Learning Labeled Robot Affordance Models by using Simulations and Crowdsourcing

Adam Allevato  
Dept. of Mech. Engr.  
The University of Texas at Austin  
Austin, TX, USA  
allevato@utexas.edu

Mitch Pryor  
Dept. of Mech. Engr.  
The University of Texas at Austin  
Austin, TX, USA

Elaine Schaertl Short  
Department of Computer Science  
Tufts University  
Medford, MA, USA

Andrea L. Thomaz  
Dept. of Elec. Engr.  
The University of Texas at Austin  
Austin, TX, USA

Abstract—Affordance models are widely used in robotics to represent a robot’s possible interactions with its environment. However, robot affordance models are inherently quantitative, making them difficult for humans to understand and interact with. To address this problem, previous works have studied grounding symbols (such as natural language) to affordance models, but have only used a limited set of expert-defined actions and effects to do so. In this paper, we use human input and randomized robot actions to construct a labeled affordance model that allows grounding simple English-language labels to robots’ internal affordance representations. We first collect label data from a combination of crowdsourced real-world human-robot interactions and online user studies. We then use this data to train classifiers predicting whether or not a particular quantitative affordance will receive a specific label from a person, achieving an average affordance prediction score of 0.87 (area under Receiver Operating Characteristic curve). Our results also show that labels are better grounded to affordance effects than affordance actions—a result that has been hypothesized in prior work but has never been directly tested. Finally, we develop a technique for automatically constructing a hierarchy of labels from crowdsourced data, discovering structure within the learned labels and suggesting the existence of a more universal affordance grounding.

I. INTRODUCTION

An intelligent robot deployed in the real world must be able to accept and understand commands from human counterparts so that it can adapt to the dynamic environments and requirements it may face. However, a robot’s internal skill representations may be disconnected from how a human would represent or describe the same skills, creating a significant communication barrier. One tool that researchers have used to break down this barrier is the affordance model [32], which represents the interactions between an agent and its environment as consisting of objects, actions[1] and the effects of those actions. The affordances available to a robot usually correspond to human-understandable concepts, such as “openable” drawers or “pushable” buttons, which raises the question of symbol grounding [11]: how should labels be assigned to affordances? Ideally, a robot’s set of symbols (i.e. affordance labels) should be learned using data from non-expert humans to build a more general in a scalable and general way [41]. However, in both affordance modeling [44, 16, 4] and symbol grounding [42, 19], prior work has focused on a fixed set of environment actions or labels, with one or both sets being chosen by a system designer and labels often being chosen without justification.

The core contribution of this paper is a data-driven method for grounding human-provided, natural-language labels to robot manipulation actions and their effects by creating labeled affordance models. In contrast to prior affordance learning and grounding works, we do not design complex actions or effect features, instead using random linear robot motions and measuring the rigid body transformation (i.e. change in pose)

---

[1]Actions are also called behaviors in some literature.
of the affordance object to reduce the amount of bias instilled by the system designer. By learning from these features along with crowdsourced label data from non-expert users (Fig. 1), we directly learn a human affordance grounding.

Our method provides insights into human perceptions of affordances, showing that humans perceive affordances primarily as a function of the effect an action has on a target object, rather than the action taken by the robot. The labels are consistent across different types of objects, pointing to a common human understanding of rigid body motions. Our collected dataset also allows us to automatically discover equivalence classes and hierarchies within the collected labels, adding additional layers of semantic meaning to the learned affordance model and providing insight into how labels are perceived by humans. These insights pave the way for a more bottom-up grounding of affordances and robot actions that will lead to more fluent and explainable human-robot interactions.

II. RELATED WORK

A. Affordances

The term affordance was originally introduced by Gibson [9] to refer to the inherent ways that a creature could perceive to interact with its environment, and based on this definition, a large body of work has studied detecting objects’ affordances directly from sensor input [25, 5, 23, 26, 35, 7, 33, 24, 31]. Affordances were adapted for robotics research by explicitly modeling them as interactions between objects, actions, and effects [32, 18]. This formulation allows a robot to more easily reason about the consequences of executing affordance actions, and has resulted in a broad set of statistical modeling techniques to learn affordance models. These techniques include Bayesian networks (BayesNets) [17, 39, 37, 38, 13], graph representations [43], Support Vector Machines (SVMs) [45, 10, 46], and conditional random fields [16, 21]. For more examples of affordance research, we refer the reader to recent surveys [50, 12, 20].

