2019 International Conference on Robotics and Automation (ICRA)
Palais des congres de Montreal, Montreal, Canada, May 20-24, 2019

Deep Multi-Sensory Object Category Recognition Using Interactive Behavioral Exploration

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

***Abstract*— When identifying an object and its properties, humans use features from multiple sensory modalities produced when manipulating the object. Motivated by this cognitive process, we propose a deep learning methodology for object category recognition which uses visual, auditory, and haptic sensory data coupled with exploratory behaviors (e.g., grasping, lifting, pushing, etc.). In our method, as the robot performs an action on an object, it uses a Tensor-Train Gated Recurrent Unit network to process its visual data, and Convolutional Neural Networks to process haptic and auditory data. We propose a novel strategy to train a single neural network that inputs video, audio and haptic data, and demonstrate that its performance is better than separate neural networks for each sensory modality. The proposed method was evaluated on a dataset in which the robot explored 100 different objects, each belonging to one of 20 categories. While the visual information was the dominant modality for most categories, adding the additional haptic and auditory networks further improves the robot's category recognition accuracy. For some of the behaviors, our approach outperforms the previous published baseline for the dataset which used handcrafted features for each modality. We also show that a robot does not need the sensory data from the entire interaction, but instead can make a good prediction early on during behavior execution.**

I. Introduction

Learning to classify objects into categories is an important skill for a wide variety of robot tasks and an open research challenge in the fields of robotics and computer vision. For example, a domestic service robot that has to clean up a dining table needs to identify semantic categories of objects, like "glass", "full", "open", etc. While some categories can be identified using visual input alone, others cannot and thus satisfactory performance in real-world applications remains a challenge [1], [2], [3], [4], [5].

Children learn to discern object categories and recognize objects through physical exploration, where they not only learn what objects look like, but also how they move, feel, and sound [6]. This knowledge is crucial for learning object semantics as the majority of the most common nouns and adjectives humans use have a non-visual component [7]. Yet, most robots today rely on pre-trained computer vision models, e.g., [8], and thus are unable to reason about semantics that cannot be detected using vision alone.

To address these limitations, we propose a deep multi-modal learning methodology that enables a robot to categorize novel objects by performing exploratory interactions and processing multi-sensory data input, shown in Figure1.

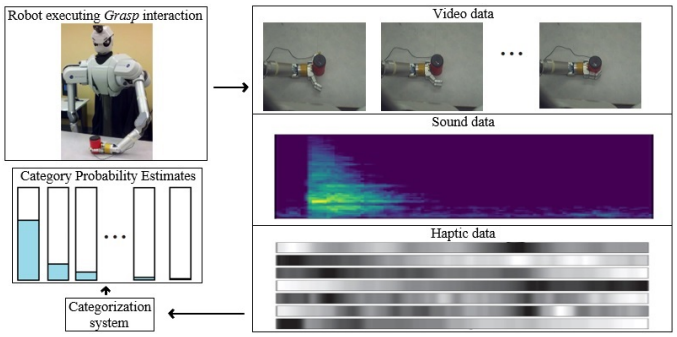

Fig. 1. Overview of the proposed categorization pipeline.

The proposed method is evaluated on a publicly available dataset in which a humanoid robot explored a set of 100 objects using 9 different exploratory behaviors while recording visual, haptic, and auditory data. For all behaviors, the proposed multi-modal network architecture either substantially outperformed the previously published baseline, or produced comparable recognition rates. Furthermore, we demonstrate that our approach can produce accurate category estimates with only a fraction of the data produced by an individual behavior, suggesting that exploratory behaviors can be designed to be shorter in duration, allowing a robot to learn multi-sensory object properties quicker in a deployed, realistic setting.

II. Related Work

Object category acquisition and recognition has been studied extensively in the visual domain, where models can be trained on large image datasets with no need for robotic interaction with objects [1], [2], [3], [4], [5]. For many semantic object categories (e.g., "soft", "empty"), visual information alone may not be sufficient as visually identical objects can differ in material, internal state, and compliance.

To address these cases, several research lines use proprioceptive, haptic, auditory, and/or tactile feedback of robot interaction with objects for category recognition [9], [10], [11], [12]. For example, Nakamura *et al.* in [9] proposed a method that enables the acquisition of object concepts from multiple modalities, such as visual, auditory, and haptic information gathered by robots. Sinapov *et al.* [10] demonstrated a category recognition framework in which

978-1-5386-6026-3/19/$31.00 ©2019 IEEE

the robot uses multiple exploratory actions (e.g., grasping, lifting, shaking, pushing) to learn object category models and categorize 100 objects. More recently, Thomason *et el.* [13], [14], [15] demonstrate how the category recognition method proposed in [11] can be deployed on a service robot to learn object semantics extracted from human-robot dialog. These examples of multi-sensory perception used hand-crafted features for different modalities and require some amount of feature engineering, especially when adding new sensory modalities.

