Robotic Affordance Learning: Old Ideas, Recent Developments, and Potential Paths Forward

Barry Ridge[1]

[1] *Humanoid and Cognitive Robotics Lab, Department of Automatics, Biocybernetics and Robotics, Jožef Stefan Institute, Ljubljana, Slovenia.* barry.ridge at ijs.si

1 Introduction

The notion of *affordances* as a concept originated in the late 70's in the field of ecological psychology as founded by Gibson and others [1], and attempts to enable robots to autonomously learn the affordances in their environments, as recent survey papers reveal [2, 3], emerged not long after in the early 90's, reaching a particularly vibrant zenith in recent years. While many of these works have borne fruit in restricted experimental settings, where the environment and robotic interactions can be controlled and guided towards certain goals, the abiding concern of designing robots capable of general and continuous affordance learning remains elusive. Meanwhile, recent experiments using deep neural networks in the fields of machine learning and computer vision [4, 5] have demonstrated that given sufficient data, computational resources, and algorithmic proficiency, impressive results can be achieved when it comes to general representations and inference. This naturally raises many questions regarding the current state of the field of robotic affordance learning and how best to push it forward.

2 Recent Developments

In recent years, many angles of attack have been exploited in the assault on the affordance learning problem, including, but not limited to, learning affordance-predictive object properties, learning to represent affordances via the effects of actions, learning affordances that emerge from different types of actions such as pushing and grasping, learning multi-object affordance relations, and so on. Many of these approaches are comprehensively discussed in the recent survey paper of Jamone *et al.* [3]. Our own recent work in the area of object push affordance learning has focused on designing object representations that marry robot trajectory information to object shape features derived from 3D point clouds, in what we refer to as *action-grounded features*. The main idea behind this approach, which relies on dynamically defining local shape features with respect to a reference frame defined by the pushing action, is that affordance learning can benefit from representations that intrinsically encode differences between objects depending on how they are interacted with.

2.1 Action-Grounded Viewpoint Feature Histogram

To this end, we have recently developed a modified form of the *viewpoint feature histogram* by Rusu *et al.* [6] specifically designed for object affordance prediction, which we call the *action-grounded viewpoint feature histogram (AGVFH)*. We make two key changes to the original VFH descriptor, modifying both the shape component and the viewpoint component respectively, in order to ground the descriptor with respect to a pushing action that can either come from recorded training data of robotic object pushes or planned pushes. In the case of the shape component, instead of using the object centroid as the central point for the SPFH computation as in the original descriptor, we now use the push contact point. In addition, we use the push direction normal as the basis for forming the Darboux frame as visualised in Figure 1.

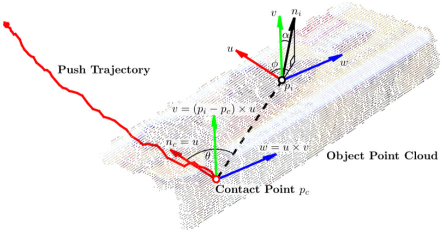

Figure 1: Action-grounded Darboux frame construction example for the shape component of AGVFH.

2.2 Action-Grounded Octree Shape Features

We have also developed an *action-grounded octree shape features (AGOSF)* representation that is similar to our original action-grounded shape feature descriptor as proposed in [7], but uses octrees to decompose the point cloud into part cells instead of separately subdividing along the axes via partitioning planes. Thus, an arbitrary level of detail may be encapsulated by the representation depending on the octree depth level. This octree subdivision process is illustrated for a segmented object point cloud sample in Figure 2. The octree subdivision is similar in nature to that of the OMS-EGI descriptor by Mar *et al.* [8], but our proposed method differs in the types of features which are encoded in each of the octree cells. In each of the cells, we derive local surface normal estimates, centroid components, point counts and local curvature estimates.

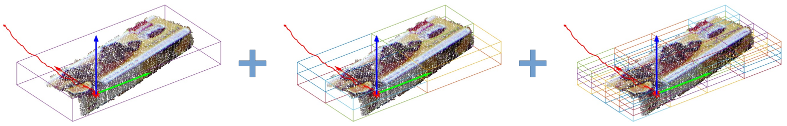

Figure 2: Action-grounded octree shape feature (AGOSF) 3-level octree decomposition of a sample segmented object point cloud.