In most affordance research, the robot is limited to a small, fixed set of actions [18, 17] or a generalized action with only a small number of parameters [47]. The effect features are generally restricted to a fixed set [8], the change in visual features [17, 45, 46], or object motion [22]. In these works, features are either named by the system designer or are unnamed. Uğur et al. [47] and Allevato et al. [2] are two works that use more generalized affordance features, but these works still rely on an expert to assign names to affordances.

B. Symbol Grounding for Affordance Models

Symbol grounding [11], which seeks to assign symbols or words to their physical meanings, is a psychology problem that is well-studied in robotics [6, 41, 28]. In this work, we consider the subproblem of grounding natural languages to affordances and manipulation contexts. Several works [30, 42, 10, 21, 35, 48] ground natural language instructions to a robot’s actions or objects in its environment, usually using a small and fixed set of actions or objects and focusing on the language modeling portion of the symbol grounding problem. Another set of works more similar to ours learn “bottom-up” rather than “top-down” affordances in an unsupervised fashion before grounding them to natural language [27, 46, 15]. Yürüten et al. [49] and Krunic et al. [17] learn a mapping from affordances to labels using visual features and BayesNets. However, they consider small sets of affordance actions chosen beforehand by the system designers and labels provided by expert annotators. Kalkan et al. [14] develop a system to ground labels to affordances and predict the effect label for a new object given the action (from a small fixed set) that will be performed on it. They assert that verbs describe only the effects of an affordance, not the action or object, but do not perform any experiments to verify this hypothesis.

C. Crowdsourcing

Crowdsourcing has been used successfully for notable robot affordance learning studies. Tellex et al. [42] use crowdsourced natural language descriptions of simulated robot action sequences to build a grounding graph and enable a robot to follow new instructions. Sung et al. [40] crowdsourced multimodal affordance demonstrations using real robot data, enabling a robot to execute affordances on novel objects. These works suggest that the “wisdom of the crowd” can be used to learn generalized groundings and affordances.

III. EXTENDING THE AFFORDANCE MODEL

This section reviews the concept of an affordance model as it is commonly used in robotics. It then introduces the notion of labels and how they can be used to augment the basic model.

A. Affordances

The affordance model “define[s] the relation between an agent and its environment through its motor and sensing capabilities” [18]. Formally, affordances are defined as the relations between the feature spaces of possible objects (∊), actions (_errno), and effects (_errno) (see Fig. 2). A triplet of features drawn from these spaces (errno,errno,errno),errno∈∐,errno∈_errno,errno∈_errno represents a single affordance. This is the same terminology used in [18] and [2] and we use it throughout the paper. One of the key strengths of the affordance framework is its simple relational structure, which allows knowledge about objects, actions, or effects to be shared across different affordances.

B. Labeled Affordance Models

The affordance representation is useful for robots to encode knowledge and plan actions, but it is not as directly useful to humans. Actions and effects, in particular, are almost always stored as software-friendly representations, such as a list of continuously-valued robot joint positions or visual features. These numbers will hold little to no semantic meaning for a human attempting to understand the model or interact with a robot that is using the model.

For many objects, such as those encountered in a household or tabletop scenario, we posit that the set of words that carry strong semantic meaning for humans is relatively constant across objects. The words tip, flip, slide, and turn describe
affordance label to a specific affordance, understandable Labels (O) of objects (A), and effects (E). (b) our representation of the model in this study, including human-understandable Labels (L) that describe the affordance.

rigid-body transformations, although it is not immediately clear how concise these labels are (i.e. move vs knock over). Each individual label may correspond to an unknown volume in one of the affordance feature spaces, and the volumes for different labels may overlap.

We formally introduce affordance labels as a fourth feature space, \( \ell \in L \), using a definition similar to \[18\] and \[17\]: a short, descriptive natural-language phrase that is associated with at least one affordance. The use of quantitative object, action, and effect spaces allows simple mathematical analysis and comparison, while the qualitative labels provide a framework for human understanding and natural language grounding.

