Affordance of Object Parts from Geometric Features

Austin Myers, Angjoo Kanazawa, Cornelia Fermuller and Yiannis Aloimonos
Department of Computer Science
University of Maryland
College Park, Maryland 20742
Email: {amyers, kanazawa, fer, yiannis}@umiacs.umd.edu

***Abstract*—As robots begin to collaborate with humans in everyday workspaces, they will need to understand the functions of tools and their parts. To cut an apple or hammer a nail, robots need to not just know the tool's name, but they must localize the tool's parts and identify their functions. In this extended abstract, we give an overview of our work on localizing and identifying object part affordance. We present a framework which provides 3D predictions of functional parts that can be used by a robot, and we introduce a new RGB-D Part Affordance Dataset with 105 kitchen, workshop, and garden tools. We analyze the usefulness of different features, and show that geometry is key for this task. Finally, we demonstrate that the approach can generalize to novel object categories, so robots like PR2, Asimo, and Baxter could use tools never seen before.**

I. Introduction

Imagine Baxter in a kitchen, trying to serve soup from a pot into a bowl. Baxter needs to find a ladle, grab the handle, dip the bowl of the ladle into the pot, and transfer the soup to the serving bowl. But what if the ladle in this kitchen has a different shape and color from the ladles that Baxter has seen before? What if Baxter has never seen any ladles at all? Today, computer vision allows robots to recognize objects from a known category, providing a bounding box around the ladle. However, in these situations Baxter needs to not just detect the ladle, but more importantly he needs to know which part of the ladle he can grasp and which part can contain the soup. As Gibson remarked, "If you know what can be done with a[n] object, what it can be used for, you can call it whatever you please" [5].

Gibson defined affordances as the latent "action possibilities" available to an agent, given their capabilities and the environment [5]. In this sense, for a human adult, stairs afford climbing, an apple affords eating, and a knife affords the cutting of another object. The last example is the most relevant to a robot using tools in a kitchen or workshop, and we use the term *effective affordance* to differentiate such affordances from those in other settings. Affordances in general have long been studied in computer vision and robotics, and most recent works have investigated grasping [9, 11, 3, 4, 6]. These approaches use real or synthetic data with grasping points provided by humans so that a robot can learn to grasp objects in the environment. Similarly, we learn affordances from RGB-D images where affordances are labeled at the pixel level by humans, but we investigate a wide range of objects and affordances. In our experiments we consider two types of grasps and five effective affordances; cut, contain, support, scoop, and pound.

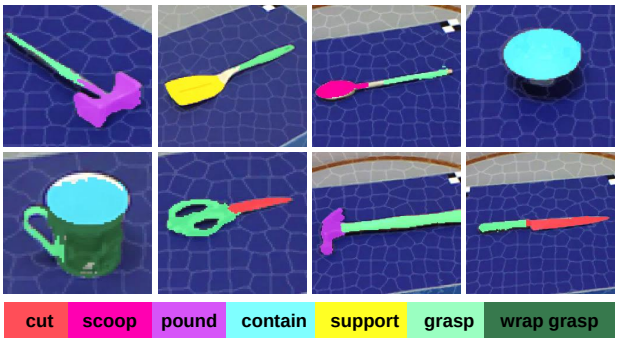

Fig. 1. Example results from our framework for objects of known categories (top row) and novel categories (bottom row).

If robots could identify the effective affordances of parts, then they could use a variety of tools, even tools they have never seen before. In this extended abstract,

- We introduce a framework for localizing and identifying part affordance, and show that it can be used for objects of known and novel categories.
- We compare different feature types for affordance identification, and show that geometric features are key for the task.
- We present a new RGB-D Part Affordance Dataset with ground truth labels for 105 kitchen, workshop, and garden tools from 17 object categories. The dataset and code from this work will be available online [1].

II. Approach

Given an object in an RGB-D image we divide it into a collection of surfaces using a superpixel segmentation algorithm. We assume that foreground objects can be obtained using background subtraction based on the table plane, motion, or more complex attention based reasoning [12]. For each superpixel of an object, we compute hierarchical sparse code features for its pixels and aggregate them using max-pooling. This gives a feature vector for each surface that can be classified with a linear SVM, and provides a prediction of each affordance for each segment. Finally, we refine the predictions and introduce pairwise information between segments by modeling the superpixel neighborhood graph as a conditional random field.

