Investigating Deep Reinforcement Learning for Grasping Objects with an Anthropomorphic Hand

Mayur Mudigonda[1], **Pulkit Agrawal**[2], **Michael R Deweese**[1], **Jitendra Malik**[2*]

Abstract

Grasping objects with high dimensional controllers such as an anthropomorphic hand using reinforcement learning is a challenging problem. In this work we experiment with a 16-D simulated version of a prosthetic hand developed for SouthHampton Hand Assessment Procedure (SHAP). We demonstrate that it is possible to learn successful grasp policies for an anthropomorphic hand from scratch using deep reinforcement learning. We find that our grasping model is robust to sensor noise, variations in object shape, position of the object and physical parameters such as the density of the object. Under these variations, we also investigate the utility of touch sensing for grasping objects. We believe that our results and analysis provide useful insights and strong baselines for future research into the exciting direction of object manipulation with anthropomorphic hands using proprioceptive and other sensory feedback.

1 Introduction

Grasping is a basic building block for object manipulation tasks. It is a well established and important problem in robotics that has garnered a lot of research interest (1; 2; 3; 4; 5; 6; 7; 8; 9; 10; 11). While past works have made excellent progress in demonstrating the ability to grasp from high-dimensional sensory observations such as images, they have employed rather simplistic two finger parallel jaw grippers. Pragmatically, it can be argued that if the end goal is simply to grasp an object of interest, parallel jaw grippers are sufficient for grasping most objects. However just grasping an object is not a very useful skill by itself and humans often perform complex in-hand manipulation of grasped objects for achieving desired goals. The ability to perform human like in-hand manipulation has generated a lot of interest in design and control of anthropomorphic hands such as the Shadow Hand, the ADROIT suite (12), the UW Hand (13), RBO 2 hand (14) and many others (15; 16).

The superiority of anthropomorphic hands to parallel jaw grippers for in-hand manipulation comes at the cost of difficulty in finding a policy to control them in a desirable fashion. Past work on controlling anthropomorphic hands has relied on imitation (15), human demonstration (17; 18) or pre-defined grasp types (15) as priors for inferring control strategies. Another line of work has exploited compliance and under-actuation (19; 20) for reducing the number of degrees of actuation which in turn simplifies the control problem. However compliance inevitably leads to loss in precise knowledge of the object state. As the observation space becomes richer and the complexity of tasks increases, constructing analytical sensorimotor models becomes progressively harder and using human demonstrations for learning control policies is not scalable because collecting demonstrations is known to be a hard and tedious process.

Reinforcement learning provides a general paradigm for determining control policies without explicitly modeling the actuator or the environment. However, most current reinforcement learning algorithms (21; 22; 23; 24) rely on random walk in the action space for exploration which makes it non-trivial to control actuators with large number of degrees of freedom. It is therefore not surprising and to the best of our knowledge, learning robust grasping behavior from scratch using model free reinforcement learning with a joint controlled anthropomorphic hand in the face of changing environmental conditions and sensory noise has not yet been demonstrated [1]. Given that it is quite challenging to manipulate objects with an anthropomorphic hand, in this work we start with learning the grasping primitive and investigate the limitations and successes of current reinforcement learning techniques towards this end.

2 Experimental Setup

Environment and Observations: We use a simulated model of the anthropomorphic hand used as part of SouthHampton Hand Assesment Procedure (SHAP) test suite (26) (see website [2]). The SHAP procedure was established for evaluating prosthetic hands and arms. With this idea in mind, prior work (27) built a prosthetic hand which could theoretically perform all useful human hand movements. Based on this hand and the DARPA Haptix challenge(28), a simulated model of a similar hand (but with fewer sensors) (29) was built using the Mujoco physics engine (30). This model was made publicly available and we use this for all our experiments.

