Affordance Prediction via Learned Object Attributes

Tucker Hermans James M. Rehg Aaron Bobick

Abstract— **We present a novel method for learning and predicting the affordances of an object based on its physical and visual attributes. Affordance prediction is a key task in autonomous robot learning, as it allows a robot to reason about the actions it can perform in order to accomplish its goals. Previous approaches to affordance prediction have either learned direct mappings from visual features to affordances, or have introduced object categories as an intermediate representation. In this paper, we argue that physical and visual attributes provide a more appropriate mid-level representation for affordance prediction, because they support information-sharing between affordances and objects, resulting in superior generalization performance. In particular, affordances are more likely to be correlated with the attributes of an object than they are with its visual appearance or a linguistically-derived object category. We provide preliminary validation of our method experimentally, and present empirical comparisons to both the direct and category-based approaches of affordance prediction. Our encouraging results suggest the promise of the attribute-based approach to affordance prediction.**

I. Introduction

A long-standing goal in robotics is the development of robot learning methods that make it possible to predict the effects of actions and support continuous improvements in task execution over time. A key task is the ability to predict the properties of objects at a distance, which can inform robot planning and action selection. As an illustrative example, consider a task in which a mobile manipulator autonomously cleans up a floor area by moving objects one at a time into their respective places. Different objects can support different manipulation strategies: grasping objects that are sufficiently small, pushing heavier objects, rolling round objects, etc. In this setting, the ability to predict the success of different actions at a distance can lead to more efficient task performance and improved robustness. The set of possible actions and their outcomes with respect to the robot and an object are referred to as *affordances*. Gibson, who developed the concept of affordance, described them as encoding the "action possibilities" latent in the environment for a given agent [1, 2].

This paper describes a novel method for learning to infer the affordances of objects based upon their visual appearance. The key insight is to leverage an intermediate level of representation — visual and physical *attributes*. The set of attributes that describe an object can be estimated from low level visual features, and these attributes can in turn be used to infer specific affordance properties. A key advantage of attribute-based prediction is the ability to leverage object properties which are shared by multiple affordances, leading to more effective generalization to novel examples and the ability to learn new affordances with limited training data. While there have been several recent works which pursue an attribute-based approach to object category prediction, we believe this is the first work to address an attribute-based approach to affordance learning.

This paper makes three contributions. First, we introduce a novel attribute-based approach to affordance prediction. Second, we describe a method for affordance learning which is fully learning-based, in the sense that all mappings can be learned from data. Third, we describe a new dataset of objects and images for affordance prediction in the context of a clean-up task. In particular, our dataset was collected autonomously by a mobile robot.

Tucker Hermans, James M. Rehg, and Aaron Bobick are with the Center for Robotics and Intelligent Machines and The School of Interactive Computing, Georgia Institute of Technology, Atlanta, Ga {thermans, rehg, afb}@cc.gatech.edu

II. Related Work

A. Affordance Learning

While a significant amount of work has been performed on the learning of affordances in robotics [3], only a limited amount addresses the fundamental problem of inferring affordance values from perceptual measurements [4–6].

Fundamentally, any agent that selects an action to perform to accomplish a task must somehow encode the expected outcomes. When these actions involve interacting with an object, these expected outcomes implicitly represent the affordances of the objects (with respect to the agent). In the early work on affordance prediction described in [7, 8], a humanoid robot learns to segment objects through actions such as poking and prodding. After interaction with a set of objects, the system could learn to predict the object motion that would result from a poking action.

Related, Stoytchev [9] describes a method for learning the functionality of a tool through observation of the effects of exploratory behaviors, a process that he termed behavioral babbling. In experiments with a mobile manipulator, the system demonstrated the ability to learn the affordances of a set of tools that could be identified by their color.

The concept of Instantiated State Transition Fragment (ISTF) is introduced in [10]. It encodes the pairing between an object and an action in the context of the state transition function for a domain-specific planner. They describe a process of learning Object Action Complexes [11] through generalization over ISTF's. Montesano et. al. [12] present a Bayesian network model that implicitly represents affordances as mappings from action to effect, which are mediated by the visual features of objects. A model for grasping,

tapping, and touching actions is learned from both self-observation and imitation of a human teacher.

In the work presented here we build upon our previous work in [13], which explicitly represents affordances as discrete valued nodes within a Bayesian network model. In that work, the notion of object category was employed as an intermediate representation to support affordance inference. Here we explore the alternative approach of using physical object attributes as an intermediate layer. We primarily benefit by eliminating the concern of being tied to a predefined set of semantic categories. This change will permit the system to generalize more effectively and efficiently during affordance learning and inference as new objects are encountered.

We note here that our approach complements many state of the art manipulation algorithms. The simple controllers used for performing the manipulation tasks in this work could be replaced by more sophisticated methods. For example, given that an object is known to be graspable, the correct grasp frame could be computed a variety of methods: [14, 15]. Similarly, much work on object pushing could be used to help guide the manipulation of objects, once it is known to be pushable [16, 17]. Generating ground truth information using these same methods would allow the affordance prediction algorithm to be tailored to the specific controllers or planners used. Additionally a task level planner or policy could make use of the affordance inference machinery in determining the most efficient or reliable way to complete the problem at hand.