A labeling function, \( f \), determines the applicability of an affordance label to a specific affordance, \( f: O \times A \times E \times L \to \mathbb{R} \). This can be used to estimate the probability that a particular label will apply to an affordance, or different labels can be compared to develop a single label \( \ell_{best} \) that is likely to be the best for a given affordance:

\[
f(o, a, e, \ell) = P(\ell | o, a, e) \tag{1}
\]

\[
\ell_{best}(o, a, e) = \arg \max_{\ell} P(\ell | o, a, e) \tag{2}
\]

In this work, we test the predictive power of both actions and effects for determining the human-provided label. However, we still wish to uncover affordances that are applicable across multiple objects, so we learn jointly over all objects, making our predictor independent of object class. Eqs. \[1\] and \[2\] then become:

\[
f(o, a, e, \ell) = P(\ell | a, e) \tag{3}
\]

\[
\ell_{best}(o, a, e) = \arg \max_{\ell} P(\ell | a, e) \tag{4}
\]

When creating a model of affordance labels from freeform crowdsourced data we are faced with the issue that the label data from humans can be noisy because of differences in perception, vocabulary, and the ambiguity of natural language. In practice, we resolve this issue by learning a set of commonly-used affordance labels from data.

C. Collecting Affordance Data

Our procedure for data collection is shown in Algorithm 1. The first step is sampling to create the dataset \( D = \{o, a, e, \ell\}_n \), \( n = 1 \ldots N \) triplets. This requires choosing \( O, A, \) and \( E \) such that the resulting data is useful for learning and executing affordances.

Even once the feature spaces are chosen, it is not immediately clear how to generate affordance samples that provide informative training data. As the number of exploration dimensions grows, we face an exponential increase in the sampling space. Prior work in affordances has used a grid search over the action space to find self-exploration actions, used human demonstrations to guide exploration in the action space, or selected from a pre-determined set of actions. However, grid search or a fixed set of actions could impose unwanted structure on the data, potentially leading to lower-accuracy models that do not exhibit all types of effects. Therefore, we sample actions from a uniform random distribution over the action space. A fully random sampling approach avoids instilling biases about which affordances are “interesting,” providing a more general set of data to measure human perception of affordances.

We also must take into account the workspace of our manipulator. Therefore, we select a working volume from our action space, \( V \in A \), and uniformly sample actions \( a_1, a_2, \ldots, a_N \sim U(V) \). Each action is performed on all objects \( o_1, o_2, \ldots, o_M \in O \) in a series of episodes, where each episode includes one action on a single object. After each episode, we store the resulting effect, completing the object-action-effect triplet \( (o, a, e_i) \).

D. Labeling the Model

The next step is to complete the dataset by collecting information from a human labeler (the “Survey” step in Algorithm 1). We select an affordance from \( D \), display the affordance to a human survey respondent, and ask them to provide the single labels. Since the mapping from effects to labels is not necessarily one-to-one, we request the set of all

**Algorithm 1** The two main steps of building a labeled affordance dataset.

1. function SAMPLE(O, A, N) ▷ Section III-C
2. \( D \leftarrow \emptyset \)
3. for all \( o \in O \) do
4. for \( i = 1 \) to \( N \) do
5. \( a_i \sim U(V \subset A) \)
6. \( D_i \leftarrow (o, a_i, \text{SIMULATE}(o, a_i)) \)
7. return \( D \)
8. function SURVEY(D, M) ▷ Section III-D
9. for \( d = (o, a, e) \in D \) do
10. SETUPSIMULATOR(o)
11. PERFORMACTION(a)
12. \( \ell, \ell^* \leftarrow \text{COLLECTLABEL()} \) ▷ User input
13. \( d \leftarrow (o, a, e, \ell, \ell^*) \) ▷ Update datapoint
labels that are appropriate and store it as the set \( \ell \). We also request the single “best” or most appropriate label, which is added to the affordance tuple as \( \ell' \).

Once we have collected label information, we train classifiers to predict affordance labels. These labels and their associated classifiers define the labeling function, and therefore can perform label prediction. We also note that by inverting the labeling function, the model could also be used for action selection, finding an action to perform given an input object and target label.

As discussed above, the final dataset \( \mathcal{D} \) may include too many unique labels for effective learning, so it is helpful to collapse the possibilities into a small subset of the most popular provided labels, which we denote by \( \mathcal{L}' \). Using a discrete-valued label space allows us to learn a single multiclass label classifier over actions or effects, and simplifies the problem of training binary classifiers for the applicability of each individual affordance label. This simplifies the natural language processing problem but maintains a data-driven approach to develop a set of labels, instead of relying on experts.

E. Uncovering Label Relations

The labeled affordance models created by our approach can be used for supervised learning, but we also wish to understand the physical meaning of different affordance labels to to uncover semantic links between them, which we propose can be done in an unsupervised fashion. Many of the terms we are interested in, such as push, have multiple meanings in English, and we are only interested in meanings related to manipulation. Therefore, lexical relationships from traditional databases such as GloVe \cite{pennington2014glove} or ConceptNet \cite{speer2017conceptnet} do not make sense for this task.

