We Can See Objects As Well As Our Hands

Alvin Sun
Department of Mechanical Engineering
Stanford University
Stanford, California 94305
Email: alvinsun@stanford.edu

I. Introduction

Precise manipulation of objects, especially in unknown environment, has long been a challenging problem that is preventing autonomous robot manipulators from reaching the intelligence of a human. There are several difficulties:

1) Computing for the dynamics of the manipulator alone could sometimes be quite difficult considering the number of DOF it could have. This is especially true if the manipulator is deployed to mobile platforms where the true state of the platform itself is hard to estimate.
2) The surrounding environment, even though for the non interacting part, is hard to model, given it can potentially includes infinite number of possible configurations.
3) Most importantly, precise manipulation of objects require some knowledge of the object itself – either its physical dynamics or its states. However, the object(s) of interests, or the ones to be manipulated, has physical properties that are most of the time unknown to the manipulator. Many commonly seen objects could even be deformable, of which the dynamics are even harder to learn.

Human naturally deal with such challenges with feedback. The control policy gets adjusted on the fly when a human approach some objects and potentially make mistakes such as misalignment or grasp inaccuracy. Vison, tacile, and proprioception are the main sensory feedbacks that aid humans to achieve such adjustment. However, if we look at how we think about the manipulation tasks, we don't really keep separate internal dynamics of ourselves and the objects. Instead, we think of the manipulation dynamics as a whole where some action changes the states of both ourselves and the objects. As pointed out by Schaal [13], the learning process of human infant rely heavily on imitation based on visual perception.

Taking the inspiration from such biological intuition, we propose a supervised learning method for obtaining joint visual representation of the dynamics of both the manipulator and the environment it is interacting with. In control literature, real-world dynamical system is usually modeled with

$$x(t) = f(\dot{x}(t), u(t)), \tag{1}$$

while sometimes discretized to the form

$$x_{k+1} = f(x_k, u_k). \tag{2}$$

Our method is effectively a forward predictive model in the form of Equation 1 where the encoded states x encapsulate both the manipulator as well as the object it is trying to manipulate. This learned dynamics can be integrated for more efficient downstream tasks such as control optimization that takes the manipulated object into account.

II. Related Work

A. Model Learning

The learning of unknown dynamics, also known as the system identification problem, has been tackled from a wide range of angles. Dynamic Mode Decomposition [14] and its nonlinear variant eDMD [15] are two of the most widely applied system identification approaches that exploited the eigen structures of (locally) linear dynamical systems. Later, Proctor et al. [12] applied control algorithms to models learned from DMD. More recently, Brunton et al. [1] proposed to SINDy, which utilizes compressed sensing techniques to learn the true underlying physics of arbitrary dynamical systems by getting a sparse sets of coefficients for a large non-linearity dictionary. Brunton et al. [2] later applied model predictive control on systems identified using SINDy. Champion et al. [3] built on top of SINDy and uses deep neural networks to generalize for high dimensional inputs such as images. However, for all of the methods mentioned, the learned models are mostly for standalone dynamics that does not account for complicated interactions such as contact. In the RL literature, Deisenroth and Rasmussen [4] proposed to model arbitrary unknown (discrete) dynamical model following Equation 2 as Gaussian processes. Though it resulted in data efficient learning of a wide range of systems, it again does not consider any dynamical interaction with the environment. Gal et al. [7] later generalizes this method to work with high dimensional image spaces, but still does not concern interaction.

B. Visual Representation

Levine et al. [11] started the area on visuomotor learning with an end-to-end learning algorithm for manipulation task. However, the learning process rely on local controller that has full observability over the states of the object. Several methods [8, 9] explicitly learn latent dynamics of the environment, but without interpretability over the learned latent space. Other work [10, 6] has come up with explicit visual representation of the environment, but the connection of those learned representations to control policies are unclear. Xu et al. [16] explicitly learns the physics of real-world items through programmed interactions, but is fairly limited to just estimating inertia properties of the objects instead of complicated manipulation dynamics.

III. Proposed Research

A. Problem Formulation

We propose a self-supervised vision-based learning method for estimating joint dynamical representation of the robotic manipulator as well as the object to be manipulated. Considering causal effect between the control inputs and states, some control input could cause some states (sometimes for both manipulator and objects) to change. The change is reflected in direction of motion, which is the time derivative term in Equation 1. However, during the reaching phase of a manipulation, when the manipulator is not in contact with the object, it is unlikely that the action of the manipulator could cause a state change of an object unless there are external interaction among the environments. If we use x_m and x_e as the states of the manipulator and environment respectively, and u as the action for the manipulator, we can first create a contact detector model c such that

$$c(x_m, x_e) = \mathbb{1}\{x_m \text{ and } x_e \text{ is in contact}\}. \tag{3}$$

Then we can formulate an explicit switching dynamics for both the manipulator and the environments,

$$\dot{x}_m = \begin{cases} f_m^{\bar{c}}(x_m, u) & \text{if } c(x_m, x_e) = 0 \\ f_m^{c}(x_m, x_e, u) & \text{otherwise} \end{cases} \tag{4}$$

$$\dot{x}_e = \begin{cases} f_e^{\bar{c}}(x_e) & \text{if } c(x_m, x_e) = 0 \\ f_e^{c}(x_m, x_e, u) & \text{otherwise} \end{cases} \tag{5}$$

Now, given all those contact / non-contact dynamics are unknown, and the inputs are high dimensional and also potentially multimodal, we would like a compressed latent representation of the dynamics.

