Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Using Tactile Senses in Multimodal Robot Environments

Dissertation
zur Erlangung des akademischen Grades
Dr. rer. nat
an der Fakultät für Mathematik, Informatik und Naturwissenschaften
der Universität Hamburg

Eingereicht beim Fachbereich Informatik
von Yannick Jonetzko

Dezember 2024

Gutachter:
Prof. Dr. Jianwei Zhang
Prof. Dr. Stefan Wermter

Tag der Disputation: 10.06.2025

Abstract

This dissertation investigates the role and integration of the sense of touch in multimodal robot environments. The objective is to utilize tactile data in different application areas and to investigate how and where it can be employed in a useful way, always in conjunction with other modalities.

In the initial phase, tactile data is used to gain deeper insights into the robot's surrounding environment. In one experiment, the robot combines tactile information with acoustic data to classify the contents of different containers. The objective is to investigate whether tactile data is as effective as acoustic information in achieving a comparable success rate and whether the combination of both modalities improves the results. Furthermore, the robustness of the classification in the presence of background noise will be investigated.

In the next step, the sense of touch is combined with proprioceptive data for the purpose of analyzing the intrinsic state of the robot. In this experiment, the robot is equipped with an underactuated gripper with an unknown joint state configuration. With the help of tactile sensors on the phalanges and position data from the motors, it is possible to reliably estimate the state of the gripper.

Once the robot has gathered information about the environment and its own condition, the human is integrated into the interaction. The third research phase is concerned with the effective and simple presentation of tactile data to humans. To this end, a robot-independent teleoperation system is developed that employs Mixed Reality and Virtual Reality glasses for the control of arm poses and gripper states and for the provision of feedback. In a study, participants control a robotic arm, perform manipulation tasks, and receive feedback on tactile data in the form of visualizations and acoustic signals.

In the final stage of this thesis, a custom haptic display is designed and constructed, which produces haptic feedback within a Virtual Reality environment. The device is capable of positioning various objects on a table in a manner that aligns with their virtual representations within the virtual environment. This enables users to perceive not only the virtual objects but also to physically interact with them through haptic feedback, thereby enhancing the immersive experience in Virtual Reality.

Zusammenfassung

Diese Dissertation untersucht die Rolle und Integration des Tastsinns in multimodalen Roboterumgebungen. Ziel ist es, taktile Daten in verschiedenen Anwendungsbereichen zu nutzen und zu erforschen, wie und wo sie sinnvoll eingesetzt werden können – stets in Kombination mit anderen Modalitäten.

Im ersten Schritt werden taktile Daten verwendet, um mehr über die Umgebung des Roboters herauszufinden. In einem Experiment kombiniert der Roboter taktile Informationen mit akustischen Daten, um den Inhalt verschiedener Behälter zu klassifizieren. Der Fokus liegt dabei auf der Frage, ob taktile Daten eine vergleichbare Erfolgsquote wie akustische Informationen erzielen und ob die Kombination beider Modalitäten die Ergebnisse verbessert. Darüber hinaus wird untersucht, wie robust die Klassifikation gegenüber Störgeräuschen ist.

Im nächsten Schritt wird der Tastsinn mit propriozeptiven Daten kombiniert, um den intrinsischen Zustand des Roboters zu analysieren. In diesem Experiment ist der Roboter mit einem unteraktuierten Greifer ausgestattet, dessen exakte Konfiguration unbekannt ist. Mithilfe taktiler Sensoren an den Fingergliedern und Positionsdaten der Motoren gelingt es, den Zustand des Greifers zuverlässig zu erfassen.

