From Perception to Action: A Pipeline for Autonomous Garment Manipulation

Julia Hohensee[1], Christian Dreher[1], Adreas Gaukel[2], and Tamim Asfour[1]

***Abstract*— Manipulation of deformable objects such as garments remains a major challenge in robotics due to their high-dimensional configurations, complex dynamics, and frequent self-occlusions. Previous studies on folding or dressing have often relied on detailed cloth models, extensive training, or large datasets, which limit their applicability. We present a perception–action pipeline for autonomous garment handling that focuses on reliably picking up and transporting garments. The system uses FoundationStereo for stereo-based depth estimation and Grounded SAM 2 for open-vocabulary segmentation to generate garment point clouds. Grasping is based on geometric heuristics, and manipulation is executed with Via-Point Movement Primitives learned from kinesthetic demonstrations, with force-based hand closure. Implemented on a humanoid robot, the system autonomously clears garments from a tabletop into a laundry basket, demonstrating that effective garment handling can be achieved without complex modeling or task-specific training.**

I. Introduction

Manipulation of deformable objects remains a fundamental challenge in robotics. Unlike rigid bodies, textiles exhibit high-dimensional, complex and often unpredictable dynamics, and frequent self-occlusions. Small interactions can produce large, nonlinear deformations, making perception, state estimation, and motion planning substantially more difficult than for rigid objects [1], [4]. Garments vary in shape, size, and material properties, and often appear crumpled or tangled, further complicating grasping and manipulation [1], [3].

Previous studies have addressed garment manipulation tasks such as folding, dressing, and flattening. Early approaches relied on deterministic sequences and precise models [8], [2], assuming full observability and accurate action execution. However, these assumptions often fail in real-world scenarios where perception is noisy due to occlusions and self-occlusions, and where manipulation actions are inherently unreliable, grasps may fail or inadvertently capture multiple entangled items [5]. While such sophisticated methods are necessary for complex reconfiguration tasks they may be overly complex for simpler manipulation scenarios. In contrast, for straightforward tasks such as picking and transporting crumpled garments into a basket, where the goal is merely to isolate and move items rather than achieve specific shape configurations, such elaborate modeling and planning frameworks are not necessary. Recent work has explored probabilistic planning frameworks, such as Partially Observable Markov Decision Processes (POMDPs), which explicitly model uncertainty in both perception and action outcomes [5]. These approaches achieve robust performance by accumulating information through repeated manipulation. However, POMDP methods require extensive empirical characterization of action success probabilities and observation noise through repeated experimental trials, representing a significant engineering investment that may only be justified for complex manipulation tasks requiring precise uncertainty quantification.

Building on these advances, we propose a pipeline for garment manipulation that focuses on robustly picking and transporting garments rather than performing complex reconfigurations or folding. Unlike prior work that relies on detailed cloth models, extensive task-specific training, or large datasets, our method leverages foundation models for perception and learned movement primitives for execution in a straightforward manner. By keeping the pipeline simple, we reduce engineering complexity and increase deployability in real-world scenarios, demonstrating that for the straightforward task of picking crumpled garments and placing them in a basket, effective garment handling can be achieved without overcomplicating perception or control.

II. Method

We implement a modular perception action pipeline for deformable garment manipulation, as illustrated in Fig. 1.

The pipeline uses stereo-based depth perception using FoundationStereo [7], open-vocabulary segmentation with Grounded SAM 2 [6], point-cloud-based object representation, heuristic grasp selection, and motion execution using Via-Point Movement Primitives (VMPs) [9].

At the beginning of each manipulation cycle, the robot orients its gaze to the tabletop and acquires stereo images of the scene. FoundationStereo generates a dense RGB-D representation, which is subsequently processed by Grounded SAM 2 [6] to segment objects corresponding to the semantic class *garment*. Two-dimensional segmentation masks are projected into three-dimensional space using the depth data, generating individual point clouds for each detected garment. To ensure consistency and feasibility, only garment instances detected within the last 20 seconds and located within predefined reachability bounds in the robot space are considered.

