What Would You Do? Acting by Learning to Predict

Adam W. Tow, Niko Sünderhauf, Sareh Shirazi, Michael Milford, Jürgen Leitner
ARC Centre of Excellence for Robotic Vision, Queensland University of Technology, Brisbane, QLD, Australia
adam.tow@qut.edu.au

***Abstract*—We propose to learn tasks directly from visual demonstrations by learning to predict the *effect* of human and robot actions on an environment. We enable a robot to physically perform a human demonstrated task without knowledge of the thought processes or actions of the human, only their visually observable state transitions. We evaluate our approach on two table-top, object manipulation tasks and demonstrate generalisation to previously unseen states.**

I. Introduction

Robots that can learn from human demonstrations are unquestionably a desire of many roboticists. However, to be useful in real world settings, some specific traits of any such approach are required. Firstly, the robot should generalise human demonstration sequences to unseen states; i.e. predict the *effect* of a humans actions in states not visited during the demonstrations. Secondly, human demonstrations should be robot-agnostic; i.e. no knowledge or access to the target robot is required to record task demonstrations. Thirdly, the approach should be task-agnostic; i.e. the robot can learn new tasks provided new demonstrations alone.

We propose to learn tasks directly from visual demonstrations by learning to predict the *effects* of a humans actions on an environment. Operating on visual demonstrations allows for a wide range of avenues for obtaining human task demonstrations and quite naturally leads to the setting of **state** as an RGB image and **task** as sequences of RGB images.

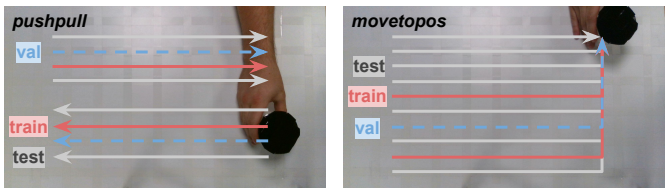
II. Learning from Demonstration with Raw Images

Let us assume a deterministic function $\boldsymbol{\pi}^H : \mathbb{S} \rightarrow \mathbb{U}$ describes the actions $\mathbf{u}_t \in \mathbb{U}$ chosen by a human demonstrator when in state $\mathbf{s}_t \in \mathbb{S}$, so that $\mathbf{u}_t = \boldsymbol{\pi}^H(\mathbf{s}_t)$. Let us also assume the deterministic function $\boldsymbol{\pi}^R : \mathbb{S} \rightarrow \mathbb{A}$ describes the actions $\mathbf{a}_t \in \mathbb{A}$ chosen by a robot when in state $\mathbf{s}_t \in \mathbb{S}$, so that $\mathbf{a}_t = \boldsymbol{\pi}^R(\mathbf{s}_t)$.

Our objective is to choose a policy $\boldsymbol{\pi}^R$ that mimics $\boldsymbol{\pi}^H$. Unlike many approaches to Learning from Demonstration, we argue that the humans' actions $\mathbf{u}_t$ are unobservable by the robot and potentially incompatible. To address this challenge, we propose the following robot policy:

$$\boldsymbol{\pi}^R(\mathbf{s}_t) = \underset{\mathbf{a}_t^{(i)}}{\operatorname{argmin}} \, \tilde{P}(\mathbf{s}_t) \ominus \tilde{Q}(\mathbf{s}_t, \mathbf{a}^{(i)}) \tag{1}$$

We define $P : \mathbb{S} \rightarrow \mathbb{S}$ as a deterministic predictive model that models the humans state transitions such that $P(\mathbf{s}_t) = \operatorname{argmax}_{\mathbf{s}_{t+1}} p(\mathbf{s}_{t+1}|\mathbf{s}_t, \boldsymbol{\pi}^E)$. We also define $Q : \mathbb{S} \times \mathbb{A} \rightarrow \mathbb{S}$ as an action-conditional deterministic predictive model for the robots state transitions such that $Q(\mathbf{s}_t, \mathbf{a}_t) = \operatorname{argmax}_{\mathbf{s}_{t+1}} q(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$.



(a) *pushpull* (b) *movetopos*

Fig. 1. (a) The *pushpull* task involves moving the object left or right based on its spatial location. (b) The *movetopos* task involves moving the object to a specific location; the upper right-hand corner in our case.

The policy $\boldsymbol{\pi}^R(\mathbf{s}_t)$ executes the optimal action $\mathbf{a}_t^*$ that minimises the difference between the *predicted effect* of the expert acting in state $\mathbf{s}_t$, and the *predicted effect* of the agent executing $\mathbf{a}_t$ in the current state. We write $\ominus$ above to indicate a suitable difference metric on the state space $\mathbb{S}$.

In this paper we utilise PredNet [2] to learn the approximations $\tilde{P}(\mathbf{s}_t)$ and $\tilde{Q}(\mathbf{s}_t, \mathbf{a}^{(i)})$ and train it directly on raw images. The state space $\mathbb{S}$ therefore is the space of RGB images, and we show that a suitable metric to implement the $\ominus$ operator is the mean squared error between the raw pixel values. We choose to operate on raw images to maintain the robot-agnostic and task-agnostic traits of our approach.

III. Experiments

We demonstrate our approach on two table-top, single-object manipulation tasks that demonstrate the desirable traits of our approach. The first task requires the target object to be moved to a target location and is referred to as *movetopos*. The second task requires the target object to be moved in a specific direction based on its spatial location and is referred to as *pushpull*.

For both tasks, the table was discretised into a 15x9 grid with the distance the human and robot could move the object per action equal. The state of the environment was captured with a fixed overhead camera for both human and robot interactions.

