Robotics: Science and Systems 2020
Corvalis, Oregon, USA, July 12-16, 2020

Learning Labeled Robot Affordance Models Using Simulations and Crowdsourcing

Adam Allevato
Department of Mechanical Engineering
The University of Texas at Austin
Austin, TX, USA
allevato@utexas.edu

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

Mitch Pryor
Department of Mechanical Engineering
The University of Texas at Austin
Austin, TX, USA

Andrea L. Thomaz
Department of Electrical Engineering
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 constructed affordance models by grounding (connecting) them to natural language, but primarily used expert-defined actions, effects, or labels to do so. In this paper, we use short text responses provided by humans and simple randomized robot manipulation actions to construct a *labeled affordance model* that defines a relationship between English-language labels and 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 more accurately predicted by 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 set of affordance primitives.**

I. Introduction

An intelligent robot deployed in the real world should possess the ability to accept commands and learn from humans via a shared understanding of how natural language can represent various tasks and skills. However, a robot's internal skill representations may be disconnected from how a human would represent or describe the same skills, which creates a significant communication barrier. One tool that researchers have used to break down this barrier is the *affordance model* [34], which represents the interactions between an agent and its environment as consisting of *objects*, *actions* (or behaviors), and the *effects* of those actions. The affordances available to a robot usually correspond to human-understandable concepts that are described by natural language, such as "openable" drawers or "pushable" buttons. However, these concepts are not direcly represented in the model, making it unclear whether a particular concept is applicable to specific affordances effects or actions. The problem of assigning concepts or affordance labels to specific primitives is an example of the *symbol grounding* problem [11]. Ideally, a robot's set of affordance labels (symbols) should be learned using data from non-expert humans, allowing the robot to build an affordance model in a more general and scalable way [43]. To enable this approach, the actions, effects, and labels used for learning should all be kept open-ended and general rather than being predetermined by a system designer, has been done by recent works in both affordance learning [46, 17, 4] and symbol grounding [44, 20].

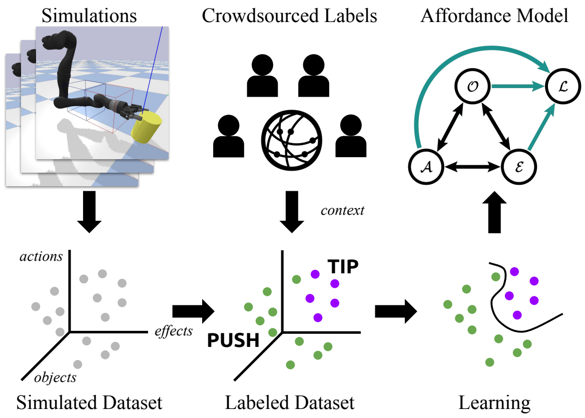


Figure 1: We apply crowdsourced labels to a dataset of robot manipulation actions to create *labeled affordance models* that are grounded to natural language via more general human understanding.

The core contribution of this paper is a data-driven method for creating *labeled affordance models* that connect human-provided, natural-language labels to robot manipulation actions and their effects. We do not dictate a fixed set of complex actions or effect features *a priori*; instead, we use random linear robot motions and measure the rigid body transformation (i.e. change in pose) of the affordance object to reduce the amount of bias instilled by the system designer. Inspired by work from the symbol grounding community, our approach learns from these simple, open-ended primitives along with crowdsourced label data from non-expert users (Fig. 1), directly developing a grounding between human-provided labels and affordance features.

Our method provides insights into the saliency of different types of affordance data, showing that effects are a much stronger predictor of the affordance label than are robot actions. The affordance labels provided by humans are also applicable 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 general, data-driven grounding for 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 [26, 5, 24, 27, 38, 7, 35, 25, 33]. Affordances were adapted for robotics research by explicitly modeling them as interactions between objects, actions, and effects [34, 19]. 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) [18, 41, 39, 40, 13], graph representations [45], Support Vector Machines (SVMs) [47, 10, 48], and conditional random fields [17, 22]. For more examples of affordance research, we refer the reader to recent surveys [52, 12, 21].

In some affordance research (especially foundational works in the field), the robot is limited to a small, fixed set of actions [19, 18] or a generalized action with only a small number of parameters [49]. The effect features have more variety, and can consist of a fixed set [8], the measured change in visual features [18, 47, 48], or object motion [23]. Uğur et al. [49] and Allevato et al. [2] are two works that use more generalized affordance feature spaces for both actions and effects, but both 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 also well-studied in robotics [6, 43, 30]. In this work, we consider the subproblem of grounding natural languages to affordances and manipulation contexts. Several works [32, 44, 10, 15, 22, 37, 50, 29] focus on the language modeling portion of the symbol grounding problem to connect natural language instructions to a robot's actions or objects in its environment. 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 [28, 38, 16].