B. Visual Attributes

The computer vision community has recently produced a number of works on learning visual attributes [18–25]. While some of these works are concerned with identifying instances of particular visual attributes in a visual scene [20, 21], we are interested in those methods which use attributes to assist in performing some other task, such as object recognition [23–25].

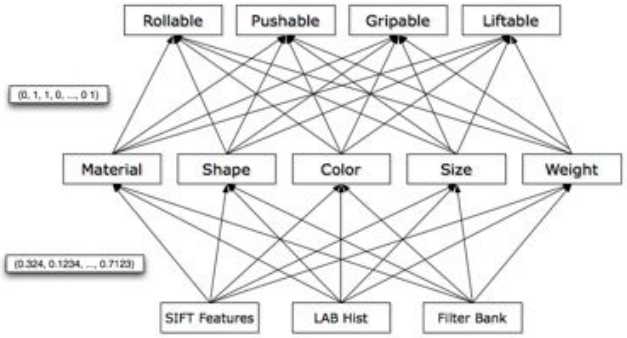


Fig. 1: Attribute affordance prediction model.

Our work most closely resembles that of [24], where visual attributes are learned discriminatively in a supervised manner. These predicted attributes form a bit-string which maps directly to an animal class; having this intermediate representation allows the system to recognize a previously unseen object class by predicting the correct set of attributes for the object of interest. Instead of this direct mapping from the predicted attributes to object class, our work uses the attribute predictions as an intermediate feature vector for performing classification of affordance values.

III. Attribute Affordance Prediction Model

Most previous approaches to vision-based affordance prediction (e.g. [26]) use a *direct perception* approach, in which visual features are measured about an object and the affordances are directly inferred from a learned mapping. Such a model is trained with a requisite number of examples consisting of a visual feature vector and a set of affordance values. A separate map is learned for each affordance by constructing an appropriate classifier.

There are several fundamental difficulties with this approach. First affordances are not actually determined — in the physical sense — by visual features, rather by the physical properties of the objects. Whether an object can roll is influenced by its shape; whether it can be pushed is influenced by its material properties. We refer to these physical object properties as semantic *attributes*. Note that attributes are qualitative properties of the object *independent of any robot capabilities or perception systems.* We argue that it is easier to learn to predict affordances from such attributes than from arbitrary visual features.

Second, the criteria for what makes a powerful visual feature are tightly coupled to imaging and viewing phenomena. Features are best chosen, for example, to be invariant or robust with respect to change in illumination or perspective. Since they will be computed over much of the image they must be computationally efficient. These constraints are entirely decoupled from the question of how predictive they might be of affordance values.

Finally, a liability of the direct perception methods is that there is no knowledge transfer between objects. Each direct map from features to affordance is constructed independently to discriminate between the affordance values for the set of training objects. In reality, there is no notion of learning how to infer the physical properties of an object from the visual ones. For example, consider the inference that an object has the attribute of cloth-material. For a given domain of objects, an inference of this attribute can be learned to be inferred from various texture features. This learned mapping is likely valid for any new object with that same semantic attribute. Such *information transfer* [24] reduces the data required to learn successfully to predict the affordances of new objects.

To address these limitations we propose a novel model for perception-based affordance prediction illustrated in Figure 1. At the top most level are the affordances, defined earlier as the action possibilities embodied in an object with respect to a particular agent - in this case a mobile robot platform with a single gripper. As the goal is to infer affordances for visual input, the bottom-most layer

is comprised of visual features, which are chosen to be relatively stable under a variety of viewing conditions.

Where our model differs from direct perception approaches is the inclusion of an intermediate layer targeting specific selected object attributes. The goal is that by explicitly training the robot in a supervised manner to map from features to attributes and then attributes to affordances, the system will capture the physical regularities present in the domain. This physically-meaningful middle layer is enforced by providing attribute valued training data and requiring the system to learn the mapping from features to attributes independent of the affordance inference. This semantically meaningful intermediate layer prevents the learning system from establishing arbitrary affordance decision surfaces in feature space.

To exploit such a model we need to solve two learning problems: visual features to attributes, and attributes to affordances.

IV. Attribute Learning

Attributes are inherent properties of an object which are independent of any sensing or cognition system perceiving them. While we choose to learn attributes which act as constraints on the affordances we wish to predict (e.g. size), we also examine semantic attributes, which may only encode affordances indirectly (e.g. green objects may be heavy withing some operating environment). Additionally we do not restrict our attribute set to be purely visual attributes (e.g. shape), but also include physical attributes, which may be only indirectly perceived visually (e.g. weight).

A. Visual Features

In order to capture information relevant to the broad nature of semantics encoded in the chosen attributes, we extract texture, color, and visual word features. Following the terminology of [18] these define the *base features*.

