Learning about Dynamics of Objects

Saket Sadani and Victoria Zhou

University of Texas at Austin
saketsadani@gmail.com
victoria.ch.zhou@gmail.com

Abstract. Advances in the field of cognitive science and developmental behavior has shown that infants typically learn via association and through visual-manual exploration [1]. To mimic this developmental learning in robotic agents, the robot must be able to interact with objects and learn features as a result of its actions. Specifically, we intend to move past solely visual cues in order to discern the characteristics and behavior of objects. In order to achieve this, we will utilize various machine learning algorithms to train the robot to be able to accurately predict future results. The process will consist of three major parts, 1) having our robotic arm manipulate different objects and recording all state information, 2) picking pertinent features to use as a basis for classification, and 3) training based on these features and then predicting the trajectory of objects depending on how the robot interacts with it.

1 Introduction

As humans, after our early ages, we are able to accurately discern what objects are like. This means being able to understand objects based on their shape, size, density and other such features. We can fairly easily predict what will happen to a half-full water bottle if we push it towards the top and contrast that with the result of pushing it at the bottom. With our understanding of the physical world, we know that pushing at the top is likely to make it tip over (because of torque), and pushing it at the bottom will at most make it slide across the surface it's on (or perhaps do nothing at all). Results of this type obviously depend on the action taken, where on the object the action is utilized, and how much force is used to complete the action. Clearly, it is not possible to "hard-code" this knowledge into any physical agent as it is highly-dimensional and not feasible to encode every combination. As such, it is necessary that we explore other avenues in order to implement such understanding.

There are many applications for this kind of understanding. In the general case, knowing how objects behave and react to certain actions will allow a robot to make more complex decisions in a complex environment. The uses for such intelligence are widespread, and we demonstrate this variety by listing some examples of what our robot could potentially do. Consider the case where we want the robot to deliver an object to someone. It may know how to get somewhere,

but it is not entirely useful if the arm cannot safely pick up an object. If we want to reach for something and there is an obstacle in the way, being able to consider the consequences of moving the obstacles without any harm would be a useful skill to have. Furthermore, understanding objects in the surroundings could allow the robot to do neat things like cleaning! The specific applications of understanding object dynamics are seemingly limitless.

2 Background and Related Work

This ability to understand object dynamics is a fairly important aspect of "intelligence" for any physical agent that must interact with its surroundings. To implement this ability in our robots in the Building Wide Intelligence (BWI) Lab, we look towards some studies in human development as well similar robotic experiments. We see from research in cognitive science that human infants typically learn in two ways - by association and by visual-manual exploration [1]. This second method occurs when children play with toys or objects and then gain some understanding of them. Similarly, we hope to train our infant-like robotic arm to manipulate objects, and glean data from said manipulation. If we can store data on the characteristics of the action and the resulting movements, perhaps we can predict what those actions will result in when applied in the future. In essence, can we use the results of previous iterations to guess the trajectory of subsequent ones.

Parts of our process are similar to the work seen in [2]. This experiment attempted to classify a large group of objects into several categories based on multiple features. Sinapov et al. grouped 100 objects into 20 groups and further clustered the groups into a hierarchy as seen in figure 1. The features Sinapov et al. extracted were from visual, proprioceptive, and audio data. Proprioception takes into account the nearby parts of the body (in our case this will be the arm) as well as the relative motions of these parts. This paper makes the point that multiple sensory modalities can provide different different viewpoints and insights on certain events. As such, we hope to use both visual and proprioceptive data to understand the dynamics of the objects we are studying. However, we will likely be using a smaller subset of items than those used in this work.

For the complete proposed work-flow to achieve this goal, see the Technical Approach section.

3 Problem Formulation and Technical Approach

Intelligent physical agents (robots) should be able to understand objects. However, currently the agents do not yet have an understanding of the dynamics of objects, i.e. how would an object react if action A was acted upon the object. If we can teach our robots to play with objects and then learn from them that, the way infants do, then we could make our robots more "intelligent". The robot

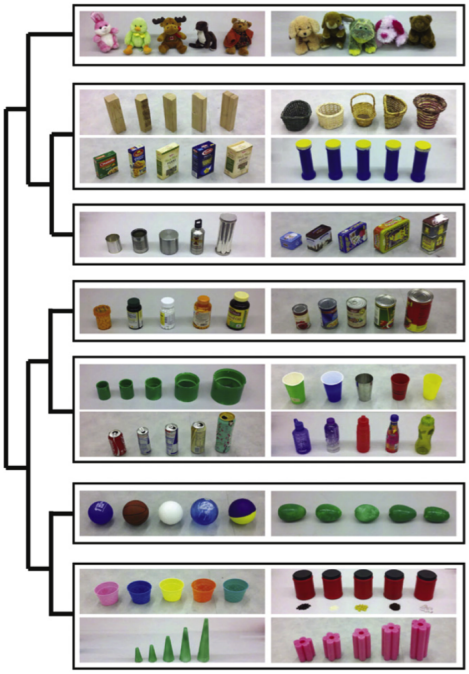

Fig. 1. A hierarchical clustering of 20 categories based on the confusion matrix encoding how often each pair of categories is confused by the robot's context-specific category recognition models [2].

we will be using to train and collect data from is the Kinova robot arm in the Building Wide Intelligence (BWI) Lab. The arm sits on top of a Segway base. The arm itself has six degrees of freedom and has 2 gripper fingers.