Yürüten et al. [51] and Krunic et al. [18] 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 specific verb 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 these verb labels 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. [44] 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. [42] 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

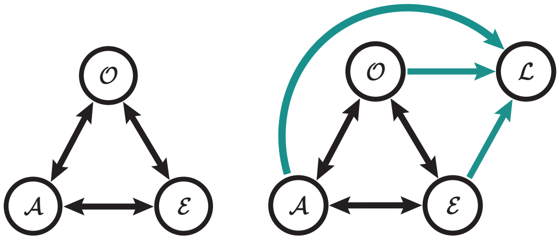
This section reviews the concept of an affordance model as it is commonly used in robotics. It then introduces the notion of labels and how they can be used to augment the basic model.

A. Affordances

The affordance model "define[s] the relation between an agent and its environment through its motor and sensing capabilities" [19]. Formally, affordances are defined as the relations between the feature spaces of possible objects ($\mathcal{O}$), actions ($\mathcal{A}$), and effects ($\mathcal{E}$) (see Fig. 2). A triplet of features drawn from these spaces $(o, a, e)_i, o \in \mathcal{O}, a \in \mathcal{A}, e \in \mathcal{E}$ represents a single affordance. This is the same terminology used in [19] 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



(a) (b)

Figure 2: (a): Lopes's original affordance representation [19] of objects ($\mathcal{O}$), actions ($\mathcal{A}$), and effects ($\mathcal{E}$). (b) our representation of the model in this study, including human-understandable Labels ($\mathcal{L}$) that describe the affordance.

human attempting to understand the model or interact with a robot that is using the model.

We posit that the set of words that carry strong semantic meaning for humans is relatively constant for many objects. *Tip*, *flip*, *slide*, and *turn* describe rigid-body transformations, although it is not immediately clear how concise these labels are (i.e. *move* vs *knock over*). These labels correspond to unknown volumes in affordance feature space, spaces which may overlap.

We formally introduce *affordance labels* as a fourth feature space, $\ell \in \mathcal{L}$, using a definition similar to [2] and [18]: 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, $f : \mathcal{O} \times \mathcal{A} \times \mathcal{E} \times \mathcal{L} \rightarrow \mathbb{R}$. 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 ℓ_{best} that is likely to be the best for a given affordance:

$$f(o, a, e, \ell) = P(\ell \mid o, a, e) \tag{1}$$

$$\ell_{best}(o, a, e) = \arg\max_{\mathcal{L}} P(\ell \mid o, a, e) \tag{2}$$

In this work, we test the predictive power of both actions and effects for determining the human-provided label. However, because we are seeking to learn object-independent affordances, we omit o's direct effect on the outcome of the probabilistic model and thereby marginalize over object class. Eqs. (1) and (2) then become:

$$f(o, a, e, \ell) = P(\ell \mid a, e) \tag{3}$$

$$\ell_{best}(o, a, e) = \arg\max_{\mathcal{L}} P(\ell \mid a, e) \tag{4}$$

Algorithm 1 Labeled affordance datasets in two steps.

```
function SAMPLE($\mathcal{O}, \mathcal{A}, N$)                ▷ Section III-C
    $\mathcal{D} \leftarrow \emptyset$
    for all $o \in \mathcal{O}$ do
        for $i = 1$ to $N$ do
            $a_i \sim \mathrm{U}(V \subset \mathcal{A})$
            $\mathcal{D}_i \leftarrow (o, a_i, \textsc{Simulate}(o, a_i))$
    return $\mathcal{D}$
function SURVEY($\mathcal{D}$, $M$)                         ▷ Section III-D
    for $d = (o, a, e) \in \mathcal{D}$ do
        SETUPSIMULATOR($o$)
        PERFORMACTION($a$)
        $\ell, \ell^* \leftarrow$ COLLECTLABEL()             ▷ User input
        $d \leftarrow (o, a, e, \ell, \ell^*)$               ▷ Update datapoint
```

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 selecting a subset of the most commonly-used affordance labels from our dataset.

C. Collecting Affordance Data

Our procedure for data collection is shown in Algorithm 1. The first step is sampling to create the dataset $\mathcal{D} = (o, a, e, \ell)_n, n = 1 \ldots N$ tuples. This requires choosing $\mathcal{O}, \mathcal{A}$, and $\mathcal{E}$ 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, $V \in \mathcal{A}$, and uniformly sample actions $a_1, a_2, \ldots a_N \sim \mathrm{U}(V)$. Each action is performed on all objects $o_1, o_2, \ldots o_M \in \mathcal{O}$ 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 $\mathcal{D}$, 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 labels that are appropriate and store it as the set $\mathcal{L}_a$. We also request the single "best" or most appropriate label, which is added to the affordance tuple as ℓ^*.

