Contextual Reinforcement Learning of Visuo-tactile Multi-fingered Grasping Policies

Visak Kumar[1]*, Tucker Hermans[2], Dieter Fox[3], Stan Birchfield[3] and Jonathan Tremblay[3]
[1]School of Interactive Computing at Georgia Institute of Technology, `visak3@gatech.edu`
[2]Robotics Center and the School of Computing at University of Utah, `thermans@cs.utah.edu`
[3]NVIDIA, `{dieterf, sbirchfield, jtremblay}@nvidia.com`

Abstract

We propose Grasping Objects Approach for Tactile (GOAT) robotic hands to learn control policies in simulation to be deployed on real robots. In our approach, we use real human hand motion demonstrations to initialize and reduce the search space for learning. We contextualize our policy with the bounding cuboid dimensions of the object of interest, which allows the policy to work on a more informative representation than directly using an image or point cloud. Finally, leveraging fingertip touch sensors in the hand allows the policy to overcome the reduction in geometric information introduced by the coarse bounding box, as well as pose estimation uncertainty. We show our simulation-only learned policy successfully runs on a real robot without any fine tuning, thus bridging the reality gap.

1 INTRODUCTION

Enabling robots to autonomously grasp objects of varying shape and size with multi-fingered hands stands as a fundamental challenge. This skill is necessary for more advanced robot tasks such as pick-and-place, human handover, dexterous tool use, *etc*. Classical solutions to this problem take a model-based planning and control approach. A typical pipeline estimates the object pose, given either a 3D point cloud or mesh of the object, then plans a set of contact locations and hand configuration to define the grasp, and finally generates a motion plan to reach and grasp the object.

In this work we explore a different approach by learning a policy to grasp objects varying in geometry and scale with a multi-finger gripper using deep reinforcement learning (RL). A few important challenges arise in formulating the multi-fingered grasping problem as an RL problem. First, how to cope with the relatively high dimensionality of the multi-fingered hand's configuration space in order to effectively explore the space of possible grasping policies? Second, how should the learner represent the object to be grasped in a way that can effectively generalize across objects of varying shape, while still being succinct enough to train efficiently? Third, how can we learn such a policy purely in simulation with no need to fine tune the policy for use in the physical world?

In order to efficiently search over the high-dimensional space of grasping policies, we leverage recent advancements in camera-based human hand pose estimation [4] and imitation learning [16] to provide human grasping demonstrations from an RGB camera. We use these grasping demonstrations as a component in our reward function, providing a prior for preferred grasping trajectories to the learner in simulation. We address the problems of object representation and sim-to-real transfer by proposing a bounding-box based state-object representation. This work makes the following contributions:

- We present a system that leverages human demonstrations of grasping, reinforcement learning and sim-to-real to accomplish a multi-finger grasp task on a real-world system. We

[*] Work performed while the first author was an intern at NVIDIA.

NeurIPS 2019 Workshop on Robot Learning: Control and Interaction in the Real World, Vancouver, Canada

demonstrate that our system generalizes to unseen shapes in the real-world without any fine tuning.

- We introduce a novel approach to combining visual and tactile information in learned grasp policies, using 3D keypoints for context variables encoding object shape and binary contact signals as part of the state input to the policy. This allows our policy to reason about the object size and orientation implicitly creating a versatile policy that can adapt locally by leveraging the sensed contact information.