Several works have explored deep learning methods for tasks like surface material classification and tactile understanding using visual and haptic modalities [16], [17], [18]. Erickson *et al.* [16] presented a semi-supervised learning approach for material recognition with Generative Adversarial Networks (GANs) that enables a robot to learn from haptic features such as force, temperature, and vibration data from interactions with everyday objects and classify them into six material categories. Gao *et al.* [17], proposed a deep learning method for tactile understanding using haptic and visual signals. First, individual visual and haptic prediction networks were trained and then they used activations from these networks to train a multimodal network. They demonstrated that combining data from both modalities improves performance. We note that further research work is necessary to use modern learning techniques, which is relatively unexplored in object category recognition. In particular, we present an architecture that uses a larger number of diverse exploratory actions, and consider three types of sensory feedback at the same time: visual, haptic, and auditory.

III. Learning Methodology

For each sensory modality, we investigated several network configurations to find ones that achieve high performance on object categorization tasks using visual, audio, and haptic data in a multimodal setting[1]. Next, we describe these networks along with notation and problem formulation.

A. Notation and Problem Formulation

Let $\mathcal{B}$ be the set of exploratory behaviors, let $\mathcal{O}$ be the set of objects, and let $\mathcal{M} = \{v, a, h\}$ be the set of modalities (vision, audio, and haptics). During each object exploration trial, the robot applies all of its exploratory behaviors on an object $o \in \mathcal{O}$ and records the 3 different sensory data signals for each modality. Thus, during the i^{th} exploration trial, for each behavior $b \in \mathcal{B}$, the robot observed features:

$$\mathrm{X}_i^v \in \mathbb{R}^{w \times h \times t_i^v}, \mathrm{X}_i^a \in \mathbb{R}^{f \times t_i^a}, \mathrm{X}_i^h \in \mathbb{R}^{d \times t_i^h} \qquad (1)$$

where w and h are the width and height of each image, f is number of frequency bins in the sound spectrogram, d is the number of channels (e.g., number of robot joint-torque sensors) in haptic data, and t_i^v, t_i^a, and t_i^h are the number of frames (e.g. number of images) produced over the course of the interaction for each of the three modalities.

Let the function *label* $(o) \rightarrow y$ be a labeling function that given an object o outputs a label $y \in Y$, where Y is the set of category labels. The task of the robot is to learn a category recognition network for each behavior $b \in \mathcal{B}$, that predicts the correct label y, given a sensory signal from modality $m \in \mathcal{M}$ detected while interacting with object o using b. In addition, for each behavior, the robot also learns a multimodal neural network that takes all the modalities of an interaction with an object as input and predicts its category label. Each of the networks estimates a probability for each of the category labels as described below:

$$\begin{aligned} &\Pr(\hat{y} = y | x_i^m), \text{ for a single modality} \\ &\Pr(\hat{y} = y | x_i^v, x_i^a, x_i^h), \text{ for all the modalities} \end{aligned} \qquad (2)$$

B. Visual Network Architecture

1) Image Sequence Pre-processing: For each behavior $b \in \mathcal{B}$, we calculated the average number of image frames per interaction and extracted that many equally-spaced frames from each interaction's image sequence, where each frame was resized to 120 x 90 pixels.[2] For example, the video of a *press* interaction took 48 frames on average for each of 500 trials (100 objects with 5 trials each), so we extracted 48 frames from all the videos of *press* interactions. These pre-processing steps were applied to all the videos of each interaction.

2) Video Network Architecture: Convolutional neural networks (CNNs) have been highly successful in image classification tasks [19], [20], [21] and Recurrent Neural Networks (RNNs) have been shown to perform well in classifying sequential data [22], [23], [24], [25]. Much work uses the combination of a CNN and an RNN by processing each frame using CNN before feeding it to RNN for video classification [26], [27], [28]. This approach turned out to be impractical for our dataset because the combination of a CNN and an RNN makes a network very deep, which requires a large number of examples to learn all the parameters of the network during training; however, our dataset is very small - there are only 20 examples per category as each object was explored 5 times and the model was trained on 4 out of the 5 objects per category.

We used Tensor-Train Gated Recurrent Unit (TT-GRU), a type of RNN, for video classification proposed by Yang *et al.* [29], which has been shown to achieve results very close to the state-of-the-art networks on various video datasets, despite having a very simple architecture. To reduce the number of weight matrix parameters to be learned, TT-GRU factorizes the input-to-hidden weight matrix using Tensor-Train decomposition which is trained with the weights at the

[1]Datasets and source code for study replication is available as Jupyter Notebooks at: `https://github.com/gtatiya/Deep-Multi-Sensory-Object-Categorization`. Development environment and network hyper-parameters details are discussed in the README file of the repository. Some alternative network configurations are also discussed with the source code.