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 both a single multiclass label classifier over actions or effects, as well as one-class binary classifiers for each label individually. 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 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 [31] or ConceptNet [36] do not make sense for this task.

We propose an approach to constructing *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. (5) ($\cap$ and $\wedge$ are logical AND and OR, respectively; recall that $\mathcal{L}_a$ is the set of all valid labels for a given datapoint).

$$\mathrm{CP}(\ell_a, \ell_b) = \frac{p(\ell_a \cap \ell_b)}{p(\ell_b)} \approx \frac{\sum\limits_{o,a,e,\mathcal{L}_a,\ell^* \in \mathcal{D}} \mathbf{1}(\ell_a \in \mathcal{L}_a \wedge \ell_b \in \mathcal{L}_a)}{\sum\limits_{o,a,e,\mathcal{L}_a,\ell^* \in \mathcal{D}} \mathbf{1}(\ell_b \in \mathcal{L}')} \tag{5}$$

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:

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

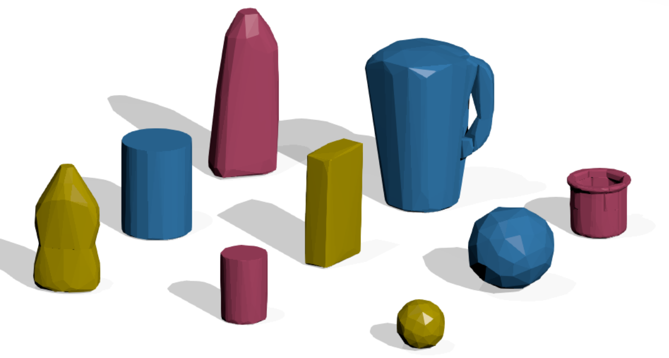


Figure 3: simplified 3D mesh models (see Section IV-C1) of the 9 YCB objects used for collecting affordance data. From left to right and top to bottom, the objects' YCB identifiers are: *mustard_bottle, master_chef_can, bleach_cleanser, tomato_soup_can, sugar_box, pitcher_base, baseball, mini_soccer_ball*, and *mug*. Color is for illustration purposes only.

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 III-C, we selected general feature spaces for objects, actions, and effects. A 9-object subset (Fig. 3) was selected from the standardized YCB dataset presented in [3].

We sampled each action as a straight-line motion in Cartesian space where the orientation of the end effector is held constant.These motions 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 $(-0.05\mathrm{m}, 0, \alpha_z/2)$ and the end state region centered on $(0.05\mathrm{m}, 0, \alpha_z/2)$. 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 $\alpha_x = \alpha_y = \alpha_z = 0.2\mathrm{m}$, 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 $\mathbf{SE}(3)$ 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

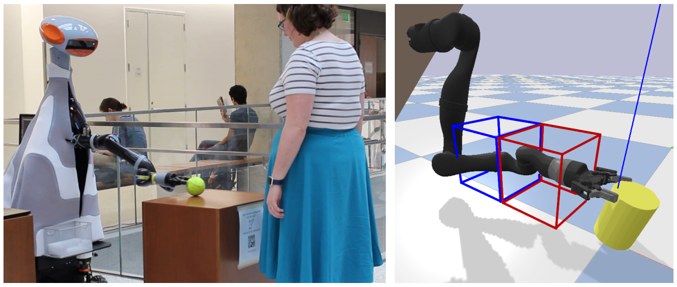


Figure 4: Left: our in-the-wild study recreated by one of the authors. The robot manipulates an object in a public area on campus and collects labels from passers-by. Right: a frame representative of the rendered simulation videos used in the online crowdsourced study. The object being manipulated is shown in yellow. Superimposed blue and red wireframes show the start and end positions of the robot's wrist joint, respectively (for visualization only, not shown to participants).

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 (see Fig. 4). The object's position was detected using a 3D point-cloud based visual object detector [1] 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 issues. Participants also provided answers such as "the robot *failed to pick up* the object," which is not useful for our study[1]. 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 $\mathcal{L}'$, 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[2]. 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 ith simulated run, the target object having class c_i was rotated about the z-axis by a random angle θ_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[3]. We fix friction values to reasonable defaults in our simulator[4] 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 list of *valid labels* derived from the first study (i.e., "The robot *pushed* the object"). Participants were asked to check boxes next to any statement that was true for the video. The second question provided the same options as the first, but only allowed a single label to be selected. We call this single label the *canonical label* (ℓ^*).To encourage data diversity, we limited each participant to a maximum of 10 surveys.

