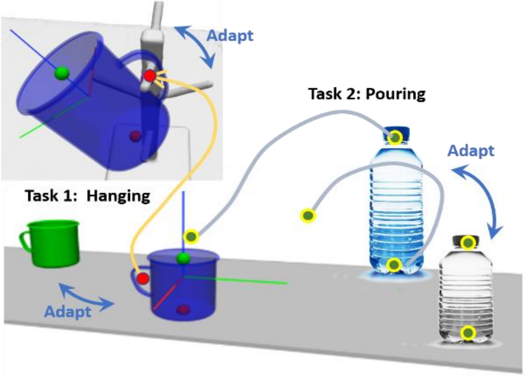
Dense-correspondence-based Visual Imitation Learning of Tool-Use Strategy



In recent years, imitation learning has been investigated as a way to efficiently and intuitively program autonomous behavior, where human demonstrates how to perform a task and a robotic system learns a policy to execute the given task by imitating the demonstrated tool-use strategies.

Visual demonstrations (RGB-D video) provides very detailed spatio-temporal information about the scene and the skill. The tool-use strategy can be described by a set of keypoints on the tool and the environment (including task constraints and goals), and the motion of the tool represented in a proper local coordinate system. In terms of task constraints, keypoints-based representation shows better intra-/extra-category generalization ability on both solid and deformable objects. The semantic meaning of each keypoint may vary across tasks. The task-awareness enables multi-modal and more versatile tool-use strategies. Explicitly extracting meaningful keypoints also gives the possibility to connect other components in symbolic (e.g. task modeling and planing) and subsymbolic level (motion generation and control).

In this work, firstly, you can use the dense correspondence model to extract dense keypoint from the visual demonstration and then use a out-of-box human/hand pose detection model to extract the motion trajectory. Secondly, utilize an optimization algorithm to find minimal keypoints or local frames to represent the task goals and constraints, which then will be used as task-parameters for movement primitive generalization. The goal is to learn from a few demonstrations and afterward successfully execute the demonstrated tool-use strategy with unsceen intra-category tools.

Relevant research questions include:

- How to improve the dense correspondence model in terms of intra-category generalization?
- How to use local-frame based representation of motion for better extrapolation in motion generation?
- How to extract task-aware tool-centric strategies from human demonstration?

This work will use the humanoid robot ARMAR-6, computer vision models and movement primitives:

- ArmarX (C++, Python): armarx.humanoids.kit.edu
- Dense Correspondence Network (Python)
- Via-point Movement Primitive (C++)

Contact: Jianfeng Gao (jianfeng.gao@kit.edu)

Institut für Anthropomatik und Robotik | Lehrstuhl Prof. Asfour (H²T) | www.humanoids.kit.edu