Nachdem der Roboter Informationen über die Umwelt und seinen eigenen Zustand gesammelt hat, wird der Mensch in die Interaktion integriert. Der dritte Forschungsabschnitt untersucht, wie taktile Daten dem Menschen effektiv und einfach präsentiert werden können. Dazu wird ein roboterunabhängiges Teleoperationssystem entwickelt, das Mixed-Reality- und Virtual-Reality-Brillen sowohl zur Steuerung von Armpositionen und Greiferzuständen als auch zur Bereitstellung von Feedback nutzt. In einer Studie steuern Teilnehmende einen Roboterarm, führen Manipulationsaufgaben aus und erhalten Rückmeldungen zu taktilen Daten in Form von Visualisierungen und akustischen Signalen.

Im letzten Schritt wird ein haptisches Display entworfen und konstruiert, das haptisches Feedback in einer Virtual-Reality-Umgebung bietet. Das Gerät kann verschiedene Objekte auf einem Tisch so positionieren, dass sie mit den virtuellen Darstellungen in der VR-Umgebung übereinstimmen. Dadurch können Nutzer die virtuellen Objekte nicht nur sehen, sondern auch physisch mit der kompletten Hand ertasten, was die realistische Erfahrung in der Virtual Reality erheblich verbessert.

Contents

1 Introduction **1**
- 1.1 Motivation . . . 1
- 1.2 Aim of this Thesis . . . 2
- 1.3 Fundamental Questions and Contributions . . . 3
- 1.4 Structure of this Thesis . . . 4
- 1.5 Publications . . . 5

2 Fundamentals **9**
- 2.1 Biological Sense of Touch . . . 9
- 2.2 Tactile Sensor Technologies . . . 10
- 2.3 ROS . . . 13
- 2.4 Unity . . . 14
- 2.5 Virtuality Continuum . . . 15

3 Multimodal Classification **17**
- 3.1 Related Work . . . 18
  - 3.1.1 Tactile Analysis and Classification . . . 18
  - 3.1.2 Acoustic Analysis and Classification . . . 20
  - 3.1.3 Multimodal Analysis and Classification . . . 20
- 3.2 Setup . . . 21
  - 3.2.1 Robot Platform . . . 22
  - 3.2.2 Tactile Sensors . . . 23
  - 3.2.3 Microphone . . . 24
  - 3.2.4 Pill Container . . . 24
  - 3.2.5 Pill Classes . . . 24
- 3.3 Approach . . . 25
  - 3.3.1 Sample Selection . . . 25
  - 3.3.2 Mel Frequency Cepstral Coefficients . . . 27
  - 3.3.3 Network Architecture . . . 27
- 3.4 Evaluation . . . 29
  - 3.4.1 Experiment . . . 29
  - 3.4.2 Results . . . 30
  - 3.4.3 Discussion . . . 32

3.5 Conclusion . . . 32

4 Robot State Estimation **35**
4.1 Related Work . . . 37
4.2 Hardware Setup . . . 38
4.2.1 3-Finger Adaptive Gripper . . . 38
4.2.2 Contact Sensor . . . 39
4.2.3 Visual Tracking . . . 40
4.3 Approach . . . 41
4.3.1 Baseline - Mathematical Analytical Approach . . . 43
4.3.2 Recurrent Neural Network Approach . . . 44
4.4 Experiments . . . 44
4.5 Results . . . 46
4.6 Discussion . . . 47
4.7 Conclusion . . . 50

5 Mixed Reality Teleoperation **53**
5.1 Related Work . . . 54
5.1.1 Teleoperation with Extended Reality . . . 54
5.1.2 Feedback Methods in Extended Environments . . . 55
5.2 Teleoperation Approach . . . 56
5.2.1 Pose Tracking . . . 56
5.2.2 Robot . . . 59
5.2.3 Extended Environment . . . 59
5.2.4 Registration . . . 60
5.2.5 Jogging . . . 61
5.2.6 Tactile sensing . . . 62
5.3 User Study . . . 62
5.3.1 Experiment Setup . . . 62
5.3.2 Tasks . . . 63
5.3.3 Conditions . . . 63
5.3.4 Measurements . . . 65
5.3.5 Conduct . . . 66
5.4 Results . . . 67
5.4.1 Quantitative Results . . . 67
5.4.2 Qualitative Results . . . 68
5.5 Discussion . . . 69
5.6 Conclusion . . . 71