To solve this problem, we plan to use three steps: collecting data, extracting features, and predicting results from machine learning algorithms.

3.1 Collecting Data

We plan on training the robot with three basic motions: push (across the table), drop, and squeeze (push from the top of an object). For each of the three basic actions we want to train the physical agent on, we want to add parameters to the actions. For example, the angle/ orientation at which the object is originally placed test each motion at different angles, the position on the object where the motion is performed (i.e. the top vs the bottom of the object), the force/ velocity that is applied to the object, etc.

In Jivko's paper, 'Grounding semantic categories in behavioral interactions: Experiments with 100 objects', Jivko et al, trained the robot on a large data

set (100 objects) with a few data points/ behaviors. However, in this project we want to focus on learning the dynamics of objects and will instead be training the arm robot on a smaller set of objects (10-20 objects) but with many variations of actions/ behaviors (dozens of actions). In addition, the original location of the object will be kept the same to remove any noise from the background.

The raw data that we will collect will include both visual and proprioceptive data. Visual data will include the color/ RGB frames, SURF to extract the main feature points, and optical flow of before, during, and after the action is performed. The proprioceptive data collected will detail the force each joint applies during the action. In addition, the position and orientation of the arm and object before and after the action is performed, and the qualities of the action performed (the parameters on actions detailed above).

3.2 Extracting Features

Once all of the raw data is collected, we will want to extract the important features from the raw signals. The main questions we want to answer from the data collected are

1. Is there any change to the position or orientation of the object?
2. How far away did the object end up?
3. How long did the object take to get to it's final position?
4. What path/ trajectory did the object take?

These questions can be answered by getting the important information/ features from SURF and optical flow.

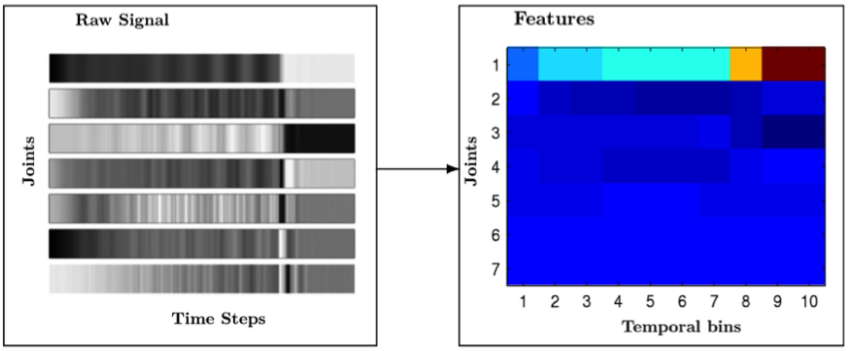

Fig. 2. Illustration of the proprioceptive feature extraction routine. The input signal is sampled during the execution of a behavior at 500Hz and consists of the raw torque values for each of the robot's seven joints. Features are extracted by discretizing time (horizontal axis) into 10 temporal bins, resulting in a 7 x 10 = 70 dimensional feature vector. [2].

The proprioceptive data collected on the forces used by each of the joints (sampled at 500 Hz) will provide many data points. Obviously, this will provide too many data points to feasibly feed and train in a machine learning model. To solve this issues and extract the main features, the time (horizontal axis) will be discretized into ten temporal bins, as shown in Figure 2. A similar technique is used to extract the important auditory signals as shown in Figure 3.

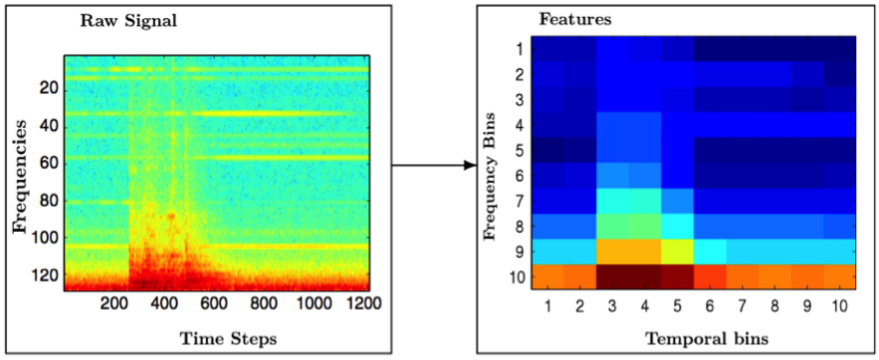

Fig. 3. Illustration of the auditory feature extraction procedure. The input consists of the discete Fourier transform spectrogram of the audio wave recorded while a behavior is executed. The spectrogram encodes the intensity of 129 frequency bins and was calculated using a raised cosine window of 25.625 ms computed every 10.0 ms. To reduce the dimensionality of the signal both the time and the frequencies were discretized into 10 bins, resulting in a 10 x 10 = 100 dimensional feature vector. [2]

