Get a grip — Accurate Robot–Human Handovers

Esranur Erturk[1]

[1]*Department of Computer Science, Lund University*

Abstract

This project addresses the precision challenges in robot–human handover processes, particularly in surgical environments. Two different approaches are explored: (1) learning from a few human demonstrations with hand and instrument tracking and (2) collecting data from remotely controlled robot movements using a joystick or VR system. The aim is to achieve precise, natural handovers without requiring the human to regrip, and this research focuses on recognizing user intentions, accurately predicting hand and tool poses, and evaluating robot responsiveness.

Keywords

Human-robot collaboration, Human-robot handovers, Surgical robotics

1. Introduction and Motivation

Precise human–robot collaboration is crucial in surgical environments, where surgeons expect instruments to be handed over without needing to adjust their grip or visually confirm the object. In human–human handovers, tools are placed directly into the hand in the correct orientation. However, existing robotic systems often deliver instruments to fixed positions or require the human to actively retrieve the object [1], failing to meet the precision requirements demanded in surgical contexts.

This project aims to overcome these limitations by developing a system that accurately detects the surgeon's hand pose and responds to intuitive voice commands to select and deliver the correct instrument. Unlike previous robotic nurse systems, which primarily relied on basic voice commands or predefined gestures[2], our approach emphasizes real-time situation awareness and dynamic precision during handovers[2][1]. Additionally, experimental comparisons between different learning models, including RT-2X and OpenVLA [1], will be conducted. Existing methods often use predefined movements and lack detailed real-time hand pose recognition. Our approach uniquely integrates detailed hand tracking and multimodal intent recognition for improved precision.

Most current imitation learning models focus on model architecture and training strategies but do not address the learning of collaborative actions. As a result, it remains unclear how well such models perform in real-world robot-human handovers and what modalities (such as force sensing or tactile feedback) are required for natural and precise collaboration. Although the overall goal of imitation learning is to simplify robot programming for non-experts, this aspect is rarely evaluated in existing research.

SAIS2025: Swedish AI Society Workshop 2025, 16-17 June 2025, Halmstad, Sweden.

esranur.erturk@cs.lth.se (E. Erturk)

0009-0004-5877-6976 (E. Erturk)

The project is carried out in collaboration with the Children's Heart Center at Skåne University Hospital in Lund and integrated into the WASP Research Arena WARA Robotics.

2. Approach

This project investigates two methodological approaches to achieving accurate robot-to-human handovers using the ABB YuMi Single Arm robot operating under ROS2. Robot motion control is performed through EGM, and gripper operations are managed via RWS.

- **Method 1 – Human to Human Demonstration:** Data is collected from human–human interactions, where YOLOv8 and SAM2 are used for accurate hand and instrument segmentation and hand pose tracking extracts detailed hand and finger configurations[2]. For each instrument, pre-trained neural networks predict its 6D pose (position and orientation), allowing the robot to imitate observed human movements during handovers[3].
- **Method 2 – Remote Joystick or VR Data Collection:** Robot movements are controlled remotely using a joystick or VR setup[4] and joint positions, end-effector poses, and control commands are recorded via ROS2. These datasets are used to train ACT models or fine-tune large Vision-Language-Action (VLA) models such as OpenVLA[4], diffusion policy methods are also explored[5].

Data will be collected in collaboration with hospital staff, including OR nurses and surgeons. The aim is to gather approximately 50 demonstrations per method and instrument. Experimental evaluation will focus on measuring handover accuracy, assessing the performance of hand pose and voice recognition, and analyzing user interaction and satisfaction across both data acquisition and deployment phases.

In both cases, the surgeon's hand pose is tracked in real time, and instruments are delivered in the correct orientation [2]. For voice command interpretation, speech signals are transcribed by an automatic speech recognition engine and subsequently processed by a transformer-based large language model (LLM) within the OpenVLA framework to yield appropriate robotic actions [1]. Surgeon user studies will assess usability, handover accuracy, and timing, and their feedback will guide iterative tuning to individual preferences.

3. Project status

The technical components have been fully implemented, and all modules are operational; however, data collection is still in progress and has not yet been completed. Next steps include finishing data collection, initial model evaluations, and structured user studies to assess performance, reliability, and user experience.

References

[1] S. Li, J. Wang, R. Dai, W. Ma, W. Y. Ng, Y. Hu, Z. Li, Robonurse-vla: Robotic scrub nurse system based on vision-language-action model, arXiv preprint arXiv:2409.19590 (2024).

[2] L. Wagner, S. Jourdan, L. Mayer, C. Müller, L. Bernhard, S. Kolb, F. Harb, A. Jell, M. Berlet, H. Feussner, et al., Robotic scrub nurse to anticipate surgical instruments based on real-time laparoscopic video analysis, Communications Medicine 4 (2024) 156.

[3] L. Wang, X. Chen, J. Zhao, K. He, Scaling proprioceptive-visual learning with heterogeneous pre-trained transformers, volume 37, 2024, pp. 124420–124450.

[4] T. Kamijo, C. C. Beltran-Hernandez, M. Hamaya, Learning variable compliance control from a few demonstrations for bimanual robot with haptic feedback teleoperation system, in: 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2024, pp. 12663–12670.

[5] J. Li, Y. Zhu, Y. Xie, Z. Jiang, M. Seo, G. Pavlakos, Y. Zhu, Okami: Teaching humanoid robots manipulation skills through single video imitation, in: 8th Annual Conference on Robot Learning, 2024.