2 Method

We formulate the task of multi-finger grasping as a contextual policy search problem [7]. This differs from the classic Markov Decision Process (MDP) [19] in that the agent (robot) observes some context variable $\boldsymbol{\kappa}$ at the beginning of the episode which parameterizes the reward function $r : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$; where $\mathcal{S}$ and $\mathcal{A}$ define the state action spaces respectively. The objective of the contextual policy search problem remains the same as standard reinforcement learning, namely to find a policy $\pi : \mathcal{S} \rightarrow \mathcal{A}$, that maximizes the expected accumulated reward, conditioned on the observed context $\boldsymbol{\kappa}$:

$$J(\pi_{\boldsymbol{\theta}}) = \mathbb{E}_{\mathbf{s}_0,\mathbf{a}_0,\ldots,\mathbf{s}_T,\mathbf{a}_T} \sum_{t=0}^{T} \gamma^t r(\mathbf{s}_t, \mathbf{a}_t; \boldsymbol{\kappa}), \tag{1}$$

where $\mathbf{s}_0 \sim p_0$, $\mathbf{a}_t \sim \pi_{\boldsymbol{\theta}}(\mathbf{s}_t; \boldsymbol{\kappa})$, and $\mathbf{s}_{t+1} = \mathcal{T}(\mathbf{s}_t, \mathbf{a}_t)$. The remaining components of the MDP also exist in our problem formulation, specifically $\mathcal{T} : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{S}$ is the transition function, p_0 is the initial state distribution and $0 < \gamma < 1$ is the discount factor. We additionally make explicit the policy parameters $\boldsymbol{\theta}$ which we seek to learn through rollouts of the system.

We define the context variable, $\boldsymbol{\kappa}$, for our multi-fingered grasping problem as the keypoints of a bounding box of the object of interest at the beginning of the episode. This defines a low dimensional feature representation to encode the object geometry. There are several ways to infer these features at runtime such as using pose estimation of known objects [21]. By providing this information of the object's pose only at the beginning of the trial, we remove the need to explicitly track the object during execution.

In addition to localizing the object, we hypothesize that contact information provides extremely useful information in learning grasps that can generalize across different object geometries. The state space includes the Cartesian palm pose (in the robot base frame) denoted by $P_{xyz} \in \mathbb{R}^3$ and orientation $u \in SO(3)$ represented as a quaternion, joint positions and velocities of the 16 DOF four-fingered hand represented as $q_h \in \mathbb{R}^{16}$ and $\dot{q}_h \in \mathbb{R}^{16}$, and contact vector $c \in \mathbb{Z}_2^4$ which contains binary contact information about the four fingertips. This results in final state space of dimension 43. The context variable κ is 24 dimensional, containing the Cartesian $x, y, z \in \mathbb{R}^3$ locations of each of the eight corners of a bounding cuboid in the robot base frame. We define the robot action space as the desired 6-DOF hand pose and the desired joint positions of the fingers. Thus, the action space has 22 dimensions.

The task of reaching and grasping a wide range of objects with a multi-fingered hand is a challenging task to learn. To solve this problem, the reward function is the sum of three terms: 1) We minimize the distance from the robot palm to the top of the object; 2) We reward the policy when the robot's fingertip locations track the fingertip locations obtained from the recorded human demonstration; and 3) Finally, we provide a binary reward for each finger that makes contact with the object. Each of the reward terms mentioned above helps in learning, hence the weight coefficients were tuned such that each term has relatively equal magnitude while learning.

We use the proximal policy optimization (PPO) [18] algorithm to learn the policy. We represent the policy as a multi-layered perceptron (MLP) with 2 hidden layers each containing 128 neurons. During training, at the beginning of each rollout, we generate a new synthetic cuboid object with dimensions and pose uniformly sampled from a pre-specified range, and the keypoint locations (after adding slight perturbations to simulate sensor noise present in the physical system) are passed as context input to the policy. These keypoint values then remain the same throughout that rollout. Since we wish to deploy the policy learned in simulation on a real robot, domain randomization is applied to the objects to account for the discrepancy between the simulator and physical world. Uniform noise is added to the object mass, friction coefficients between the fingers and object, PD gains of