We propose an approach to constructing \emph{ad hoc} relationships within a set of labels. Taking each pairwise combination of valid labels \((\ell_a, \ell_b), \ell_a \neq \ell_b\), we calculate the probability of one label conditioned on the other (conditional probability or CP), over the discrete dataset using Eq. \ref{eq:cp} (\& and \& are logical AND and OR, respectively; recall that \( \ell \) is the set of all valid labels for a given datapoint).

\[
\text{CP}(\ell_a, \ell_b) = \frac{p(\ell_a \cap \ell_b)}{p(\ell_b)} \approx \frac{\sum_{a, a', c, \ell, \ell' \in \mathcal{D}} 1(\ell_a \in \ell \land \ell_b \in \ell)}{\sum_{a, a', c, \ell, \ell' \in \mathcal{D}} 1(\ell_b \in \ell)}
\]

The CP scores can be used to generate a label hierarchy, connecting labels via equality and parent-child relationships. To do this, we select a threshold \( t \in [0, 1] \) and apply the following rules to determine the relationship between the labels:

\[
\begin{cases}
\ell_a \equiv \ell_b & \text{CP}(\ell_a, \ell_b) > t, \text{CP}(\ell_b, \ell_a) > t \\
\ell_b = \text{child}(\ell_a) & \text{CP}(\ell_a, \ell_b) > t, \text{CP}(\ell_b, \ell_a) \leq t \\
\ell_a \neq \ell_b & \text{otherwise}
\end{cases}
\]

These rules can generate more fine-grained or broader label hierarchies depending on the chosen threshold \( t \).

IV. EXPERIMENTAL SETUP

To validate our approach we conducted two user studies, collecting datasets of labeled affordance models using in-the-wild interaction with humans and online crowdsourcing of simulated videos.

A. Objects, Actions, and Effects

To build a dataset as described in Section \ref{sec:method} we selected general feature spaces for objects, actions, and effects A 9-object subset (Fig. \ref{fig:ycb_objects}) was selected from the standardized YCB dataset presented in \cite{Silva2018}.

We sampled each action as a straight-line motion in Cartesian space where the orientation of the end effector is held constant. In the rest of the paper, this general action will be referred to as “push”, although the actual label of the affordance and associated action depends on the parameterization, target object, and ultimately the observer. Pushes were parameterized by the starting and ending \((x, y, z)\) position of the end effector in a reference frame centered on the object being explored. This resulted in a 6-dimensional action parameterization: \( a_i = (x_0, y_0, z_0, x_f, y_f, z_f) \in \mathcal{A} \). We define the sampling volume \( V \in \mathcal{A} \) by cubic regions of size \( \alpha_x \times \alpha_y \times \alpha_z \), with the start state region centered on \((-0.05m, 0, \alpha_z/2)\) and the end state region centered on \((0.05m, 0, \alpha_z/2)\). This choice of start and end volumes ensures that the robot will not begin touching the object, but will often touch the object by the end of the action. We also let \( \alpha = \alpha_y = \alpha_z = 0.2m \), dictated by the workspace of the robot arm. As discussed above, actions were sampled randomly from this space to avoid further biasing the actions to work better for specific objects in the dataset.

Effects \( \mathcal{E} \) encode \( \text{SE}(3) \) transformations, representing the change in position and rotation of the object’s centroid relative to its starting position as a translation \((\Delta x, \Delta y, \Delta z)\) and a quaternion \((q_x, q_y, q_z, q_w)\). This 7-dimensional effect space
has two desirable properties: it is object-independent and easy to measure in simulation.

B. Crowdsourcing In-the-Wild

We first conducted a user study to test the feasibility of collecting crowdsourced labels using “in-the-wild” human-robot interactions. Our robot is equipped with a Kinova Jaco 7-degree of freedom arm with a Robotiq 2-finger gripper (the gripper is held in a fixed open position for this study), mounted at a 90 degree angle so it is parallel to the ground plane. During each interaction, the robot used its arm and gripper to perform an action on an object as a human participant watched. The object’s position was detected using a 3D point-cloud based visual object detector and actions were translated to be centered on the object’s position (but not orientation). After watching the action, the participant would complete the fill-in-the-blank sentence “the robot ____ the object” using a nearby computer. The study was conducted in the public area of a campus building and we observed spontaneous interactions with approximately N=20 passers-by. Many participants completed multiple interactions during this study, resulting in 73 crowdsourced labels.