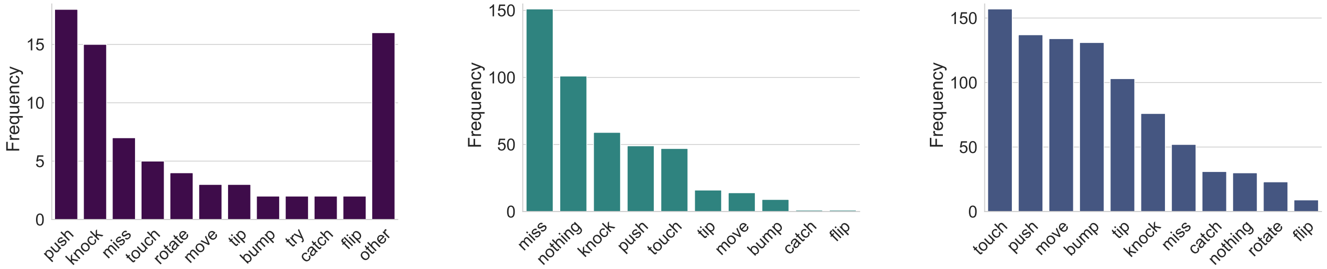
Part 2 asked for a participant's gender, whether or not English was their first language, and how many times they

[1] We did not restrict response lengths, so "failed to pick up" was a valid way to fill in the blank.

[2] https://bulletphysics.org

[3] https://github.com/kmammou/v-hacd

[4] The friction values used were $\mu_{lateral} = 1.0, \mu_{spinning} = 0.001$, based on the PyBullet documentation and examples and tuned empirically.



(a) Histogram of labels provided by in-the-wild study participants (N=79). Labels that only occurred once are grouped in the "other" category.

(b) Histogram showing the number of canonical labels provided in the online crowdsourced study.

(c) Histogram showing the number of valid labels provided by participants (excluding single-label responses).

Figure 5: Histograms of label frequency for the collected datasets.

Table I: Details of the fields in our crowdsourced survey

Field name	**Survey prompt**	**Data type**
Freeform label	Complete the sentence (fill in the blank) to describe what happened in the video.	String
Valid labels	Answer the following true/false questions: The robot [label]ed the object	Set of Boolean values for 11 labels
Canonical label	Select the single most specific description of the video	Categorical: Choice from set of 11 possible labels
Robot Familiarity	How many times have you interacted with a robot arm before?	Categorical: Never, less than 10 times, more than 10 times
Sex	—	Categorical: Male, female, prefer not to answer
English	Is English your first language?	Boolean

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 $f(o, a, e, \ell)$ for a given label. One set of SVMs was trained using only action-label pairs, $(a, \mathbf{1}(l))$, individually for each label $l \in \mathcal{L}'$. Another set of SVMs were trained using only effect-label pairs, $(e, \mathbf{1}(l))$, and a third set was trained using both the action and effect features as inputs, $((a, e), \mathbf{1}(l))$. 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.

Table II: Statistics on crowdsourced data.

Metric	**Result**
Participants	319
Responses	448
Gender	51.3% F, 47.8% M
Familiarity with robots	None: 80.3% (360) Less than 10 interactions: 12.3% (55) More than 10 interactions: 7.3% (33)
English as first language	Yes: 95.1% (426), No: 4.9% (22)

B. Label Analysis

The frequency of the **canonical labels** (when only one label could be selected) are shown in Fig. 5b. 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. 5c. 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

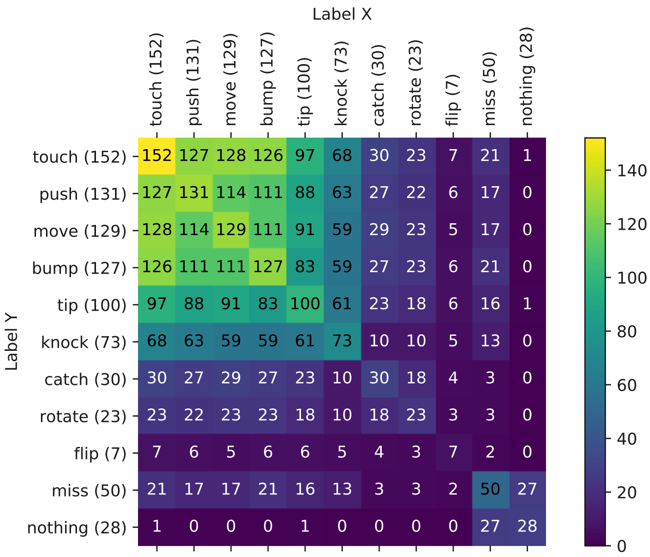


Figure 6: Correlation between valid labels—the number of times that each pair of labels appeared together. Brighter colors indicate higher correlation. The value in each cell is the total number of Label X that coincided with Label Y. The plot is symmetric along the central diagonal.

