Title

Decoupling Vision and Motion: Object-Centric Representations for Enhanced Manipulation

Abstract

We introduce a novel approach for manipulating articulated objects with the ability to generalize across object translations, scaling, and robot kinematics. In traditional imitation learning scenarios, the learned policy fails if camera perspectives or the relative positions of objects change during inference, and is specific to certain robot configurations. To tackle these challenges, we propose an object-centric representation for both observations of object motions and robot actions, which decouples vision-based decision making from robot execution. The center here refers to a canonical frame of the articulated object that remains invariant regardless of the contact point. We further use object-centric imitation learning to make stable predictions under varying camera views, object placements, and robot kinematics. Experiments and analysis demonstrate that our method achieves state-of-the-art performance in articulated object manipulation and dramatically improves generalization performance.

Overview

Question: The core problem in robot imitation learning is the strong dependency between a robot’s learned policy and its specific configuration, camera perspectives, and the relative positions of objects. This makes it difficult to generalize the robot’s learned skills to new situations.

Motivation: This problem is crucial in enabling robots to perform tasks in real-world environments. For example, a service robot may need to open doors of varying sizes and placements within its environment, and we want the policy to work cross embodiment. Coupling learning of visual perception and robot control makes generated action highly overfit to noise and task-irrelevant inputs. By developing methods for object-centric learning, we decouple observed object motions and robot actions to make our policy generalize over various object translation, scaling and robot configurations.

Challenge: The current state-of-the-art robot learning methods often rely on pixel-based visual inputs and trajectories anchored in the robot's base frame. However, we want to decouple vision-based decision making and robot execution.

Approach: Inspired by the grasp detection algorithm, we want to "detect" object-centric executable trajectories for door opening. Object-centric means the trajectories are attached to the object itself, irrelevant to specific robot configurations, and we study observations of object motions irrelevant to the object’s placements or camera views.

Related Work

Learning from Human Demonstrations

Recent studies focus on imitation learning including utilizing Transformer architectures [16, 17, 18, 20, 21, 22] to predict robot actions. To break the limitations of 2D observations, the 3D policy models C2F-ARM [19] and PerAct [16] voxelize the robot's workspace and are trained to identify the specific 3D voxel that contains the next key pose for the end effector. ACT3D [23] represents the scene as a continuous resolution 3D feature field rather than voxel. Another method is to apply diffusion models to robotic applications. Chi et al. [24] introduced the groundbreaking Diffusion Policy (DP), while 3D-DP [25] learns to predict full trajectories by employing a modified MLP as an efficient backbone for processing point clouds. A combination can be seen in ChainedDiffuser, which [26] first predicts high-level key points using ACT3D, then generates low-level trajectories that connect these key points through standard trajectory diffusion techniques. We show in our paper that the actions can be directly generated as the point features based on our object-centric representation.

Articulated Object Manipulation

Manipulating articulated objects is challenging, with some approaches using analytical methods [2, 3, 4, 5, 6] and data-driven learning-based methods [7,8], including techniques to model manipulation with explicit parameters [9, 10, 11, 12], visual affordances [7, 13, 14] and articulation flows [1, 15]. Different from previous learning-based works which tend to learn the representation through a pure object model, we incorporate contact point concept and leverage the point for prediction.

Problem Formulation

The task we aim to solve is to open an articulated object from a single-view depth sensor. At a high level, our objective is to decouple agents' observed object motions with its actions in imitation learning, and then to use the learned policy to perform generalization across the object translations, scaling and robot configuration. We focus on opening an articulated pivot-hinge door, supposing there is a handle in the door and the door can be directly pulled or pushed to open.

We propose an object-centric representation to learn the robot's observation and action. To be more specific, in the door-opening task, the robot first establishes and maintains contact with the door before manipulating the door's 1 degree of freedom (DoF) rigid plane along its pivot axis. The object's configuration and geometry property provides a possible frame definition, and the contact point serves as an anchor for this frame, together formulating our object-centric here.