This work has been supported by the European Union's Horizon Europe Widening Program through the HERON project, and the German Federal Ministry of Research, Technology and Space (BMFTR) under the Robotics Institute Germany (RIG).

The authors are with the High Performance Humanoid Technologies Lab, Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology (KIT), Germany

[1]E-mails: {julia.hohensee,c.dreher,asfour}@kit.edu

[2]E-mail: wf9665@partner.kit.edu

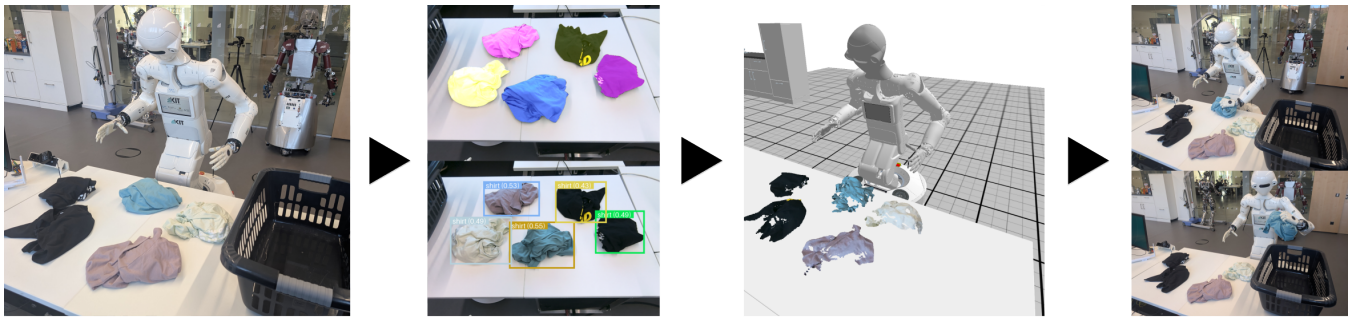

Fig. 1. The system uses FoundationStereo and Grounded SAM 2 for perception and segmentation of individual garments from point cloud data. Manipulation actions are generated using VMPs learned from kinesthetic demonstrations. Grasp execution is triggered based on force feedback.

Among all reachable garments, the next target is selected based on the minimal distance to the robot's camera. The grasping hand is chosen adaptively based on the lateral position of the garment relative to the robot: garments located to the right of the robot are grasped with the right hand, and vice versa. For garments outside the reachability of the selected arm, the robot performs a repositioning of itself to reduce the offset between the lateral shoulder and the object before grasping. After the robot repositioned itself, the position of the selected garment in the local robot frame changed. Thus, the new global garment position is recalculated using the robot's new pose.

Manipulation actions are executed using VMPs learned through kinesthetic teaching. The approach trajectory is initialized at the current end-effector pose and parameterized toward a grasp pose defined at the center of the garment point cloud. Hand closure is triggered using force feedback from a force-torque sensor mounted inside the wrist of the robot. After successfully grasping the garment, the arm is retracted via a second VMP, bringing the garment in front of the robot's torso. The robot moves to the laundry basket, whose location is currently fixed in the environment, places its hand over the basket, releases the garment, and returns to the table. It repeats this process until no more garments are detected. Bimanual manipulation is supported by alternating grasping actions between the left and right arms.

III. Preliminary Results

The garment manipulation pipeline was tested on the ARMAR-7 robot in two environments: the Automatica 2025 exhibition and our laboratory. While no systematic metrics were recorded, these deployments provided practical insights into system behavior.