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.

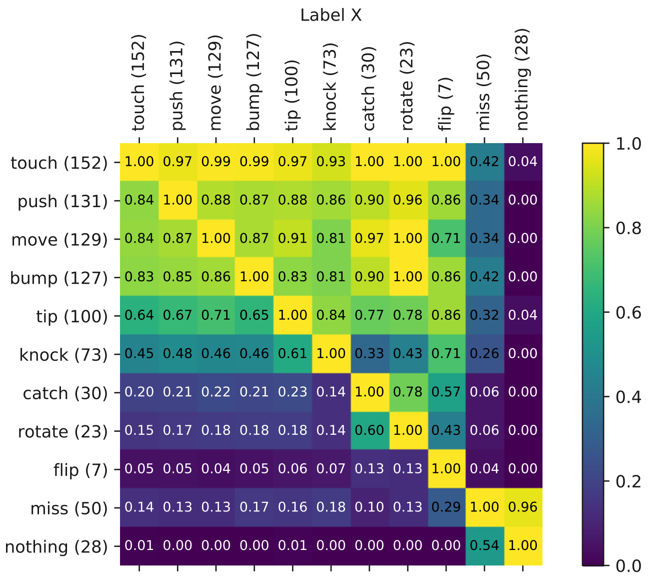


Figure 7: Conditional probability matrix for the set of valid labels. Brighter colors indicate higher co-occurrence. The value shown in each cell is CP(X, Y). For example, for X=*knock*, Y=*touch*, the score is 0.93. *Knock* appeared less than half as often than *touch*, but 93% of the time the *knock* label was provided, *touch* was also provided.

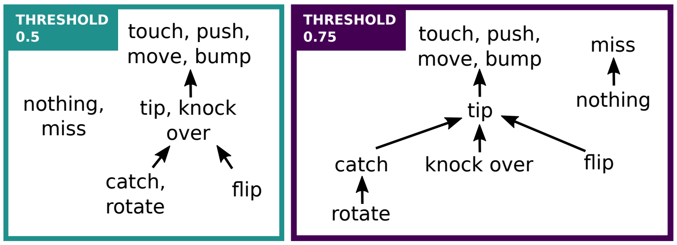
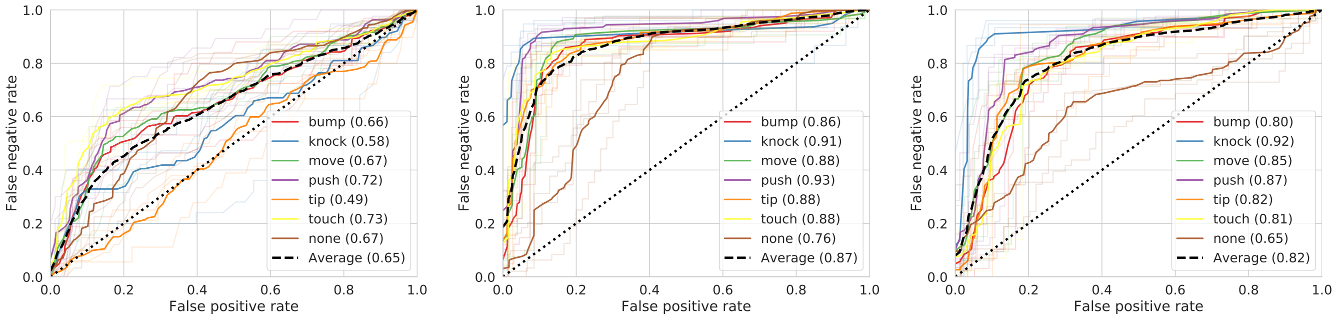


Figure 8: Two label hierarchies constructed from the CP scores shown in Fig. 7, using different thresholding values. Arrows point from child labels to their parents (hypernyms), which are more general than their children.

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 $\mathcal{L}'$ 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 chance (dashed line); the same could not be said for action-



(a) Action features

(b) Effect features

(c) Combined action and effect features

Figure 9: *k*-fold macro-average curves for SVM classifiers trained on different features. The numbers in the legend are the Area Under the Receiver Operating Characteristic curve (AUROC).

Table III: Percent of correct Canonical Label predictions for the top 7 labels in the dataset. The "Most common" result in the table is a baseline comparison: the frequency of the *none* label in the canonical label dataset. The best result for each feature is shown in bold.

Predictor	**Actions**	**Effects**	**Combined**
Random chance	0.14	0.14	0.14
Most common (*none*)	0.33	0.33	0.33
Multiclass RBF SVM	**0.37**	0.47	0.30
Nearest Neighbor	0.34	0.44	0.39
Random Forest	0.34	**0.49**	**0.40**

based predictors. Learning from actions and effects combined as a single input feature reduced performance compared to only using effects, as can be seen in Fig. 9c.