6 Haptic Feedback **73**
6.1 Related Work . . . 74
6.1.1 Encountered-Type Haptic Displays . . . 74
6.1.2 Magnetic Control . . . 76
6.2 Encountered-Type Haptic Display . . . 76
6.2.1 Requirements . . . 76
6.2.2 Design . . . 77
6.3 Technical Evaluation . . . 83
6.3.1 Payload, Velocity, and Magnet Size . . . 83
6.3.2 Friction Compensation . . . 84
6.3.3 Object Recovery . . . 86
6.3.4 Interaction Detection . . . 87
6.3.5 Results . . . 87
6.3.6 Discussion . . . 88
6.4 User Study . . . 89
6.4.1 Setup . . . 90
6.4.2 Task . . . 91
6.4.3 Measurements . . . 92
6.4.4 Conduct . . . 93
6.4.5 Results . . . 94
6.4.6 Discussion . . . 96
6.5 Conclusion . . . 97

7 Conclusion and Perspectives **99**
7.1 Future Work . . . 100

8 Appendix **103**

List of Figures **125**

List of Tables **127**

Acronyms **127**

1 Introduction

1.1 Motivation

Most people do not realize the importance of the 'sense of touch' in everyday life. It is only when the sense of touch is no longer functioning that we become aware of the extent to which it enables us to interact with our environment, to perceive objects, to operate machines, and even to perform simple tasks such as getting dressed in the morning. This experience can be replicated by placing the hands in ice-cold water until they become numb. Even simple actions which were previously carried out without difficulty, such as operating a smartphone or opening a sweet wrapper, suddenly become challenging [1]. In 1984, Westling and Johansson [2] conducted an experiment in which they anesthetized the skin of participants, thereby ensuring the absence of the sense of touch. They then had the participants perform simple tasks, such as holding objects. Despite the participants' ability to utilize all their other senses and observe their actions, it was observed that the candidates had difficulty in making precise movements and maintaining the stability of the objects.

The same issue also applies, to a certain extent, to robots. In contrast to humans, robots can be customized and designed for specific purposes. For example, a robot can be equipped with an infinite joint for stirring coffee, which is not a feasible feature for a human. The human visual system comprises two eyes, which enable the perception of space. However, as shown in [2], this is not stable without the additional input of the sense of touch. A robot can be equipped with a greater number of visual sensors, including depth cameras, infrared cameras, or laser scanner, independent of the location as in the robot's head or in the wrist. With a camera in the wrist, a grasp can be monitored with millimeter precision. Nevertheless, it may be the case that this sense is of no benefit, for example due to occlusion. Robots, just like humans, benefit from the tactile sense in order to sense contacts with greater precision, for example when grasping, in order to observe features such as pressure, contact point, and slipping. Furthermore, object characteristics can also be perceived with greater accuracy with the sense of touch, such as weight, stiffness, or how the surface feels and how it behaves [1].

The significance of tactile sensors in robotics was first acknowledged in the early 1980s [3], with initial surveys indicating considerable promise for their application in this field [4, 5, 6]. Meanwhile, tactile sensors have reached a point where they integrate multiple sensors to perceive not only contacts but also other properties, such as temperature or stiffness [7]. The development of a sensor that is capable of matching human performance in all aspects remains a challenge. While individual requirements can be met or surpassed, a sensor that outreaches the human sense

of touch is not yet developed, to the best of our knowledge. It is unclear whether this will ever be possible.