[2]Experimentation with the original image resolution (320 x 240) was also performed, but there was no improvement in accuracy. However, training took a longer time.

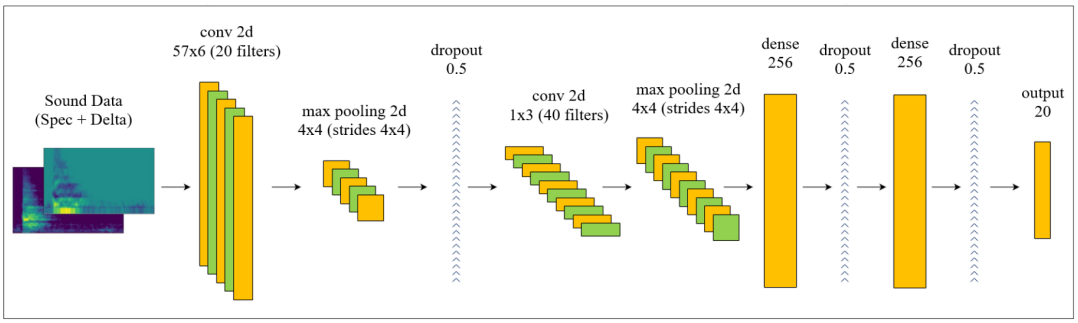

Fig. 2. The architecture of CNN used for sound classification.

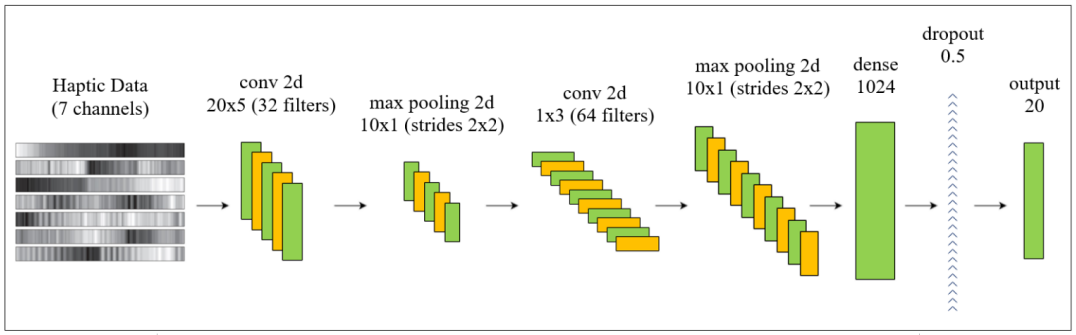

Fig. 3. The architecture of CNN used for haptic classification.

same time. For each frame, a large group of pixel inputs are mapped to the RNN as a latent vector, which is usually lower in dimensionality. This latent vector is then enriched by its predecessor at the last time step recurrently for hidden-to-hidden mapping. In this manner, the RNN is able to learn the inter-frame transition patterns to extract the representation of the entire sequence of frames, and captures the correlation between spatial and temporal patterns because the input-to-hidden and hidden-to-hidden mappings are trained simultaneously. For more details on tensor factorization models and tensor train-decomposition, see [30], [31].

C. Auditory Network Architecture

1) Sound Pre-processing: We used librosa 0.6.0 [32], a python package for music and audio analysis, to generate log-scaled mel-spectrograms of the wave files with FFT window length of 1024, hop length of 512 and 60 mel-bands. In addition to the spectrogram, we computed its derivative as a second channel using the default librosa settings. To get the fixed length input, we interpolated both channels of the spectrogram, so that the rate of the audio frames was consistent with that of the visual frames. Specifically, for each frame in a video, we interpolated 5 frames for the corresponding audio file. For example, the video of a *press* interaction has 48 frames, so we interpolated 240 (48 x 5) frames from its audio data.

2) Sound Network Architecture: While CNNs are largely used on image data, they have also shown strong performance in speech [33], [34] and music analysis [35], [36]. There is abundant research that demonstrates that the ability of finding local features can be successfully applied in sound classification [37], [38], [39]. Therefore, we used CNN[3] for the sound dataset depicted in Figure 2 and described as follows. The CNN consisted of a total of 6 learned layers including 2 convolutional ReLU, 2 max-pooling and 2 fully connected layers. The first convolutional ReLU layer consisted of 20 filters of kernel size 57 x 6 and stride 1 x 1, and max-pooling with a pool shape of 4 x 4 and stride of 4 x 4. The second convolutional ReLU layers consisted of 40 filters of kernel size 1 x 3 and stride 1 x 1, with max-pooling of shape 4 x 4 and 4 x 4 pool stride. Both the first and the second fully connected layer consisted of 256 nodes.

