Speech Enhanced Imitation Learning and Task Abstraction for Human-Robot Interaction

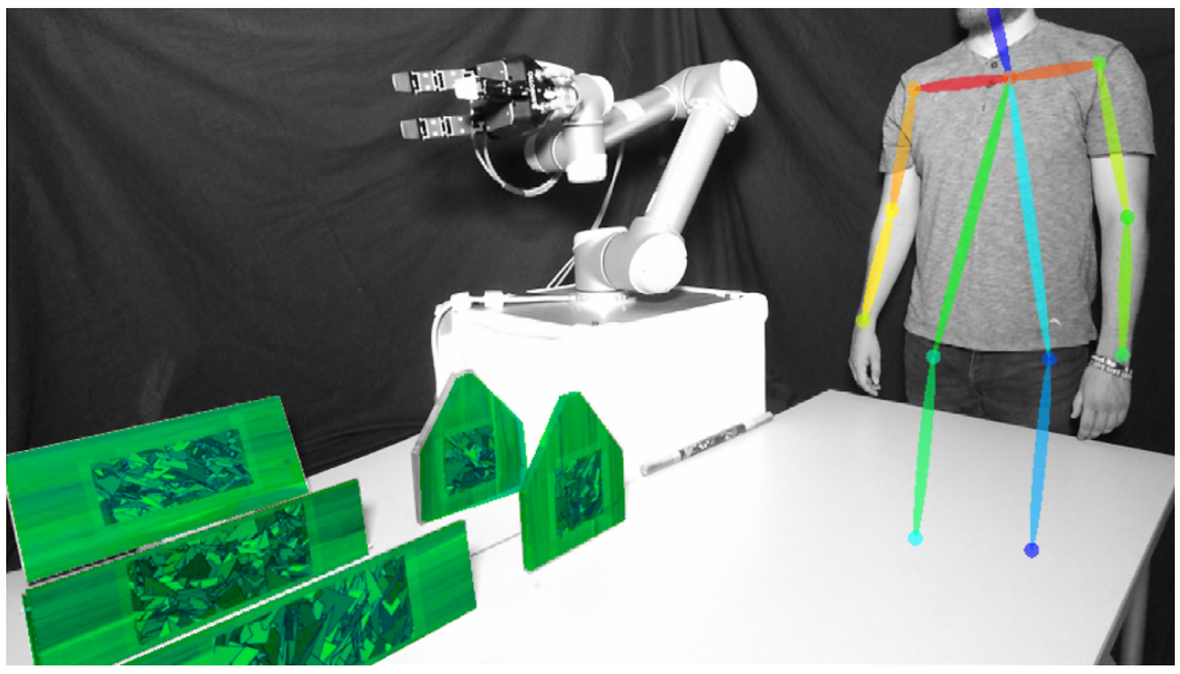
Simon Stepputtis, Chitta Baral and Heni Ben Amor

School of Computing, Informatics and Decision Systems Engineering
Arizona State University
{sstepput / chitta / hbenamor} @ asu.edu

Abstract

In this short paper, we show how to learn interaction primitives [1] and networks from long interactions by taking advantage of language and speech markers. The speech markers are obtained from free speech that accompanies the demonstration. We perform experiments to show the value of using speech markers for learning interaction primitives.

Concept



In this work we investigate how verbal instructions extracted from human speech can be used to segment and semantically annotate human demonstrations. We show that this information can be used to learn both

1. low-level interaction primitives, as well as
2. higher-level interaction networks that encode the transition model among primitives.

A feature of this approach is that verbal cues are used to outline the object of interest to ease the training process. Additionally, the robot can learn multiple skills from a single demonstration.

Experimental Results

Based on multiple demonstrations, synonyms for different keywords can be inferred. The following table shows the basic actions and their discovered synonyms (similar list for goals).

