Learning Robot Activities from First-Person Human Videos Using Convolutional Future Regression

Jangwon Lee and Michael S. Ryoo
School of Informatics and Computing, Indiana University, Bloomington, IN
mryoo@indiana.edu

Abstract

We design a new approach that allows robot learning of new activities from unlabeled human example videos. Given videos of humans executing an activity from their own viewpoint (i.e., first-person videos), our objective is to make the robot learn the temporal structure of the activity as its future regression network, and learn to transfer such model for its own motor execution. We present a new fully convolutional neural network architecture to regress the intermediate scene representation corresponding to the future frame, thereby enabling explicit forecasting of future hand locations given the current frame. The full version of the paper is available as [2].

1. Introduction

Recent progress in robotics include new deep learning algorithms for robot manipulation, which directly learn motor control policies given visual inputs [3, 1]. However, many of these deep approaches have been limited to relatively simple actions such as object grasping and pushing. This is because a large amount of 'robot' data is necessary for the direct training of their models with millions of parameters, and this is a limiting aspect particularly when we want to teach a robot new (i.e., previously unseen) activities.

In this paper, we present a new convolutional neural network (CNN)-based approach that enables robot learning of its activities from 'human' example videos. We extend the state-of-the-art object detection network (SSD [4]) to learn the intermediate scene representation abstracting object-hand information in an image frame, and newly introduce the concept of using a fully convolutional network to regress the intermediate representation corresponding to the future frame (e.g., 1-2 seconds later). Combining these allows direct prediction of (ideal) future locations of human hands and objects during the activities. Our robot takes advantage of such future location forecasts to infer its motor control. We experimentally confirm that our approach enables learning of robot activities from unlabeled human interaction videos, and demonstrate that our robot is able to execute the learned activities in real-time.

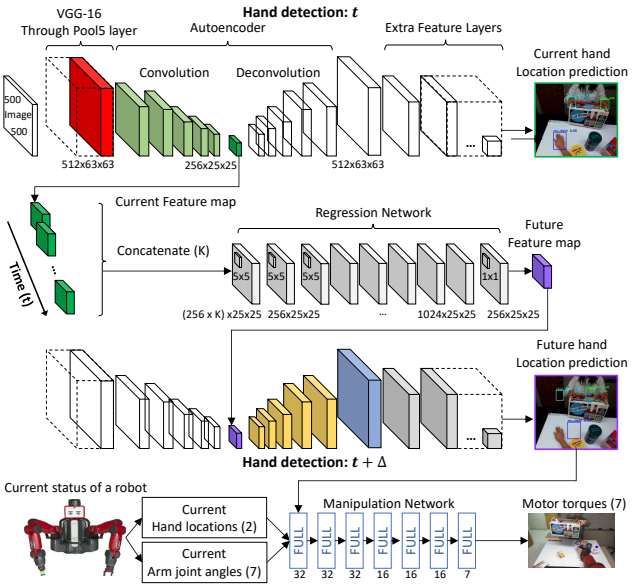

Figure 1. Overview of our approach: Our perception component consists of two networks: We have two copies of an identical network in the 1st row and the 3rd row (i.e., an extended version of SSD [4]), one for the current frame t and the other for the (hallucinated) future frame $t + \Delta$. The fully convolutional future regression network (in the 2nd row) regresses the intermediate scene representation of the future frame based on that of the current frame (256x25x25-D). This regression network does not require activity labels or hand/object labels in videos for its training. The manipulation network (the last row) generates robot control commands given current robot joint state, current robot hand locations, and predicted future robot hand locations.

2. Approach

Given a sequence of frames, the goal is to (i) predict the robot's ideal future hand locations and (ii) generate robot control commands to move the robot hands to such loca-

Table 1. Evaluation of predicting 1-sec future hand locations.