D. Haptic Network Architecture

1) Haptic Pre-processing: In our dataset, the haptic signals from 7 joints were sampled at 500 Hz. To get the fixed size input and to synchronize the haptic signals with video and sound data, we interpolated each haptic feedback to 50Hz[4]. For example, the *press* interaction takes 4.8 seconds, so we interpolated 240 (4.8 x 50) frames for each haptic signal of a *press* interaction.

2) Haptic Network Architecture: Several works in the literature have used CNNs to exploit the haptic signal for material classification [18], [40]. CNN performed very well because haptic feedback is expected to have temporal correlations with repeating local features in a hierarchical order of scales. For this reason, we used a CNN illustrated

[3]Experiments were also performed using an RNN as well as a CNN-RNN combination, but both produced lower accuracy recognition rates.

[4]Experimentation with the original sampling rate (500Hz) was also performed, but there was no improvement in accuracy. However, training took a longer time.

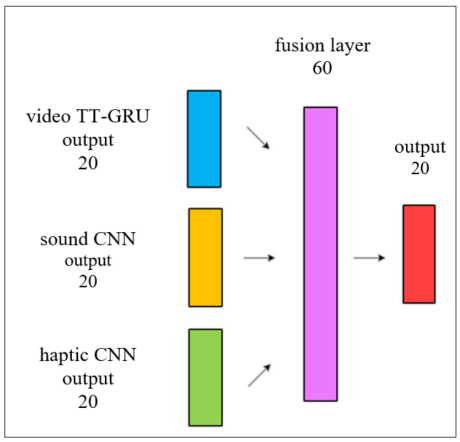

Fig. 4. The architecture multimodal network.

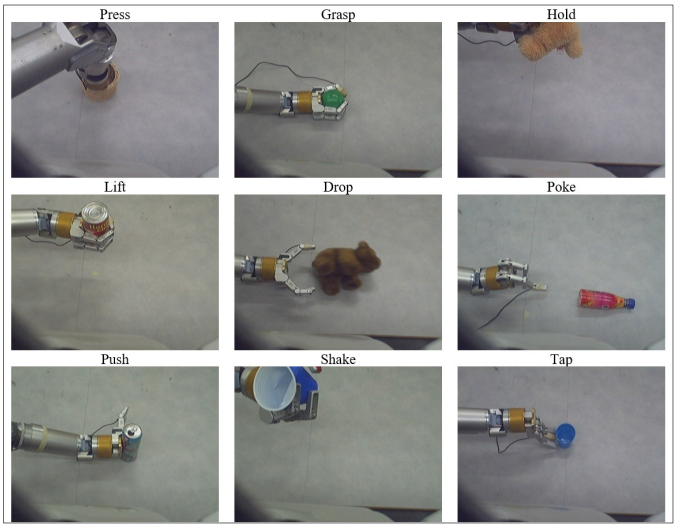

Fig. 5. The exploratory interactions that the robot performed on all objects. From top to bottom and from left to right: (1) Press, (2) Grasp, (3) Hold, (4) Lift, (5) Drop, (6) Poke, (7) Push, (8) Shake and (9) Tap.

in Figure 3 for the haptic data, which consists of 5 layers that includes 2 convolutional ReLU, 2 max-pooling and 1 fully connected layers. The first convolutional ReLU layer's kernel dimensions are 20 x 5 with 32 filters, and the second convolutional ReLU layer has kernel size 1 x 3 and 64 filters. Both first and second max-pooling layers have a pool size of 10 x 1 and stride of 2 x 2. The fully connected layer has 1024 neurons.

E. Multimodal Network Architecture

The multimodal network inputs the same pre-processed video, audio and haptic data as described above. We used the same network architecture for each modality and in addition, added a fusion layer shown in Figure 4. For each modality-specific network, the last layer outputs 20 values for the 20 categories in the dataset. We activated these 20 outputs for each network using ReLU activation and concatenated them to get a layer of 60 neurons. We again activated these 60 neurons using ReLU activation and connected it to a linear layer of 20 outputs for final predictions. ReLU activation function gives a non-linear component to the network and lets the network find useful patterns, while suppressing the irrelevant features. For example, a *hold* interaction does not produce relevant sound, so the network learns to give more importance to vision and haptic feedback than audio. The multimodal network was trained from scratch which produced better results than training the modality-specific networks first, and then only training the fusion layer. We also considered combining the outputs of the modality-specific networks using a uniform combination, but using a fusion layer increased category recognition accuracy.