Using spontaneous real-world interaction resulted in slow collection rates—hundreds of hours would be required to reach our desired dataset size. It also introduced several sources of error. The real-world behavior that we hoped to capture by using a physical robot resulted in inconsistent object effects, even for the same objects and actions tested repeatedly, introducing significant noise into the effect data. Some of this noise was due to the inherent variability of real-world physics, including surface friction and offset centers of mass, with other noise coming from perception noise. Participants also provided answers such as “the robot failed to pick up the object,” which is not useful for our study. Finally, we had no ability to control for factors such as poor vocabulary or English language skill, which further degraded our data quality.

C. Crowdsourcing Online

To collect a larger dataset, we next conducted a large-scale crowdsourced online survey. We simulated videos of a simulated version of our robot completing the same types of affordance actions used in the first study. These videos were then used in an Amazon Mechanical Turk (MTurk) survey as described below.

We used the data from the first study to develop the label subset \( L' \), which we refer to as the set of valid labels, by taking all labels that occurred more than once in that study, along with a “nothing happened” label. These labels’ frequencies are shown in Fig. 5a. This list represents a fixed set of possible affordance labels that were learned from data provided by non-expert humans.

1) Simulated Affordances: To produce affordances for the online study, we rendered videos of a simulated robot in the Bullet simulator. The simulated hardware was chosen to closely match the arm and gripper of the physical robot used in the first study (see Fig. 4).

At the beginning of the \( i \)th simulated run, the target object having class \( c_i \) was rotated about the z-axis by a random angle \( \theta_i \) to encourage data diversity. To account for any possible dynamic behaviors, such as objects sliding, rolling, or tipping, we allowed a settling time before measuring the effect on the object, determined empirically based on the behavior of objects in the testing set. Each simulation episode was 4 seconds long plus 1 second of settling time, and resulted in one object-action-effect triplet.

Each object in the set was represented by its YCB mesh model, decomposed into convex shapes by the V-HACD library. We fix friction values to reasonable defaults in our simulator and assume that objects have uniform density when calculating center of mass. Each run was rendered into a video from a fixed viewpoint in the scene (see Fig. 4).

2) Crowdsourcing survey: Our survey consisted of 2 parts. In part 1, MTurk participants were shown the rendered video of a manipulation action and asked two questions to describe what happened. The first question presented statements constructed using the valid labels from the first study (i.e., “The robot pushed the object”). Participants were asked to check boxes next to each statement that was true for the video. The second question did the same as the first, but required exactly one label to be selected. We call this label the canonical label (\(^*\)). The second question required the participant to be more decisive about their label choice than the first, and so we prevented participants from returning to the first question to change their answer. To encourage data diversity, we limited each participant to a maximum of 10 surveys.

Part 2 asked for a participant’s gender, whether or not English was their first language, and how many times they

---

https://bulletphysics.org
https://github.com/kmamou/v-hacd

The friction values used were \( \mu_{\text{friction}} = 1.0, \mu_{\text{spinning}} = 0.001 \). These values were based on the PyBullet documentation and examples and tuned empirically.
had interacted with a robot arm before. Table I summarizes the survey questions.

We compensated participants 0.10 USD for each task, which was usually completed in under one minute. We limited our study to participants located in the United States (to improve English skill) and with a 85%+ MTurk approval rating.

D. Training Predictors

After completing our data collection via the online survey, we trained C-Support Vector Machines (SVMs) over the labeled dataset to analyze the predictive power of action and effect data. The goal for each SVM (one per label in \( \mathcal{L'} \)) was to estimate the labeling function \( f(a, a, c, \ell) \) for a given label. One set of SVMs was trained using only action-label pairs, \((a, l(1))\), individually for each label \( l \in \mathcal{L'} \). Another set of SVMs were trained using only effect-label pairs, \((e, l(1))\), and a third set was trained using both the action and effect features as inputs, \(((a, e), l(1))\). In all three cases, the inputs (actions and effects) were scaled to have mean 0 and variance 1 before training, but no other pre-processing was applied. Both sets of SVMs used radial basis function kernels, \( \gamma = 2, C = 1 \), and were tested via 4-fold cross-validation.

V. RESULTS

A. Crowdsourced Data

After discarding responses that failed the quality checks described in Section IV-C, we collected a total of 448 responses from 319 participants. Table II summarizes the participants.

