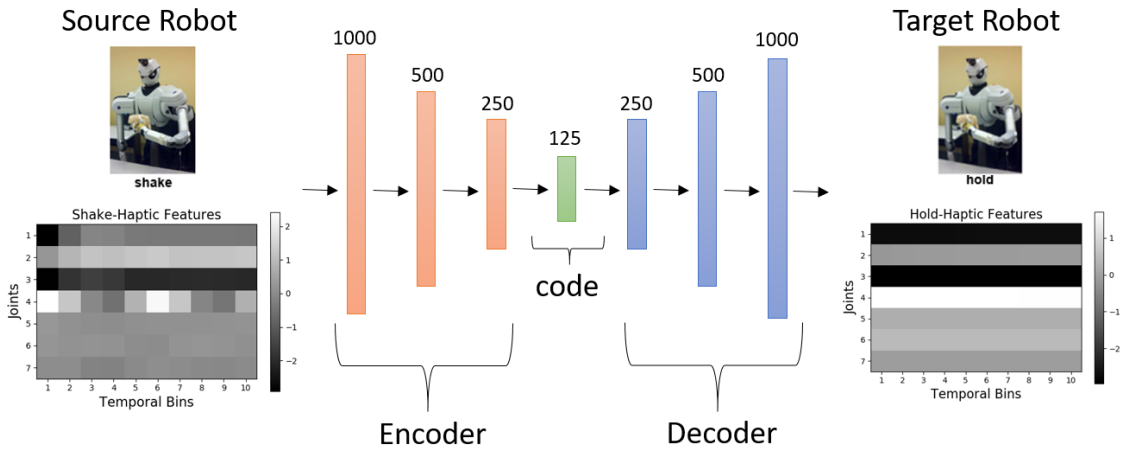

Fig. 3. Encoder-decoder network architecture and an example of a *shake-haptic* to *hold-haptic* projection.

sensory signals for the other 5 object categories in $\mathcal{Y}_{\text{source-only}}$ that the target robot did not interact with. Each sensory signal from objects in these categories experienced by the source robot is thus "transferred" to a target feature vector.

We consider two possible category recognition approaches: our proposed transfer-learning pipeline using the projected data from the source context (i.e., how well it would do if it transferred knowledge from the source robot), and a non-transfer ideal baseline using ground truth features produced by the target robot (i.e., the best the target robot could do if it had explored all the objects itself during the training phase). In both cases, real features observed by the target 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 target robot did not interact with and the test set consisted of the remaining objects. Since the robot explored each object 5 times, there were 100 (4 objects x 5 categories x 5 trials) examples in the training set, and 25 (1 objects x 5 categories x 5 trials) examples in the test set. This procedure was repeated 5 times, such that each object was included 4 times in the training set and once in the test set.

We used two metrics to evaluate the category recognition performance of the target robot on the object categories it did not explore. First, we consider accuracy, defined as $A = \frac{\text{correct predictions}}{\text{total predictions}}$ (often reported as a percentage). The process of selecting 15 categories randomly to train encoder-decoder network, generating the features of the other 5 categories, training two classifiers using projected and ground truth features, and computing accuracy for both classifiers on ground truth features by 5-fold cross validation is repeated 10 times to compute statistics for each projection.

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 for the source robot to project its sensory features in the target robot feature space, and the target robot can use these projected features to learn a classifier that can achieve comparable performance as if the target robot actually explored the objects.

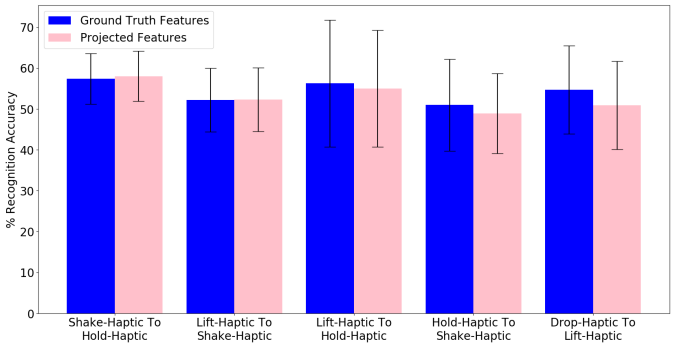

Fig. 4. Projections where the Accuracy Delta (SVM) is minimum.

E. Results

1) Illustrative Example: Consider the case where the source robot performs *shake* behavior and the target robot performs *hold* behavior. Projecting *haptic* features from *shake* to *hold*, enables the target robot to achieve 58% recognition accuracy[3], compared with 57.36% when using features from real interactions (shown in Fig. 4). In other words, the target robot's category recognition model is as good as it would have been had it been trained on real data.