For texture features we extract a set of filter responses to create textons as described in [27]. The texton centers were trained using a subset of the CUReT texture dataset [28] and grouped into 256 clusters using k-means clustering [29]. For a given image we then extract the filter responses densely over the image and quantize the responses to the closest of the cluster center creating a 256 element histogram. To encode color we extract a $12 \times 6 \times 6$ dimension LAB color histogram over the entire image. SIFT features are extracted densely at four scales in 8 pixel x and y spacing over the entire image [30]. The raw SIFT features are encoded using a *bag of visual words* model into a 512 element histogram of visual words. We create the clusters using a subset of the training data, again using k-means clustering. Concatenation of all feature descriptors results in a single feature of 1200 elements.

B. Attributes

For our current work we wish to learn attributes describing: size, shape, color, material, and weight. We denote each of the m attributes for a specific object as $\alpha_j \in \beta_j$, where β_j is the distribution of possible values for attribute j. Specifically, size is described by the object height and diameter of its footprint in centimeters. Shape attributes are represented by the binary labels of spherical, cylindrical, 2D-boxy, and 3D-boxy. These shape attributes were chosen following [18]. Color attributes are members of the set: blue, red, yellow, purple, green, orange, black, white, and gray. Colors are not mutually exclusive, allowing a single object to express multiple color attributes. Materials are cloth, ceramic, metal, paper, plastic, rubber, and wood. Weight is the object weight in kilograms.

Binary labels encode the shape, color, and material attributes, while the raw floating point values of height, diameter, and weight are used. Using the visual features described in Section IV-A we compare the use of a support vector machine (SVM) and a simple k-nearest neighbor (k-nn) classifier for attribute prediction. We use standard binary SVMs for the binary labels and SVM regression for the real valued attributes. In the case of the k-nearest neighbor classifier, the highest weighted label of the k neighbors gives the binary prediction, while a weighted average of the neighbors is used to predict real valued attributes. Thus to predict an attribute vector for a given test image with appearance x^i, we must extract the base features once, run m attribute classifiers, and concatenate the responses into a single attribute vector α^i.

The multi-channel χ^2 kernel has been shown to work well for the classification of image data, where multiple, unrelated, features are extracted and combined [31, 32]. For our purposes we extend the multi-channel χ^2 kernel to be used as a distance metric with k-nearest neighbors for direct comparison with the SVM kernel implementation. If we denote each of these feature vectors as $F = (f_1, f_2, ..., f_p)$ then we can compute the multi-channel χ^2 distance between feature vectors x and y using the following:

$$\chi^2_{mc}(x, y) = \sum_{i=1}^{p} \left[w_i \cdot \sum_{j \in f_i} \frac{(x_j - y_j)^2}{x_j + y_j} \right] \tag{1}$$

Where $j \in f_i$ is a slight abuse of notation, denoting the indexes of the elements of the feature vectors x and y, that correspond to feature f_i. Additionally, the weights w_i of each channel are taken to be the inverse of the average distance across all pairs of feature vectors in the training set D:

$$w_i = 1/\mathbb{E}\left[\sum_{j \in f_i} \frac{(x_j - y_j)^2}{x_j + y_j} \right]_{x,y \in D} \tag{2}$$

Following [32] we take the exponential of the negative multi-channel χ^2 distance for use as the SVM kernel $K_{mc}(x, y) = \exp\{-\chi^2_{mc}(x, y)\}$. In the case of our visual features the multiple channels correspond to: the LAB color histogram, the texton filter bank response, and the SIFT codeword feature.

V. Affordance Learning

A. Attribute Based Affordance Learning

Our task now falls to predicting affordance values based on the attribute model described in the previous section. We desire to use the attribute values to share information across affordance classification tasks; however, we do not wish to hard code a mapping from attributes to affordances, as was done for animal classification in [24]. Instead we learn the second layer mapping from attribute values to affordance values.

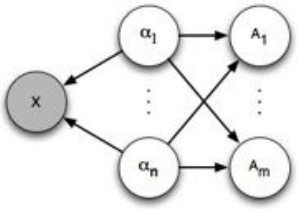


Fig. 2: The Attribute-Affordance graphical model.

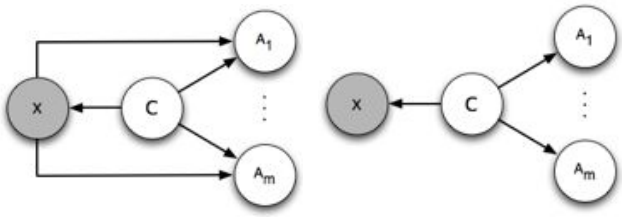
Affordance prediction in the attribute context requires estimation of the function $a_i = f(\alpha^j)$, $a_i \in \{0, 1\}$ for each of the m affordances. Here again we implement both SVM and k-nn classifiers for comparison. For the k-nearest neighbor classifier, we use a more general form of the multi-channel distance to better match the attribute encoding. As with the multi-channel χ^2 distance, we compute a weighted sum of the component channel distances. The difference here is that we use the standard euclidean distance for the real-valued attributes and the Hamming distance for the binary components [33]. As before the weights are the inverse of the average distance between members of the training set. We build a kernel from this distance following the same procedure as before. Specifically, we compute the exponential of the negative of the normalized distance.