B. Modeling with Deep Neural Networks

Following Lee et al. [10], we can extract some multimodal representation of the sensor readings through neural network. To extent upon their representation, the multimodal fusion module could be implemented with RNNs such as GRU or LSTM to account for temporal effects of dynamics. Let $\mathcal{I}$ denotes the collection of input sensors, we can learn the latent dynamics with neural networks parametrized by $\{\theta, \phi\}$ as follows,

$$x_m = f_\theta(\mathcal{I}) \tag{6}$$

$$x_e = f_\phi(\mathcal{I}) \tag{7}$$

Similarly, the four dynamical models from Equations 4 and 5 can also be learned as GRU or LSTM units due to their ease of back propagation. The supervision training signal will be provided by trying to reconstruct the image space derivatives through the latent space derivatives. More specifically, $\dot{x}_m$ and $\dot{x}_e$ should contain information for reconstructing the optical flow that reflects movement of the manipulator and the objects respectively. The contact model, $c(x_m, x_e)$, can also be learned as a binary classifier considering we have some touch and torque feedback from the manipulator. Since this representation learns a continuous dynamical model in a latent space, we can also use it to roll out longer horizon trajectories with discretization. Simple forward Euler integration could be implemented for simplicity as well as for being back propagation friendly. With the rolled out trajectory, we can make training harder by using longer time horizon. The model can then be trained with progressively longer prediction horizon and equivalently higher difficulties.

C. Automatic Data Generation

RGBD camera sensor can be used to generate ground truth optical flows. Since we have a manipulator which relatively robust pose estimates, we can use [5] to segment out the moving parts (i.e. the manipulator and / or the objects) and calculate flows with camera back projections. For obtaining an interpretable learned latent dynamics, we can provide some supervision on it using knowledge from proprioception.

TODO: more details here.

Acknowledgments

References

[1] Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113 (15):3932–3937, 2016. ISSN 0027-8424. doi: 10.1073/pnas.1517384113. URL https://www.pnas.org/content/113/15/3932.

[2] Steven L Brunton, Joshua L Proctor, and J Nathan Kutz. Sparse identification of nonlinear dynamics with control (sindyc). *IFAC-PapersOnLine*, 49(18):710–715, 2016.

[3] Kathleen Champion, Bethany Lusch, J. Nathan Kutz, and Steven L. Brunton. Data-driven discovery of coordinates and governing equations. *Proceedings of the National Academy of Sciences*, 116(45):22445–22451, 2019. ISSN 0027-8424. doi: 10.1073/pnas.1906995116. URL https://www.pnas.org/content/116/45/22445.

[4] Marc Deisenroth and Carl E Rasmussen. Pilco: A model-based and data-efficient approach to policy search. In *Proceedings of the 28th International Conference on machine learning (ICML-11)*, pages 465–472. Citeseer, 2011.

[5] Ross Finman, Thomas Whelan, Michael Kaess, and John J. Leonard. Toward lifelong object segmentation from change detection in dense rgb-d maps. In *2013 European Conference on Mobile Robots*, pages 178–185, 2013. doi: 10.1109/ECMR.2013.6698839.

[6] Peter R. Florence, Lucas Manuelli, and Russ Tedrake. Dense object nets: Learning dense visual object descriptors by and for robotic manipulation. *CoRR*, abs/1806.08756, 2018. URL http://arxiv.org/abs/1806.08756.

[7] Yarin Gal, Rowan McAllister, and Carl Edward Rasmussen. Improving pilco with bayesian neural network dynamics models. In *Data-Efficient Machine Learning workshop, ICML*, volume 4, page 25, 2016.

[8] Danijar Hafner, Timothy P. Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. *CoRR*, abs/1811.04551, 2018. URL http://arxiv.org/abs/1811.04551.

[9] Danijar Hafner, Timothy P. Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *CoRR*, abs/1912.01603, 2019. URL http://arxiv.org/abs/1912.01603.

[10] Michelle A. Lee, Yuke Zhu, Krishnan Srinivasan, Parth Shah, Silvio Savarese, Li Fei-Fei, Animesh Garg, and Jeannette Bohg. Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8943–8950, 2019. doi: 10.1109/ICRA.2019.8793485.

[11] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *J. Mach. Learn. Res.*, 17(1):1334–1373, January 2016. ISSN 1532-4435.

[12] Joshua L Proctor, Steven L Brunton, and J Nathan Kutz. Dynamic mode decomposition with control. *SIAM Journal on Applied Dynamical Systems*, 15(1):142–161, 2016.

[13] Stefan Schaal. Is imitation learning the route to humanoid robots? *Trends in Cognitive Sciences*, 3(6):233–242, 1999. ISSN 1364-6613. doi: https://doi.org/10.1016/S1364-6613(99)01327-3. URL https://www.sciencedirect.com/science/article/pii/S1364661399013273.

[14] Peter J Schmid. Dynamic mode decomposition of numerical and experimental data. *Journal of fluid mechanics*, 656:5–28, 2010.

[15] Matthew O Williams, Ioannis G Kevrekidis, and Clarence W Rowley. A data–driven approximation of the koopman operator: Extending dynamic mode decomposition. *Journal of Nonlinear Science*, 25(6):1307–1346, 2015.

[16] Zhenjia Xu, Jiajun Wu, Andy Zeng, Joshua B. Tenenbaum, and Shuran Song. Densephysnet: Learning dense physical object representations via multi-step dynamic interactions. *CoRR*, abs/1906.03853, 2019. URL http://arxiv.org/abs/1906.03853.