Unsupervised Learning of Affordance Coordinate Frame for Robotic Task Generalization

Zhen Zeng
Electrical and Computer Engineering
University of Michigan
Ann Arbor, United States
zengzhen@umich.edu

Pranav Suhas Joshi
Electrical and Computer Engineering
University of Michigan
Ann Arbor, United States
pranavsj@umich.edu

Odest Chadwicke Jenkins
Computer Science, Robotics Institute
University of Michigan
Ann Arbor, United States
ocj@umich.edu

***Abstract*—It is important to generalize robotic tasks over novel object instances. We represent robotic tasks by end-effector manipulation trajectories in an object-centric coordinate frame. Given demonstrations of a particular task, we aim to learn an object-centric coordinate frame that can be registered on point cloud of novel object instances, such that the demonstrated trajectories can be easily transferred. An example of task transfer based on object-centric coordinate frame is as shown in Figure 1, such object-centric coordinate frame helps establish the correspondences of affordances across object instances, and we call it *Affordance Coordinate Frame (ACF)*. It is super expensive to manually label *ACF* for supervised learning, thus we propose to learn *ACF* through deep unsupervised learning. Our loss function is defined based on the underlying assumption that objects or object parts that afford the same task are similarly operated through similar manipulation trajectories. At task execution time, *ACF* is first registered on the novel object instance, then the manipulation trajectory can be generated by sampling from the probability distribution of the demonstrated trajectories, which can be modeled by Gaussian Mixture Models.**

***Index Terms*—affordance learning, task generalization, deep unsupervised learning**

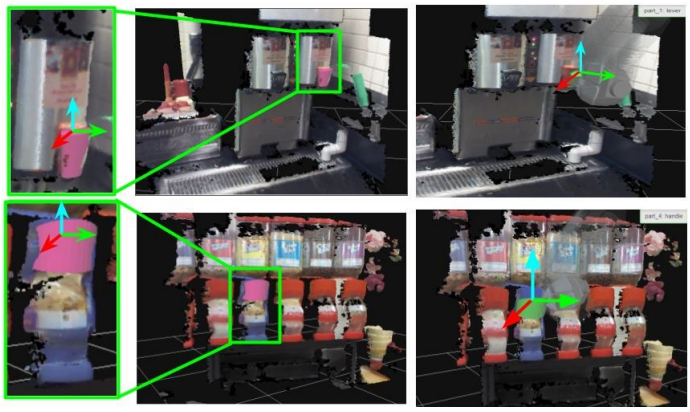

Fig. 1: Example of *Affordance Coordinate Frame (ACF)* of a coffee machine (*upper*) and a cereal dispenser (*lower*) for a handle pulling task. *left*: object point cloud with *ACF* attached; *middle*: full scene point cloud; *right*: end-effector trajectory for pulling handle collected using Robobarista platform [5].

I. Introduction

Given demonstrations of a particular task, robots should be able to generalize it over novel object instances. In our work, we represent tasks in the form of robot end-effector manipulation trajectories. These trajectories are represented in an object-centric frame such that these trajectories stay invariant to object poses. Our key to generalize tasks to a novel object instance is to predict the object-centric frame given the object point cloud, then the demonstrated manipulation trajectories can be transferred to the the novel object instance.

Figure 1 illustrates an example of object-centric frame for a handle pulling task. As we can see, the object-centric frame establishes the correspondences of affordances between different object instances, thus we call such object-centric frame the *Affordance Coordinate Frame* (*ACF*). Our contribution is predicting *ACF* given object point cloud for robotic task generalization based on deep unsupervised learning, without requiring expensive manual labeling of object affordances.

Our approach to automatically extracting *ACF* from object point cloud is based on deep neural networks. The input to the network is object point cloud, the output is an *ACF* registered on the object point cloud. At training time, we also provide demonstrated end-effector trajectories for each object point cloud in the training set. The loss function is designed such that demonstrated trajectories (each represented in the associated *ACF* as predicted by the network) are similar to each other, i.e., small distance measure. This is based on the underlying assumption that objects (or object parts) that afford the same task are similarly operated, leading to similar end-effector trajectories. Note that the distance measure of the end-effector trajectories should be robust to noise, and we rely on dynamic time warping for a reasonable distance measure. At test time, the robot first predicts the *ACF* given the object point cloud, and then a manipulation trajectory can be generated again based on techniques from trajectory LfD, such as sampling from a probability distribution of the demonstrated trajectories modeled via Gaussian Processes, Gaussian Mixture Models, and etc.