When we talk about 'multimodal robot environments', on the one hand, the environment in which robots interact is meant. This contains a range of scenarios, including the operation of service robots. The robots are required to move around in a domestic environment and interact with the same objects and tasks as humans do. They must respond to a wide range of stimuli and fulfill resulting tasks, including untidy apartments, dirty dishes in the kitchen, meal preparation and setting the table, and, last but not least, interacting with people or even animals. To accomplish these tasks, robots must utilize all their senses to perceive the complexity of the environment and carry out the necessary actions. On the other hand, the term 'multimodal robot environments' contains the numerous sensor possibilities that can be integrated into a robotic system, enabling the perception and interaction with the aforementioned stimulus-rich environment. In this context, sensors are designed to emulate the functions of human senses, such as cameras, tactile sensors, and microphones. However, modalities that are not available to humans, such as laser scans or thermal images, can also be used.

Tactile sensors can assist in perceiving a multimodal environment, whether in the analysis of objects, grasps, the intrinsic values of the robot, or in interactions with humans or other robots. In an ideal scenario, this is done together with other modalities. What happens when other senses are unavailable? Can the robot rely exclusively on its sense of touch in such circumstances? These are the questions that motivates this thesis and are addressed by scientists in recent years.

1.2 Aim of this Thesis

As we have just seen, tactile sensors have been getting more attention in the field of robotics in recent years. The value that these sensors offer is recognized by scientists, who are integrating them into their systems. In the domain of robotics, the applications of these sensors can be broadly classified into five categories:

1. **Grasp and manipulation tasks** – Here, tactile sensors are used for two purposes [8]. Firstly *perception for action*, tactile sensors are used to control the quality of the grasp, in terms of force applied, contact points, slip detection, or even dexterous manipulation [9, 10]. Secondly, in *action for perception* to determine the properties of grasped objects, such as shape, classification, surface properties, stiffness, or temperature [11].

2. **Self-perception** – Tactile sensors can be used to sense the robot's proprioception. Furthermore, the sensors can also be used to detect unwanted external factors, such as collisions with obstacles and the stress placed on individual robot components [12].

3. **Feedback** – To enable better interaction between humans and robots or robots and robots, tactile sensors can be used to provide feedback. The robot can do this through direct

contact with the communication partner, in the form of haptic feedback, or through non-contact provision of information such as verbal communication or visualization [13]. Furthermore, this is also used for user interfaces, such as teleoperation in surgical applications.

4. **Quality assurance** – Tactile sensors can also be used for quality assurance, for example for material inspection, the measurement of gap dimensions in manufactured products, or alignment and positioning in manufacturing lines [14].

5. **Navigation** – Another area is the use of tactile sensors for navigation purposes. However, it should be noted that these applications primarily use contact or bumper sensors, which typically incorporate binary sensors. A simple example is the bumper of robot vacuum cleaners. In such cases, a contact sensor is often sufficient to fulfill the mapping and localization tasks required for navigation [15].

The aim of this thesis is to explore some of these fields of application and investigate how tactile sensors can be used in combination with other modalities and whether tactile data can provide added value to the application or experiment. To narrow the focus a bit, this thesis will concentrate on the first three categories. The subject area of navigation will be excluded, as this subject area relates less to tactile sensing and more to methods for mobile robotics as of navigation, localization, and path planning. However, for the sake of completeness, this topic should be mentioned. Furthermore, the issue of quality control is not addressed in this work. While the potential of tactile sensors is significant, they are not currently suited to the existing robots in our laboratory, but would be an interesting aspect for future work.

This thesis will examine the potential of tactile data in object analysis, with a particular focus on object content classification (category 1). Furthermore, tactile sensors have been employed for the purposes of grasp control [9] and slip detection [10] (category 1), as I was involved in these papers as a co-author, these works are mentioned here but not considered in more detail in the rest of the work. Tactile sensors are also utilized for proprioception in robotic grippers (category 2). In order to provide feedback, our robots are integrated into Mixed Reality (MR) and Virtual Reality (VR) environments. This integration provides visual and acoustic feedback of tactile readings on the one hand, and direct haptic feedback to the user on the other (both category 3). The resulting fundamental questions and contributions are explained in more detail in the following section.