B. Label Analysis

The frequency of the canonical labels (when only one label could be selected) are shown in Fig. 5a. In this paper, “nothing” represents “nothing happened,” and the other labels correspond to verb phrases of the form “The robot X (past tense version) the object.” For example, “push” was displayed as “The robot pushed the object.” “knock” was shown to participants as “The robot knocked over the object.”

By analyzing the set of non-exclusive valid labels collected in Part 2 of the study, we can see which labels are the most prevalent and the relationships between them. The frequency of valid labels is shown in Fig. 5b. Note the differences compared to Fig. 5a. 57% of participants (255/448) provided one or zero valid labels, even when explicitly allowed to select more than one by the survey. MTurk workers may tend to hurry through the task and select fewer labels than are appropriate, as their compensation is directly tied to their task completion.
rate. However, we cannot control for this type of behavior, since it is possible that only one or zero labels are appropriate. To avoid being biased by this effect, we only considered the responses with at least 2 valid labels provided.

After removing single-label responses, how often different labels coincide allows us to uncover hierarchy and relationships between labels. Fig. 6 shows the correlation between different labels. We can see that labels such as touch and move often coincided, but it is hard to draw stronger conclusions because the labels’ frequencies are not equal. Fig. 7 is a graphical depiction of the CP calculated between all pairs of valid labels, which balances for label frequency. This figure provides new information, such as the fact that miss and nothing are closely correlated with each other while being largely disjoint from the rest of the labels in the set.

In a similar way, we can see a relatively strong correlation (high CP values) between the affordances catch and rotate, two words which do not appear to be semantically similar. One possible explanation for this particular similarity is that the word catch is perceived by observers to refer to “getting caught” in the gripper, rather than the more commonly-used meaning of the word (as in the phrase “catching a baseball”).

Fig. 8 shows the label hierarchies discovered in our dataset by using the CP thresholding procedure described in Section III-E with two different values of the threshold $t$. These hierarchies uncover sensible semantic relationships between the labels, including that tip implies touch and that rotate and catch are, depending on the threshold used, equivalent or similar labels.

C. Valid Label Prediction

To characterize the performance of our SVM labeling function estimates, we calculated their Receiver Operating Characteristic (ROC) curves. The predictors using affordance actions as the input are shown in Fig. 9a and the predictors using affordance effects as the input are shown in Fig. 9b. Fig. 9c shows predictor performance using both inputs. The results shown are for the 7 most common affordance labels in our data; the other labels in $L'$ were not common enough to provide data for meaningful learning. The predictions are across all objects in the dataset.

The area under the ROC (AUROC) was higher for all predictors when using effects for prediction (mean 0.87 across all labels) than when using actions (mean 0.65). The effect-based predictors all performed significantly better than random.
Our approach allows a robot to build a labeled affordance model in a data-driven fashion, ensuring that it is grounded using the perceptions of non-expert users. The results also generalized across objects and study subjects, suggesting that labels are perceived via a set of consistent, salient features.

Our results show that effects were a more informative feature for label prediction than actions. This encouraging result allows researchers to optimize actions with respect to other variables (such as safety, speed, minimizing joint motion, etc.), as long as the resulting effect on the object being manipulated remains the same. In this study, the space of actions was continuous and general, but each individual action was simple. Larger-scale computational affordance datasets (whether real or simulated) that include more complex actions and richer environments would allow conducting more in-depth studies on this particular topic.

Our label analysis in Section V-B revealed hierarchies within our dataset’s labels. As opposed to general-purpose lexical databases such as WordNet, which sometimes prove too general for use in robotics, our hierarchies are grounded in a robot’s actions, forming a domain-specific lexicon for manipulation. However, we note that there is no clear consensus among our participants as to which level of specificity to use for a canonical label, even if they would agree on the set of valid labels. This highlights the importance of collecting and analyzing both types of labels in this study. In addition, the proper threshold value to use for parent/child relationships will likely depend on the dataset and the desired level of detail.

VII. Conclusions

This paper introduced a procedure for collecting a dataset of labeled affordances and provided insights into how affordances are generalized and perceived by humans. The labels provided for simulated affordances were consistent enough to train predictive classifiers to act as a labeling function. The label data could also be analyzed to discover synonyms and other relationships, adding structure to the affordance model’s label groundings. A labeled affordance model helps define the part of an affordance model that is “human-facing” (the labels) and the part that is “robot-facing” (the rest of the model). Grounded models such as the one developed in this study will improve human-robot communication and become increasingly important as our interactions with robots become more frequent and complex.