We aim to train an imitation learning policy $\pi^*(a|o)$ based on the demonstration dataset $D = \{(o_t, a_t)\}_{t=1}^{N}$ .To obtain our object-centric representation of both the object motions and robot actions, we define a frame $A \in SE(3)$ attached to the contact point $p^* \in P$. When the robot opens the door, we detect the contact points $p^* = (x^*, y^*, z^*)$, and the rotation

matrix ${}^{A}_{C}R$ based on the object-centric definition. Later the robot actions are transformed to the frame A. During inference, we predict per-point binary contact possibility and robot actions in frame A. The robot actions predicted on contact points are then transformed in robot base frame $B \in SE(3)$ for execution.

Subproblems(solution)

We collect a demonstration dataset $D = \{(o_t, a_t)\}_{t=1}^{N}$ where ${}^{o}{}_{t}$ is in the camera frame $C \in SE(3)$ and ${}^{a}{}_{t}$ is in the robot base frame $B \in SE(3)$. In this work, O_t is a 3D point cloud $P = \left\{ p_k \in \mathbb{R}^3 \right\}_{k \in [n]}$ captured by depth cameras and ${}^{a}{}_{t}$ is the end effector pose.

In the door-opening task, we define the contact frame A as following: Z-axis is aligned with the door's pivot hinge axis. X-axis is perpendicular to the manipulated surface of the door, which is also the surface to which the handle is attached. The Y-axis is computed as the cross product of the X and Z axes, i.e., $Y = X \times Z$.

For each point ${}^{p}{}_{k} \in P$, we identify its frame ${}^{F}{}_{k}$ and obtain actions ${}^{a^k}_{t:t+H}$ under it. A per-point flow ${}^{f}{}_{k} = \left({}^{C}{}_{k}, {}^{a^k}{}_{t:t+H} \right)$ is defined, where ${}^{C}{}_{k} \in \{0, 1\}$ indicates whether the point is a contact point. Thus, we would like to find a function ${}^{f}{}_{\theta}\left({}^{o}{}_{t} \right)$ that predicts the 3D trajectory flow directly from point cloud observations. We define the objective of minimizing the L_2 error of the predicted flow: $L_{MSE} = \sum_k \left\| {}^{f}{}_{k} - {}^{f}{}_{\theta}\left({}^{o}{}_{t} \right)_k \right\|_2$.

During inference, we consider $\hat{a}^{k}_{t:t+H}$ where $k \in \left\{ k \middle| {}^{C}{}_{k} = 1 \right\}$ and transform the action from ${}^{F}{}_{k}$ to base frame B for execution.

References

[1] B. Eisner, H. Zhang, and D. Held. Flowbot3d: Learning 3d articulation flow to manipulate articulated objects. arXiv preprint arXiv:2205.04382, 2022.

[2] R. Mart´ın-Mart´ın, S. H¨ofer, and O. Brock. An integrated approach to visual perception of articulated objects. In 2016 IEEE international conference on robotics and automation (ICRA), pages 5091–5097. IEEE, 2016.

[3] K. Desingh, S. Lu, A. Opipari, and O. C. Jenkins. Factored pose estimation of articulated objects using efficient nonparametric belief propagation. In 2019 International Conference on Robotics and Automation (ICRA), pages 7221–7227. IEEE, 2019.

[4] B. Abbatematteo, S. Tellex, and G. Konidaris. Learning to generalize kinematic models to novel objects. In Proceedings of the 3rd Conference on Robot Learning, 2019.

[5] R. Staszak, M. Molska, K. Młodzikowski, J. Ataman, and D. Belter. Kinematic structures estimation on the rgb-d images. In 2020 25th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA), volume 1, pages 675–681. IEEE, 2020.

[6] A. Jain, R. Lioutikov, C. Chuck, and S. Niekum. Screwnet: Category-independent articulation model estimation from depth images using screw theory. In 2021 IEEE International Conference on Robotics and Automation (ICRA), pages 13670–13677. IEEE, 2021.

[7] F. Xiang, Y. Qin, K. Mo, Y. Xia, H. Zhu, F. Liu, M. Liu, H. Jiang, Y. Yuan, H. Wang, et al. Sapien: A simulated part-based interactive environment. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 11097–11107, 2020.