The hand has five fingers and 25 joints out of which many are tendon coupled, i.e. actuating one joint actuates a set of other joints. For example, curling the tip of a finger automatically actuates the other two joints on the finger so that the finger moves towards the palm resulting in complex and articulated dynamics. Out of the 25 joints, thirteen are actuated. Out of these thirteen, ten joints control the motion of fingers and the other three control the rotation of the hand. Additionally, there are three degrees of motion along the (x, y, z) axis and therefore overall 16 degrees of actuation. The hand is controlled by setting the position of these 16 actuators. The

[*1] Redwood Center for Theoretical Neuroscience, UC Berkeley, mudigonda,deweese@berkeley.edu

[†2] Department of Electrical Engineering and Computer Science UC Berkeley, pulkitag,malik@eecs.berkeley.edu

[1] our work is concurrent with (25)

[2] https://rctn.github.io/deephapticsgrasp/

observation space (input to the model) consisted of the internal position and velocity of the 25 joints and the object position resulting in 53 dimensions (25*2 + 3 = 53). In some experiments, we made use of the touch sensors (19 in number); in these cases, the total number of dimensions were 72. In all experiments, this hand was tasked to grasp a single object kept on a table.

Reward Structure: order to grasp an object the hand must reach the object and then lift it. One way to specify the reward for grasping is the height of the object from the table ($\mathcal{D}_{ot}$). However, this reward is sparse because random exploration is unlikely to result in object grasps by chance. Unsurprisingly, with this sparse reward, learning was unsuccessful. In order to encourage the hand to grasp the object we shaped the reward to penalize the distance from the palm of the hand to the object (i.e. $-\mathcal{D}_{op}$). Our reward function therefore looks like: $r(t) = -0.1 \times \mathcal{D}_{op}$ where $\mathcal{D}_{op} \geq \epsilon$ for reaching and $r(t) = \mathcal{D}_{ot}$ where $\mathcal{D}_{op} < \epsilon$ for grasping. Where ϵ is the parameter that encourages the hand to reach the object. We set ϵ to be 0.15, which was about approximately two times the radius of the sphere object used in our experiments. A positive value signifies that the hand has reached the object and is manipulating it. A value of approximately 120 and above implies the object was grasped and lifted off of the table successfully.

3 RESULTS

We used trust region policy optimization (TRPO) (22) as the policy learning algorithm. We found that it was hard to train the off-policy methods such of deep deterministic policy gradient (DDPG) (21) for this task. For TRPO, the neural network policy was represented by a multi-layer perceptron (MLP) with two hidden layers with sizes of 32 units and 32 units which serve to model a Gaussian distribution from which we sample actions. We experimented with varying the size of the network by testing networks of size 64 and 16 but found that they performed comparably (64 units) or worse (16 units). Based on these preliminary results we used two layer neural networks of size (32, 32) for the rest of our experiments.

We found that three parameters were critical to training: (a) batch size (b) the initial standard deviation of the Gaussian policy and (c) normalizing the observations. We varied the initial standard deviation of the Gaussian across multiple experiments and settled on a value of 3.0. We fixed the episode length to be 500 time steps long. We empirically found that this was a long enough trajectory to reach to almost any part of the table with the actions that were being generated by the policy. To normalize the observations, we unrolled 1000 random episodes and computed their means and standard deviations. These were then used to normalize the observations during learning. We tried experiments without normalizing the observations and the model failed to learn. Given the seed sensitivity of many Deep-RL algorithms, we were cautious and ran multiple (at least three seeds) for each experiment.

3.1 BATCH SIZE

To test the robustness of training to batch size, we used batch sizes of - 40k, 60k, 80k and 120k. As batch size was increased the model learnt a more robust policy for grasping and lifting the object. The performance scaled linearly with the number of trajectories TRPO saw. The numbers in Table 1 describe the cumulative reward (undiscounted) averaged across multiple trajectories.

batch size	feature space
40k	169.499 +/- 96.98
60k	332.57 +/- 21.04
80k	281.887 +/- 36.65
120k	349.90 +/- 38.60

Table 1: Batch size

Object Type	Sphere Success Rate	Multi-Obj Success Rate
cuboid	0.92	1.0
sphere	0.85	0.99
ellipsoid	0.01	0.98
cylinder	0.17	1.0
can	0.0	0.06
coin	0.46	0.47
screwdriver	0.29	0.97

Table 2: Multiple Objects - Generalization

Table 1 comparing the mean and standard deviation (across three seeds) average reward per episode during training for different batch sizes. **Table 2** describing the success of grasping an object and lifting it off the table. Success rate is defined here as the number of times the object was at a distance of 0.7 or greater across all episodes for that object. In our simulation this usually meant the object must have been grasped reasonably for the object to be that far from the table. We simulated 1000 episodes and this led approximately to 125 episodes per category.