IV. Evaluation and Results

A. Dataset Description

We used the publicly available dataset of the experiment performed by Sinapov *et al.* [10], in which an upper-torso humanoid robot (shown in Figure 1) explored 100 different household objects belonging to 20 different categories (shown in Figure 6) using 9 exploratory behaviors performed with its left arm: Press, Grasp, Hold, Lift, Drop, Poke, Push, Shake and Tap (shown in Figure 5). During each interaction, the robot recorded visual feedback in the form of RGB images at 10 fps, auditory feedback in the form of a waveform at 44.1 KHz, and haptic feedback consisting of the joint-torque values sampled at 500Hz. Each behavior was performed 5 times on each object, resulting in a total of 9 x 5 x 100 = 4,500 interactions.

Fig. 6. The robot along with the 100 objects, grouped in 20 object categories.

B. Evaluation

We evaluated how well the trained networks perform when recognizing the category of objects that are not found in the training set, via 5-fold object-based cross validation. During each round of evaluation, the training set consisted of the data from 4 objects from each category and the test set consisted of the remaining object for each category. Since the robot explored each object 5 times, there were 400 (80 x 5) examples in the training set, and 100 (20 x 5) examples in the test set. This procedure was repeated 5 times, such that each object was included four times in the training set and once in the test set. We used two metrics to evaluate the category recognition performance. The first metric was

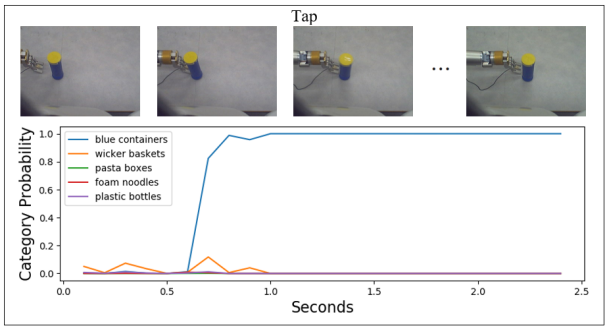

Fig. 7. An illustrative example of the multimodal network category probability estimates as the robot performs the *tap* behavior on one of the blue container objects. The robot's category estimates converges to the correct category after about 0.7 seconds of interaction.

TABLE I

CATEGORY RECOGNITION ACCURACY (%) RATES FOR EACH BEHAVIOR

Behavior	SVM Baseline [10]	Multimodal Network
Grasp	65.2	**71.4**
Hold	67.0	**76.8**
Lift	**79.0**	77.8
Drop	71.0	**78.0**
Poke	**85.4**	73.8
Push	**88.8**	67.4
Shake	76.8	**83.6**
Tap	**82.4**	81.6
Press	**77.4**	58.8

accuracy (%) as defined below:

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

The second metric was the F-score, which is defined as the harmonic mean between the precision and recall for a given category label. The F-score is given by:

$$F = 2 \times \frac{precision \times recall}{precision + recall}.$$

The F-Score is always in the range of 0.0-1.0. For a given category, a high value of the F-Score indicates that the category is easy to recognize, while a low value shows the opposite.

C. Results

1) Illustrative Example: An example of the multimodal network category probability estimates as the robot performs a behavior on an object is shown in Figure 7. The robot's category estimate converges to the correct category after about 0.7 seconds of interaction. The figure plots the estimates for only 5 of the 20 categories to prevent clutter.

2) Accuracy Results of Category Recognition: Table I shows the accuracy for each behavior, compared with the baseline Support Vector Machine (SVM) machine learning approach presented by Sinapov *et al.* [10], which used hand-crafted auditory, haptic features, and visual features (bag-of-word SURF and a histogram of optical flow). In general, the multimodal network yields comparable performance to the baseline (chance accuracy is 5%).

In addition, we tested the accuracy of networks trained on individual sensory modalities as a function of time over the course of each interaction. For example, the *hold* behavior's duration was 1.2 seconds but we hypothesized that the robot would not need all 1.2 seconds of sensory signals to make a good prediction. Figure 8 shows the accuracy curve for every combination of interaction and sensory modality. The results show that for many behaviors, accurate predications can be made without needing to execute the entire behavior. This result is important as behavioral exploration of objects can be costly in terms of time and suggests that in future work, exploratory behaviors can be designed not only to maximize accuracy but also to minimize their duration such that a robot can learn object properties quicker.

3) F-Score Results of Category Recognition: F-scores shown in Figure 9 indicate which modality and behavior work better for each category. For categories in which all the objects have similar shape and color, the visual modality network performs better than the auditory and haptic models. For *hold* and *lift* interactions, the haptic network detects categories better than the sound network. Overall, the results show that different modalities and behaviors are relevant for different categories and suggests that robots need to purposefully select relevant actions when learning new categories.

