Visual Imitation Learning of Manipulation Tasks for Humanoid Robots

Jianfeng Gao, Sebastian Rietsch and Tamim Asfour

***Abstract*— Imitation learning offers an efficient and intuitive way to teach humanoid robots new manipulation skills. However, acquiring generalizable task representations from sparse visual demonstrations remains challenging. This extended abstract summarizes our work on developing a novel Keypoints-based Visual Imitation Learning (KVIL) framework for humanoid robots. KVIL, a bottom-up approach, focuses on identifying invariant task features and extracting subsymbolic and symbolic task representations from scarce human demonstration videos. It comprises four key components: Uni-KVIL for unimanual task learning, Bi-KVIL for bimanual coordination, Seq-KVIL for learning action sequences, and Pro-KVIL for probabilistic task representation and inter-category generalization. These approaches collectively enable object-centric, viewpoint-invariant, and embodiment-independent task representations, allowing humanoid robots to generalize learned manipulation skills across object instances and categories. The framework is evaluated through real-world experiments, showcasing its effectiveness in learning daily tasks.**

I. INTRODUCTION

Observational learning is a fundamental mechanism by which humans develop new skills by watching others and understanding the consequences of their actions. This ability enables skill acquisition through demonstration, mitigating the need for costly trial-and-error processes. Humans achieve this by identifying invariant task features, such as keypoints, from high-dimensional visual inputs, leveraging statistical evidence. Insights from cognitive psychology have motivated research in robotics to develop Visual Imitation Learning (VIL) techniques that mimic human observational learning mechanisms. However, acquiring generalizable symbolic and subsymbolic task representations from only sparse human demonstration videos remains a significant challenge.

This work adopts a bottom-up approach, extracting essential invariant task features from demonstrations without relying on ground-truth labels or the prior knowledge embedded in large language models, as is common in top-down approaches. At the subsymbolic level, we extract object-centric, keypoints-based geometric constraints as invariant features that represent object functional parts. Additionally, we employ neural object descriptors to facilitate the transfer of learned tasks to novel object instances or categories. At the symbolic level, our approach identifies common object parts that afford specific types of actions, enabling the transfer of the learned task representations across object categories sharing the same affordance. Furthermore, we model spatial relations between objects and their functional parts and generalize these constraints to novel tasks. Throughout the work, we employ variance-based statistical analysis to extract invariant task features, including keypoints, affordance regions, common viewpoints, and spatial and temporal constraints from sparse human demonstrations.

This work has been supported by the Carl Zeiss Foundation through the JuBot project, by the European Union's Horizon Europe Framework Programme under grant agreement No 101070596 (euROBIN), and (partially) by the German Federal Ministry of Education and Research (BMBF) under the Robotics Institute Germany (RIG).

The authors are with the Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology, Karlsruhe, Germany. E-mails: {jianfeng.gao, sebastian.rietsch, asfour}@kit.edu

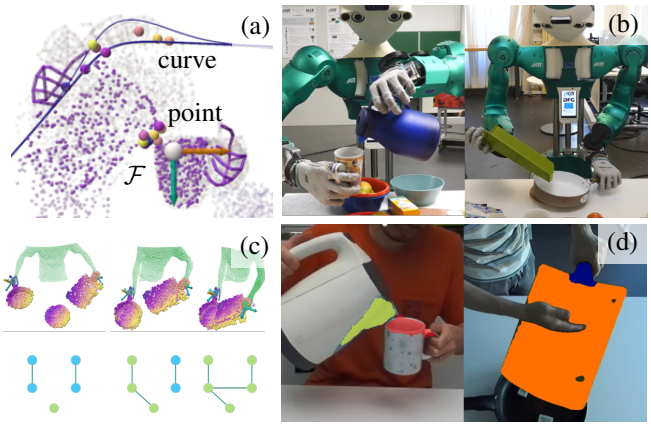

Fig. 1: Overview of the KVIL framework: (a) depicts the geometric constraints of Uni-/Bi-KVIL, (b) shows robot executions of pouring and transporting tasks, (c) illustrates the sub-actions and scene graph changes of Seq-KVIL, and (d) demonstrates extracted affordance regions of Pro-KVIL.

The research addresses the following key questions: i) How can task-relevant keypoints be effectively detected and extracted from visual input? ii) How can keypoint-based subsymbolic task representations be modeled to enable intra-category generalization? iii) How to extract affordance regions and spatial relations to facilitate inter-category generalization? iv) How can uni- and bimanual task representations be unified in compliance controllers?

II. CONTRIBUTION

The contributions of the work are as follows. i) Uni-KVIL for learning subsymbolic unimanual task representations with intra-category generalization; ii) Bi-KVIL for learning bimanual coordination strategies and control policies; iii) Seq-KVIL for learning a sequence of actions using hierarchical motion segmentation; and iv) Pro-KVIL for learning probabilistic affordance and spatial relations with inter-category generalization. See Fig. 1 for an overview of the work.

A. Uni-KVIL: KVIL of Unimanual Tasks

Uni-KVIL [1] focuses on learning generalizable subsymbolic task representations from human demonstration videos

of unimanual manipulation tasks. To this end, we proposed a perception pipeline for VIL that integrates and optimizes various pre-trained state-of-the-art computer vision models. This pipeline detects humans and objects, tracks their status, and detects dense correspondence between object instances – essential for addressing viewpoint mismatches and variations in object shape, pose, and appearance.