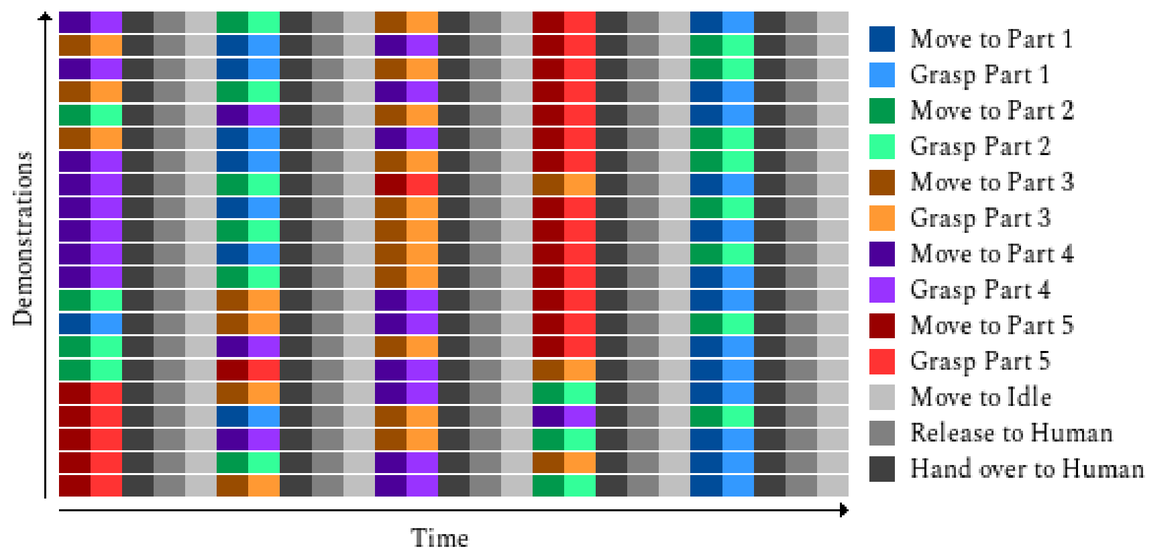
- **lift**: lift, hand over, handover, give, lift up, bring, hand
- **grasp**: grasp, pick, pick up, pickup
- **move**: move, goto, get, collect, move to, go to, moving, moved
- **release**: release, free

We performed three experiments to show the benefit of using speech markers during training in comparison to traditional methods.

1. **Speech assisted segmentation** is demonstrated in the first experiment. Based on five demonstrations, our system succeeds 80% of the time for the assembly task.
2. **Manual segmentation** was used in the second demonstration. Without the ability to outline the object of interest, the task was accomplished with a success rate of 0%.
3. **Generalizing towards different positions** is shown in the third experiment, where the robot was able to successfully (80%) grasp objects from random positions.

Approach

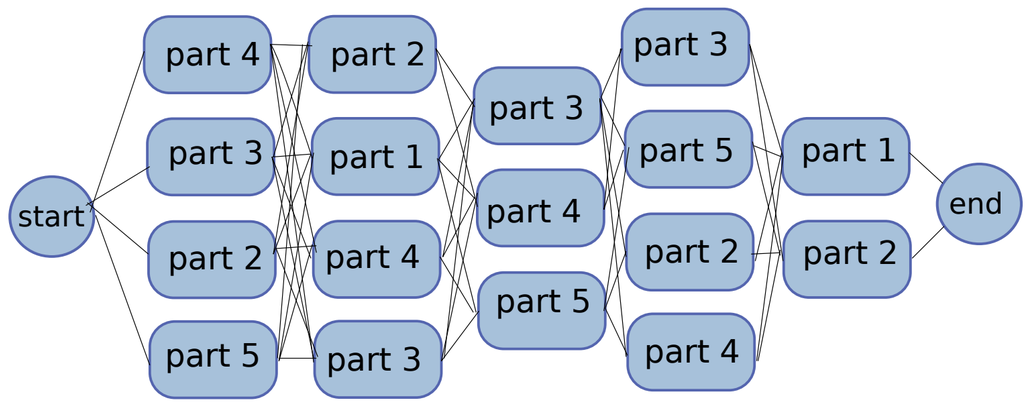
Our research goal is to observe only a few demonstrations of a given task and from that, learn the interaction primitives (such as, moving, grasping, lifting and releasing) as well as the interaction network (collaborative assembly).



Leveraging free speech gives us multiple benefits:

- Extraction of *actions* and *goals* (as well as their synonyms)
- Reduction of problem dimensionality (object of interest)
- Exploration of subtasks (training multiple actions)
- Partial ordering of explored subtasks (interaction network)

Based on the labeled data, Interaction Primitives are used to train the robot's behavior for the identified subtasks. Additionally, subtasks are used to create the interaction network as shown below.



At run time, the interaction network is transitioned independently by the robot. It chooses one of the actions in each layer while maintaining its ability to finish the task by not selecting actions twice or blocking itself from further execution.

Conclusion

In this paper we showed the following

1. Verbal clues can be used to outline the object of interest and greatly improve robot training.
2. Multiple tasks can be trained from a single demonstration by using speech markers to identify subtasks.

Future work will focus on using broader knowledge from speech to enable direct feedback and contextual questions from the robot to allow more natural interactions with the presented system.

References

[1] H. B. Amor, G. Neumann, S. Kamthe, O. Kroemer, and J. Peters, "Interaction primitives for human-robot cooperation tasks," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2831–2837, May 2014.