3.2 MULTIPLE OBJECTS AND GENERALIZATION

Once we found that it was possible to grasp a single object (sphere), we sought to investigate if it was possible to grasp previously unseen objects. This resulted in poor generalization as seen in column 1 of Table 2. Notably, although the model was trained on spheres it generalized better to cuboids due to their relative simplicity in grasping but did poorly on all other objects.

Given the poor generalization with a single object during training, we then setup an experiment with multiple objects - a sphere, cuboid and an ellipsoid. We gave the learnt model novel objects such as a screw driver, a can, an ellipsoid and a cylinder. We show that we are able to learn a policy to grasp for multiple objects as well. Some visualizations from this experiment can be seen in Figure 1. We see that it learns to grasp multiple objects it had not seen during training. Nevertheless, it was not always successful; often failing with objects that did not share any common geometric properties with objects in training.

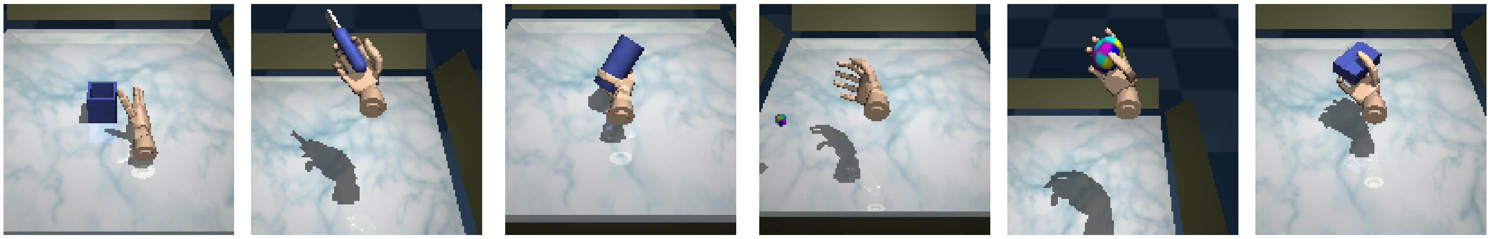

Figure 1: Figure showing some of the objects and their grasps by the model (snapshot at different time intervals). Left to right, a can, screw driver, cylinder, coin, ellipsoid and cuboid are shown. The table is bounded by four walls (green in color) that can be seen

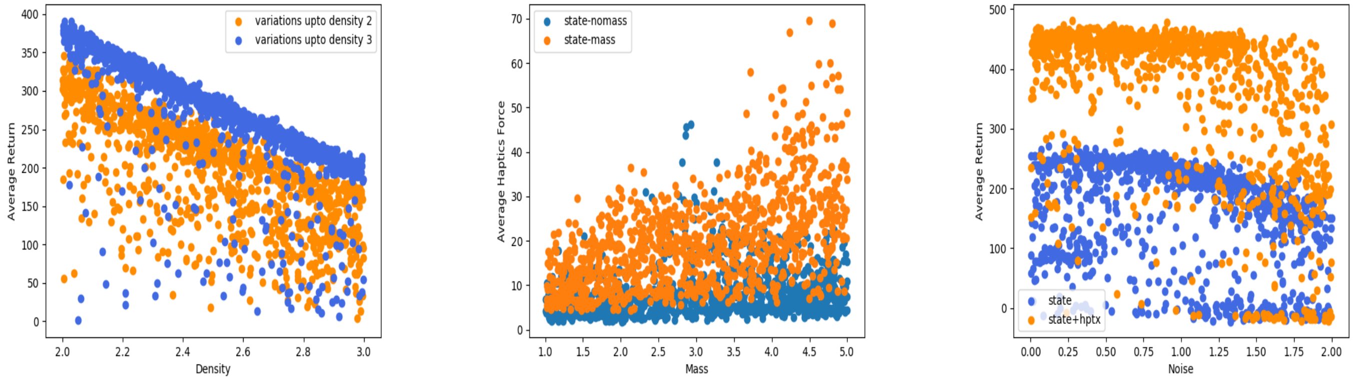

Figure 2: Scatter plots based on approximately 1000 trials. **Left (a):** comparing the average reward (y-axis) and the density (x-axis) for models trained on density variation only up to 2 (blue) versus one that is trained on density variation (orange) up to 3. The density was then varied between 2 and 3 for each condition to populate the graph. This graph shows the effect of training on more general cases of density variations and their effect on rewards. **Middle (b):** comparing the average force (y-axis) and the density (x-axis) for models trained on density variation (blue) versus one that is not trained on density variation (orange). **Right (c):** We examine the relationship between varying observer noise level and average return. Plotted on the x-axis is the observer noise sampled from a uniform distribution with range 0 to 2 on the x-axis. On the y-axis is the average return across an episode. Two feature spaces were compared - with (orange) and without (blue) haptics as feature spaces.

