Early Failure Detection in Humanoid Manipulation: A Comparison of In-Hand Sensing Modalities

Janna Nefzer and Tamim Asfour

***Abstract*— Early detection of manipulation failures enables timely recovery before task completion. We compare three in-hand sensing modalities–hand Proprioception, Force, and Time-of-Flight–for failure detection during grasping and lifting actions on the humanoid robot ARMAR-DE. Using 165 labeled executions on unknown objects, we trained Decision Tree classifiers with varying temporal sampling strategies for each modality. Hand Proprioception achieves the highest accuracy (96.97%) at the earliest detection point (50% execution time), substantially outperforming other modalities. Our findings show that Finger Encoder readings enable reliable failure detection at mid-execution when recovery actions remain feasible.**

I. INTRODUCTION

For the deployment of robots in real-world applications, it is crucial that they are not only able to function in normal situations but also when abnormalities and failures occur. To achieve this, failures should be detected early and with high accuracy. In the context of failure detection for robotic manipulation, several approaches have been proposed; however, most of them focus primarily on vision data [1], [2], [3], while humans rather use tactile sensory information from their hand to monitor manipulation processes [4]. To address this, we propose a failure detection approach for grasping and lifting actions that leverages multiple in-hand sensor modalities. We further investigate how early within an action each modality can be used to detect failure by training Decision Tree models for every single modality, as well as for their combination at different Points-in-Time. The Points-in-Time are sampled uniformly, and we additionally employ Points-in-Time that mark relevant phase transitions in human manipulation.

II. RELATED WORK

Many publications in the field of failure detection and recovery have shown that vision data from head-mounted cameras and robot proprioception data can be used to detect failures, such as [1], [2], [3]. However, investigations in humans show that the tactile system plays a central role in the prediction and monitoring of manipulation actions while vision only provides indirect information about mechanical interactions [4]. For this reason, we want to investigate failure detection based on in-hand sensor modalities. Further, enhancing vision data with additional sensor readings, such as audio or tactile sensors, increases the detection performance [1], [5], [6], [7]. With the exception of [6], which employs a three-finger robotic hand, all of the previously discussed works use grippers. Prior work by Hegemann et al. [8] uses a humanoid hand and further incorporates a Force-Torque sensor, in addition to RGB-D camera data and proprioception to learn task models capable of predicting action transitions, which are used to detect failures in manipulation tasks of humanoid robots. Compared to the approach of Hegemann et al. [8], we leverage additional proprioceptive information from the hand and complement it with a palm-mounted Time-of-Flight sensor. Further, their approach primarily focuses on identifying failures via transitions between actions rather than detecting failures within an ongoing action. Therefore, failure is only detected at the end of an action, and recovery is attempted by repeating the failed action. However, this is not suitable for all failure cases and thereby induces a delay between the occurrence of a failure and its detection. To reduce detection latency, we analyze how early in an action failures can be detected using multiple in-hand sensor modalities.

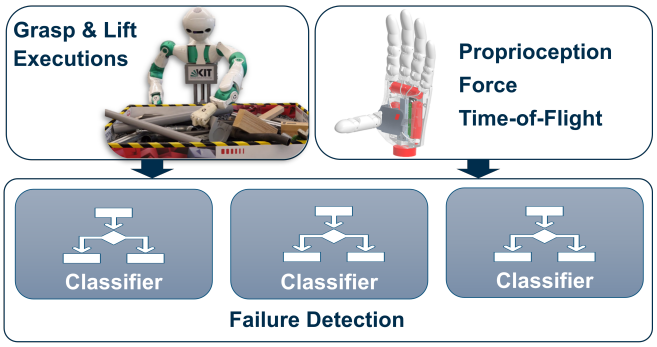

Fig. 1: Failure detection for manipulation actions based on three in-hand sensor modalities. Grasp and lift actions on unknown objects in cluttered scenes are executed on the humanoid robot ARMAR-DE, during which in-hand-sensor data is recorded with 10 Hz. This data is pre-processed according to different strategies and used to train Decision Tree-based classifiers to detect failures.

III. APPROACH

We employ the humanoid robot ARMAR-DE to grasp and lift previously unknown objects of different sizes, shapes, and

The research leading to these results has received funding from the German Federal Ministry of Research, Technology and Space (BMFTR) under the Robotics Institute Germany (RIG) and the competence center ROBDEKON, as well as from the Baden-Württemberg Ministry of Science, Research and the Arts (MWK) as part of the state's "digital@bw" digitization strategy in the context of the Real-World Lab "Robotics AI".

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

materials from cluttered environments. This setting is challenging for grasping and lifting actions, making it suitable for studying failure detection. To plan and execute grasps on previously unknown objects, the MAkEable-Framework [9] is employed. Further, the robot is equipped with the ARMAR-DE hand [10] including the following in-hand sensor modalities: (1) Force sensor in the wrist, (2) Time-of-Flight sensor in the palm, (3) relative encoder for thumb flexion, (4) relative encoder for index finger flexion and (5) relative encoder for the mechanism combining middle, ring, and pinky flexion.