1.3 Fundamental Questions and Contributions

The objective of this work is to utilize the tactile sense in as many areas of the robot's environment as possible, both for the robot in isolation and in interaction with humans. This resulted in the following research areas and fundamental questions we want to address in this thesis:

FQ1 Environmental Sensing – Can tactile sensors add value to the exploration of the environment and contribute to its robust recognition? How do they perform in comparison to and in combination with other modalities?

FQ2 Intrinsic State Analysis – Is it possible to use tactile sensors to infer intrinsic information about the robot? In a scenario in which the state of the robot is partially unknown, it would otherwise have to be equipped with sensors that can measure the joint state. Can this be compensated by the use of tactile information?

FQ3 Augmenting Human Perception – To what extent can humans be supported in interaction scenarios with the robot by communicating tactile information? What is the optimal way of presenting this tactile data and and whether a multimodal approach would be beneficial?

FQ4 Haptic Feedback – In a human-robot interaction scenario, is it possible to let the robot provide humans with meaningful haptic feedback to improve usability and the quality of the experience?

1.4 Structure of this Thesis

The remainder of the thesis is structured as follows:

Chapter 2 introduces the fundamentals of this work. It explains how the human sense of touch is structured and the most important receptors. The most common tactile sensor technologies will be introduced, including their functions and applications. In all experiments and studies, robots are used, which are operated with the Robot Operating System (ROS) which is introduced. Furthermore, two studies employ the use of Mixed Reality (MR) and Virtual Reality (VR). The taxonomy of the Virtuality Continuum (VC) is explained, as well as Unity, the development environment used for the creation of applications for MR and VR.

Chapter 3 describes the interplay between the tactile and acoustic modality in a classification task. A robot classifies the contents of eight visually indistinguishable containers by grasping and shaking them. The contents are classified on the basis of the vibrations and acoustic signals generated in the process. The influence of the individual modalities on the result is determined and evaluated.

Chapter 4 presents an approach for state estimation of a robotic gripper, integrating tactile sensor data and proprioceptive data from the actuators. The gripper used is underactuated, thereby the state of the states of the fingers are not known. A neural network is employed to train a model on previously recorded ground truth training data, enabling the estimation of the state of the hand from grasping sequences.

Chapter 5 presents a MR teleoperation approach. Extended Reality (XR) devices are employed to track the pose of controllers or the user's hands and transmit this data to the end effector of the robot to control it remotely. One or two manipulators can be controlled simultaneously to perform collision-free tasks. The performance, usability, and user experience are being evaluated in a pilot study.

Chapter 6 introduces an Encountered-Type Haptic Display (ETHD), which provides the user with haptic feedback in VR in order to perceive the virtual world even more realistically and thus improve immersion. The design and construction of this display is presented and evaluated in a user study. With this device it is possible to place different objects on a conventional table for the user to explore.

Chapter 7 provides a summary of the thesis and highlights the key contributions. Furthermore, ideas for future research questions and experiments are given.

1.5 Publications

The following list of peer-reviewed publications contains the main contributions to this dissertation. The first author was responsible for the main research, conceptualization, and writing of these publications.

- **Yannick Jonetzko**, Niklas Fiedler, Manfred Eppe, and Jianwei Zhang. "Multimodal Object Analysis with Auditory and Tactile Sensing using Recurrent Neural Networks" In: *Cognitive Systems and Signal Processing: 5th International Conference, ICCSIP 2020*. Zhuhai, 2020, pp. 253–265.

  In a shaking experiment, we tested to what extent the tactile modality provides added value to acoustic signals in the classification of pills in visually indistinguishable containers. Not only the improvement of the accuracy performance was investigated, but also the robustness against acoustic noise. (Chapter 3)

- **Yannick Jonetzko**, Judith Hartfill, Niklas Fiedler, Fangwei Zhong, Frank Steinicke, and Jianwei Zhang. "Evaluating Visual and Auditory Substitution of Tactile Feedback during Mixed Reality Teleoperation". In: *International Conference on Cognitive Computation and Systems (ICCCS)*. Beijing, 2022, pp. 331–345.

