From Perception to Planning, modified version for CVPR

Yu

November 1, 2022

Abstract

This draft concludes the work that we are going to submit at CVPR2023, with some of my own thoughts about how the new project will be like in the future to follow this line to further build the gap between perception and manipulation level of robotics.

1 The Improvements from eccv to cvpr

1.1 Dataset

We've added two new datasets:

EGO4D [1], which is a very large dataset containing egocentric videos proposed by Kristen Grauman in 2022, and the dataset that we've been collecting:

AI2THOR-EGO, a new dataset that we are going to propose, in which we use agent to immitate the human activities in the robotic simulator ai2thor [2] and use policy to collect 1200 videos, containing four different kinds of scenarios: kitchen, bathroom, bedroom, livingroom.

The advantage of why we use simulator to collect dataset:
1.In order to build the domain gap between sim and real, between vision and robotic vision, the use of simulator is very necessary to give ground truth and test robotic planning on it.
2.Simulator provides a platform for us to take our own videos, give groundtruth about new task using hand-carfted policy(our code), which often have more long-term relations in it and is beneficial for our proposed pattern extractor to extract more meaningful patterns to reveal human intensions.

Details of dataset and how we collect it:
We use hand-crafted policy in the taken of frames, and the labeling process, to make sure the randomness and sequence of collected dataset. We use topological sort to make sure our collected activities are reasonable.
The dataset contains 300*4 videos, including 4 different categories of scenes. We set different random seed to give each video different initialization, each scene has different settings, ornament, different distributions of the objects.

1.2 Tasks

We've replaced the task of 'long-term action recognition' with **'context-based environment affordance'**, please see details in fig 1. The input will be a clip of video and a frame of a new scene, the output will be based on that clip of video, what kind of action the agent is likely to do in this new scene.

The supriority of our proposed task:
Environment affordance is a concept about what the kind of action that environment can afford. For example, take a frame about a hearth, with a closed fridge and a coffeemachine beside it. The environment affordance about that scene will be 'turn on stoveknob, open fridge', 'turn on coffeemachine'.

Unlike environment affordance in a frame, the context-based environment affordance requires more reasoning ability of the relationships between actions in the given video. It requires more ability of

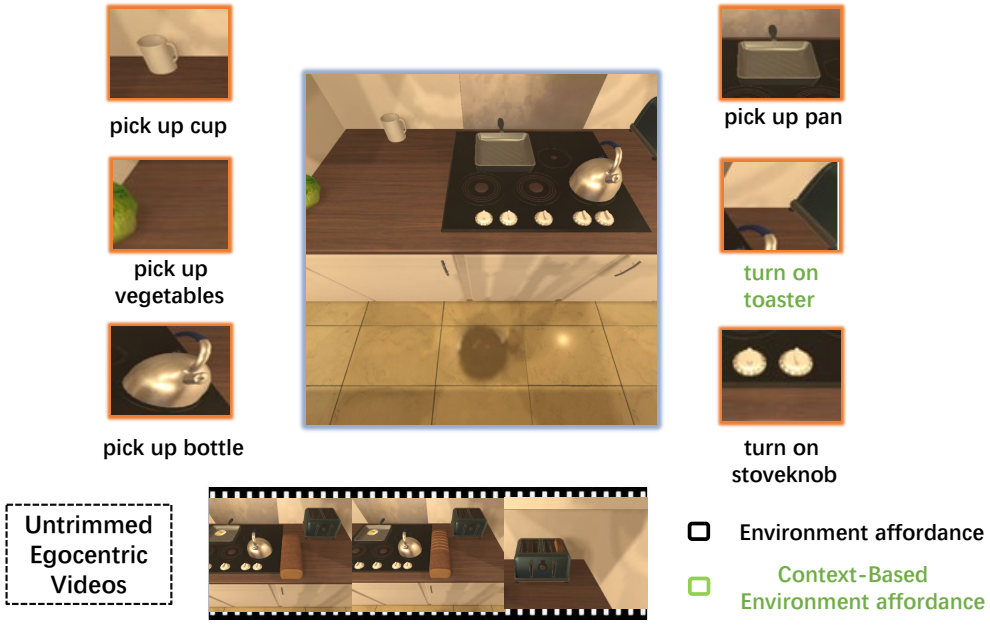

Figure 1: Our proposed context based environment affordance

'to select more reasonable actions that is more likely to happen in that scene, on the condition of the actions that have already happened'. In previous case, if the video about context based environment affordance will be

1.3 Methodology

The improvement of methodology contains the adding of the new branch in the long-short transformer in the anticipation task of EGO4D. The method on our new dataset doesn't contain any methodology-improvement.

Unfortunately, the transformer part is not finished by me, so I will leave out further details here because of confidentiality.

1.4 What I have done

Apart from the group meeting in designing of the direction of the project, I am responsible for dataset collection, took part in implementing hand-crafted policy with the help of a junior undergrad, and the label for two tasks: long-term anticipation and context-based environment affordance is done by myself, using code to automatically give labels, including modify dataset interface for the network training. I am responsible for showing that our proposed method can guide agent's movement in high-level planning.

2 My thoughts about the future

It seems that egocentric video, representing the intersection of robotic vision and manipulation, will be the hot spot of general artificial intelligence field, and it can greatly improve the development of robotic.