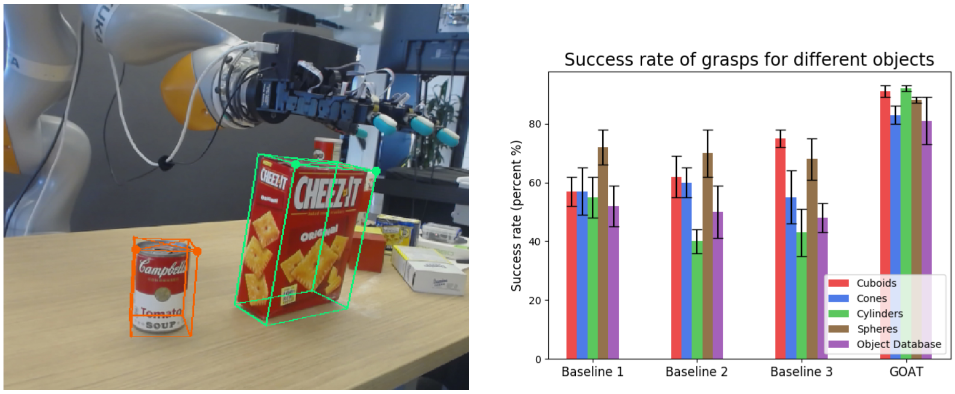

Figure 1: Left: Robotics setup showing the pose estimation from DOPE overlaid. Right: Grasp success rate of trained policies in simulation. The experiment was done for four different seeds.

the robot, and damping coefficients of the robot joints. The range of the uniform distribution is manually specified based on initial experiments on the robot. The DART [8] physics engine was used to simulate the robot and the objects. The simulation runs at 500 Hz, while the actions from the policy are applied at 30 Hz. Our method takes about 6 hours of training time using four threads collecting 1.2×10^6 samples across 500 iterations. These numbers are consistent across four different seeds.

3 Experiments

Our experiments seek to answer the following questions: 1) To what extent does hand demonstration data improve grasp policy learning? 2) Can our policy successfully transfer to a real robot without adaptation?

3.1 Simulation

In order to evaluate the proposed method in simulation, we compare it to three baselines:

Baseline 1. The policy does not use any contact information; we hypothesize that local contact information is important in adapting to non-cuboid shapes and for identifying stable grasps once the robot hand makes contact with the object.

Baseline 2. We include contact information, however, we do not reward the policy for tracking the human hand demonstrations. This allows us to test the importance of demonstration data in learning in this high-dimensional action space, which, combined with the sparse nature of the reward, makes it a difficult reinforcement learning problem.

Baseline 3. We change the context variable κ to a single 6-DoF pose vector of the object's center. This tests our hypothesis that using keypoint information as the context variable provides a coarse representation of the object geometry, enabling the policy to adapt to objects of varying shape.

To compare the effectiveness of our method to that of the policies trained using the baseline methods we perform two different tests. First, we generate 100 random objects unseen by the policies during training and test grasps for each object from 5 random poses on the table. We compare the number of successful grasps out of these 500 resulting trials.

Figure 1-right illustrates the number of successful grasps achieved by each method on different object types. Each bar represent the average of four different trained seeds on a specific category. Object Database refers to the open source dataset of 3D objects and grasps [6] from which we randomly selected 20 objects. We can clearly see that our proposed method outperforms all the baselines for the different object types. Interestingly the baselines all perform somewhat similarly and thus suggests that our method provides the most detailed information for accomplishing this task.