  We present an approach to remotely control robots with XR devices. The possibility of tracking the user's hands with these devices and mapping the pose to robot arms is utilized in order to control them. In a study, the influence of feedback on the measured tactile signals of the robotic grippers in MR on the performance of teleoperation is investigated, as well as how the usability and user experience change. (Chapter 5)

- **Yannick Jonetzko**, Oscar Ariza, Susanne Schmidt, Niklas Fiedler, and Jianwei Zhang. "Encountered-Type Tabletop Haptic Display for Objects On-Demand in Virtual Environments". In: *2023 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. Samui, 2023, pp. 1–7.

  In this work, we present an Encountered-Type Haptic Display (ETHD) for haptic feedback in Virtual and Augmented Reality. An X-Y-Z-Yaw plotter is mounted below a regular table, equipped with four magnets at the end effector. The mechanism is capable of moving a magnetic object on the top of the tabletop and place it with an accuracy of 0.5 cm in a workspace of 47.2×26 cm. (Chapter 6)

- **Yannick Jonetzko**, Theresa Alexandra Aurelia Naß, Niklas Fiedler, and Jianwei Zhang. "State Estimation of an Adaptive 3-Finger Gripper using Recurrent Neural Networks". In: *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. Abu Dhabi, 2024, pp. 8739–8745.

  Grasping is one of the essential functionalities in robotics. With adaptive and soft grasping, one can achieve robust and reliable grasps, often without precise knowledge of the exact state of the hand. However, understanding this state is crucial for grasp analysis to determine successful grasping or the quality of the grasp. Using tactile sensors and deep learning, we employ two approaches to assess the state of the underactuated 3-Finger Gripper from Robotiq. We compare these methods with an existing analytical approach. (Chapter 4)

The following publications are only marginally part of this dissertation. As a co-author, I contributed to text writing, implementation, conceptualization, or advice and expertise.

- Philipp Ruppel, **Yannick Jonetzko**, Michael Görner, Norman Hendrich, and Jianwei Zhang. "Simulation of the SynTouch BioTac Sensor". In: *Intelligent Autonomous Systems 15: Proceedings of the 15th International Conference IAS-15*. Baden-Baden, 2018, pp. 374–387.

- Dennis Krupke, Frank Steinicke, Paul Lubos, **Yannick Jonetzko**, Michael Görner, and Jianwei Zhang. "Comparison of Multimodal Heading and Pointing Gestures for Co-Located Mixed Reality Human-Robot Interaction". In: *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. Madrid, 2018, pp. 1–9.

- Zhen Deng, **Yannick Jonetzko**, Liwei Zhang, and Jianwei Zhang. "Grasping Force Control of Multi-Fingered Robotic Hands through Tactile Sensing for Object Stabilization" In: *Sensors 20.4*. 2020, pp. 1050.

- Niklas Fiedler, Philipp Ruppel, **Yannick Jonetzko**, Norman Hendrich, and Jianwei Zhang. "A Low-Cost Modular System of Customizable, Versatile, and Flexible Tactile Sensor

Arrays". In: *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. Prague, 2021, pp. 1771–1777.

- Niklas Fiedler, Philipp Ruppel, **Yannick Jonetzko**, Norman Hendrich, Jianwei Zhang. "Low-cost fabrication of flexible tactile sensor arrays". In: *HardwareX 12*. 2020, pp. e00372.
- Niklas Fiedler, **Yannick Jonetzko**, Jianwei Zhang. "A Multimodal Pipeline for Grasping Fabrics from Flat Surfaces with Tactile Slip and Fall Detection". In: *2023 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. Samui, 2023, pp. 1–6.

