Learning to Grasp Objects in Virtual Environments through Imitation

Alexandre Filipe
Thesis student,
Instituto Superior Técnico

Alexandre Bernardino, Plinio Moreno
IST Robotics Department,
Instituto Superior Técnico

Abstract

To improve the quality of robot grasps we propose the use of human demonstrations as a guide for the robot to follow, a process often referred to as imitation learning. To do so a human subject would perform grasps and manipulation tasks on a virtual environment, using a glove with sensors capable of capturing the entirety of the hand motions. Our goal is to develop a predictive model, which uses past and current joints' information to estimate the forthcoming joints' positions. Our model is a Recurrent Neural Network that generates the joint positions for a virtual robot, replicating the demonstrated task. To ensure the objects are well grasped, the task is segmented in two phases, after and before the object is considered grasped, avoiding the model continuing a task with an object not well grabbed, dropping it or not even lifting as a result.
Using this model, trained with the recorded demonstrations, we guide the virtual robot to perform a series of simple manipulation tasks, and manage to do so with an good success rate.
Also, to allow anyone intending to try and test their own imitation algorithms, we will provide our virtual environment and complete dataset of demonstrations freely.

1 Introduction

Even with today's standards of what robots can achieve, robotic manipulation still stands as a huge challenge, due to the great number of degrees of freedom present in a human hand, leading to the community using a claw/gripper or a 3-finger hand instead [1][2].

To train a robot to grasp and manipulate an object, several approaches have been used, from physics based to trial and error methods, but a group that has shown some promise are imitation based methods, where a human subject demonstrates the grasping task and, using some form of machine learning, the robot tries to replicate the same task.

To record said demonstrations, again, several methods are employed, from depth-image [1] to video [2][3], but we propose the use of a virtual environment, as some works before did [3], as it is easier to capture data and does not suffer the common issues of having objects, including the hand, obstructing the view of the camera. To capture the hand movements, in case the recording isn't made through video, a range of tools are used, including motion controllers included with VR headsets [1] and keyboard/console controllers [2] but, to fully capture the complexity of the human hand, we are going to use gloves with sensors [4][5].

2 Process

With our objective of teaching a robot through demonstrations, being these demonstrations performed in a virtual environment using a glove with sensors, we must first have a virtual environment capable of capturing these. To do so we used Unity3D, a well-known game engine, to create a simple virtual environment that consists on a table with several objects to interact with (Figure 1). The physics interactions would be processed by Unity3D's own physics engine.

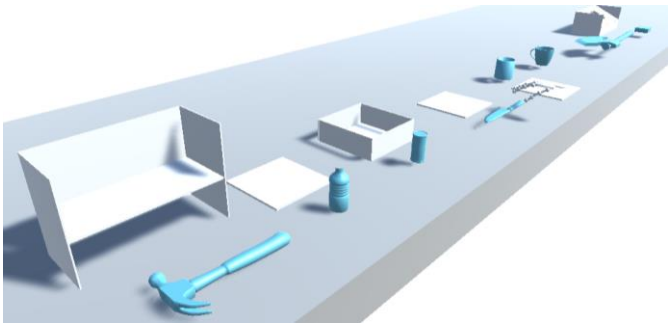

Figure 2 - Virtual Environment

To interact with this virtual environment we used the VMG 35 Haptic glove (Figure 2), a glove containing sensors to measure how much each finger joint is bent and 2 gyroscopes, one on the hand and another on the wrist, to read the rotational position of the hand.

Figure 1 - VMG 35 Haptic glove

With everything set it was now possible to record the demonstrations. Such is done by recording the hand and object positions every fifth of a second, creating a sequence of stages/iterations that represent the demonstration, and then writing this data to a txt file. For each task, having each one consisting of grasping an object and, in some, interacting with a second object, 200 demonstrations were recorded. To create some variance the initial position of the objects is randomized at the start of each recording, inside a plausible area.

Having a considerable set of demonstrations, it is now needed to use this to allow a robot to use their information to perform the task. To achieve this we used an LSTM [6], a recurrent neural network that had already been shown to be successful in robot manipulation [2]. To be precise, we found it better to use 2 LSTMs, segmenting the demonstration in 2 parts, before the object is considered to be grabbed (the reaching segment), and after (the manipulation segment), having each LSTM training a single segment. This was done as a single LSTM would sometimes lead to the hand almost grabbing the object and them proceeding with the task without the object in hand. With 2 LSTMs the hand will only advance to the manipulation segment after the grab condition is met, continuing to try to grab the hand until then.