A key component of our approach is to use the current state of the environment as input for predicting the next state *if* the human were to act. In practice, we found that our predictor performed poorly under this restrictive setting. To remedy this, we instead input the full sequence of robot-visited states and found this greatly improved prediction performance. Using the full sequence presumes that the sequence of human-demonstrated states, and so the action spaces of the human and robot, are compatible. For this reason, we set the action spaces of the human and robot as $\mathbb{U} = \mathbb{A} = \{up, down, left, right\}$.

A. Next-frame prediction

Setting state as images allowed for leveraging work from the computer vision community on next-frame video prediction [2, 1, 3, 4, 5, 6, 7, 8]. The PredNet architecture was used herein [2]. PredNet is comprised of a number of stacked modules that attempt to predict the input to that module. PredNet is shown to perform well on both synthetic and real world tasks and can support variable length inputs at test time due to internal recurrent layers.

B. Predicting what the human would do

We wish to approximate human task demonstrations with PredNet. A small number of demonstrations were recorded for each task as per Figure 1. Using the collected images, one PredNet was trained for each task using similar hyper-parameters to those reported by the networks authors in [2].

C. Predicting what the robots actions would do

We wish to approximate the results of a robot's primitive actions on the environment with PredNet. Specifically, we use a separate PredNet to approximate each of the action-specific predictive models $Q(\mathbf{s}_t, \mathbf{a}^{(i)})$ of the robot. Two demonstrations were performed for each action from different starting locations. Each demonstration involved the robot moving the object across the table, repeatably applying the same action.

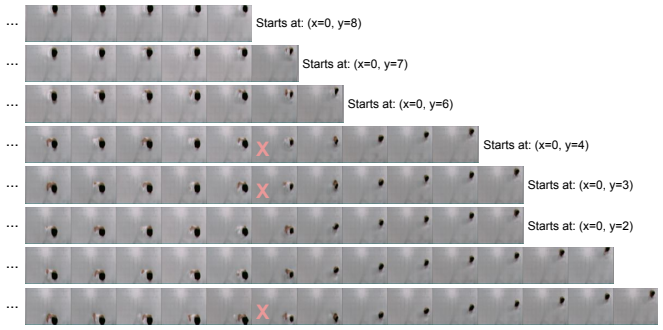
Note for both collecting the robot primitive training data and implementing the approach on the robot, we require that the robot arm be removed from the image of the scene. By removing the robot arm, we remove any bias of the predictions prescribing the robots joint configuration while performing the task.

IV. Results

We report the overall performance of our approach on our two proposed tasks. For each task, we tested the system from every possible start location, excluding their goal locations. We define a trajectory as a sequence of steps the robot is allowed to move the object within the task environment. A trajectory is successful if the object arrives at the ground truth final location, in alignment with the demonstrations.

100% of the 135 trajectories for the *movetopos* task successfully arrived at the goal location in the top right-hand corner of the task-space. 74.1% of the 112 trajectories for the *pushpull* task successfully arrived at the goal locations. Under the more restrictive single-image prediction setting, the percentage of successful trajectories reduced to 10.4% for *movetopos* and 37.5% for *pushpull*.

We show a number of full sequences of predicted images as exemplars in Figure 2. These sequences used the full history of prior states as input for future predictions. As can be seen, the predictions move the block across the task-space in alignment with the human-performed demonstrations. Note that a number of the states visited in these exemplars were not visited by the human, demonstrating generalisation to unseen states.



(a) *movetopos*

Fig. 2. Full sequences of predicted images as exemplars are shown. The red x's mark predicted images that resulted in an incorrect action selection. This caused these sequences to take one additional step over the ground truth. Overall, 50 of the 134 successful *movetopos* trajectories contained an additional step at the transition point from moving rightwards to upwards. The MSE between action primitive prediction right and up at the failure locations were very close.

V. Conclusion

We presented a Learning from Demonstration approach that operates on raw images. Our approach is task, robot and human agnostic. These traits are desirable for two key reasons. Firstly, the human demonstrator does not require knowledge of the target robot a priori, allowing for large, freely available video databases such as YouTube to be used. Secondly, the robot can perform the task correctly on its first attempt by predicting the *effect* of all its actions before choosing to act at every state.

Acknowledgements

This research was supported by an Australian Government Research Training Program (RTP) Scholarship and the Australian Research Council Centre of Excellence for Robotic Vision (project number CE140100016).

References

[1] Ross Goroshin, Michael F Mathieu, and Yann LeCun. Learning to linearize under uncertainty. In *Advances in Neural Information Processing Systems*, pages 1234–1242, 2015.

[2] William Lotter, Gabriel Kreiman, and David Cox. Deep predictive coding networks for video prediction and unsupervised learning. *arXiv preprint arXiv:1605.08104*, 2016.

[3] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440*, 2015.

[4] Randall C O'Reilly, Dean Wyatte, and John Rohrlich. Learning through time in the thalamocortical loops. *arXiv preprint arXiv:1407.3432*, 2014.

[5] Rasmus Berg Palm. Prediction as a candidate for learning deep hierarchical models of data. *Technical University of Denmark*, 5, 2012.

[6] Viorica Patraucean, Ankur Handa, and Roberto Cipolla. Spatio-temporal video autoencoder with differentiable memory. *arXiv preprint arXiv:1511.06309*, 2015.

[7] William R Softky. Unsupervised pixel-prediction. *Advances in Neural Information Processing Systems*, pages 809–815, 1996.

[8] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhutdinov. Unsupervised Learning of Video Representations using LSTMs. In *ICML*, pages 843–852, 2015.