Method	Evaluation		
	Precision	Recall	F-measure
Hand-crafted features	0.30 ± 0.37	0.15 ± 0.19	0.20 ± 0.25
Hands only	4.78 ± 3.70	5.06 ± 4.06	4.87 ± 3.81
SSD w/ future Annot.	27.53 ± 23.36	9.09 ± 8.96	13.23 ± 12.62
Deep Regressor (ours): K=1	27.04 ± 16.50	21.71 ± 14.71	23.45 ± 14.99
Deep Regressor (ours): K=5	29.97 ± 15.37	23.89 ± 16.45	25.40 ± 15.51
Deep Regressor (ours): K=10	**36.58** ± 16.91	**28.78** ± 17.96	**30.90** ± 17.02

tions. We employ two components for achieving the goal. The first component is the perception component that consists of two fully convolutional neural networks: (1) an extended version of the Single Shot MultiBox Detector (SSD) [4] for human hand detection and (2) a future regression network to predict the intermediate scene representation corresponding to the future frame. The second component is the manipulation component that maps 2-D hand locations in the image coordinate to the actual motor control using fully connected layers. Fig. 1 illustrates our approach.

The key idea of our approach is that our perception component can predict the future (1-2 seconds later) hand locations given current video input. Such future prediction can be learned based on humans' first-person activity videos by using them as training data, with the assumption that the robot camera has similar viewpoint as the first-person videos. This allows the robot to infer the ideal future locations of its hands. The manipulation component generates robot control commands to move robot's hand to their desired positions based on the prediction results.

3. Experimental Results

In order to provide quantitative comparisons, we compared our perception component with three different baselines: **(i) Hand-crafted representation** uses a hand-crafted state representation based on explicit object and hand detection. It encodes relative distances between all interactive objects in our two scenarios, and uses it to predict the future hand location using neural network-based regression. **(ii) Hands only** baseline uses frame-based hand detection results for the future regression. It predicts future hand locations solely based on hand detection results of the current frame. **(iii) SSD with future annotations** is a baseline that uses the original SSD model [4] trained based on EgoHands dataset. Instead of training the model to infer the current hand locations given the input frame, we fine-tuned this model on EgoHands dataset with "future" locations of hands as their ground truth labels.

Table 1 shows quantitative results of our future hand prediction. Here, K represents number of frames we used as an input for our regression network. We can clearly observe that our approach significantly outperforms all baselines, including the state-of-the-art object detector SSD modified for the hand prediction.

Table 2. Mean pixel distance between ground truth and predictions. The video resolution was 1280x960.

Method	Mean Pixel Distance
Hand-crafted features	143.85 ± 48.77
Hands only	247.88 ± 121.94
SSD w/ future Annot.	58.58 ± 36.76
Deep Regressor (ours): K=1	51.31 ± 39.10
Deep Regressor (ours): K=5	51.41 ± 38.46
Deep Regressor (ours): K=10	**46.66** ± 36.92

Table 3. Experimental results evaluating the success level of our human-robot collaboration.

Method	Task 1	Task 2	Average
Base SSD + Base control	1.25 ± 0.43	2.21 ± 1.41	1.72 ± 0.92
Base SSD + Our control	1.5 ± 0.96	2.33 ± 1.60	1.92 ± 1.28
Our perception + Base control	2.33 ± 1.18	2.25 ± 1.36	2.29 ± 1.27
Ours	**3.17** ± 1.40	**3.42** ± 1.61	**3.29** ± 1.50

In our second evaluation, we measured mean pixel distance between ground truth locations and the predicted positions of hands. We measured this only when both the ground truths and the predictions are present in the same frame. Table 2 shows the mean pixel distance errors. Once more, we can confirm that our approaches greatly outperform the performances of the baselines.

Finally, we conducted a user study to evaluate the success level of robot activities performed based on our proposed approach, with a total of 12 participants. They were asked to perform two different types of collaborative activities together with our robot and complete a questionnaire to evaluate the success level of our human-robot collaboration with scales from 1 (bad) to 5 (good). Table 3 shows the results.

More experimental results are presented in [2]. Our real-time robot demonstration video is also available at:

https://youtu.be/OCnp_eduA6Q

Acknowledgement: This work was supported by the Army Research Laboratory under Cooperative Agreement Number W911NF-10-2-0016.

References

[1] C. Finn, I. Goodfellow, and S. Levine. Unsupervised learning for physical interaction through video prediction. In *NIPS*, 2016. 1

[2] J. Lee and M. S. Ryoo. Learning robot activities from first-person human videos using convolutional future regression. *arxiv preprint arXiv:1703.01040*, 2017. 1, 2

[3] S. Levine, C. Finn, T. Darrell, and P. Abbeel. End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research*, 17(39):1–40, 2016. 1

[4] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Fu, and A. Berg. SSD: Single shot multibox detector. In *ECCV*, 2016. 1, 2