As a baseline comparison to our method we also build binary classifiers for each affordance trained directly from the base features. We call this method direct perception (DP) of affordances, following the convention of Gibson and others [2, 13]. As with the attribute learning we implement both SVM and k-nn approaches, using the same multi-channel χ^2 measures.

As an alternative intermediate representation we learn affordance classifiers conditioned on the object class. Here we can build a classifier (multi-class SVM or k-nn) to perform object class prediction using the same base features as above to encode the image. Using this class label we then perform DP affordance prediction using binary classifiers for each affordance trained again on the base features. The difference with the DP approach lies in the fact that, for a given affordance, a separate classifier is built for each object class, rather than one for all objects. This method equates to the Category-Affordance Full model of [13]. Additionally, we implement a majority vote classifier for each affordance, conditioned on object class, corresponding to the Category-Affordance Chain model. We now briefly review these Category-Affordance models.

B. Category Affordance Models

The Category-Affordance full (CA-full) and Category-Affordance chain (CA-chain) models link object categorization with affordance prediction, relying on the assumption, that appearance based object categorizations provide useful cues for affordance prediction [13]. For the CA-full model, this is exploited by factoring the joint probability distribution of the affordance a, appearance x, and category c as



(a) (b)

Fig. 3: The Category-Affordance Full (a) and Chain (b) models.

$$p(x, a, c) = p(c)p(x|c)\prod_{i=1}^{m} p(a_i|x, c). \quad (3)$$

It is depicted as a Bayesian network in Fig. 3a. From the perspective of affordance prediction, the key term in the CA-full model is $p(a_i|x, c)$, which relates the probability of an affordance to the object category and appearance. Given a trained model and an input x, we can compute the posterior distribution over an affordance a_i by marginalizing out the unknown category label:

$$p(a_i|x) = \sum_c p(a_m|c, x)p(c|x), \quad (4)$$

where $p(c|x)$ is the posterior distribution over the category.

Alternatively, the CA-chain model, makes a simplifying assumption that the presence of a specific affordance a_i is conditionally independent of appearance x given the object class label c given the equation:

$$p(x, a, c) = p(x|c)p(c)\prod_{i=1}^{m} p(a_i|c). \quad (5)$$

While these methods were designed to give probabilistic interpretations we have not preserved the probabilistic predictions in our current implementation. The implementation of the CA-chain and CA-full models for this work use a k-nearest neighbor classifier with multi-channel χ^2 to perform the object category classification. The CA-full model also uses k-nearest neighbor with multi-channel χ^2 to compute the affordance, conditioned on the object class. The CA-chain model uses a majority vote classifier for each affordance, given the object category. The motivation for such an implementation was to more directly compare the independence assumptions of each of the CA models to the attribute and DP approaches.

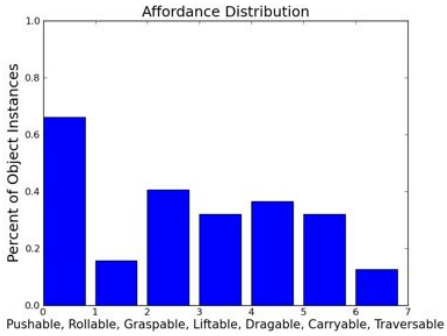


Fig. 4: Distribution of affordance values across all data.

VI. Experimental Validation

We collected data from six object categories: balls, books, boxes, containers (mugs, bottles, and pitchers), shoes, and towels. For each category we collected 55 to 67 images of 8 to 12 object instances per class, giving a total of 375 frames. We show the distribution of positive examples of the seven affordances across all classes in Figure 4. We examined learning visual classification of the following seven affordances: pushable, rollable, graspable, liftable, dragable, carryable, and traversable.

We collected all data from an autonomous mobile robot with a Pan-Tilt-Zoom (PTZ) camera. The collected data was used offline to train all attribute and affordances classifiers. We perform offline validation of our approach, comparing to direct perception as a baseline method for affordance classification. Finally, we perform classification online to inform behavior selection for performing a cleanup task of placing objects within a specified region of the floor.

A. Data Collection and Learning Procedure

All data was collected autonomously by the robot. We use a Mobile Robots Pioneer 3 DX equipped with a PTZ (768×480) and 2-DOF (pinch and lift) gripper. An overhead camera provides localization information of the robot and objects. For each object detected in the overhead view, the robot turns to center its camera on the object and uses the distance estimate to the object to zoom and tilt the camera. Tilting allows for the object to be relatively centered in the frame, while zooming, so that all objects are viewed at the same magnification, removes scale issues, so that size may be estimated from the image (alternatively, the use of depth sensing could remove any scale ambiguities).

Fig. 5: Overhead camera view of the robot in its operating environment.

To generate the ground truth affordance labels, the robot was commanded to perform each of its atomic behaviors of: push, shove (i.e. push-roll), grasp, lift, drag, carry, and traverse (attempt to drive over). A human recorded the success or failure of each action attempt. For cases where the action could be successfully completed in only a constrained situation (e.g. grabbing the object only when facing its narrowest side), we labeled the affordance as not present, since our current system has no mechanism for dealing with object pose or, alternatively, a probabilistic affordance value.

