Learning Egocentric Video Representation Using Cross-video Activity Patterns

Yu

November 1, 2022

Abstract

This draft concludes the work that we got rejected by ECCV2022, with some of the reviewers' comments and our future plans. Moreover, it includes what I have done and learnt in this research, with a few of my comments and thoughts.

1 Introduction

1.1 Egocentric Videos

Egocentric video, another name of first person video, is captured by a werable camera typically worn on the head or the chest of people. It arouses great interest of computer science academy and brings a lot of challenging tasks including human-object interactions, gaze detection, body-pose estimation.

1.2 Why We Study It

Recently, because of the arise of EGO4D [2], which is a large dataset containing egocentric videos, the area of egocentric videos is like a rising star in the intersection of vision and robotic. The study of it can greatly benefit the development of robotic vision, building the gap between perception-level and planning or manipulation-level of robotic.

1.3 Proposed Activity Patterns

Pattern is an old concept in computer vision studies for a long time before the times of deep learning. It is used to describe something with regular occasions.

Activity Pattern is a new concept that we firstly use in egocentric videos, to discover tuples with unusually high frequency of occurrence, which is often meaningful and reveal long-term human intensions. For example, *('open fridge', 'close fridge')* and *('pick up tomato', 'wash tomato', 'chop tomato')* are two meaningful activity patterns, they reveal the intention of human to 'cook a dish' in the future.

In fact, you can see fig 1 for further understanding of long-term patterns.

The supriority of our pattern tuples

They extract features that are often hides when only considering temporarily adjacent frames.
Take *('turn on coffeemachine', 'turn off coffeemachine')* as an example, there is a time gap until the coffee is cooked, so humans can go to do something else, for example, peel a potato, and wait for the coffee. With only tools like *RNN*, it will be very hard to find the hidden relationships between two temporarily-distant frames. But because we use statistical estimation methodologies to extract those patterns, the distance between them will no longer be obstacles, which greatly benefits the relation-extraction of distant frames and representation learning of egocentric video.

2 Selected Methodologies

2.1 A Pattern Miner

The pattern miner uses statistical T methodologies to calculate pattern sequence that has unusually high frequent occurance.

2.2 Cross-video Activity Pattern Mining

We use the dataset EPIC-kitchen [1] to get meaningful patterns by abstract all videos into sequences and using our pattern miner on it.

2.3 Graph Construction

Apart from the location-node in EGO-TOPO [**?**], we additionally add pattern-node in it, with each node contains our extracted patterns, with GRU used for the extraction of temporal relations.

3 Dataset

We use the dataset : EPIC-kitchen [1] and EGTEA+, because the occurance of EGO4D is behind us so we didn't use it at first.

4 Tasks and Why We Design Them

We bring up two tasks, including long-term action anticipation and long-term action recognition.

long-term action anticipation
Taken a sequence of frames(a clip of video) as input, anticipate all possbile actions that are likely to happen in the future.

long-term action recognition
Taken a sequence of frames(a clip of video) as input, recognize all actions that are happening in that sequence of frames.

why we design them
As is shown in 1.3 (the supriority of proposed pattern), we put emphasis on the extraction of long-term relations. So we design two tasks that require the observation of relations between two distant frames. For example, long-term action anticipation requires the whole observation at the long sequence of input frames, which is often failed by the RNN because although it extract temporal relations, it puts more emphasis on frames that is more closely happening.

We find the deficiency of our designed 'long-term action recognition' part because it doesn't quit make sense. And it got criticized by the reviewers

5 What I have been done and learnt in this

I took part in each group meeting to discuss the direction of the project, meanwhile I am responsible for the graph construction and the pattern extraction. I also help with the training of neural network, and the initial draft of the paper is done by me.

Apart from the hard and joy in doing research and the friendships with each other, I realize that deep learning can sometimes have obstacles that only requires classical statistic tools to solve them. It doesn't absolute supriority over statistical methods. For example, the extraction of every relation between two distant frames can bring very dense computational burden to the neural network, so current methods like RNN put emphasis on more temporarily-related ones. However, classical pattern mining methods in the area of *database* can solve problems in $O(n)$, which covers the shortage of current deep learning methods.

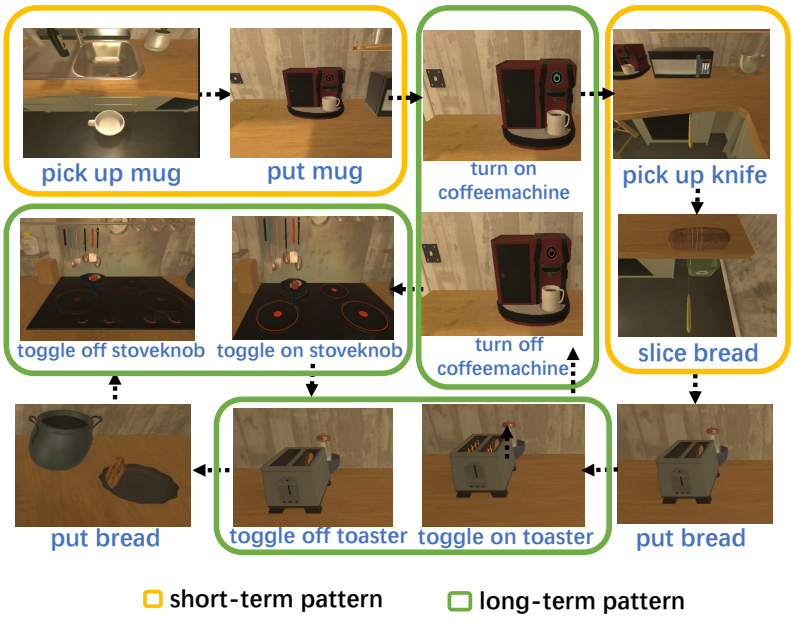

Figure 1: Two categories of activity patterns
The traditional deep learning methodologies like RNN can only extract the information of short-term patterns, which have very close temporal relationships. However, our proposed method can explicitly discover not only short-term pattern but also and more importantly, long-term pattern, which often reveals more human intension.

Meanwhile, it makes me realize that to do research we can often have the opportunity to learn something in interdisciplinary area, which is so good because I always like to learn new things.

6 Reviewers' Comments

We got rejected by ECCV2022

The reviewers comments can be concluded as the methods is too simple, and the 'long-term action recognition task' is not very much reasonable, and they also point out something that needs to pay attention to when writing papers.

7 Future Plans and My Outlook

For future plans, please see the CVPR23 version of the draft.

The problems we are trying to solve, in my point of view, is a little step ahead of others and **touching the core of video representation learning**. We are the first to bring up the existance of activity pattern in video in the academy of computer vision. Our proposed activity pattern can reveal long-term human activity intension in a much more direct way, and human intension is exactly the meaning of the studying of videos. However, current methodologies fail to pay more attention to the extraction of long-term relation between frames, our additional design using the classical method can greatly enhance their effectiveness.

References

[1] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *European Conference on Computer Vision (ECCV)*, 2018.

[2] 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.