To visualize *shake-haptic* to *hold-haptic* projection, we reduced the dimension of the ground truth and the projected features of the 5 categories the target robot did not interact with into 2D space (shown in Fig. 5) by Principal Component Analysis implemented in scikit-learn [33]. As shown in Fig. 5, the clusters of projected features look very similar to the ground truth features indicating that the "reconstructed" features generated by the source robot are realistic.

[3]Chance accuracy for 5 categories is 20%. Note that accuracy can be boosted to nearly 100% by combining multiple behaviors and sensory modalities [35] but this is out of scope for this paper.

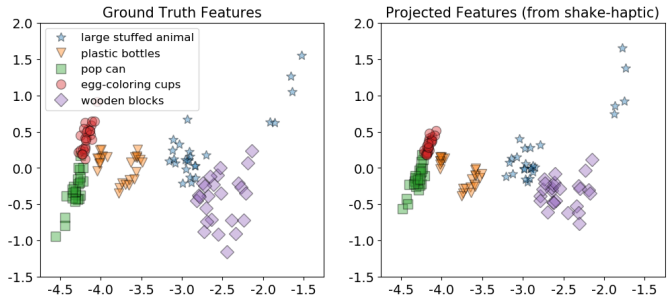

Fig. 5. Target robot's *hold-haptic* ground truth features (left) and the projected features (right) in 2D space using Principal Component Analysis.

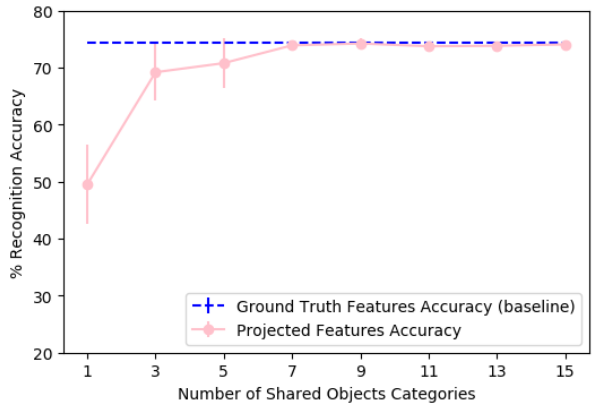

Fig. 6. Accuracy (SVM) achieved by the target robot for different number of shared objects classifier for *shake-haptic* to *hold-haptic* projection.

To find the minimum number of object categories both robots need to interact with to train an encoder-decoder network that achieves good performance, we varied the number of shared categories for *shake-haptic* to *hold-haptic* projection. As shown in Fig. 6, performance saturates at about 7 shared object categories (i.e., using 5 objects per class, the robot needs 35 shared objects out of 100 possible).

2) Accuracy Results of Category Recognition: Since there are 2 modalities (*audio* and *haptic*) there are 4 possible mappings from the source to the target robot: *audio to audio*, *audio* to *haptic*, *haptic* to *audio*, and *haptic* to *haptic*. Each of the 9 behaviors are projected to all of the other 8 behaviors, so for each mapping, there are 72 (9 x 8) projections. Fig. 4 shows the 5 projections where the accuracy delta is minimum among all 288 (4 x 72) projections.

Overall, mappings within same modality (*audio* to *audio* and *haptic* to *haptic*) achieve higher accuracy than mapping to a different modality. This is intuitive, as knowing what an object feels like when performing a behavior can inform what it would feel like better than what it will sound like given another behavior.

3) Accuracy Delta Results: Fig. 7 shows the accuracy delta for all 4 possible modality mappings. Darker color indicates smaller accuracy delta, thus the diagonal is black as there is no accuracy drop when both robots perform the same behavior. Comparatively, *haptic* to *haptic* projections achieve smallest accuracy delta. *Audio* to *audio* is the second best performing mapping, indicating that mappings within the same modality achieve less accuracy delta. Some specific projections that support this observation are shown in Fig 4. However, when both robots perform actions using different modalities, the accuracy delta is relatively higher. For example, *drop-haptic* to *tap-audio* and *hold-haptic* to *tap-audio* are the two projections where the accuracy delta is highest.

When both robots perform behaviors that capture similar object properties, the projected features are more realistic. For example, lifting an object provides a good idea how it would feel to hold that object as indicated by smaller accuracy delta. Producing *hold-audio* features from most of the source robot's features is an easy task, possible because holding an object does not produce much sound.