D. Canonical Label Prediction

We also sought to predict the single canonical label provided for each affordance using a single predictor using various machine learning techniques. This is a multiclass prediction problem, which is significantly harder than the binary classification problems being predicted in the previous section. As can be seen in Table III, effects again proved to be a better input feature than actions or a combination of the two.

Overall, the results for this experiment were mixed. The best prediction score of 0.49 was achieved by using a random forest (10 estimators, forest depth of 5) over effect features only. Random forests may perform better in this case because they are not as sensitive to outliers as SVMs, but the difference in prediction accuracy was too small to make strong claims about the best solution for this prediction problem.

VI. Discussion

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 and learning from a dataset of labeled affordances and provided insights into how affordances are generalized and perceived by humans. Our results suggest that humans may perceive affordances primarily as a function of the effect an action has on a target object, rather than the action taken by the robot. We also showed that our labeled affordances could be analyzed to discover synonyms and other relationships, adding semantic structure to the set of collected labels. A labeled affordance model helps define which part of the model is "human-facing" (the labels) and which part is "robot-facing" (the rest of the model). We expect that data-driven labeled affordance models will become increasingly important for facilitating human-robot communication as our interactions with robots continue to become more frequent and complex.

References

[1] Adam Allevato. *An object recognition and pose estimation library for intelligent industrial automation (Master's Thesis)*. University of Texas, 2016.

[2] Adam Allevato, Andrea Thomaz, and Mitch Pryor. Affordance Discovery using Simulated Exploration. In *International Conference on Autonomous Agents and Multiagent Systems*, Stockholm, Sweden, 2018.

[3] Berk Calli, Arjun Singh, Aaron Walsman, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M. Dollar. The YCB object and Model set: Towards common benchmarks for manipulation research. In *2015 International Conference on Advanced Robotics (ICAR)*, pages 510–517. IEEE, July 2015. ISBN 978-1-4673-7509-2. doi: 10.1109/ICAR.2015.7251504.

[4] Vivian Chu and Andrea L Thomaz. Analyzing differences between teachers when learning object affordances via guided exploration. *The International Journal of Robotics Research*, 36(5-7):739–758, March 2017. ISSN 0278-3649. doi: 10.1177/0278364917693691.

[5] Vivian Chu, Baris Akgun, and Andrea L Thomaz. Learning haptic affordances from demonstration and human-guided exploration. In *Haptics Symposium (HAPTICS), 2016 IEEE*, pages 119–125. IEEE, 2016.

[6] Silvia Coradeschi, Amy Loutfi, and Britta Wrede. A short review of symbol grounding in robotic and intelligent systems. *KI-Künstliche Intelligenz*, 27(2):129–136, 2013.

[7] Thanh-Toan Do, Anh Nguyen, and Ian Reid. Affordancenet: An end-to-end deep learning approach for object affordance detection. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–5. IEEE, 2018.

[8] Leni K Le Goff, Oussama Yaakoubi, Alexandre Coninx, Stéphane Doncieux, Leni K Le Goff, Oussama Yaakoubi, Alexandre Coninx, and Stéphane Doncieux. Building an Affordances Map with Interactive Perception. *CoRR*, abs/1903.0, 2019.

[9] James G Greeno. Gibson's affordances. *Psychological Review*, 101(2):336–342, 1994.

[10] Sergio Guadarrama, Erik Rodner, Kate Saenko, Ning Zhang, Ryan Farrell, Jeff Donahue, and Trevor Darrell. Open-vocabulary Object Retrieval. In *Robotics: Science and Systems*, volume 2, page 6, 2014.

[11] Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346, 1990.

[12] L Jamone, E Ugur, A Cangelosi, L Fadiga, A Bernardino, J Piater, and J Santos-Victor. Affordances in Psychology, Neuroscience, and Robotics: A Survey. *IEEE Transactions on Cognitive and Developmental Systems*, 10(1):4–25, March 2018. ISSN 2379-8939. doi: 10.1109/TCDS.2016.2594134.

[13] Esteban Jaramillo-Cabrera, Eduardo F Morales, and Jose Martinez-Carranza. Enhancing object, action, and effect recognition using probabilistic affordances. *Adaptive Behavior*, page 1059712319839057, 2019.

[14] Sinan Kalkan, Nilgün Dag, Onur Yürüten, Anna M Borghi, and Erol Şahin. Verb concepts from affordances. *Interaction Studies*, 15(1):1–37, 2014.

[15] Casey Kennington and David Schlangen. Simple Learning and Compositional Application of Perceptually Grounded Word Meanings for Incremental Reference Resolution. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 292–301, Beijing, China, jul 2015. Association for Computational Linguistics.