3.2 Real Robot

The ultimate test for GOAT is to verify that the learned policy can be deployed onto a real world robot without additional tuning of the policy. We test our policies on an Allegro robotic hand mounted on a 7-DoF Kuka LBR iiwa 7 R800 arm. The Allegro robot hand has 4 fingers and each finger has 4 degrees of freedom. Each fingertip is equipped with a BioTac sensor. We use this sensor only for binary contact detection. The object localization and bounding-box keypoint location is estimated by DOPE [21]. Figure 1-left shows our robotics set up with projected cuboids overlaid on the graspable objects. We use the 5 objects DOPE can detect from the YCB dataset [2]: cracker box, meat, mustard, soup, and sugar box. Other methods could be used here to fit a bounding box around the object, similar to [12], we could leverage point cloud sensing to fit a bounding box on points above the work surface assuming a non cluttered environment. During the experiment, the object was placed randomly within the robot's workplace five times with a random in plane orientation between $-30°$ and $30°$, where $0°$ means the object's axis is aligned with the robot base. For each pose detection we sample a normal distribution with standard deviation of 1 mm and 10 mm to perturb the object location. With a standard deviation of 1 mm our policy succeeds in grasping the object of interest in 22 of 25 attempts, when the standard deviation is increased to 10 mm our policy succeeds on 14 of 25 attempts. These results are encouraging, especially when compared against a hand-written baseline that succeeds in only 12 of 25 attempts for 10 mm injected noise.

4 Related Work

Reinforcement Learning (RL) has been gaining prominence for robotic manipulation in recent years; many of these works have focused on learning grasping, but the majority focus on the simpler 2D gripper problem [5, 20, 23, 24, 9, 17, 1, 3]. Andrychowicz *et al.* have trained a multi-finger robotic hand policy to reposition a cube in-hand to match a desired pose [14]. Similar to our work they leverage simulation to train a policy to be deployed in the real world, however they do not focus on grasping, instead assuming the object already rests in the robot's hand.

The closest previous work to ours by Osa *et al.* which also learns a grasping policy as contextual policy search [15] initialized by human demonstrations. Another work with similar goals to ours uses supervised learning coupled with analytical planning to plan multi-fingered grasps of different types, *i.e.*, precision and power [12]. They achieve this by explicitly modeling the grasp type as a decision variable in the grasp optimization. Similar to previous robotics work [22, 13, 11, 10, 6], they learn a grasp success predictor from data.

5 Conclusion and Discussion

In this work we have presented a contextual policy search approach to learning policies for grasping unknown objects with multi-fingered hands using bounding box features and contact sensing. We validate that our approach can train purely in simulation and be successfully deployed in the real world on a physical robot. We show that coupling this keypoint representation with contact sensing in the policy allows the robot to adapt to previously unseen shapes and overcome uncertainty in object pose estimation arising from noisy visual sensing. This allows our method to handle objects with shape deviating greatly from that of a bounding box (*e.g.*, a cone).

Th method presented in this work is not currently able to replace classic approaches to multi-finger robotic grasping. Many aspects of the system can cause grasping failure: error in pose estimation, hand placement above the object is quite sensitive, the weight of the object, and slippage during grasp. This makes our proposed system unstable sometimes, however we believe the results show the ability to use low-dimensional visual features (e.g., a bounding box) coupled with contact sensing to enable sim-to-real results that are comparable to a simple hand written grasping policy.

For future work, there are many ways to improve the stability and robustness of this work. For one, the trained policy has no awareness of a power grasp, as such, we could provide new demonstrations to enable more stable power grasps. Moreover we used binary contact information, leveraging richer tactile information such as normal force or explicit slip detection will likely alleviate some limitations of our system. Finally, exploring online learning while manipulating the object might help our policy to learn how to react to prediction error.

References

[1] S. Caldera, A. Rassau, and D. Chai. Review of deep learning methods in robotic grasp detection. *Multimodal Technologies and Interaction*, 2(3):57, 2018.

[2] B. Calli, A. Singh, A. Walsman, S. Srinivasa, P. Abbeel, and A. M. Dollar. The YCB object and model set. In *IEEE Int. Conf. on Advanced Robotics*, pages 510–517, 2015.

[3] K. Fang, Y. Zhu, A. Garg, V. Mehta, A. Kuryenkoy, L. Fei-Fei, and S. Savarese. Learning task-oriented grasping for tool manipulation with simulated self-supervision. In *Robotics Science and Systems*, 2018.

[4] U. Iqbal, P. Molchanov, T. Breuel, J. Gall, and J. Kautz. Hand pose estimation via latent 2.5D heatmap regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 118–134, 2018.