To build the SVM classifiers we use the SVM^{light} package kernel [34]. We use a custom implemented multi-channel χ^2 kernel.

B. Attribute Prediction

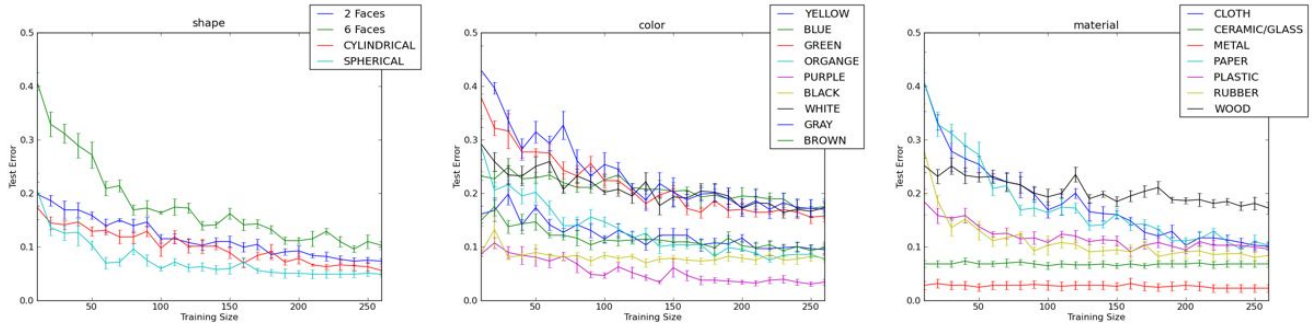
We first examine the performance of our attribute learning procedure. We performed experiments for training set sizes ranging from 10 to 260 images, testing on the remaining 115 images. For each prediction method we performed 5 random testing/training splits of the data and report the results averaged across the 5 splits. Figure 6 compares the performance of SVM and k-nn based attribute prediction on the binary labeled attributes (shape, color, and material). The SVM outperforms the nearest neighbor approach for all attributes, producing lower error rates with fewer training examples. However, the SVM approaches find little benefit in having more than 100 training examples. This early convergence is likely an artifact of the low diversity in training examples in the dataset. Conversely, the nearest neighbor regression outperformed the SVM regression for the size and weight attributes, producing marginally lower error rates. The nearest neighbor approach continued to improve performance with larger training sizes, while the SVM approach converged quickly.

C. Affordance Learning Comparison

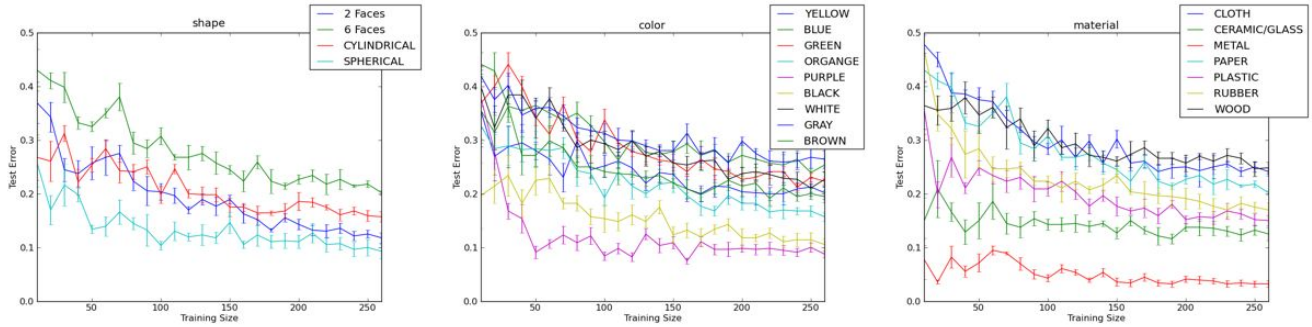
In order to examine the benefits of attribute based learning of affordances, we constructed an experiment to perform affordance prediction on a set of known objects. For this closed set of objects we compare with a direct perception implementation of affordance classification, as well as methods leveraging object categories.

All methods use the same base set of visual features as well as the same choice of classifiers. As such we can directly evaluate the utility of the auxiliary semantic attribute labels. The training was performed with varying training sizes on 5 random splits in the manner as described above for attribute prediction.

Figure 8 summarizes the percentage of correctly predicted affordances for the various methods at the largest training size using the SVM approach, which outperformed the nearest neighbor approach. Figure 7 shows the error rates as a function of training size for three of the SVM methods. The attribute based approach quickly approaches its final

