1st German Robotics Conference
March 13-15, 2025. Nuremberg, Germany

Scene Understanding Through Visual and Haptic Perception

Marko Pavlic[1], Timo Markert[2] and Darius Burschka[1]

***Abstract*— The industrial sector continuously demands efficient methods to enable non-expert operators to reprogram robots in a timely and cost-effective manner. Advances in task-level programming (TLP), robotic skill acquisition, and Learning from Demonstration (LfD) have yielded promising outcomes. Nonetheless, many existing approaches remain dependent on extensive datasets or necessitate prior user expertise in robotic systems. This paper introduces a framework for deriving parameterized skill sequences from passive observation of human demonstrations. These skill sequences reflect human behavior and enable the design of a task plan to execute on the robot. Since passive observation alone does not provide information about the physical properties of objects, which are critical for effective manipulation, our approach integrates robotic tactile and kinesthetic sensing to estimate both static and dynamic physical properties of the manipulated objects.**

I. Introduction

A new paradigm in robotics has emerged, emphasizing the flexibility of robotic systems and the transferability of robot programs across different and potentially novel tasks. A significant challenge in this context is empowering non-expert factory personnel to reprogram robots, which is traditionally a resource-intensive and time-consuming process during the setup of new tasks on production lines. The concepts of task-level programming (TLP), robot skills, and Learning from Demonstration (LfD) have demonstrated potential in providing adaptable solutions that are intuitive and accessible to shop-floor workers without necessitating specialized programming expertise [1] [2]. There are three primary approaches to demonstration: kinesthetic teaching, teleoperation, and passive observation [3]. Kinesthetic teaching enables the user to demonstrate by physically moving the robot through the desired motions. The demonstration quality depends on the user's dexterity and smoothness, often requiring smoothing or post-processing, even with expert input. Demonstrations become more challenging as the hardware's degrees of freedom (DOF) increase. Teleoperation involves controlling the robot via an external input, such as a joystick, graphical user interface, or similar devices. Unlike kinesthetic teaching, teleoperation allows LfD techniques to be used remotely without requiring the user to be physically present. However, it may involve extra effort for interface development, extended user training, and ensuring the availability of input hardware. In the third approach, the robot learns by passively observing a user performing a task. This method requires no training for the demonstrator and is well-suited for high-DOF or non-anthropomorphic robots where kinesthetic teaching is challenging. It offers convenience for factory workers, as it does not require wearing sensing devices and allows tasks to be demonstrated without concern for robotic execution. However, it necessitates encoding or learning a mapping from human actions to robot-executable commands, and solely relying on visual data makes it difficult to accurately determine the physical properties of the manipulated objects.

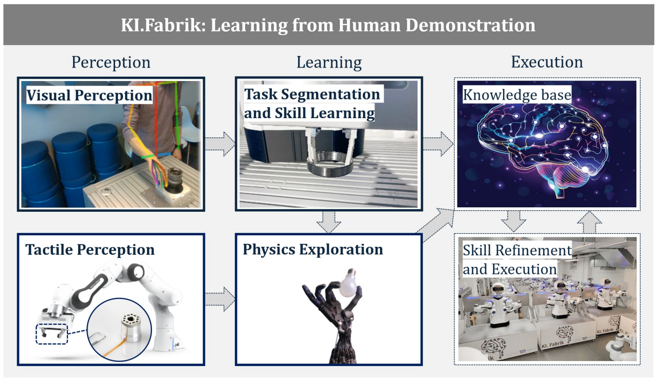
KI.Fabrik: Learning from Human Demonstration
Perception
Learning
Execution
Visual Perception
Task Segmentation and Skill Learning
Knowledge base
Tactile Perception
Physics Exploration
Skill Refinement and Execution

Fig. 1. A learning from human demonstration framework is essential for the factory of the future, where robots must autonomously acquire new tasks to enable flexible and adaptive assembly lines.