Table 1: 10-Fold CV Random Forests Classifier Results

Features	**F-Score**
VFH (contact point as viewpoint) [6] + action features	0.4725
AGVFH (cf. Sec. 2.1)	0.8889
OMS-EGI [8]	0.6480
AGOSF (cf. Sec. 2.2)	**0.9526**

2.3 Results

Table 1 shows some recent results comparing the two feature descriptors described above against the original VFH descriptor [6] and the recently proposed OMS-EGI descriptor [8] on an expanded version (one additional object) of the object push affordance learning dataset described in [7]. These results are encouraging in the sense that they appear to support the idea that tight coupling of actions and object properties in feature representations may, in turn, be advantageous to affordance representation and learning, which is an idea that is promulgated in many affordance formalisms [3].

3 Looking to the Future

Although the above methods provide an interesting basis for the development and evaluation of certain ideas as they pertain to affordance learning, e. g. action-grounding and multi-scale representation via octrees, there are clear possibilities for improvement going forward. For instance, rather than hand-designing the features, as was the case in the above, it would likely be beneficial to take a more data-driven approach to the feature representation using deep neural networks, particularly convolutional neural networks (CNNs), which would learn to represent features inherent in the data at multiple scales. This would, however, present the challenge of using a sufficient amount of the right kind of data to train these networks, and given the additional complication of requiring both object point cloud data and robot trajectory data in the action-grounded setting described above, such data could prove to be difficult to acquire.

3.1 Deep Learning

Although there have been a number of high-profile recent successes in the development and application of deep learning approaches both in vision [4] and robotics [5], most of these have been restricted to 2D image domains. Thus, given the fact that affordance learning occurs in a 3D world where the relationships between actions, objects and effects might best be described using 3D representations, perhaps one of the most promising recent developments comes in the guise of volumetric shape representations via CNNs such as *3DShapeNets* [9]. If such representations could be effectively action-grounded in a similar way as described above, they have the potential to be quite powerful, not least because they allow for the recovery of full 3D shape from partial views of objects, thus potentially alleviating problems with pushes coming from occluded object sides.

3.2 Alleviating the Dearth of Data

The high cost of entry in terms of engineering expertise, hardware resources and time, invariably restricts the capacity of researchers to gather the quantities of data of the appropriate nature that might be required to push the field of affordance learning forward beyond potential stagnation. This situtation is changing however. The advent of multi-robot data-gathering farms means that unprecedentedly large datasets that include RGB, depth, and trajectory data of robots performing exploratory grasps and pushes are now becoming available to researchers worldwide [10, 5]. Encouraging the gathering and use of such large-scale datasets could be of significant benefit if the field wishes to move forward beyond the more simplified, restricted experimental settings in which it found its origins.

References

[1] J. J. Gibson, *The Ecological Approach to Visual Perception*. Houghton Mifflin, 1979.

[2] T. E. Horton, A. Chakraborty, and R. S. Amant, "Affordances for robots: a brief survey," *Avant*, vol. 3, no. 2, pp. 70–84, 2012.

[3] L. Jamone, E. Ugur, A. Cangelosi, L. Fadiga, A. Bernardino, J. Piater, and J. Santos-Victor, "Affordances in psychology, neuroscience and robotics: a survey," *Accepted for publication in IEEE Transactions on Cognitive and Developmental Systems*, 2016.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[5] C. Finn, I. Goodfellow, and S. Levine, "Unsupervised Learning for Physical Interaction through Video Prediction," in *arXiv preprint arXiv:1605.07157 (Accepted for publication at NIPS 2016)*, 2016.

[6] R. Rusu, G. Bradski, R. Thibaux, and J. Hsu, "Fast 3d recognition and pose using the Viewpoint Feature Histogram," in *Proc. 23rd IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, (Taipei, Taiwan), pp. 2155–2162, Oct. 2010.

[7] B. Ridge and A. Ude, "Action-grounded push affordance bootstrapping of unknown objects," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2791–2798, Nov. 2013.

[8] T. Mar, V. Tikhanoff, G. Metta, and L. Natale, "Multi-model approach based on 3d functional features for tool affordance learning in robotics," in *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*, pp. 482–489, Nov. 2015.

[9] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao, "3d ShapeNets: A deep representation for volumetric shapes," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1912–1920, June 2015.

[10] "Google Brain Robotics Data." `https://sites.google.com/site/brainrobotdata/home`. Accessed: 2016-08-23.