3.3 Machine Learning Algorithm and Predicting Results of Actions

Once the input and output signals have been collected, they will be fed into a machine learning models to train the physical agent to be able to predict the effects of the actions it performs on various objects, i.e. be able to predict the trajectory the object will take as the result of a push at certain angle and velocity. For example, Sinapov showed that it was possible to train the physical agent to predict and classify possible trajectories and object will take, as shown in Figure 4 [3].

There are several possible machine learning models that could be used to train the robot. Some examples include category recognition model, k-nearest neighbor model, decision tree learning, reinforcement learning, deep learning, and neural networks. WEKA (Waikato Environment for Knowledge Analysis) may be used as a simple and efficient way to test and use multiple machine learning models. As we further know our specific inputs and outputs, we will be able to determine which machine learning will be best to use.

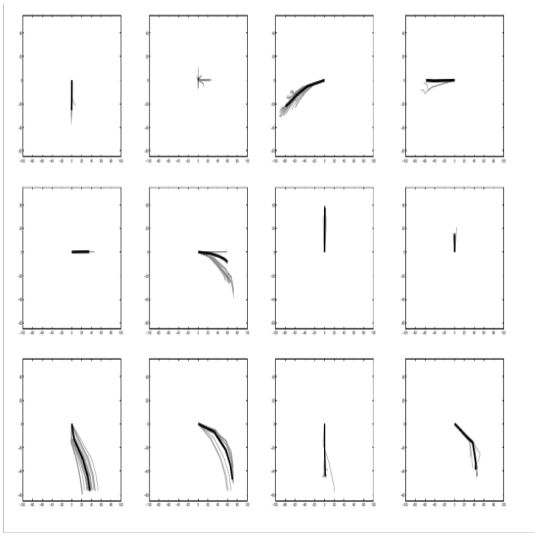

Fig. 4. All twelve leaf outcome classes of the learned taxonomy for the L-Stick tool. The dark trajectory shows the outcome prototype for each leaf class in the learned taxonomy, while the lighter trajectories visualize the observed outcomes that fall within v_j [3].

4 Metric for Success and Evaluation

We will measure our success in three main ways: Accuracy of prediction from the machine learning model, ability to classify objects, and ability of judging behavior of new objects it has not seen before.

4.1 Machine Learning Prediction

We will want to see how accurately the robot can predict the trajectory objects will take. For example, we will compare the predicted trajectory (as mentioned above) with the actual trajectory.

4.2 Classification

We want to see how accurately the robot can classify objects into separate categories (i.e. empty box, full box, empty bottle, etc.). This can be calculated from

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

.

4.3 Novel Objects

Last, but not least, we will want to see the the trained robot will be able to predict the trajectories of objects and classify new objects it has not yet seen or been trained on yet.

5 Expected Contribution and Extensions

As mentioned above, we hope to be able to predict the trajectories of objects given a motion/action that the robot performs. That, in essence, is the overarching goal of this project. The applications of this feature will follow after. Therefore, if we can provide some basis for learning about the dynamics of objects, we could apply this to the many possibilities enumerated in the Applications section, as well as other extensions. For example, we would like to be able to generalize our understanding to more motions, not just the ones we trained and tested on. Similarly, removing the dependency of the type of object would also be a logical next step as we want to be able to make estimates about all objects, not just a tiny subset. Lastly, we could demonstrate true understating if we were able to make observations in the reverse direction. For example, if we are able to predict that a bottle of a certain weight will roll a certain way, it would be impressive to see a bottle rolling and classify its characteristics. The scope is this area of inquiry is fairly large, and we hope to explore as much of it as possible.

6 Acknowledgments

Thank you Jivko for giving us advice and answering all of our questions.

References

1. Johnson, S. P. (2010). How infants learn about the visual world. Cognitive Science, 34(7), 1158-1184.
2. Sinapov, J., Schenck, C., Staley, K., Sukhoy, V., & Stoytchev, A. (2014). Grounding semantic categories in behavioral interactions: Experiments with 100 objects. Robotics and Autonomous Systems, 62(5), 632-645.
3. Sinapov, J., & Stoytchev, A. (2008, August). Detecting the functional similarities between tools using a hierarchical representation of outcomes. In 2008 7th IEEE International Conference on Development and Learning (pp. 91-96). IEEE.