[1] www.umiacs.umd.edu/~amyers/part-affordance-dataset

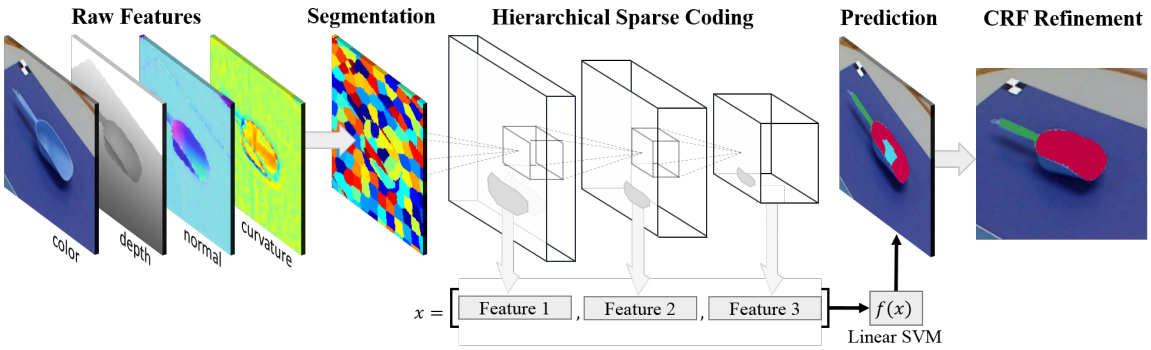

Fig. 2. Our framework for part affordance localization and identification. An RGB-D image is segmented into superpixels, where each segment serves as a candidate part surface (left). For each superpixel, hierarchical sparse code features are extracted from color, depth, normal, and curvature information (middle). Superpixels are classified using a linear SVM, and the final labeling is refined using a CRF (right).

A. Superpixel Segmentation

Man-made tools are typically composed of parts, where each part is a collection of *surfaces* that can provide an effective affordance. We define a surface's effective affordance by the way it comes in contact with the objects that they affect. For example, the inner surface of a cup is "contain" since it contacts the liquid that it holds. The outer surface of the cup on the other hand does not "contain" liquids, but it can be held by a hand using a "wrap-grasp". Since we consider the surfaces that make up object parts, we take a segmentation based approach to affordance identification. We use a modified SLIC [1], incorporating depth and surface normal information, to divide objects in the RGB-D image into small surface fragments. Using color, depth, and surface normals is important to achieve a good segmentation, since affordance parts are usually connected to other surfaces with different affordances but with some properties in common.

B. Geometric Features for Affordance Identification

Our goal is to predict the affordances of these surfaces from their features, such as those from a Kinect sensor. We hypothesize that *there is a deep relationship between effective affordance and geometry* of a part, since the geometric and physical properties of objects are closely tied to the ways they can interact with the environment. To test this, we use feature learning to extract useful representations from each of several feature types; color, grayscale, depth, surface normal, and curvature. In addition to providing an equal footing to compare feature types, recent feature learing methods like [2, 10] have achieved state-of-the-art performance on computer vision tasks. We use a hierarchical sparse coding technique, M-HMP [2], to extract representations for each of the pixel-level features.

For each superpixel, we compute the M-HMP features of its pixels and aggregate them using max-pooling. This gives a feature vector for each surface that can be classified with a linear SVM. To refine the predictions we introduce pairwise information between segments $\vec{S}$ by modeling assignments $\vec{c}$ over the superpixel neighborhood graph $G(S, E)$ as a conditional random field [7]. We model the posterior distribution

$$-\log P(c|G) = \sum_{s_i \in S} \Phi(c_i|s_i) + w \sum_{(s_i,s_j)\in E} \Psi(c_i, c_j|s_i, s_j), \tag{1}$$

where the unary potential Φ is determined by the SVM, and w is a weight on the pairwise potential. The pairwise term is

$$\Psi(c_i, c_j|s_i, s_j) = \left(\frac{B(s_i, s_j)}{1 + ||s_i - s_j||}\right) \delta(c_i \neq c_j). \tag{2}$$

where δ is an indicator function, and $B(s_i, s_j)$ is the length of the shared boundary between s_i and s_j.

III. RGB-D Part Affordance Dataset

We developed a new dataset tailored to everyday tools and the affordances of their parts. The dataset contains 105 kitchen, workshop, and garden tools, and provides pixel-level affordance labels for more than 10,000 RGB-D frames covering a full 360° range of views. These objects were collected from 17 different object categories with 7 affordances: grasp, wrap-grasp, cut, contain, support, scoop, and pound. Examples of the five effective affordances are shown in figure 3. The dataset is also designed so that each affordance is represented by objects from several categories, which permits zero-shot or novel category test settings.

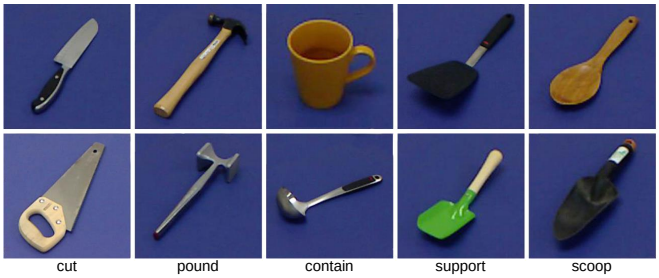

Fig. 3. Objects from the RGB-D Part Affordance Dataset. Each column shows example objects with parts that share the same affordance. The top and bottom rows show example training and testing objects for the novel category setting, respectively.