II. Related Work

Existing works on object affordance detection has made use of deep learning for more robust detection. Kokic et al. [2] deployed deep neural networks for object affordance detection on partial or complete point cloud. They were able

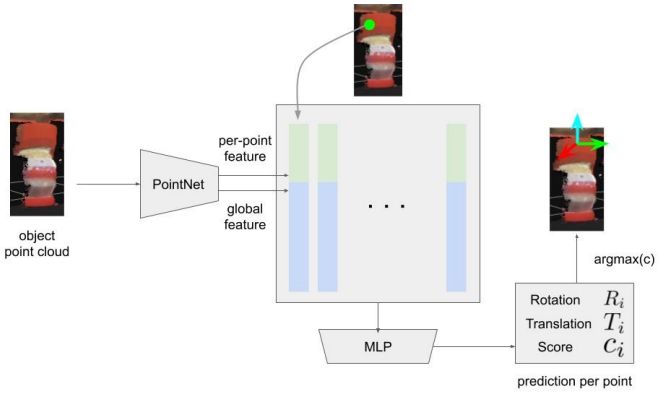

Fig. 2: Network architecture for predicting *ACF*.

to determine task dependent grasps based on the affordance detection. Myers et al. [3] proposed two methods for affordance detection on kitchen, workshop and garden tools. These methods use geometric features of image to predict per-pixel affordance scores. Above mentioned methods use supervised learning methods, which require labelled data that is expensive to acquire. We do not require manual labeling, instead we rely on unsupervised learning for learning object affordances towards task generalization.

Although object affordance detection works provide pixel-wise or voxel-wise object affordance labels, such as "cut", "poke", theses labels cannot be directly translated into how a robot should perform a cutting or a poking task. Instead, manipulation trajectories or policy can be used for robot to perform tasks. Sung et al. [5] learned the compatibility between object point cloud, manipulation trajectory and language, in order to select a manipulation trajectory for robotic task execution. Similar to our work, the manipulation trajectory is defined in an object-centric frame estaliblished through PCA analysis on the 3D point cloud, and this severely limits the generalizability of the manipulation trajectory across different object instances. Dang et al. [1] transferred learned manipulation trajectories through ICP based shape matching technique, which is vulnerable to 3D shape variances across object instances. Instead, our work relies on neural networks to extract the task-relevant features of object point cloud for establishing the object-centric frame.

Zech et al. [7] provides a survey on the computational models of affordances, they made an important observation that affordance models at *local* level exhibit better generalization ability, where object parts are being modeled for representing affordances. Our approach is capable of registering an object-centric frame for manipulation on 3D sub-structure of object point cloud, thus associating object parts with affordances.

We use PointNet as proposed by Qi et al. [4] to extract 3D features given point cloud data. PointNet has shown capability on classification and segmentation, and object pose estimation [6] of object point cloud, and the network is invariant to the order of point clouds, can capture local interactions between points and is invariant to coordinate transformations.

III. Affordance Coordinate Frame

We introduce *Affordance Coordinate Frame (ACF)* as a way to transfer manipulation trajectory for robotic task generalization. The *ACF* is an object-centric frame that is registered on a given object point cloud P_i, and the robotic task is represented by end-effector trajectories $\tau_i = \{\tau_i^k\}_{k=1}^{l_i}$ operated on the object point cloud P_i. Robot end-effector trajectory τ_i is recorded in a global reference frame in the workspace during demonstration.

IV. Deep unsupervised Learning of Affordance Coordinate Frame

We use deep neural networks for unsupervised learning of *ACF* . As shown in Figure 2, the input is the object point cloud, and the output is the predicted *ACF* . In order to deal with point cloud data, we use PointNet [4] for extracting 3D features. We output a prediction of the *ACF* for each point with a score, similarly to [6]. The loss function is defined as

$$L = \sum_{\tau_i, \tau_j \in T} D(M_i \tau_i, M_j \tau_j) \tag{1}$$

where T is the set of all demonstrated trajectories for a particular task, and M_i, M_j are the transformation matrix representing the predicted *ACF* for input point cloud P_i, P_j respectively. Note that τ_i, τ_j are trajectories associated with P_i, P_j respectively, thus $M_i\tau_i, M_j\tau_j$ are demonstrated trajectories transformed into the *ACF* . We use the distance function $D(\cdot, \cdot)$ between two trajectories as defined by Sung et al. [5], which is based on dynamic time warping.

References

[1] H. Dang and P. K. Allen. Robot learning of everyday object manipulations via human demonstration. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1284–1289. IEEE, 2010.

[2] M. Kokic, J. A. Stork, J. A. Haustein, and D. Kragic. Affordance detection for task-specific grasping using deep learning. In *2017 IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids)*, pages 91–98. IEEE, 2017.

[3] A. Myers, C. L. Teo, C. Fermüller, and Y. Aloimonos. Affordance detection of tool parts from geometric features. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1374–1381. IEEE, 2015.

[4] C. R. Qi, H. Su, K. Mo, and L. J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 652–660, 2017.

[5] J. Sung, S. H. Jin, and A. Saxena. Robobarista: Object part based transfer of manipulation trajectories from crowd-sourcing in 3d pointclouds. In *Robotics Research*, pages 701–720. Springer, 2018.

[6] C. Wang, D. Xu, Y. Zhu, R. Martn-Martn, C. Lu, L. Fei-Fei, and S. Savarese. Densefusion: 6d object pose estimation by iterative dense fusion. *arXiv preprint arXiv:1901.04780*, 2019.

[7] P. Zech, S. Haller, S. R. Lakani, B. Ridge, E. Ugur, and J. Piater. Computational models of affordance in robotics: a taxonomy and systematic classification. *Adaptive Behavior*, 25(5):235–271, 2017.