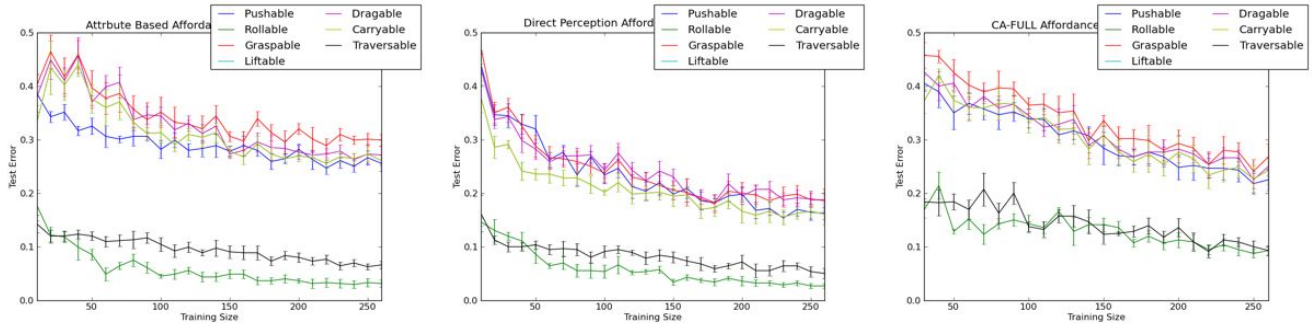


(a) Shape (SVM) (b) Color (SVM) (c) Material (SVM)



(d) Shape (k-nn) (e) Color (k-nn) (f) Material (k-nn)

Fig. 6: Shape, color, and material attribute prediction error versus training size.



(a) Attribute Based (b) Direct Perception (c) CA-Full

Fig. 7: Affordance prediction (SVM) versus training size.

error rate, while the CA-Full approach continues to improve with more training data. Direct perception also increases performance as a function of training size, although the slope flattens near the largest training sizes. We examined this claim by testing affordance prediction on ground truth attribute labels. Figure 9 shows the results for this test, which converges to much lower error rates, with many fewer training examples than any of the current methods.

D. Affordance Prediction on Novel Object Classes

One theoretical benefit of using semantic attributes to guide affordance prediction, as opposed to object class labels, lies in the ability to generalize to unseen object categories. Therefore we introduce an experiment to quantitatively compare the performance of attribute based affordance classification with direct perception of affordances on previously unseen objects. Only the SVM based classifiers are examined on this task, since they outperformed the nearest neighbor method.

We examine the novel object affordance classification problem by training on five of the six object categories and testing on the sixth. For example, to examine the inference capabilities on unseen boxes, we train our direct perception and attribute based classifiers on the data from the object classes of balls, books, containers, shoes, and towels. We repeat this procedure for each of the six object classes in turn. Our findings our summarized in Figure 10.

The attribute approach, on average, only outperforms direct perception on two of the novel object categories. Note, some affordances could not be predicted on unseen object classes, (i.e. only balls were rollable). While the average test

Affordance	Attribute	DP	Full	Chain
PUSHABLE	74.43	83.75	77.50	65.56
ROLLABLE	96.87	97.32	90.71	84.14
GRASPABLE	70.09	81.25	73.21	55.48
LIFTABLE	73.91	83.93	75.71	67.48
DRAGABLE	72.87	81.43	75.00	60.00
CARRYABLE	73.91	83.93	75.71	67.48
TRAVERSABLE	93.39	95.00	90.71	86.61
TOTAL	81.12	85.46	79.21	68.57

Fig. 8: Percent of correctly classified affordances compared between the attribute based method (Attribute), direct perception (DP), Category-Affordance full, and Category-Affordance chain models. Results are for the SVM implementations with the largest (260) training size.

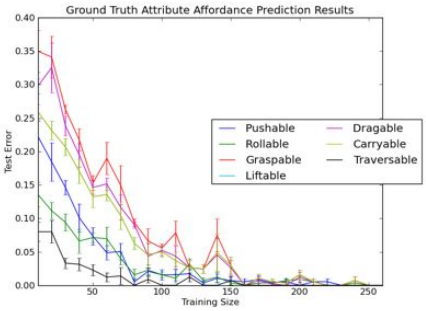


Fig. 9: Affordance prediction for ground truth attributes.

error across all affordances is relatively high for both DP and the attribute based approach, examining a specific object class shows that some affordance predictions better transfer to new object classes.

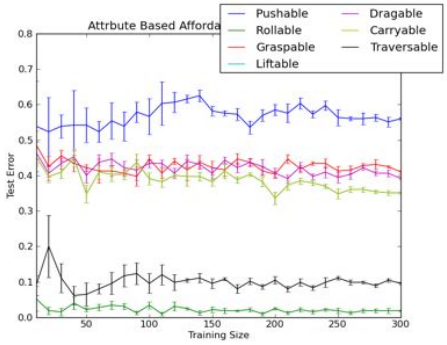
The superior performance of direct perception can be attributed to the limited number of object classes, which provided limited information transfer when one was removed from the training set. The proposed attribute approach was not exposed to a spanning set of attribute vectors capable of scaling to some object categories. In contrast, direct perception makes use of features which may not express any specific semantic quality, which may be more informative when specific examples of attributes arising in the new object category are not provided at training time. Figure 11 shows results for testing on the novel object category of boxes. The relatively flat error rates are representative of other novel categories, informing us that little information is shared between the known object classes and the novel category.

