Robot Learning Collaborative Manipulation Plans from YouTube Cooking Videos

Hejia Zhang
Department of Computer Science
University of Southern California
Los Angeles, USA
Email: hejiazha@usc.edu

Stefanos Nikolaidis
Department of Computer Science
University of Southern California
Los Angeles, USA
Email: nikolaid@usc.edu

***Abstract*—People often watch videos on the web to learn how to cook new recipes, assemble furniture or repair a computer. We wish to enable robots with the very same capability. Previous work has shown that the space of human manipulation actions has a linguistic, hierarchical structure that relates actions to manipulated objects and tools. Building upon this theory of language for action, we propose a framework for understanding and executing demonstrated action sequences from full-length, unconstrained cooking videos on the web. We demonstrate performance of the system in three full-length Youtube videos that include collaborative actions between two participants. We additionally propose an open-source platform for executing the learned plans in a simulation environment as well as with an actual robotic arm.**

I. INTRODUCTION

We focus on the problem of learning collaborative action plans for a robot. Our goal is to have the robot "watch" unconstrained videos on the web, extract the action sequences shown in the videos and convert them to an executable plan that it can perform either independently, or as part of a human-robot or robot-robot team.

We build upon the theory of language for action [4] to propose a framework for understanding both *individual* and *collaborative* actions from *full-length* YouTube videos. Our key insight is that *hands contain both spatial and temporal information of the demonstrated actions*. This allows using hand trajectories to temporally segment full-length videos to short clips, derive hand-object and object-object associations and infer the demonstrated actions.

The current framework is focused on cooking videos assigning to objects properties such as "tools", "ingredients" and "containers." We hypothesize that these properties are easily transferable to other domains as well, such as furniture manufacturing, and we leave investigating this for future work.

II. FRAMEWORK

The input to the framework is a full-length, unconstrained video from the web. We assume that objects in the video are labeled and a bounding box is provided for each object.

A. Hand Detection

We use OpenPose [1], which detects jointly the human body and hands. We use the detected hands to (1) segment videos by tracking the hand trajectory, and (2) detect which objects are manipulated at a given point in time.

B. Video Segmentation

We temporally segment the video to short clips using the trajectories of the detected hands as time-series data. We use a greedy approach [2], which formulates the segmentation as a covariance-regularized maximum likelihood problem of finding the segment boundaries.

C. Object Association

We extract objects that are relevant to actions in each segment by associating objects with hands and with other objects based on their relative positions in the frame. We introduce a semantic hierarchy of objects, by assigning them to three classes: *tools*, *containers*, and *ingredients* based on their functions.

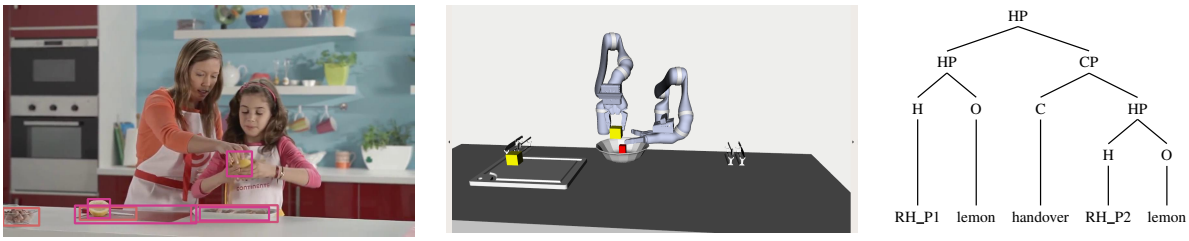
Hand-Object Association. We associate detected hands with objects whose bounding boxes overlap with the box of the hand. In the case of multiple overlaps, we associate the hand with the object that has the largest overlap.

Object-Object Association. We then associate the grasped object with other objects based on the bounding box overlaps. We finally associate container objects with ingredients, if there is an overlap in the bounding boxes of the two.

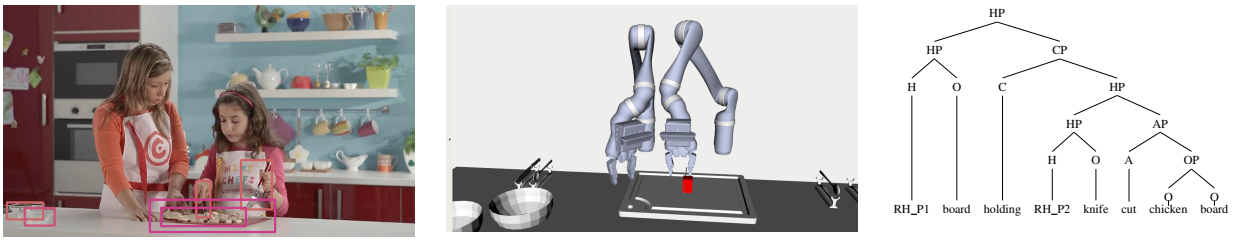
D. Action Recognition

We have two types of actions to recognize, actions performed by a single person, namely *individual actions*, and *collaborative actions* performed by a pair of humans in the video. As a special case of individual actions, we introduce *transfer* actions, which occur when an object moves from one container to another.