[16] George Konidaris, Leslie Pack Kaelbling, and Tomas Lozano-Perez. From skills to symbols: Learning symbolic representations for abstract high-level planning. *Journal of Artificial Intelligence Research*, 61:215–289, 2018.

[17] Hema S. Koppula and Ashutosh Saxena. Anticipating Human Activities using Object Affordances for Reactive Robotic Response. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(1):14–29, January 2016. ISSN 0162-8828. doi: 10.1109/TPAMI.2015.2430335.

[18] Verica Krunic, Giampiero Salvi, Alexandre Bernardino, Luis Montesano, and José Santos-Victor. Affordance based word-to-meaning association. In *2009 IEEE International Conference on Robotics and Automation*, pages 4138–4143. IEEE, 2009.

[19] Manuel Lopes, Francisco S. Melo, and Luis Montesano. Affordance-based imitation learning in robots. In *Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1015–1021. IEEE, IEEE, October 2007. ISBN 978-1-4244-0911-2. doi: 10.1109/IROS.2007.4399517.

[20] Cynthia Matuszek, Liefeng Bo, Luke Zettlemoyer, and Dieter Fox. Learning from unscripted deictic gesture and language for human-robot interactions. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

[21] Huaqing Min, Chang'an Yi, Ronghua Luo, Jinhui Zhu, and Sheng Bi. Affordance research in developmental robotics: A survey. *IEEE Transactions on Cognitive and Developmental Systems*, 8(4):237–255, 2016.

[22] Dipendra K Misra, Jaeyong Sung, Kevin Lee, and Ashutosh Saxena. Tell me dave: Context-sensitive grounding of natural language to manipulation instructions. *The International Journal of Robotics Research*, 35(1-3):281–300, 2016.

[23] Bogdan Moldovan, Plinio Moreno, Martijn van Otterlo, José Santos-Victor, Luc De Raedt, Martijn Van Otterlo, and Luc De Raedt. Learning Relational Affordance Models for Robots in Multi-Object Manipulation Tasks. *2012 IEEE International Conference on Robotics and Automation*, pages 4373–4378, May 2012. doi: 10.1109/ICRA.2012.6225042.

[24] Austin Myers, Ching L Teo, Cornelia Fermüller, and Yiannis Aloimonos. Affordance detection of tool parts from geometric features. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages

1374–1381. IEEE, 2015.

[25] Tushar Nagarajan, Christoph Feichtenhofer, and Kristen Grauman. Grounded Human-Object Interaction Hotspots from Video. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8688–8697, 2019.

[26] Lorenzo Natale, Giorgio Metta, and Giulio Sandini. Learning haptic representation of objects. In *International Conference on Intelligent Manipulation and Grasping*, 2004.

[27] Anh Nguyen, Dimitrios Kanoulas, Darwin G Caldwell, and Nikos G Tsagarakis. Detecting object affordances with convolutional neural networks. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2765–2770. IEEE, 2016.

[28] Hanna M Pasula, Luke S Zettlemoyer, and Leslie Pack Kaelbling. Learning symbolic models of stochastic domains. *Journal of Artificial Intelligence Research*, 29: 309–352, 2007.

[29] Rohan Paul, Jacob Arkin, Nicholas Roy, and Thomas M. Howard. Efficient Grounding of Abstract Spatial Concepts for Natural Language Interaction with Robot Manipulators. In *Robotics: Science and Systems XII*. Robotics: Science and Systems Foundation, 2016. ISBN 9780992374723. doi: 10.15607/RSS.2016.XII.037.

[30] David Paulius and Yu Sun. A Survey of Knowledge Representation in Service Robotics. *Robotics and Autonomous Systems*, 118:13–30, 2019.

[31] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. ISBN 9781937284961. doi: 10.3115/v1/D14-1162.

[32] Raquel Ros, Severin Lemaignan, E. Akin Sisbot, Rachid Alami, Jasmin Steinwender, Katharina Hamann, and Felix Warneken. Which one? Grounding the referent based on efficient human-robot interaction. In *19th International Symposium in Robot and Human Interactive Communication*, pages 570–575. IEEE, September 2010. ISBN 978-1-4244-7991-7. doi: 10.1109/ROMAN.2010.5598719.

[33] Eduardo Ruiz and Walterio Mayol-Cuevas. Scalable Real-Time and One-Shot Multiple-Affordance Detection. *2nd ICRA International Workshop on Computational Models of Affordance in Robotics*, 2019.

[34] Erol Şahin, Maya Çakmak, Mehmet R. Doğar, Emre Uğur, and Göktürk Üçoluk. To afford or not to afford: A new formalization of affordances toward affordance-based robot control. *Adaptive Behavior*, 15(4):447–472, 2007. doi: 10.1177/1059712307084689.