VII. Conclusions and Future Work

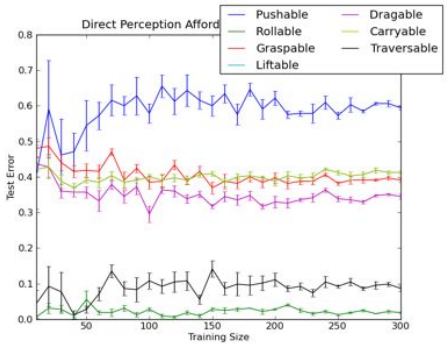
We have presented a novel method for learning object affordances through an intermediate prediction of physical attributes. We believe this method should outperform a direct-perception approach on affordance prediction on novel objects in ways that previous object-classification methods can not. Specifically, semantic attributes form an expressive vocabulary, which can be used as a basis for information transfer to unseen object classes. Our preliminary results show that attribute based affordance prediction can perform

	Ball	Book	Box	Cont.	Shoe	Towel
Att	52.03	39.39	69.01	76.28	60.97	53.63
DP	57.99	65.58	67.69	58.96	67.86	67.91

Fig. 10: Percent of correctly classified affordances of unseen object classes comparing between SVM based direct perception (DP) and the attribute based method (Att). Class labels represent the testing class. Results are the average of five training-testing splits with a training size of 300 images.



(a) Attribute Based



(b) Direct Perception

Fig. 11: SVM based affordance prediction as a function of training size testing on novel object category of boxes.

on par with direct perception and object based affordance classification methods.

Poor attribute prediction appears to be the bottle neck in achieving effective affordance inference in the current work. Attribute prediction could be improved by training attribute classifiers on a much larger auxiliary data set, which need not be comprised solely of images taken from the robot of interest. Such a training set could be collected from annotated web images or standard computer vision databases. Additionally the small set of object classes provided little information sharing when performing inference on novel objects. As such, we intend to extend this approach to data sets with a greater diversity in objects. Beyond this, object class labels still provide much utility; we believe that merging the attribute and object models into a single framework, capable of making use of all information could produce better results, than either method in isolation.

While our work has focused purely on visual features, we believe that higher accuracy shape and size information

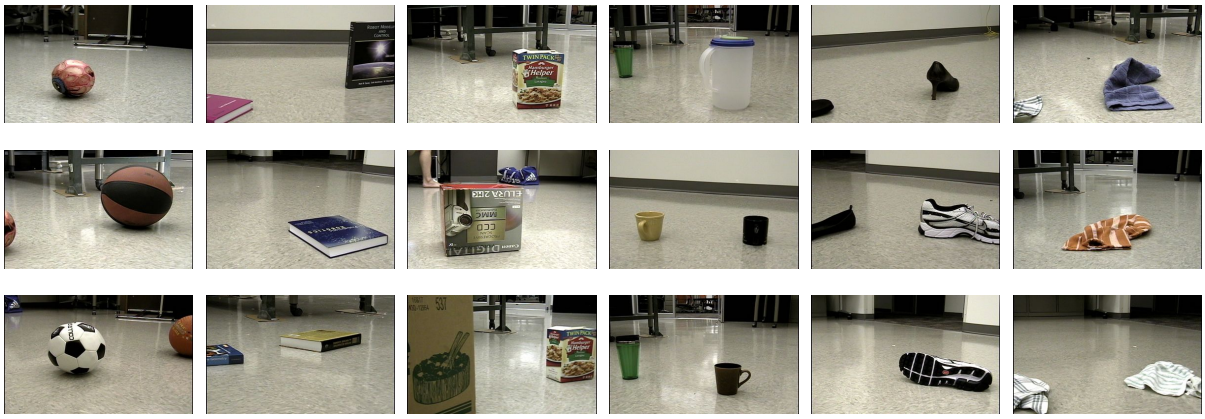


Fig. 12: Examples of images from the six object classes used: Balls, Books, Boxes, Containers, Shoes, and Towels.

could be obtained by incorporating depth sensing, such as stereo imaging, laser scanners, or RGB-D cameras. Our active research is focusing on this use of depth information as well as extending these methods to more complex domains and environment elements other than objects.

REFERENCES

[1] J. J. Gibson, "The theory of affordances," in *Perceiving, Acting, and Knowing: Toward an Ecological Psychology*, R. Shaw and J. Bransford, Eds. Hillsdale, NJ: Lawrence Erlbaum, 1977, pp. 67–82.

[2] J. J. Gibson, *The Ecological Approach to Visual Perception*. Boston, MA: Houghton Mifflin, 1979.

[3] E. Rome, J. Hertzberg, G. Dorffner, and P. Doherty, "06231 executive summary – towards affordance-based robot control," in *Towards Affordance-Based Robot Control*, ser. Dagstuhl Seminar Proceedings, E. Rome, P. Doherty, G. Dorffner, and J. Hertzberg, Eds., no. 06231. Dagstuhl, Germany: Internationales Begegnungs- und Forschungszentrum für Informatik (IBFI), Schloss Dagstuhl, Germany, 2006. [Online]. Available: http://drops.dagstuhl.de/opus/volltexte/2006/725

[4] Fritz, Paletta, Breithaupt, and Rome, "Learning predictive features in affordance based robotic perception systems," in *Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on*, Oct. 2006, pp. 3642 –3647.