Individual Actions. We recognize commonsense actions, using a trained language model from natural language corpus [3]. Given a set of candidate actions and a set of candidate objects, we extract $P(Object|Action)$ for each possible bigram consisting of one object word and one action word in corpus. We then compute the probabilities of each action given the involved objects such as tool used, ingredient manipulated as follows:



a The woman is handing over a lemon to the girl.



b The woman is holding the chopping board for the girl to cut the meat.

Fig. 1: Example frames, snapshots in simulation environment of two robots executing the same actions with humans and generated action trees of 2 successful cases. The captions depict the ground-truth descriptions of each successful case.

$$P(A|O_1, ..., O_k) \sim \prod_k P(O_k|A)P(A)$$

where A is the performed action and $O_1, ..., O_k$ are the objects involved in the action respectively. We then select the most likely action.

Collaborative Actions. Following previous work [5], we detect a *handover* when the object ownership switches between two humans, or a *holding* when the object grasped by one person is used as a tool to manipulate an object grasped by another person.

E. Action Grammar Parsing

We use a collaborative manipulation action grammar [5] to represent the structure of the recognized actions for a robot to execute them. We also introduce an object phrase, which we use to indicate container - ingredient relationships between objects as well as transfer actions from one container to another.

F. Action Graph Generation and Execution

We generate an action graph that combines the generated action trees to executable action sequences.

We implement the action graph as an open-source platform [1], that enables collaborative task execution in the cooking domain.

III. EXPERIMENTS

We show the applicability of the entire framework, in two public YouTube videos [2,3] including a total of 13401 frames and 67 executed actions of 12 different action types.

We evaluate the performance of the framework with respect to the percentage of correctly learned action-trees. We define a correct action tree when the structure and all nodes of the tree are identical to the ground-truth, and the segment corresponding to that tree has a non-zero temporal overlap with the ground-truth segment. We observe that the framework achieved average 0.63 precision and 0.43 recall. Fig. 1 shows the example action-trees and snapshots of the action tree executions by two robotic arms in the open-sourced platform.

To further demonstrate the applicability of our framework, we selected an "easy" video of 2421 frames,[4]. In the accompanying video[5], we show the execution of the complete action graph by two simulated robotic arms as weel as a robot-human team in the real world with the open-source platform.

IV. CONCLUSION

We have presented a framework that takes as input an unconstrained cooking video with annotated object labels and outputs a human-interpretable plan. We demonstrate the execution of the plan in a simulation environment with two robotic arms as well as in a real world environment with a human and a robotic arm and show that we can fully reproduce the actions of a simple cooking video. We find that this work brings us closer to the goal of robots executing a variety of manipulation plans by watching cooking videos online.

REFERENCES

[1] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In *arXiv preprint arXiv:1812.08008*, 2018.

[1]https://github.com/icaros-usc/wecook
[2]https://www.youtube.com/watch?v=1p2wBBmhPmk&t=138s
[3]https://www.youtube.com/watch?v=jAhQfH1PspU&t=119s
[4]https://www.youtube.com/watch?v=d3SZH7NFDjc&list=PL4C3C1C9AB9931360&index=75
[5]https://www.youtube.com/watch?v=XOGuTy-9clc&t=153s

[2] David Hallac, Peter Nystrup, and Stephen Boyd. Greedy gaussian segmentation of multivariate time series. *Advances in Data Analysis and Classification*, pages 1–25, 2018.

[3] Javier Marin, Aritro Biswas, Ferda Ofli, Nicholas Hynes, Amaia Salvador, Yusuf Aytar, Ingmar Weber, and Antonio Torralba. Recipe1m+: A dataset for learning cross-modal embeddings for cooking recipes and food images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2019.

[4] Katerina Pastra and Yiannis Aloimonos. The minimalist grammar of action. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1585):103–117, 2012.

[5] Hejia Zhang, Po-Jen Lai, Sayan Paul, Suraj Kothawade, and Stefanos Nikolaidis. Learning collaborative action plans from youtube videos. In *Proceedings of the International Symposium on Robotics Research (ISRR 2019)*, Hanoi, Vietnam, 2019.