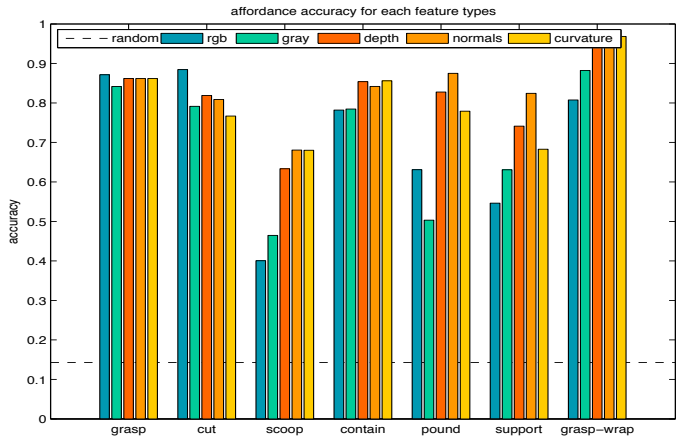

Fig. 4. Comparison of different raw features for each affordance type in the known category setting.

IV. Experiments and Results

We first analyze the effectiveness of different raw feature types in order to test our hypothesis that geometry is related to part affordance. As shown in figure 4, we found that geometric features significantly outperform appearance features for predicting most affordances. This differs from recent results for RGB-D object recognition, which found that visual features outperform geometric features in instance and category recognition [8].

Following these results, we evaluate our framework by testing on objects from known and novel categories. We can see from table I that *Geometry* (depth, normal, and curvature) is superior to *Appearance* (RGB and gray) for both known and novel settings. Even more telling, combining all features does not provide significant improvement, indicating that geometry is key for this task. While the CRF does not give a quantitative improvement, we found that it is an important step for producing an output useful to a robot. As shown in figure 1, the output of this framework provides a robot with precise 3D regions corresponding to affordances, which can be used for further reasoning and manipulation.

TABLE I
Overall results for known and novel settings.

	Appearance	Geometry	All	All + CRF
Known	73.2 ± 3.5	86.5 ± 6.6	86.2 ± 5.6	86.5 ± 5.0
Novel	46.0	63.6	64.8	64.8

V. Conclusions

We introduced a novel problem of localizing and identifying part affordances, and a new dataset designed to address it. We then proposed a framework to predict the affordance of parts for objects of known and completely novel categories. Finally, we showed that geometry is critical for predicting affordance. This new dataset and the results from our experiments open many avenues for future research.

References

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and Sabine Susstrunk. SLIC superpixels compared to state-of-the-art superpixel methods. *PAMI*, 34(11):2274–2282, 2012.

[2] Liefeng Bo, Xiaofeng Ren, and Dieter Fox. Multipath sparse coding using hierarchical matching pursuit. In *CVPR*, 2013.

[3] Jeannette Bohg and Danica Kragic. Grasping familiar objects using shape context. In *Int. Conf. on Advanced Robotics*, 2009.

[4] Abdeslam Boularias, Oliver Kroemer, and Jan Peters. Learning robot grasping from 3-d images with markov random fields. In *Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on*, pages 1548–1553. IEEE, 2011.

[5] James J. Gibson. The theory of affordances. *Perceiving, Acting, and Knowing: Toward and Ecological Psychology*, 1977.

[6] Oliver Kroemer, Emre Ugur, Erhan Oztop, and Jan Peters. A kernel-based approach to direct action perception. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pages 2605–2610. IEEE, 2012.

[7] John Lafferty, Andrew McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, 2001.

[8] Kevin Lai, Liefeng Bo, Xiaofeng Ren, and Dieter Fox. A large-scale hierarchical multi-view rgb-d object dataset. In *ICRA*, 2011.

[9] Ashutosh Saxena, Justin Driemeyer, and Andrew Y Ng. Robotic grasping of novel objects using vision. *The International Journal of Robotics Research*, 27(2):157–173, 2008.

[10] Richard Socher, Brody Huval, Bharath Bhat, Christopher D. Manning, and Andrew Y. Ng. Convolutional-recursive deep learning for 3d object classification. In *NIPS*, 2012.

[11] Michael Stark, Philipp Lies, Michael Zillich, Jeremy L. Wyatt, and Bernt Schiele. Functional object class detection based on learned affordance cues. In *International Conference on Computer Vision Systems (ICVS)*, May 2008.

[12] Ching L Teo, Austin Myers, Cornelia Fermuller, and Yiannis Aloimonos. Embedding high-level information into low level vision: Efficient object search in clutter. In *ICRA*, pages 126–132. IEEE, 2013.