The proposed learning approach to overcome these challenges is illustrated in Fig. 1. The corresponding robot setup is shown in Fig. 2. A vision system in the robot's head observes and tracks human motions and interaction objects to establish a foundational understanding of the scene and the demonstrated task. Based on this visual input, the demonstration is segmented into discrete subtasks, referred to as skills, along with geometric grounding of the interaction objects. However, successful object manipulation and environmental interaction require knowledge of physical properties, which is not extractable from pure passive observation. In previous work [4], we showed how a robot manipulator is used to acquire tactile information during robotic manipulation. This physical information enhances the completion of learned skills and improves task execution. For adaptability in new environments, a skill refinement process is conducted through collective learning [5]. All acquired information is stored within a centralized knowledge base. This paper focuses on task segmentation from a passive observation described in

The authors acknowledge the financial support by the Bavarian State Ministry for Economic Affairs, Regional Development and Energy (StMWi) for the Lighthouse Initiative KI.FABRIK, (Phase 1: Infrastructure as well as research and development program, grant no. DIK0249).

[1]Marko Pavlic and Darius Burschka are with the Machine Vision and Perception Group, Chair of Robotics, Artificial Intelligence and Real-time Systems, Technical University of Munich, Munich, Germany. Darius Burschka is also a principal investigator of the KI.FABRIK project.

[2]Timo Markert is with Resense GmbH, Klingenberg, Germany and with the Semantic Information Systems Group, Osnabrück University, Osnabrück, Germany.

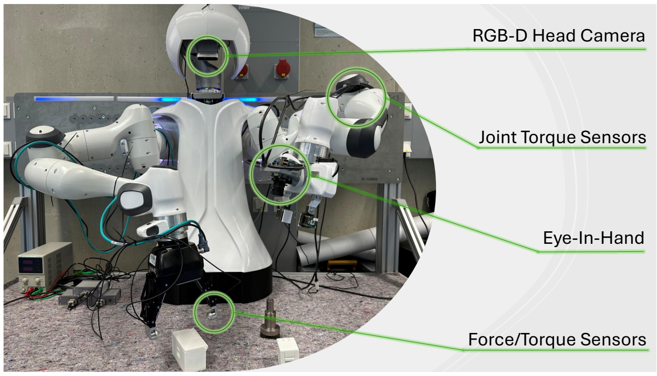

Fig. 2. A dual-arm manipulation platform utilizing two Franka Emika Panda robots is equipped with an RGB-D camera (Intel RealSense) mounted on the head and an eye-in-hand camera for scene observation. Resense HEX21 F/T sensors are embedded in the fingertips for haptic feedback.

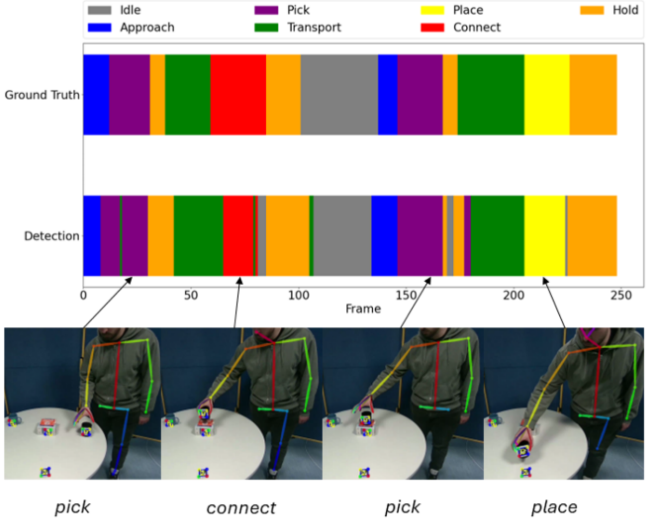

Fig. 3. Task segmentation from a human demonstration for a transportation task.

Sec. II and the physics exploration of manipulation objects described in Sec. III.

II. Task Segmentation

The current framework is designed for transportation tasks and enables the detection of the six skills: approach, pick, transport, place, connect, and hold. When none of these skills is detected, the status ”idle” is assigned. For each video frame, a state vector is used to concisely represent the current world state, which is necessary to recognize the skills from the demonstration. The state vector comprises features computed from estimated 6D poses from human keypoints and objects. Each robotic skill is characterized by two critical elements: preconditions and postconditions [1]. Preconditions define the criteria that must be satisfied before the skill can be executed, while postconditions verify whether the skill execution was successful. These conditions ensure consistency and reliability throughout the process. Our approach manually defines the pre- and postconditions for all six skills in the state vector space. It compares the extracted current state vector with the designed pre- and postconditions vectors to detect the correct skill. We defined the state vector so that pre- and postconditions of the six skills are unique. Fig. 3 shows the results of a transportation task demonstration. The ground truth was created by manual video inspection. The framework can reproduce the correct sequence with minor time delays. However, this is not a problem as the execution on the robot is not dependent on this. For the execution, we employed Behavior Trees (BTs) to implement these six robotic skills. BTs offer a modular, hierarchical framework well-suited for dynamic environments, enabling seamless mapping of the observed skill sequence to the robot platform.