[5] T. Johannink, S. Bahl, A. Nair, J. Luo, A. Kumar, M. Loskyll, J. A. Ojea, E. Solowjow, and S. Levine. Residual reinforcement learning for robot control. In *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, 2019.

[6] D. Kappler, J. Bohg, and S. Schaal. Leveraging big data for grasp planning. In *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, pages 4304–4311, 2015.

[7] J. Kober, E. Oztop, and J. Peters. Reinforcement learning to adjust robot movements to new situations. In *International Joint Conference on Artificial Intelligence*, pages 2650–2655, 2011.

[8] J. Lee, M. X. Grey, S. Ha, T. Kunz, S. Jain, Y. Ye, S. S. Srinivasa, M. Stilman, and C. K. Liu. DART: Dynamic animation and robotics toolkit. *The Journal of Open Source Software*, 3(22):500, Feb 2018.

[9] S. Levine, P. Pastor, A. Krizhevsky, J. Ibarz, and D. Quillen. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *Int. J. Robot. Res.*, 37(4-5):421–436, 2018.

[10] M. Liu, Z. Pan, K. Xu, K. Ganguly, and D. Manocha. Generating grasp poses for a high-DOF gripper using neural networks. In *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019.

[11] Q. Lu, K. Chenna, B. Sundaralingam, and T. Hermans. Planning multi-fingered grasps as probabilistic inference in a learned deep network. In *International Symposium on Robotics Research (ISRR)*, 2017.

[12] Q. Lu and T. Hermans. Modeling grasp type improves learning-based grasp planning. *IEEE Robotics and Automation Letters*, 4(2):784–791, 2019.

[13] J. Mahler, F. T. Pokorny, B. Hou, M. Roderick, M. Laskey, M. Aubry, K. Kohlhoff, T. Kröger, J. Kuffner, and K. Goldberg. Dex-Net 1.0: A cloud-based network of 3D objects for robust grasp planning using a multi-armed bandit model with correlated rewards. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 1957–1964, 2016.

[14] OpenAI et al. Learning dexterous in-hand manipulation. *arXiv preprint arXiv:1808.00177*, 2018.

[15] T. Osa, J. Peters, and G. Neumann. Hierarchical Reinforcement Learning of Multiple Grasping Strategies with Human Instructions. *Advanced Robotics*, 32(18):955–968, 2018.

[16] X. B. Peng, P. Abbeel, S. Levine, and M. van de Panne. DeepMimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM Transactions on Graphics (TOG)*, 37(4):143, 2018.

[17] D. Quillen, E. Jang, O. Nachum, C. Finn, J. Ibarz, and S. Levine. Deep reinforcement learning for vision-based robotic grasping: A simulated comparative evaluation of off-policy methods. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 6284–6291, 2018.

[18] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[19] R. S. Sutton and A. G. Barto. *Reinforcement Learning : An Introduction*. MIT Press, 1998.

[20] G. Thomas, M. Chien, A. Tamar, J. A. Ojea, and P. Abbeel. Learning robotic assembly from CAD. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2018.

[21] J. Tremblay, T. To, B. Sundaralingam, Y. Xiang, D. Fox, and S. Birchfield. Deep object pose estimation for semantic robotic grasping of household objects. In *Conference on Robot Learning (CoRL)*, 2018.

[22] J. Varley, J. Weisz, J. Weiss, and P. Allen. Generating multi-fingered robotic grasps via deep learning. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4415–4420, 2015.

[23] W. Yu, V. C. Kumar, G. Turk, and C. K. Liu. Sim-to-real transfer for biped locomotion. *arXiv preprint arXiv:1903.01390*, 2019.

[24] A. Zeng, S. Song, S. Welker, J. Lee, A. Rodriguez, and T. Funkhouser. Learning synergies between pushing and grasping with self-supervised deep reinforcement learning. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4238–4245, 2018.