[8] K. Mo, S. Zhu, A. X. Chang, L. Yi, S. Tripathi, L. J. Guibas, and H. Su. Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 909–918, 2019.

[9] L. Yi, H. Huang, D. Liu, E. Kalogerakis, H. Su, and L. Guibas. Deep part induction from articulated object pairs. arXiv preprint arXiv:1809.07417, 2018.

[10] X. Li, H. Wang, L. Yi, L. J. Guibas, A. L. Abbott, and S. Song. Category-level articulated object pose estimation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 3706–3715, 2020.

[11] X. Wang, B. Zhou, Y. Shi, X. Chen, Q. Zhao, and K. Xu. Shape2motion: Joint analysis of motion parts and attributes from 3d shapes. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8876–8884, 2019.

[12] R. Hu, W. Li, O. Van Kaick, A. Shamir, H. Zhang, and H. Huang. Learning to predict part mobility from a single static snapshot. ACM Transactions on Graphics (TOG), 36(6):1–13, 2017.

[13] Z. Xu, H. Zhanpeng, and S. Song. Umpnet: Universal manipulation policy network for articulated objects. IEEE Robotics and Automation Letters, 2022.

[14] R. Wu, Y. Zhao, K. Mo, Z. Guo, Y. Wang, T. Wu, Q. Fan, X. Chen, L. Guibas, and H. Dong. Vat-mart: Learning visual action trajectory proposals for manipulating 3d articulated objects.arXiv preprint arXiv:2106.14440, 2021.

[15] H. Zhang, B. Eisner, and D. Held. Flowbot++: Learning generalized articulated objects manipulation via articulation projection. arXiv preprint arXiv:2306.12893, 2023.

[16] M. Shridhar, L. Manuelli, and D. Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. In Conference on Robot Learning, pages 785–799. PMLR, 2023.

[17] P.-L. Guhur, S. Chen, R. G. Pinel, M. Tapaswi, I. Laptev, and C. Schmid. Instruction-driven history-aware policies for robotic manipulations. In Conference on Robot Learning, pages 175–187. PMLR, 2023.

[18] H. Liu, L. Lee, K. Lee, and P. Abbeel. Instruction-following agents with jointly pre-trained vision-language models. arXiv preprint arXiv:2210.13431, 2022.

[19] S. James, K. Wada, T. Laidlow, and A. J. Davison. Coarse-to-fine q-attention: Efficient learning for visual robotic manipulation via discretisation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 13739–13748, 2022.

[20] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. arXiv preprint arXiv:2212.06817, 2022.

[21] N. M. Shafiullah, Z. Cui, A. A. Altanzaya, and L. Pinto. Behavior transformers: Cloning kmodes with one stone. Advances in neural information processing systems, 35:22955–22968, 2022.

[22] S. Reed, K. Zolna, E. Parisotto, S. G. Colmenarejo, A. Novikov, G. Barth-Maron, M. Gimenez, Y. Sulsky, J. Kay, J. T. Springenberg, et al. A generalist agent. arXiv preprint arXiv:2205.06175, 2022.

[23] T. Gervet, Z. Xian, N. Gkanatsios, and K. Fragkiadaki. Act3d: 3D feature field transformers for multi-task robotic manipulation. In Conf. on Robot Learning, 2023.

[24] C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. Burchfiel, and S. Song. Diffusion policy: Visuomotor policy learning via action diffusion. In Robotics: Science and Systems, 2023.

[25] Y. Ze, G. Zhang, K. Zhang, C. Hu, M. Wang, and H. Xu. 3D diffusion policy: Generalizable visuomotor policy learning via simple 3D representations. In Robotics: Science and Systems, 2024.

[26] Z. Xian, N. Gkanatsios, T. Gervet, T.-W. Ke, and K. Fragkiadaki. ChainedDiffuser: Unifying trajectory diffusion and keypose prediction for robotic manipulation. In Conf. on Robot Learning, 2023.