V. CONCLUSION AND FUTURE WORK

Recognizing the category of objects is an important task for robots operating in human inhabited environments. We proposed deep learning techniques for object categorization using visual, auditory and haptic data acquired through behavioral interactions that a humanoid robot can perform on objects. We demonstrated how the robot learns to detect an object's category using a neural network for each of the sensory modalities individually. In addition, we propose a novel strategy that efficiently combines sensory modalities in a single classifier. Furthermore, unlike previous work, we showed that a robot does not need data from the entire interaction, but instead can make a good prediction early on during behavior execution.

In ongoing and future work, we are investigating the spectrum of early vs. late sensory integration in the context of category learning. In our experiments, we found that adding a fusion module, consisting of one layer, increased performance as compared to training separate modality-specific networks and combining their outputs; yet, it is an open question how deep the fusion module should be to achieve optimal performance. Another open question to be pursued in future work is how to incrementally learn new categories instead of learning all categories at the same time. The ability to acquire new categories on the fly would enable this approach to be used in grounded language learning settings, where a robot in a human inhabited environment encounters new words describing objects over time as it interacts with the people around it. Finally, to test the proposed method in a more complex scenario, we can keep multiple objects in the working space of the robot.

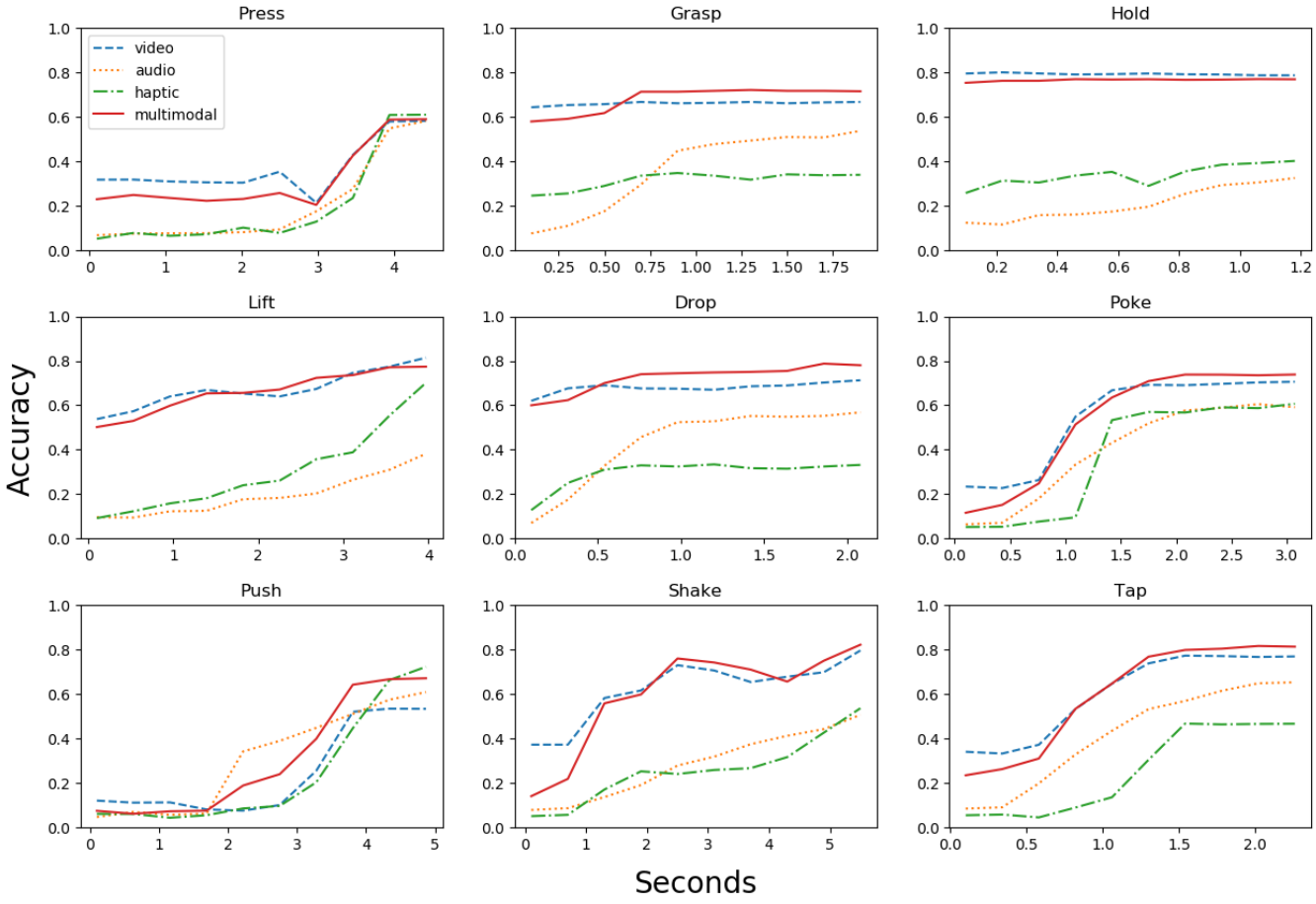

Fig. 8. Accuracy curve for all the interactions and sensory modalities. The x-axis is duration (seconds) and the y-axis is accuracy.