The relation between the RMSE loss of features used to train the encoder-decoder network and the accuracy delta is shown in Fig. 8 for all of the mappings. RMSE is the Euclidean distance between the ground truth and the projected features. Each dot in the plot corresponds to a projection from the source to the target robot. Generally, the accuracy delta increases with the increase in RMSE loss. This means when the "reconstructed" features are more realistic, the accuracy delta is expected to be smaller, and as the reconstruction gets worse, the accuracy delta increases.

V. Conclusion and Future Work

Non-visual sensory object knowledge is specific to each robot and depends on its unique embodiment, sensors, and actions. We proposed a framework for knowledge transfer that uses an encoder-decoder network to project sensory features from one robot to another robot across different behaviors. The framework enables a target robot to use knowledge from a source robot to classify objects into categories it has never seen before. In this way, the target robot does not have to learn a classifier from scratch, but instead starts immediately with a model nearly as accurate as what can be achieved if the target robot could afford to collect its own labeled training set via exploration. This result addresses some of the biggest challenges in deploying behavior-grounded multi-sensory perception models, namely that they require a lot of interaction data to train and cannot be easily transferred from one robot to another.

In future work, we will test our proposed framework on robots that not only perform different actions, but also are morphologically different and use unique feature representations. Extending the framework to allow for more than two robots to share information is also an outstanding challenge which has the potential to enable any new robot to use multi-sensory knowledge transferred from other robots that had previously interacted with a shared set of objects.

References

[1] T. G. Power, *Play and exploration in children and animals*. Psychology Press, 1999.

[2] L. Shams and A. R. Seitz, "Benefits of multisensory learning," *Trends in cognitive sciences*, vol. 12, no. 11, pp. 411–417, 2008.

[3] D. Lynott and L. Connell, "Modality exclusivity norms for 423 object properties," *Behavior Research Methods*, 2009.

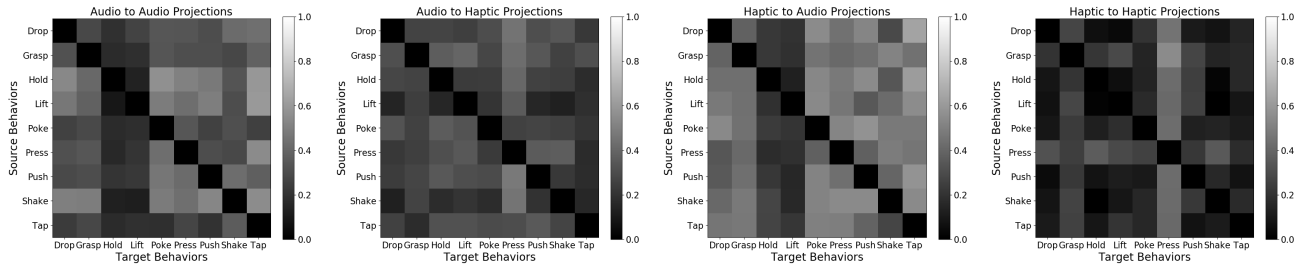

Fig. 7. Accuracy Delta (SVM) for 4 mappings: *audio* to *audio*, *audio* to *haptic*, *haptic* to *audio*, *haptic* to *haptic*. Darker color means smaller Accuracy Delta (better) and lighter color means larger Accuracy Delta (worse).

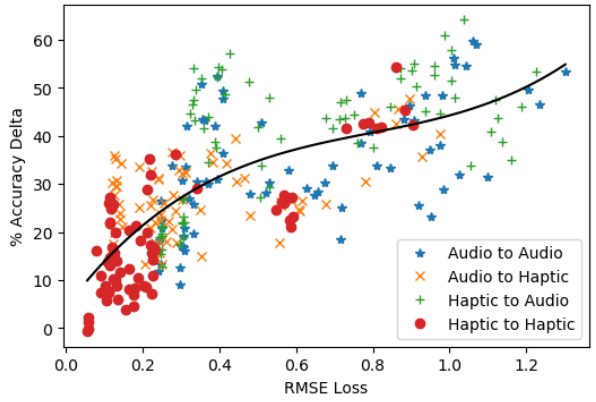

Fig. 8. Relation between RMSE Loss of the features on the training set and Accuracy Delta (SVM) computed using the trained encoder-decoder network. The solid line represents a polynomial with degree 3 that fits all the dots.

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

[5] J. Sinapov, T. Bergquist, C. Schenck, U. Ohiri, S. Griffith, and A. Stoytchev, "Interactive object recognition using proprioceptive and auditory feedback," *The International J. of Robotics Research*, 2011.