The input of the LSTM consists of the current state along with the previous 9 states. The state consists of hand position, object position and hand's limbs angular positions. The output of the LSTM is the prediction of the next state. By performing this prediction in a recurrent way, the model will be able to recreate the training tasks.

The data format consists, as mentioned before, of a sequence of iterations, being each iteration represented by 20 values that contain the values of each finger joint bend and the object and hand position. To streamline and minimize the number of input values for the LSTMs we record the relative position (spatial and angular), on the reaching segment the relative position of the hand to the object and on the manipulation segment the relative position of the hand (with the object grabbed) to another object we will interact with, or a generic position, in case no second object is used.

To estimate the forthcoming joint position, we use a kind of a regression predictive model, where the LSTM is trained with the mean squared error loss.

To make the trained LSTMs guide the virtual hand to perform task first we choose a starting position, from which the reaching LSTM will recurrently output the following iteration, having the virtual hand in Unity3D following these values. The instant the object is considered to be grabbed (this condition is verified if the thumb and at least one opposable finger is in contact with the object), a flag is sent from Unity3D informing that from that point on we will continue the reproduction using the manipulation LSTM, continuing this one predicting the following iterations recurrently.

3 Simulation Results

To test the quality of our method we trained and tested for 4 tasks: grabbing a bottle and placing it on a base, grabbing a can and placing it sideways on a box, grabbing and bringing a mug to a base and grabbing and placing a hammer on a shelf.

At the start of each test the hand and the object to be grabbed starts at a random positions (inside plausible values). Afterwards the LSTMs will try to predict the following iterations, using the process mentioned before, and try to perform the task they were trained for. If the objective of the task is achieved (that is, if the hammer is placed on the shelf, for example), then the reproduction is considered a success.

After 20 trials of testing for each task the success rate was as follows:

Table 1 - Test results

Bottle	85%
Can	80%
Mug	75%
Hammer	75%

These results show promise for this method, achieving the common values for this kind of work.

4 Next Steps

Having the model capable of guiding the virtual hand we think it would also be of interest to try to export the model to guide a real life robot to perform the tasks that has been demonstrated on the virtual environment.

To anyone that wants to test their own manipulation training methods our virtual environment and our complete dataset of demonstrations can be found in github.com/alexamor/thesis.

Acknowledgements

This work was supported by FCT with the LARSyS - FCT Project UIDB/50009/2020.

References

[1] T. Zhang, Z. McCarthy, O. Jow, D. Lee, X. Chen, K. Goldberg, and P. Abbeel, “Deep imitation learning for complex manipulation tasks from virtual reality teleoperation,” in2018 IEEE InternationalConference on Robotics and Automation (ICRA), May 2018, pp. 5628–5635

[2] R. Rahmatizadeh, P. Abolghasemi, A. Behal, and L. Boloni, “Learning real manipulation tasks from virtual demonstrations using lstm,” arXiv preprint arXiv:1603.03833, 2016.

[3] Y. Liu, A. Gupta, P. Abbeel, and S. Levine, “Imitation from observation: Learning to imitate behaviors from raw video via context translation,” in 2018 IEEE International Conference on Robotics and Automation (ICRA), May 2018, pp. 1118–1125.

[4] A. Rajeswaran, V. Kumar, A. Gupta, J. Schulman, E. Todorov, and S. Levine, “Learning complex dexterous manipulation with deep reinforcement learning and demonstrations,” inRobotics Proceedings - Robotics: Science and Systems XV, 2017.

[5] Sundaram, S., Kellnhofer, P., Li, Y. et al. Learning the signatures of the human grasp using a scalable tactile glove. Nature 569, 698–702 (2019). https://doi.org/10.1038/s41586-019-1234-z

[6] Felix A. Gers, Nicol N. Schraudolph, and Jurgen Schmidhuber. Learning precise timing with LSTM recurrent networks. Journal of Machine Learning Research (JMLR), 3(1):115–143, 2002