Human-Conditioned Models for Robotic Manipulation: A Preliminary Investigation

Adrian Vecina Tercero, Praminda Caleb-Solly, Liz Felton, and Nikhil Deshpande
School of Computer Science, University of Nottingham, UK
✉ psyav2@nottingham.ac.uk | 💻 www.adrianvecinatercero.co.uk

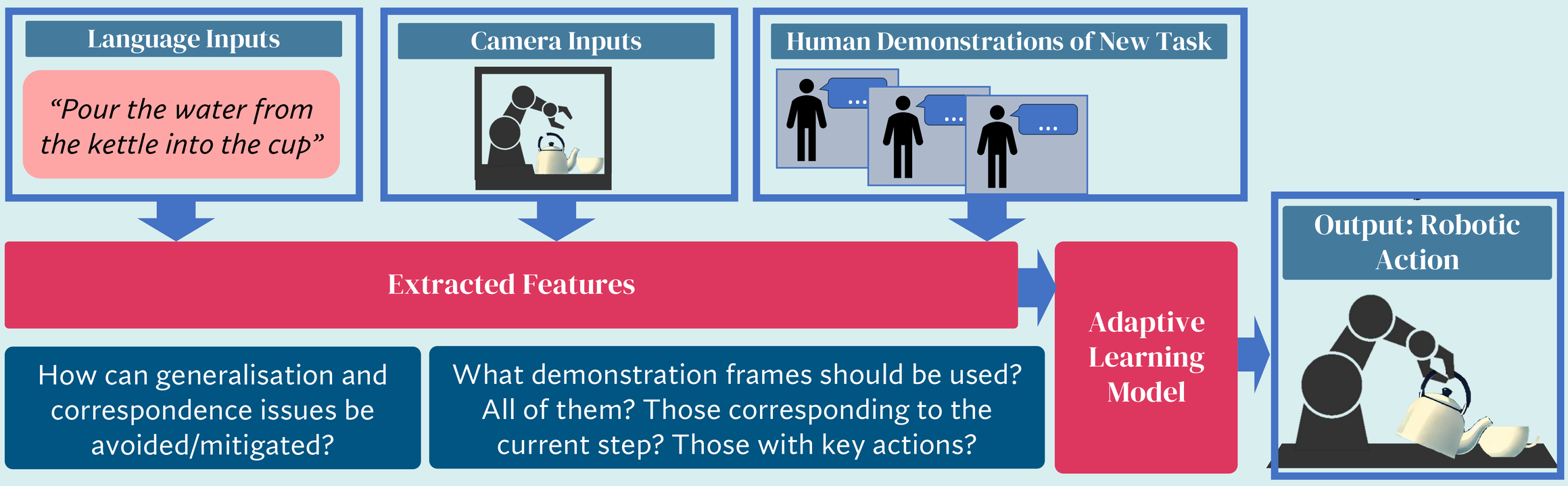
Motivation

- Robots deployed in dynamic environments need to adapt to new manipulation tasks quickly and effectively
- Adaptation in real-world settings can require complex updates to workflows during and/or after deployment
- Existing foundational models:
  - Struggle to generalise to new tasks/environments without computationally expensive re-training
  - Often require hundreds of expert demonstrations to learn tasks, which is impractical in dynamic scenarios

Human-Conditioned Models

Research Premise: Instead of relying solely on robot-centric demonstrations, which are hard to gather, robots should also be able to learn from "observing" and direct input from human teachers

The goal of our approach is to develop *"human-conditioned"* models able to learn tasks using few-shot observational learning, minimising the need for large task-specific datasets



How can generalisation and correspondence issues be avoided/mitigated?

What demonstration frames should be used? All of them? Those corresponding to the current step? Those with key actions?

A Human Observation Dataset

What input modalities are needed for effective task learning?

Observation Modalities

Visual Data
- Dynamic Body Pose
- Gaze Direction
- Hand Movements
- Object Segments
- Grasping Points

Verbal Data
- Spoken Guidance
- Task Rationale
- Object Description
- Non-Visual Elements (e.g. Force)

Sensor Data

Future Work

- Data Collection
  - We will record a dataset of human teachers performing manipulation tasks including multiple modalities of visual and verbal data
  - This dataset would contain multi-view, 3D recordings of demonstrations
  - Wearable sensors would also be included, to learn relationships between semantic descriptions, motion kinematics, and non-visual elements
- Feature Selection
  - Pre-trained models can be used to extract important features from the multimodal data
- Model training, conditioning and evaluation

Key Research Questions

- Can verbal instruction complement visual information for semantic learning?
- Can models conditioned through few-shot human observation perform new tasks without re-training?
- How can common issues in Learning from Demonstration such as correspondence be avoided?

Acknowledgments

This work was supported by CHART research, University of Nottingham. AI generated elements were used in this poster.

The University of Nottingham
UNITED KINGDOM • CHINA • MALAYSIA

Cyber-physical Health and Assistive Robotics Technologies