[6] T. Araki, T. Nakamura, T. Nagai, K. Funakoshi, M. Nakano, and N. Iwahashi, "Online object categorization using multimodal information autonomously acquired by a mobile robot," *Adv. Robotics*, 2012.

[7] J. Thomason, J. Sinapov, M. Svetlik, P. Stone, and R. J. Mooney, "Learning Multi-Modal Grounded Linguistic Semantics by Playing "I Spy"." in *Proceedings of the Intl. Joint Conf. on AI*, 2016.

[8] J. Bohg, K. Hausman, B. Sankaran, O. Brock, D. Kragic, S. Schaal, and G. S. Sukhatme, "Interactive perception: Leveraging action in perception and perception in action," *IEEE Transactions on Robotics*, vol. 33, no. 6, pp. 1273–1291, 2017.

[9] G. Calvert, C. Spence, B. E. Stein *et al.*, *The handbook of multisensory processes*. MIT press, 2004.

[10] D. M. Stack and M. Tsonis, "Infants haptic perception of texture in the presence and absence of visual cues," *British Journal of Developmental Psychology*, vol. 17, no. 1, pp. 97–110, 1999.

[11] T. Wilcox, R. Woods, C. Chapa, and S. McCurry, "Multisensory exploration and object individuation in infancy." *Dev. Psy.*, 2007.

[12] M. O. Ernst and H. H. Bülthoff, "Merging the senses into a robust percept," *Trends in cognitive sciences*, vol. 8, no. 4, pp. 162–169, 2004.

[13] J. Sinapov, M. Wiemer, and A. Stoytchev, "Interactive learning of the acoustic properties of household objects," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2009.

[14] M. Eppe, M. Kerzel, E. Strahl, and S. Wermter, "Deep neural object analysis by interactive auditory exploration with a humanoid robot," in *IEEE Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2018.

[15] T. Bhattacharjee, J. M. Rehg, and C. C. Kemp, "Haptic classification and recognition of objects using a tactile sensing forearm," in *IEEE RSJ*, 2012.

[16] J. Sinapov, V. Sukhoy, R. Sahai, and A. Stoytchev, "Vibrotactile recognition and categorization of surfaces by a humanoid robot," *IEEE Transactions on Robotics*, vol. 27, no. 3, pp. 488–497, 2011.

[17] J. Lin, R. Calandra, and S. Levine, "Learning to identify object instances by touch: Tactile recognition via multimodal matching," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2019.

[18] 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.

[19] V. Högman, M. Björkman, A. Maki, and D. Kragic, "A sensorimotor learning framework for object categorization," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 8, no. 1, pp. 15–25, 2016.

[20] Z. Erickson, S. Chernova, and C. C. Kemp, "Semi-supervised haptic material recognition for robots using generative adversarial networks," in *Conference on Robot Learning*, 2017.

[21] G. Tatiya and J. Sinapov, "Deep multi-sensory object category recognition using interactive behavioral exploration," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2019.

[22] J. Sinapov, P. Khante, M. Svetlik, and P. Stone, "Learning to order objects using haptic and proprioceptive exploratory behaviors." in *IJCAI*, 2016, pp. 3462–3468.

[23] B. Richardson and K. Kuchenbecker, "Improving haptic adjective recognition with unsupervised feature learning," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2019.

[24] G. E. Hinton and R. S. Zemel, "Autoencoders, minimum description length and helmholtz free energy," in *NIPS*, 1993.

[25] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[26] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, and F. E. Alsaadi, "A survey of deep neural network architectures and their applications," *Neurocomputing*, vol. 234, pp. 11–26, 2017.

[27] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[28] K. Kavukcuoglu, P. Sermanet, Y.-L. Boureau, K. Gregor, M. Mathieu, and Y. L. Cun, "Learning convolutional feature hierarchies for visual recognition," in *Advances in neural info. processing systems*, 2010.

[29] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," in *International Conference on Learning Representations*, 2016.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[31] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean *et al.*, "Tensorflow: A system for large-scale machine learning," in *12th Symposium on Operating Systems Design and Implementation*, 2016.

[32] C. J. Burges, "A tutorial on support vector machines for pattern recognition," *Data mining and knowledge discovery*, 1998.

[33] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, 2011.

[34] D. W. Aha, D. Kibler, and M. K. Albert, "Instance-based learning algorithms," *Machine learning*, vol. 6, no. 1, pp. 37–66, 1991.

[35] J. Sinapov and A. Stoytchev, "The boosting effect of exploratory behaviors." in *AAAI*, 2010.