[35] Johann Sawatzky, Yaser Souri, Christian Grund, and Jurgen Gall. What Object Should I Use?-Task Driven Object Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7605–7614, 2019.

[36] Robert Speer and Catherine Havasi. Representing General Relational Knowledge in ConceptNet 5. *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3679–3686, 2012.

[37] Francesca Stramandinoli, Vadim Tikhanoff, Ugo Pattacini, and Francesco Nori. Grounding speech utterances in robotics affordances: An embodied statistical language model. In *2016 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, pages 79–86. IEEE, 2016.

[38] Francesca Stramandinoli, Davide Marocco, and Angelo Cangelosi. Making sense of words: a robotic model for language abstraction. *Autonomous Robots*, 41(2):367–383, 2017.

[39] Francesca Stramandinoli, Vadim Tikhanoff, Ugo Pattacini, and Francesco Nori. Heteroscedastic regression and active learning for modeling affordances in humanoids. *IEEE Transactions on Cognitive and Developmental Systems*, 10(2):455–468, 2017.

[40] Francesca Stramandinoli, Alessandro Roncone, Olivier Mangin, Francesco Nori, and Brian Scassellati. An Affordance-based Action Planner for On-line and Concurrent Human-Robot Collaborative Assembly. *2nd ICRA International Workshop on Computational Models of Affordance in Robotics*, 2019.

[41] Yu Sun, Shaogang Ren, and Yun Lin. Object–object interaction affordance learning. *Robotics and Autonomous Systems*, 62(4):487–496, 2014.

[42] Jaeyong Sung, Seok Hyun Jin, and Ashutosh Saxena. Robobarista: Object Part based Transfer of Manipulation Trajectories from Crowd-sourcing in 3D Pointclouds. In Antonio Bicchi and Wolfram Burgard, editors, *International Symposium on Robotics Research*. Springer International Publishing, 2015. ISBN 978-3-540-48110-2. doi: 10.1007/978-3-319-60916-4{_}40.

[43] Tadahiro Taniguchi, Takayuki Nagai, Tomoaki Nakamura, Naoto Iwahashi, Tetsuya Ogata, and Hideki Asoh. Symbol emergence in robotics: a survey. *Advanced Robotics*, 30(11-12):706–728, 2016.

[44] Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew R. Walter, Ashis Gopal Banerjee, Seth Teller, and Nicholas Roy. Understanding Natural Language Commands for Robotic Navigation and Mobile Manipulation. In *Twenty-Fifth AAAI Conference on Artificial Intelligence*, 2011.

[45] Alexia Toumpa and Anthony G Cohn. Relational Graph Representation Learning for Predicting Object Affordances. *NeurIPS Workshop on Graph Representation Learning*, 2019.

[46] Emre Uğur and Justus Piater. Emergent Structuring of Interdependent Affordance Learning Tasks Using Intrinsic Motivation and Empirical Feature Selection. *IEEE Transactions on Cognitive and Developmental Systems*, 9(4):328–340, December 2017. ISSN 2379-8920. doi: 10.1109/TCDS.2016.2581307.

[47] Emre Uğur, Erhan Oztop, and Erol Şahin. Going beyond the perception of affordances: Learning how to actualize them through behavioral parameters. In *2011 IEEE International Conference on Robotics and Automation*, pages 4768–4773. IEEE, 2011.

[48] Emre Uğur, Erhan Oztop, and Erol Şahin. Goal emulation and planning in perceptual space using learned affordances. *Robotics and Autonomous Systems*, 59(7-8): 580–595, July 2011. doi: 10.1016/j.robot.2011.04.005.

[49] Emre Uğur, Yukie Nagai, Erol Sahin, and Erhan Oztop. Staged development of robot skills: Behavior formation, affordance learning and imitation with motionese. *IEEE Transactions on Autonomous Mental Development*, 7(2): 119–139, 2015.

[50] Tatsuro Yamada, Shingo Murata, Hiroaki Arie, and Tetsuya Ogata. Representation learning of logic words by an RNN: From word sequences to robot actions. *Frontiers in neurorobotics*, 11:70, 2017.

[51] Onur Yürüten, Erol Sahin, and Sinan Kalkan. The learning of adjectives and nouns from affordance and appearance features. *Adaptive Behavior*, 21(6):437–451, 2013.

[52] Philipp Zech, Simon Haller, Safoura Rezapour Lakani, Barry Ridge, Emre Ugur, and Justus Piater. Computational models of affordance in robotics: a taxonomy and systematic classification. *Adaptive Behavior*, 25 (5):235–271, October 2017. ISSN 1059-7123. doi: 10.1177/1059712317726357.