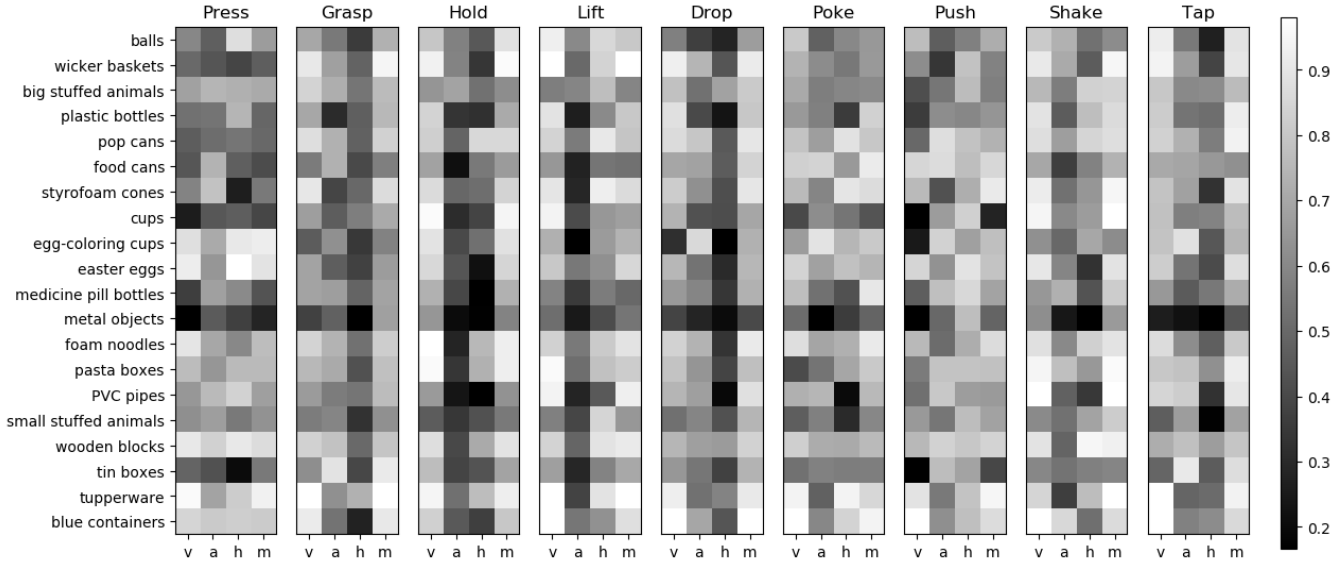

Fig. 9. Recognition F-score for each category behavior, and sensory modality: (v)isual, (a)uditory, (h)aptic and (m)ultimodal.

REFERENCES

[1] J. Zhang, J. Zhang, S. Chen, Y. Hu, and H. Guan, "Constructing dynamic category hierarchies for novel visual category discovery," in *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*. IEEE, 2012, pp. 2122–2127.

[2] L.-J. Li, C. Wang, Y. Lim, D. M. Blei, and L. Fei-Fei, "Building and using a semantivisual image hierarchy," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 3336–3343.

[3] L. Lin, R. Zhang, and X. Duan, "Adaptive scene category discovery with generative learning and compositional sampling," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 2, pp. 251–260, 2015.

[4] H. Zhang, "Building and leveraging category hierarchies for large-scale image classification," Ph.D. dissertation, Carnegie Mellon University Pittsburgh, PA, 2016.

[5] I. Chakraborty, *Object category recognition through visual-semantic context networks*. Rutgers The State University of New Jersey-New Brunswick, 2014.

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

[7] D. Lynott and L. Connell, "Modality exclusivity norms for 423 object properties," *Behavior Research Methods*, vol. 41, no. 2, pp. 558–564, 2009.

[8] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[9] T. Nakamura, T. Araki, T. Nagai, and N. Iwahashi, "Grounding of word meanings in latent dirichlet allocation-based multimodal concepts," *Advanced Robotics*, vol. 25, no. 17, pp. 2189–2206, 2011.

[10] 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.

[11] J. Sinapov, C. Schenck, and A. Stoytchev, "Learning relational object categories using behavioral exploration and multimodal perception," in *Robotics and Automation (ICRA), 2014 IEEE International Conference on*. IEEE, 2014, pp. 5691–5698.