3.3 Generalization to variation in object density and sensory noise

The mainstream mechanism for obtaining generalization is to train a policy on different instances of the same problem, where one or more parameters (such as density, sensory noise) of interest are varied. The hope is that finding a policy that works across variations in training set will also work for test set. While, this might be reasonable if the test set merely requires interpolation between two points in the training set, it is unlikely to be adequate for extrapolative generalization. We tested this in two setting of varying the density of the object and the amount of sensory noise.

In the first set of experiments, we trained one model with mass variation from 1 to 2 (blue) and other model with variation from 1 to 3 (orange). We tested these models on objects whose densities were sampled between 2 and 3 (see Figure 2a). If simply varying the density is sufficient for learning a robust policy, the performance gap between the two trained models should be small. Figure 2a shows that this is not the case. Further analysis reported in Figure 2b suggests that the average force exerted by the hand while lifting the object is higher when the model is trained with larger density variations (orange) as compared to training with smaller density variations (blue), suggesting that the trained policy can adapt to heavier objects to some extent. However, as witnessed from Figure 2a the extent of this adaption is insufficient.

In a second set of experiments, we trained these models with a noise standard deviation $\sigma = 1$. During test time, we sampled noise from a uniform distribution $\eta \in (0, 2)$. Figure 2c shows that the the performance deteriorates gracefully with more noise. We next sought to investigate if including touch sensing would increase the robustness of grasping. We re-ran this experiment with the observation space expanded to include the reasings of 19 pressure sensors on the SHAP MPL hand used in our experiments. Figure 2c shows that the policy using touch sensing (orange) performs significantly better across all noise levels.

4 Conclusion

In this work we show that it is possible to learn robust grasping policies from anthropomorphic hands using model free reinforcement learning algorithm known as trust region policy optimization (TRPO). We found that the learned policies were not object specific and when trained with three objects, our system was able to grasp four novel objects with quite different geometries as compared to the training set of objects. Further investigation revealed that the learned policies were robust to some extent to noise in sensory observations and variation in object properties such as density. The results pointed that simple randomization of parameters is insufficient for extrapolative generalization. Finally, we found that policies trained with touch sensing were more robust to sensory noise than policies trained without touch sensing.

References

[1] R. M. Murray, Z. Li, S. S. Sastry, and S. S. Sastry, *A mathematical introduction to robotic manipulation*. CRC press, 1994.

[2] A. M. Dollar and R. D. Howe, "Simple, robust autonomous grasping in unstructured environments," in *Robotics and Automation, 2007 IEEE International Conference on*. IEEE, 2007, pp. 4693–4700.

[3] C. Goldfeder, M. Ciocarlie, H. Dang, and P. K. Allen, "The columbia grasp database," in *Robotics and Automation, 2009. ICRA'09. IEEE International Conference on*. IEEE, 2009, pp. 1710–1716.

[4] T. Lampe and M. Riedmiller, "Acquiring visual servoing reaching and grasping skills using neural reinforcement learning," in *Neural Networks (IJCNN), The 2013 International Joint Conference on*. IEEE, 2013, pp. 1–8.

[5] A. Rodriguez, M. T. Mason, and S. Ferry, "From caging to grasping," *The International Journal of Robotics Research*, vol. 31, no. 7, pp. 886–900, 2012.

[6] A. Herzog, P. Pastor, M. Kalakrishnan, L. Righetti, J. Bohg, T. Asfour, and S. Schaal, "Learning of grasp selection based on shape-templates," *Autonomous Robots*, vol. 36, no. 1-2, pp. 51–65, 2014.

[7] D. Kappler, J. Bohg, and S. Schaal, "Leveraging big data for grasp planning," in *Robotics and Automation (ICRA), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4304–4311.

