Hessisches Kompetenzzentrum
für Hochleistungsrechnen

Self-Supervised Multisensory Pretraining for Contact-Rich Robot Reinforcement Learning



Project Manager
Rickmer Krohn

Researchers
Noah Farr

Principal Investigator
Prof. Dr. Georgia Chalvatzaki

Project Term
2024 - 2025

Clusters
Lichtenberg II Cluster Darmstadt

Software
PyTorch

Additional Software
PyBullet

Institute
Interactive Robot Perception & Learning

University
Technische Universität Darmstadt

Introduction

Contact-rich manipulation remains one of the most challenging tasks in robotics, requiring precise control under noisy, dynamic conditions. The usage of multiple sensors enhances the robot's ability to understand the environment compared to vision only approaches. While Deep Reinforcement Learning (RL) enables robots to learn complex behaviors, effectively combining different sensory inputs, such as vision, force, and proprioception, remains a major challenge. Our work tackles this challenge by developing a scalable RL framework that fuses heterogeneous sensory data to enable robust manipulation. Inspired by advances in multimodal learning, we introduce Multisensory dynamic pretraining (MSDP) that simplifies policy learning and improves performance under sensory noise and changing object dynamics. To evaluate the approach, we conducted extensive experiments and ablation studies across multiple Task, sensor combinations and neural network architectures.

Methods

In this work, we introduce a new framework called MSDP (Multisensory Dynamic Pretraining) to improve robot manipulation through multisensory reinforcement learning. The framework learns a shared representation from multiple sensors using a self-supervised pretraining objective based on masked autoencoding and forward prediction. Each sensor input is processed through a dedicated encoder, with visual data passed

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN
TECHNISCHE UNIVERSITÄT DARMSTADT
UNIKASSEL VERSITÄT
Philipps Universität Marburg

through a CNN and low-dimensional sensors through linear layers. The encoded sensor features are fused using a transformer-based architecture trained to predict future sensor observations, enabling the model to understand the interaction between different modalities and their dynamics through time. For downstream reinforcement learning, we extract task-relevant features using a single cross-attention layer that maps the pretrained multisensory embeddings into a compact, stable input for policy learning. A key aspect of MSDP is the modular design that allows flexible use of sensor combinations allowing for efficient multisensory reinforcement learning.

Results

We successfully trained multiple model variations and baselines across three contact-rich manipulation tasks: Peg Insertion, Push Cube and Close Drawer. The results highlight how our method, compared to competitive baselines, extracts a rich multisensory representation for downstream RL. Our novel Cross-attention extraction has shown to be crucial for task-related feature extraction. Furthermore, we have shown how the usage of all sensors boosts performance compared to the vision-only setting. Additionally, multiple sensors can be used to enrich the vision representation even if not present during task learning. Our Results underline how the proposed MSDP framework exploits multiple sensors for efficient contact-rich robot reinforcement learning.

Discussion

The results of our project demonstrate the effectiveness of the MSDP framework to extract a rich multisensory representation for Robot RL. The use of HPC was crucial in achieving these results, as it enabled the parallel simulation and training of multiple neural network architectures and ablations, offering fast implementation cycles crucial for task and model tuning. Next to our proposed framework, this project has shown how RL can benefit from multiple sensor modalities. A promising research direction is to incorporate other sensor modalities e.g. tactile or sound. Our simulation results still need to be validated in real-world experiments. Future Research will focus on different settings like bimanual and multitask leveraging multisensory representations from MSDP.

Publications

Rickmer Krohn, Self-Supervised Multisensory Pretraining for Contact-Rich Robot Reinforcement Learning, German Robotics Conference / GRC, Nuremburg, Germany, March 13-15, 2025

Krohn, R.; Prasad, V.; Tiboni, G.; Chalvatzaki, G., Self-Supervised Multisensory Pretraining for Contact-Rich Robot Reinforcement Learning, German Robotics Conference, 2024

Last Update: 2025-09-08 12:11

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN
TECHNISCHE UNIVERSITÄT DARMSTADT
UNIKASSEL VERSITÄT
Philipps Universität Marburg