Based on high-quality data from this pipeline, we propose the *Principal Constraint Estimation* (PCE) algorithm that extracts sparse yet semantically meaningful keypoints on object functional parts, utilizing statistical variances of their spatial distribution across multiple demonstrations. PCE simultaneously extracts keypoint-based geometric constraints on principal manifolds, their associated local frames, movement primitives [2], and task-oriented grasps [3] as subsymbolic task representations. While most existing approaches learn only a subset of these representations, our approach provides a more comprehensive understanding of task constraints. We develop a novel keypoint-based admittance controller to reproduce the learned task on the robot. Our key insight is that the sparse object-centric representation, combined with dense correspondence detection, significantly enhances intra-category generalization. This allows KVIL to learn various daily tasks within 10 demonstration videos and to reproduce them in novel and cluttered scenes.

B. Bi-KVIL: Bimanual Coordination and Control

Bimanual manipulation presents unique challenges due to its complexity, which involves intricate object relationships, fine-grained motion details, and diverse coordination strategies between arms. Similarly to the unimanual case, bimanual manipulation tasks exhibit invariant features across multiple demonstrations. Bi-KVIL [4] extends Uni-KVIL's task representation to bimanual tasks by introducing a novel hybrid master-slave object relationship, unifying uni- and bimanual task representations. Various coordination strategies covering a complete bimanual manipulation taxonomy were extracted from human demonstrations. Fine-grained keypoint-based geometric constraints enable Bi-KVIL to capture detailed motion styles from demonstrations, paving the way for modeling personalized task representations. Based on the extracted bimanual coordination categories, we develop suitable real-time compliance controllers to address bimanuality, motion synchronization [5], compliance adaptation, and hybrid master-slave object relationships [4].

C. Seq-KVIL: Learning Action Sequences

Humans often demonstrate tasks as a sequence of actions, making the detection of common motion segments across demonstrations crucial for learning task representations. Building on hierarchical motion segmentation, we propose Seq-KVIL, a keypoint-based motion segmentation algorithm for VIL leveraging contact relation changes and keypoint motion characteristics in object-centric local frames. By leveraging dense point tracking, the algorithm precisely identifies temporal segment points and object relationships. Motion segments that share scene graph topologies across demonstrations are grouped to learn Bi-KVIL's task representations of spatial coordination, while temporal coordination is derived from temporal segments. By integrating the spatio-temporal coordination of Bi-/Seq-KVIL, we achieve a robust representation of bimanual coordination.

D. Pro-KVIL: A Probabilistic Representation of KVIL

Subsymbolic task representations alone are insufficient for fully modeling complex manipulation tasks, particularly those requiring semantic understanding. To address this limitation, we reformulate KVIL in a probabilistic framework, named Pro-KVIL, integrating subsymbolic and symbolic task representation such as affordance regions and spatial relations. Rather than extracting a set of deterministic sparse keypoints, we propose a novel probabilistic geometric constraint. The probability distribution of keypoints reflects the object's functional parts for a given task, effectively capturing affordance information when labeled in natural language. This extension enables Pro-KVIL to generalize subsymbolic constraints to novel object categories that share the same affordance type, facilitating inter-category generalization. Additionally, we develop a probabilistic spatial relation representation between objects or their functional parts. This representation accounts for distance, directional, and topological spatial relations while considering object size and shape. Modeling invariant view perspectives further enhances the generalization of subsymbolic constraints across different objects and tasks, significantly improving inter-category generalization.

III. CONCLUSION

Through these four interconnected parts, this research aims to provide a bottom-up keypoint-based visual imitation learning framework that derive both subsymbolic and symbolic representations from sparse human demonstration videos for both uni- and bimanual manipulation tasks. The main focus was on using state-of-the-art vision algorithms and statistical analysis tools to identify invariant features in human demonstration videos, which are key to achieving generalization at both the intra-category and inter-category levels. The developed framework has been evaluated in various daily tasks on humanoid robots.

REFERENCES

[1] J. Gao, Z. Tao, N. Jaquier, and T. Asfour, "K-VIL: Keypoints-Based Visual Imitation Learning," *IEEE Trans. on Robotics*, vol. 39, no. 5, pp. 3888–3908, 2023.

[2] Y. Zhou, J. Gao, and T. Asfour, "Learning via-point movement primitives with inter- and extrapolation capabilities," in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2019, pp. 4301–4308.

[3] Y. Cai*, J. Gao*, C. Pohl, and T. Asfour, "Visual imitation learning of task-oriented object grasping and rearrangement," in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2024, pp. 364–371.

[4] J. Gao, X. Jin, F. Krebs, N. Jaquier, and T. Asfour, "Bi-kvil: Keypoints-based visual imitation learning of bimanual manipulation tasks," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2024, pp. 16 850–16 857.

[5] J. Gao, Y. Zhou, and T. Asfour, "Projected force-admittance control for compliant bimanual tasks," in *IEEE/RAS Intl. Conf. on Humanoid Robots (Humanoids)*, 2018, pp. 607–613.