[8] L. Pinto and A. Gupta, "Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours," *ICRA*, 2016.

[9] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-end training of deep visuomotor policies," *arXiv preprint arXiv:1504.00702*, 2015.

[10] M. Gualtieri, A. ten Pas, K. Saenko, and R. Platt, "High precision grasp pose detection in dense clutter," in *Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on*. IEEE, 2016, pp. 598–605.

[11] J. Mahler, J. Liang, S. Niyaz, M. Laskey, R. Doan, X. Liu, J. A. Ojea, and K. Goldberg, "Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics," *arXiv preprint arXiv:1703.09312*, 2017.

[12] V. Kumar, Z. Xu, and E. Todorov, "Fast, strong and compliant pneumatic actuation for dexterous tendon-driven hands," in *Robotics and Automation (ICRA), 2013 IEEE International Conference on*. IEEE, 2013, pp. 1512–1519.

[13] Z. Xu, V. Kumar, and E. Todorov, "The uw hand: A low-cost, 20-dof tendon-driven hand with fast and compliant actuation," *The International Journal of Robotics Research*, 2013.

[14] R. Deimel and O. Brock, "A novel type of compliant and underactuated robotic hand for dexterous grasping," *The International Journal of Robotics Research*, vol. 35, no. 1-3, pp. 161–185, 2016.

[15] F. Rothling, R. Haschke, J. J. Steil, and H. Ritter, "Platform portable anthropomorphic grasping with the bielefeld 20-dof shadow and 9-dof tum hand," in *Intelligent Robots and Systems, 2007. IROS 2007. IEEE/RSJ International Conference on*. IEEE, 2007, pp. 2951–2956.

[16] D. Rus and M. T. Tolley, "Design, fabrication and control of soft robots," *Nature*, vol. 521, no. 7553, p. 467, 2015.

[17] A. Gupta, C. Eppner, S. Levine, and P. Abbeel, "Learning dexterous manipulation for a soft robotic hand from human demonstrations," in *Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on*. IEEE, 2016, pp. 3786–3793.

[18] V. Kumar, A. Gupta, E. Todorov, and S. Levine, "Learning dexterous manipulation policies from experience and imitation," *arXiv preprint arXiv:1611.05095*, 2016.

[19] L. U. Odhner, L. P. Jentoft, M. R. Claffee, N. Corson, Y. Tenzer, R. R. Ma, M. Buehler, R. Kohout, R. D. Howe, and A. M. Dollar, "A compliant, underactuated hand for robust manipulation," *The International Journal of Robotics Research*, vol. 33, no. 5, pp. 736–752, 2014.

[20] R. Deimel, P. Irmisch, V. Wall, and O. Brock, "Automated co-design of soft hand morphology and control strategy for grasping."

[21] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *ICLR*, 2016.

[22] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, 2015, pp. 1889–1897.

[23] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, *et al.*, "Human-level control through deep reinforcement learning," *Nature*, 2015.

[24] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. P. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," in *ICML*, 2016.

[25] A. Rajeswaran, V. Kumar, A. Gupta, J. Schulman, E. Todorov, and S. Levine, “Learning complex dexterous manipulation with deep reinforcement learning and demonstrations,” *arXiv preprint arXiv:1709.10087*, 2017.

[26] C. M. Light, P. H. Chappell, and P. J. Kyberd, “Establishing a standardized clinical assessment tool of pathologic and prosthetic hand function: normative data, reliability, and validity,” *Archives of physical medicine and rehabilitation*, vol. 83, no. 6, pp. 776–783, 2002.

[27] B. PJ, “The cosmesis: A social and functional interface,” *Johns Hopkins APL Tech Dig*, vol. 30, no. 3, p. 250255, 2011.

[28] D. Weber. Hand proprioception and touch interfaces. [Online]. Available: http://www.darpa.mil/program/hand-proprioception-and-touch-interfaces

[29] V. Kumar. (2016) Shap arm implementation in mujoco. [Online]. Available: http://www.mujoco.org/forum/index.php?resources/modular-prosthetic-limb-shap-test-suites.19/

[30] E. Todorov, T. Erez, and Y. Tassa, “Mujoco: A physics engine for model-based control,” in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2012, pp. 5026–5033.