Perception relies on Grounded SAM 2 for segmentation and FoundationStereo for depth reconstruction. Grounded SAM 2 is not fully reliable and occasionally misclassifies or misses garments, affecting grasp selection. Grasping is executed in open loop using pre-learned VMPs, which limits the robot's ability to handle garments lying flat on the table, as the current heuristics do not adapt to cloth shape. Despite these constraints, the pipeline effectively handles garments in crumpled configurations on the table, demonstrating a simple and robust approach to autonomous textile handling.

IV. Conclusion and Future Work

This work presented an integrated perception–action pipeline for autonomous manipulation of deformable garments on a humanoid robot. By combining stereo-based depth reconstruction, open-vocabulary garment segmentation using foundation models, simple geometric heuristics for grasp selection, and execution via movement primitives learned from demonstration, the system enables robust garment pickup and placement in unstructured household environments. The modular architecture allows the robot to generalize to previously unseen garment types without requiring garment-specific models or extensive task training. Experiments in our lab and at the Automatica 2025 demonstrated reliable operation across varying garment configurations and support continuous autonomous clearing of a tabletop into a laundry basket.

While the proposed approach achieves robust performance in a controlled setup, several limitations remain and motivate future work. First, grasp selection is based on geometric heuristics and does not explicitly reason about garment topology or grasp stability. Incorporating learned grasp quality estimation or cloth-specific representations could improve robustness in challenging configurations. Second, more fine-grained garment manipulation tasks, such as folding, placing garments on hangers, flattening them on a surface, or removing wrinkles, require further garment models and more sophisticated motion generation strategies than those employed in the current pipeline. In this context, integrating visual and tactile feedback into motion execution would enable more adaptive handling of complex cloth dynamics. Finally, the current pipeline relies on a hard-coded basket location. Future work could extend the system to perceive the basket using the same vision-based methods as for garments, enabling dynamic navigation toward the basket and even moving the basket itself, which would increase flexibility and robustness.

REFERENCES

[1] David Blanco-Mulero, Yifei Dong, Julia Borras, Florian T. Pokorny, and Carme Torras. T-DOM: A Taxonomy for Robotic Manipulation of Deformable Objects, December 2024. arXiv:2412.20998 [cs].

[2] Frederik Ebert, Chelsea Finn, Sudeep Dasari, Annie Xie, Alex Lee, and Sergey Levine. Visual Foresight: Model-Based Deep Reinforcement Learning for Vision-Based Robotic Control, 2018. Version Number: 1.

[3] Alberta Longhini, Yufei Wang, Irene Garcia-Camacho, David Blanco-Mulero, Marco Moletta, Michael Welle, Guillem Alenyà, Hang Yin, Zackory Erickson, David Held, Júlia Borràs, and Danica Kragic. Unfolding the Literature: A Review of Robotic Cloth Manipulation, 2024. Version Number: 2.

[4] Xiao Ma, David Hsu, and Wee Sun Lee. Learning Latent Graph Dynamics for Visual Manipulation of Deformable Objects, 2021. Version Number: 2.

[5] Pol Monso, Guillem Alenya, and Carme Torras. POMDP approach to robotized clothes separation. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1324–1329, Vilamoura-Algarve, Portugal, October 2012. IEEE.

[6] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded SAM: Assembling Open-World Models for Diverse Visual Tasks, 2024. Version Number: 1.

[7] Bowen Wen, Matthew Trepte, Joseph Aribido, Jan Kautz, Orazio Gallo, and Stan Birchfield. FoundationStereo: Zero-Shot Stereo Matching, 2025. Version Number: 4.

[8] Changshi Zhou, Haichuan Xu, Jiarui Hu, Feng Luan, Zhipeng Wang, Yanchao Dong, Yanmin Zhou, and Bin He. SSFold: Learning to Fold Arbitrary Crumpled Cloth Using Graph Dynamics from Human Demonstration, 2024. Version Number: 2.

[9] You Zhou, Jianfeng Gao, and Tamim Asfour. Learning Via-Point Movement Primitives with Inter- and Extrapolation Capabilities. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4301–4308, Macau, China, November 2019. IEEE.