III. Phyiscs Exploration

From the robot’s visual perception, the geometric grounding of the objects can be extracted. However, grasping the object based purely on the geometric features is often unreasonable. Parts with variable mass usually require different grasping forces to avoid unsuccessful or destructive grasping attempts [6]. So, physical parameters must be extracted for unknown objects. In our work [4], we model the forces acting on a rigid object in the end-effector and use the measurements of the torque sensors in the robot joints (kinesthetic) and dedicated force/torque (F/T) sensors in the fingertips (tactile) to solve for the unknown physical parameters using a least squares approach. From static measurements, where the object is grasped and moved to different positions and orientations, the mass, the center of mass, and the sensor offset are estimated. The performed motion significantly influences the results when identifying dynamic parameters. A finite sum of harmonic sine and cosine functions as the excitation trajectory for each joint shows advantages in terms of maximum-likelihood parameter estimation. We can identify the six unknowns of the inertia tensor from the dynamic measurements. Object property estimation using joint torques achieves sufficiently accurate results for rather heavy objects, underlining the results of existing works [7]. This approach requires no additional hardware but is limited to torque-controlled robots. The accuracy of estimating lighter objects is significantly affected by noise in the measurements since the forces acting on the rigid object are not directly measured but derived from the joint torques and the robot kinematics. In contrast, fingertip F/T sensors measure the interaction forces directly and provide highly precise estimates for light objects only constrained by the sensor’s measurement range. Previous studies have demonstrated the application of fingertip F/T sensors for determining additional object properties, such as stiffness, compliance [8], and surface texture [9].

IV. Conclusion

We present a scene-understanding framework capable of extracting parameterized skill sequences through passive observation of human demonstrations. These sequences can be directly transferred to a robotic platform for task execution in previously unseen environments. The system leverages kinesthetic and tactile sensing to infer critical, previously unknown physical properties of the interaction objects.

References

[1] M. R. Pedersen, L. Nalpantidis, R. S. Andersen, C. Schou, S. Bøgh, V. Krüger, and O. Madsen, "Robot skills for manufacturing: From concept to industrial deployment," *Robotics and Computer-Integrated Manufacturing*, vol. 37, pp. 282–291, 2016.

[2] F. Steinmetz, V. Nitsch, and F. Stulp, "Intuitive task-level programming by demonstration through semantic skill recognition," *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 3742–3749, 2019.

[3] H. Ravichandar, A. S. Polydoros, S. Chernova, and A. Billard, "Recent advances in robot learning from demonstration," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 3, no. 1, 2020.

[4] M. Pavlic, T. Markert, S. Matich, and D. Burschka, "Robotscale: A framework for adaptable estimation of static and dynamic object properties with object-dependent sensitivity tuning," in *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pp. 668–674, 2023.

[5] S. Schneider, Y. Wu, L. Johannsmeier, F. Wu, and S. Haddadin, "A scalable platform for robot learning and physical skill data collection," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5925–5932, IEEE, 2024.

[6] C. Wang, X. Zhang, X. Zang, Y. Liu, G. Ding, W. Yin, and J. Zhao, "Feature sensing and robotic grasping of objects with uncertain information: A review," *Sensors*, vol. 20, no. 13, p. 3707, 2020.

[7] C. Gaz and A. De Luca, "Payload estimation based on identified coefficients of robot dynamics — With an application to collision detection," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, (Vancouver, BC), 2017.

[8] T. Markert, S. Matich, E. Hoerner, J. Pfannes, A. Theissler, and M. Atzmueller, "Comparing human haptic perception and robotic force/torque sensing in a simulated surgical palpation task," *IEEE/RSJ International Conf. on Intelligent Robots and Systems (IROS)*, 2022.

[9] T. Markert, S. Matich, E. Hoerner, A. Theissler, and M. Atzmueller, "Fingertip 6-axis force/torque sensing for texture recognition in robotic manipulation," *2021 IEEE 26th International Conference on Emerging Technologies and Factory Automation (ETFA)*, 2021.