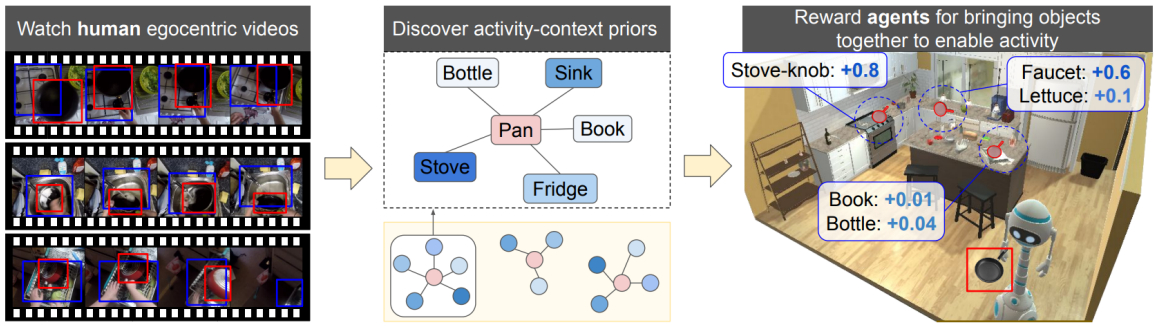

Figure 2: Shaping embodied agent behavior with activity-context priors from egocentric video

2.1 A new dataset to reduce domain gap, or just a policy?

It seems that, the shortage of the dataset in robotic vision in simulator becomes a little problem for robotic researchers to build the gap between perception and planning level of robotics.

Current ego4D [1] dataset is good, but it is in the domain of reality. And transferation from real to sim may be problematic. There are works that try to transfer the knowledge in vision prior to guide robotics planning. *prof. Kristen Grauman* and her students are trying to do this in recent years, the work 'Shaping embodied agent behavior with activity-context priors from egocentric video' [3]in Fig 2 have been using extracted activity-context priors between objects using egocentric videos to give reinforcement learning an auxiliary reward to facilitate agents visual semantic planning. It uses the Dataset Epic-kitchen, which contains different cooking videos in real worlds, and train the agents in simulator ai2thor. The work makes successful steps at first glance, however, it tries to avoid the domain gap problems of sim2real, rather than solving them directly. Because it extracts a kind of topological map to represent the relations between different objects, rather than vision features of it. Meanwhile, the design of the task: visual semantic planning is very simple(only COOL STORE HEAT CLEAN SLICE PREP TRASH), the most complex one only has 4 steps. So in conclusion, the work doesn't solve the domain gap, but avoid them, and the guidance on robotics will be short-term and very naiive, but the advantage of reinforcement learning will be more at long-term planning.

So I think the domain gap between real world feature and simulator could still be a problem for transfering knowledge in real world to simulator. In this work we build a dataset in simulator:ai2thor, but it is not a very large-scale dataset, and the complexity of it is not high. We aim to stimulate academy to pay more attention to this gap, if in the future this blank could be filled, it will be easier for researchers to use vision prior in planning level of robotics, no matter it is the modification of rewarding mechanism in Reinforcement Learning, or other traditional learning ways.

Currently there are no very handy ways to collect first-person videos in simulator, because unfortunately all very popular simulators, such as ai2thor [2] and iGibbson [4], doesn't have the interactive interface for volunteers to manipulate on embodied agents to collect frames. So we only use code to design the policy to collect the data, add some randomization to make sure the diversity of the dataset. Our step is the step from zero to one, hopefully there will be more handy ways to collect dataset.

2.2 Language prior? Or other prior?

Currently, we only use vision representation to guide agents' movement in simulator. The agent trains from current videos, take a frame, and then knows that what's more likely for it to do in this frame. But with only one dimension of sensor (vision dimension) is absolutely not enough. For example, if the agent gets training by videos of a coffeemachine that's always in the indoor kitchen, it knows that at what time it will need to turn on that coffeemachine, however the vision features change when that coffeemachine is outdoor, which means the generalization ability will become a big problem, and we can't always wait for the researchers in computer vision society to resolve that domain gap. Moreover,

agents will always see objects that it never sees. So another dimension of input information really matters, for example, the activity can be seen as a language prior, such as 'turn on the coffeemachine', I think little additional help with the language prior can better aid agents' planning.

In conclusion, I am very looking forward to see the multimodality based on egocentric video, and its applications in aiding robotics' planning.

2.3 Learning methodologies? v.s. RL?

As illustrated in 2.1, the reinforcement learning methodologies will be more suitable for the long-term planning tasks for robotics. And learning-based methods are more suitable for short-term manipulation tasks, for example, the [5] uses learning method to learn visual affordance for dual-gripper object manipulation.

Unlike learning methods, the reward for reinforcement learning is not sparse. So it will be very natural to use egocentric vision prior to give reward in reinforcement learning. Meanwhile, the long-term priority of reinforcement learning better suits the long-term intention which egocentric video reveals. However, I don't know very much about reinforcement learning, so I will leave out further details.

References

[1] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022.

[2] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*, 2017.

[3] Tushar Nagarajan and Kristen Grauman. Shaping embodied agent behavior with activity-context priors from egocentric video. *Advances in Neural Information Processing Systems*, 34:29794–29805, 2021.

[4] Bokui Shen, Fei Xia, Chengshu Li, Roberto Martín-Martín, Linxi Fan, Guanzhi Wang, Claudia Pérez-D'Arpino, Shyamal Buch, Sanjana Srivastava, Lyne Tchapmi, et al. igibson 1.0: a simulation environment for interactive tasks in large realistic scenes. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7520–7527. IEEE, 2021.

[5] Yan Zhao, Ruihai Wu, Zhehuan Chen, Yourong Zhang, Qingnan Fan, Kaichun Mo, and Hao Dong. Dualafford: Learning collaborative visual affordance for dual-gripper object manipulation. *arXiv preprint arXiv:2207.01971*, 2022.