[5] G. Metta and P. Fitzpatrick, "Early integration of vision and manipulation," *Adaptive Behavior*, vol. 11, pp. 109–128, 2003.

[6] M. Stark, P. Lies, M. Zillich, J. Wyatt, and B. Schiele, "Functional object class detection based on learned affordance cues," in *6th International Conference on Computer Vision Systems (ICVS)*, Santorini, Greece, 2008 2008, oral presentation.

[7] P. Fitzpatrick, G. Metta, L. Natale, S. Rao, and G. Sandini, "Learning about objects through action - initial steps towards artificial cognition," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, vol. 3, Sept 2003, pp. 3140–3145.

[8] G. Metta and P. Fitzpatrick, "Early integration of vision and manipulation," *Adaptive Behavior*, vol. 11, no. 2, pp. 109–128, 2003, special Issue on Epigenetic Robotics.

[9] A. Stoytchev, "Behavior-grounded representation of tool affordances," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2005.

[10] C. Geib, K. Mourao, R. Petrick, N. Pugeault, M. Steedman, N. Krueger, and F. Wörgötter, "Object action complexes as an interface for planning and robot control," in *Proc. IEEE/RAS Intl. Conf. on Humanoid Robots (Humanoids)*, Genova, Italy, Dec 4-6 2006.

[11] B. Hommel, J. Müsseler, G. Aschersleben, and W. Prinz, "The theory of event coding (TEC): A framework for perception and action planning," *Behavioral and Brain Sciences*, vol. 24, pp. 849–878, 2001.

[12] L. Montesano, M. Lopes, A. Bernardino, and J. Santos-Victor, "Learning object affordances: From sensory-motor coordination to imitation," *IEEE Trans. on Robotics*, vol. 24, no. 1, pp. 15–26, Feb 2008.

[13] J. Sun, J. L. Moore, A. Bobick, and J. M. Rehg, "Learning Visual Object Categories for Robot Affordance Prediction," *The International Journal of Robotics Research*, vol. 29, no. 2-3, pp. 174–197, 2010.

[14] A. Saxena, J. Driemeyer, and A. Y. Ng., "Robotic grasping of novel objects using vision," *International Journal of Robotics Research (IJRR)*, vol. 27, no. 2, pp. 157–173, Feb 2008.

[15] D. K. Quoc Le and A. Y. Ng, "Learning to grasp objects with multiple contact points," in *International Conference on Robotics and Automation (ICRA)*, 2010.

[16] M. T. Mason, "Mechanics and planning of manipulator pushing operations," *The International Journal of Robotics Research*, vol. 5, no. 3, pp. 53–71, September 1986.

[17] J. Scholz and M. Stilman, "Combining motion planning and optimization for flexible robot manipulation," in *International Conference on Humanoid Robotics (ICHR)*, 2010.

[18] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *CVPR*, 2009, pp. 1778–1785.

[19] A. Farhadi, I. Endres, and D. Hoiem, "Attribute-centric recognition for cross-category generalization," in *CVPR*, 2010, pp. 2352–2359.

[20] V. Ferrari and A. Zisserman, "Learning visual attributes," in *NIPS*, 2007, pp. 433–440.

[21] G. Wang and D. Forsyth, "Joint learning of visual attributes, object classes and visual saliency," in *ICCV*, 2009, pp. 537 –544.

[22] N. Loeff, A. Farhadi, I. Endres, and D. Forsyth, "Unlabeled data improves word prediction," in *ICCV*, 2009, pp. 956–962.

[23] A. Opelt, A. Pinz, and A. Zisserman, "Learning an alphabet of shape and appearance for multi-class object detection," *IJCV*, vol. 80, no. 1, pp. 16–44, 2008.

[24] C. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *CVPR*, 2009, pp. 951 –958.

[25] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, "Attribute and simile classifiers for face verification," in *ICCV*, 2009.

[26] D. Kim, J. Sun, S. M. Oh, J. M. Rehg, and A. F. Bobick, "Traversability classification using unsupervised on-line visual learning for outdoor robot navigation," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, May 2006, pp. 518–525.

[27] T. Leung and J. Malik, "Representing and recognizing the visual appearance of materials using three-dimensional textons," *IJCV*, vol. 43, no. 1, pp. 29–44, 2001.

[28] K. Dana, B. Van-Ginneken, S. Nayar, and J. Koenderink, "Reflectance and Texture of Real World Surfaces," *ACM Transactions on Graphics (TOG)*, vol. 18, no. 1, pp. 1–34, Jan 1999.

[29] J. A. Hartigan and M. A. Wong, "A K-means clustering algorithm," *Applied Statistics*, vol. 28, pp. 100–108, 1979.

[30] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Intl. Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[31] J. Zhang, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: a comprehensive study," *International Journal of Computer Vision*, vol. 73, p. 2007, 2007.

[32] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, June 2008, pp. 1–8. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2008.4587756

[33] R. W. Hamming, "Error detecting and error correcting codes," *Bell System Technical Journal*, vol. 26, no. 2, pp. 147–160, 1950.

[34] T. Joachims, "Making large-scale support vector machine learning practical," pp. 169–184, 1999.