[12] H. Guan and J. Zhang, "Multi-sensory based novel household object categorization system by using interactive behaviours," in *Robotics and Biomimetics (ROBIO), 2016 IEEE International Conference on*. IEEE, 2016, pp. 1685–1690.

[13] 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 International Joint Conference on Artificial Intelligence*, 2016, pp. 3477–3483.

[14] J. Thomason, A. Padmakumar, J. Sinapov, J. Hart, P. Stone, and R. J. Mooney, "Opportunistic active learning for grounding natural language descriptions," in *Proceedings of the 1st Annual Conference on Robot Learning (CoRL-17)*, vol. 78. Proceedings of Machine Learning Research, November 2017, pp. 67–76.

[15] J. Thomason, J. Sinapov, R. Mooney, and P. Stone, "Guiding exploratory behaviors for multi-modal grounding of linguistic descriptions," in *Proceedings of the 32nd Conference on Artificial Intelligence (AAAI-18)*, February 2018.

[16] Z. Erickson, S. Chernova, and C. C. Kemp, "Semi-supervised haptic material recognition for robots using generative adversarial networks," *arXiv preprint arXiv:1707.02796*, 2017.

[17] Y. Gao, L. A. Hendricks, K. J. Kuchenbecker, and T. Darrell, "Deep learning for tactile understanding from visual and haptic data," in *Robotics and Automation (ICRA), 2016 IEEE International Conference on*. IEEE, 2016, pp. 536–543.

[18] H. Zheng, L. Fang, M. Ji, M. Strese, Y. Özer, and E. Steinbach, "Deep learning for surface material classification using haptic and visual information," *IEEE Transactions on Multimedia*, vol. 18, no. 12, pp. 2407–2416, 2016.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[21] A. T. Lopes, E. de Aguiar, A. F. De Souza, and T. Oliveira-Santos, "Facial expression recognition with convolutional neural networks: coping with few data and the training sample order," *Pattern Recognition*, vol. 61, pp. 610–628, 2017.

[22] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[23] 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.

[24] A. Graves, N. Jaitly, and A.-r. Mohamed, "Hybrid speech recognition with deep bidirectional lstm," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, 2013, pp. 273–278.

[25] N. Srivastava, E. Mansimov, and R. Salakhudinov, "Unsupervised learning of video representations using lstms," in *International conference on machine learning*, 2015, pp. 843–852.

[26] Z. Xu, J. Hu, and W. Deng, "Recurrent convolutional neural network for video classification," in *Multimedia and Expo (ICME), 2016 IEEE International Conference on*. IEEE, 2016, pp. 1–6.

[27] A. Montes, A. Salvador, S. Pascual, and X. Giro-i Nieto, "Temporal activity detection in untrimmed videos with recurrent neural networks," *arXiv preprint arXiv:1608.08128*, 2016.

[28] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.

[29] Y. Yang, D. Krompass, and V. Tresp, "Tensor-train recurrent neural networks for video classification," *arXiv preprint arXiv:1707.01786*, 2017.

[30] I. V. Oseledets, "Tensor-train decomposition," *SIAM Journal on Scientific Computing*, vol. 33, no. 5, pp. 2295–2317, 2011.

[31] A. Novikov, D. Podoprikhin, A. Osokin, and D. P. Vetrov, "Tensorizing neural networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 442–450.

[32] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, 2015, pp. 18–25.

[33] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Transactions on audio, speech, and language processing*, vol. 22, no. 10, pp. 1533–1545, 2014.

[34] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, and G. Penn, "Applying convolutional neural networks concepts to hybrid nn-hmm model for speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 4277–4280.

[35] S. Dieleman, P. Brakel, and B. Schrauwen, "Audio-based music classification with a pretrained convolutional network," in *12th International Society for Music Information Retrieval Conference (ISMIR-2011)*. University of Miami, 2011, pp. 669–674.

[36] A. Van den Oord, S. Dieleman, and B. Schrauwen, "Deep content-based music recommendation," in *Advances in neural information processing systems*, 2013, pp. 2643–2651.

[37] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *Machine Learning for Signal Processing (MLSP), 2015 IEEE 25th International Workshop on*. IEEE, 2015, pp. 1–6.

[38] I. Ozer, Z. Ozer, and O. Findik, "Noise robust sound event classification with convolutional neural network," *Neurocomputing*, vol. 272, pp. 505–512, 2018.

[39] H. B. Sailor, D. M. Agrawal, and H. A. Patil, "Unsupervised filterbank learning using convolutional restricted boltzmann machine for environmental sound classification," *Proc. Interspeech 2017*, pp. 3107–3111, 2017.

[40] M. Kerzel, M. Ali, H. G. Ng, and S. Wermter, "Haptic material classification with a multi-channel neural network," in *Neural Networks (IJCNN), 2017 International Joint Conference on*. IEEE, 2017, pp. 439–446.