A. Data Collection

In total, the obtained dataset contains 165 grasp and lift actions. Each execution is labeled as either *successful* or *unsuccessful*. For unsuccessful grasps, a failure reason (e. g., "hand closed too early" or "grasp unstable") is added by a human operator; for successful grasps, the reason "none" is automatically selected. The in-hand sensor modalities described above are recorded with approximately $10\,\mathrm{Hz}$ for all actions. Additionally, information on the manipulation process, such as timestamps of relevant Points-in-Time during execution. The selection of relevant Points-in-Time is inspired by the evaluations of human-grasping actions from Johansson et al. [4]. The authors divide the process of grasping and lifting an object into the phases *Reach*, *Load*, and *Lift*. Similarly, the following relevant timestamps are recorded:

- $T_{pre-pose}$: The hand is at a pre-pose close to the object and starts reaching for the object
- $T_{contact}$: The hand is in contact with an object, and the fingers start closing
- T_{lift}: The finger movement is completed, and lifting the object starts
- T_{end}: The hand reached the goal height

B. Data Pre-Processing and Models

The recordings are divided into a training and a test set, with the training set comprising $80\,\%$ of the data. The partitioning is carried out with respect to the failure reason labels, which also takes the binary failure classification into account. Each action recording is limited to the interval between $T_{pre-pose}$ and T_{end}. The Finger Encoder values are not further pre-processed, as they are inherently in $[0, 1]$. Pre-processing of the Time-of-Flight sensor data produces an 8×8 matrix in which each element records the measured depth value whenever an object falls within the sensor's field of view. If the field of view contains no object, a maximum default value is used. From the matrix containing depth values, the 0.1 quartile is used as input for the model. The captured Force sensor measurements contain a three-dimensional force vector, whose norm is computed for further analysis. The Time-of-Flight and Force data are normalized before being used as input to the model. Johansson et al. [4] describe that human grasp phases can be divided by simple decision rules. For example, the ending of the reach phase is characterized by the digits having contact with the object. To achieve a human-interpretable, simple, rule-based model, we employ Decision Trees. To identify the optimal model configuration, the tree-depth hyperparameter was tuned over the range of 2 to 10, and 5-fold cross-validation was employed during the training process. To assess the effectiveness of different input modalities for failure detection and to analyze how early they provide relevant information about failures, we vary the model inputs. First, we train the models using either individual sensor modalities or a combination of all five sensor readings. Second, we examine the impact of sensor measurement timing during the manipulation process. The following variations are investigated:

- Sensor readings at each single relevant Point-in-Time
- Sensor readings at combinations of consecutive relevant Points-in-Time
- Sensor readings at a percentage p of the whole execution time with $p \in \{0, 5, 10, \ldots, 95, 100\}\%$
- Sensor readings at n Points-in-Time uniformly sampled from the whole execution time with $n \in [1, 2, 4, 5, 10, 20, 50, 70, 100, 120, 150]$

IV. RESULTS

Both the individual sensor evaluations and the combined evaluation of the three sensors achieve their maximum accuracy when sensor measurements are evaluated at a specific percentage of the total execution time. Evaluations performed at single or multiple consecutive relevant Points-in-Time, as well as sampling over the entire execution, did not achieve higher accuracy on the test set. Best detection accuracy on the test set with corresponding evaluation time percentage for single modalities and combined modalities is shown in Table I. When comparing the three modalities and their combination, the highest accuracies over all time variations are achieved by the proprioception modality, reaching the maximum of $96.97\,\%$ on the test set at the timestamps at $50\,\%$ of the time needed for the entire execution. This is also the best prediction result over all evaluations.

Sensor Modality	Max Accuracy (%)	Execution Time (%)
Proprioception	**96.97**	50
Time-of-Flight	84.85	55
Force	93.94	75
Combination	93.94	**35**

TABLE I: Highest detection accuracy and corresponding execution time percentage for individual sensor modalities and their combination.

V. CONCLUSION AND FUTURE WORK

All in all, the proposed detectors can reach detection rates up to $96.97\,\%$ based on in-hand modalities. In our dataset, the Finger Encoders reached the best detection accuracies and achieved this already after half of the execution time. In future work, we want to add more recordings to our dataset to enable more detailed evaluation, and also include other manipulation actions, such as placing.

REFERENCES

[1] A. Inceoglu, E. E. Aksoy, A. C. Ak, and S. Sariel, "FINO-Net: A deep multimodal sensor fusion framework for manipulation failure detection," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 6841–6847.

[2] S. Thoduka, J. Gall, and P. G. Plöger, "Using visual anomaly detection for task execution monitoring," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 4604–4610.

[3] S. Thoduka, S. Houben, J. Gall, and P. G. Plöger, "Enhancing video-based robot failure detection using task knowledge," in *2025 European Conference on Mobile Robots (ECMR)*. IEEE, 2025, pp. 1–6.

[4] R. S. Johansson and J. R. Flanagan, "Coding and use of tactile signals from the fingertips in object manipulation tasks," *Nature Reviews Neuroscience*, vol. 10, pp. 345–359, 2009.

[5] F. Zhu, L. Wang, Y. Wen, L. Yang, J. Pan, Z. Wang, and W. Wang, "Failure handling of robotic pick and place tasks with multimodal cues under partial object occlusion," *Frontiers in Neurorobotics*, vol. 15, p. 570507, 2021.

[6] P. Gohil, S. Thoduka, and P. G. Plöger, "Sensor fusion and multimodal learning for robotic grasp verification using neural networks," in *2022 26th International Conference on Pattern Recognition (ICPR)*. IEEE, 2022, pp. 5111–5117.

[7] S. Thoduka, N. Hochgeschwender, J. Gall, and P. G. Plöger, "A multimodal handover failure detection dataset and baselines," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 17 013–17 019.

[8] P. Hegemann, T. Zechmeister, M. Grotz, K. Hitzler, and T. Asfour, "Learning symbolic failure detection for grasping and mobile manipulation tasks," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 4302–4309.

[9] C. Pohl, F. Reister, F. Peller-Konrad, and T. Asfour, "Makeable: Memory-centered and affordance-based task execution framework for transferable mobile manipulation skills," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 3674–3681.

[10] J. Starke, F. Hundhausen, P. Weiner, S. Rader, E. Hyseni, and T. Asfour, "The KIT robotic hands – a scalable humanoid hand platform with multi-modal sensing and in-hand embedded processing," in *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2025, pp. 8479–8486.