2 Fundamentals

2.1 Biological Sense of Touch

Before looking into tactile sensors, it is important to understand the various terms of the human sense of touch. Loomis and Lederman [16] and Klatzky and Lederman [1] distinguish between cutaneous and kinesthetic senses. The cutaneous sense involves stimuli of the receptors and the associated nervous system in the skin. In contrast, the kinesthetic sense perceives the static and dynamic body posture of muscles, tendons, and joints. According to Loomis and Lederman [16], Klatzky and Lederman [1], and Dahiya et al. [8] these two senses provide the basis for our sense of touch specified as tactual perception. This tactual perception is divided in tactile, kinesthetic, and haptic perception. Tactile perception refers to stimulation solely of the cutaneous sense allowing to perceive information like temperature, force, or vibrations. Kinesthetic perception, as the name suggests, refers to the kinesthetic sense alone, allowing to perceive the body position and joint torques. This information can be used to collect knowledge about the shape, stiffness, and weight of objects. The combination of cutaneous and kinesthetic sense forms haptic perception. A visual graph of the human haptic system can be seen in the left part of Figure 2.1 [17]. When talking about the human sense of touch, most actions fall into this category.

To understand how the human anatomy of the skin, which is responsible for the cutaneous part of the haptic system, is organized, we can take a look at Figure 2.2 from Dahiya et al. [8]. A cutaneous receptor is a sensory receptor located in the skin all over the body, either in the dermis or epidermis. These receptors are integral to the somatosensory system. They include mechanoreceptors, which detect pressure or distortion; nociceptors, responsible for sensing pain; and thermoreceptors, which respond to temperature changes. We are mainly interested in the mechanoreceptors, which can be further subdivided into four categories: Pacinian Corpuscle, Ruffini Corpuscle, Merkel Corpuscle, and Meissner's Corpuscle. Meissner's Corpuscles are characterized by their ability to react to changes in pressure and enabling them to perceive small vibrations and slips. This quality has led to their classification as Fast Adaptive I (also Rapid Adaptive I). These corpuscles are exclusively present in the hairless regions of the skin, with a high density observed in the fingertips. The second type of corpuscle that responds to rapid changes (Fast Adaptive II or Rapid Adaptive II) in pressure is the Pacinian Corpuscle (also known as Vater-Pacinian Corpuscle or Lamellar Corpuscle). These are even more sensitive to vibrations but have large reception fields. They are primarily located in the palms of the hands and soles of the feet, as well as the proximal phalanges of the fingers and toes. On the other hand, there are slow adaptive corpuscle. Merkel Corpuscle (also Merkel Cells or Merkel Discs) belong

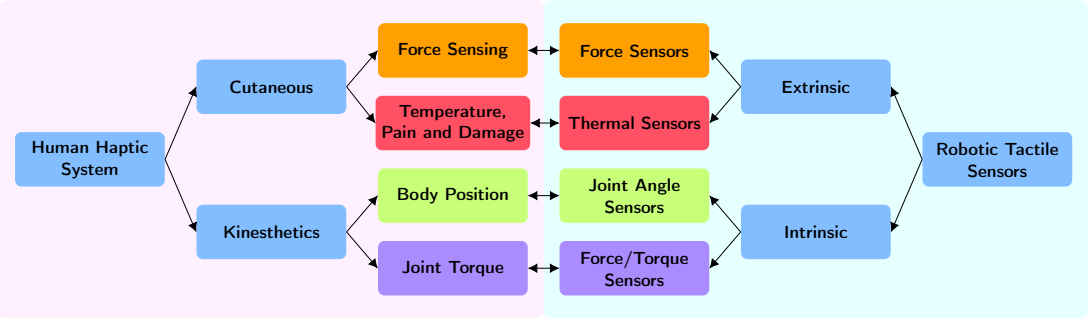


Figure 2.1: Haptic-Tactile-Process [17]

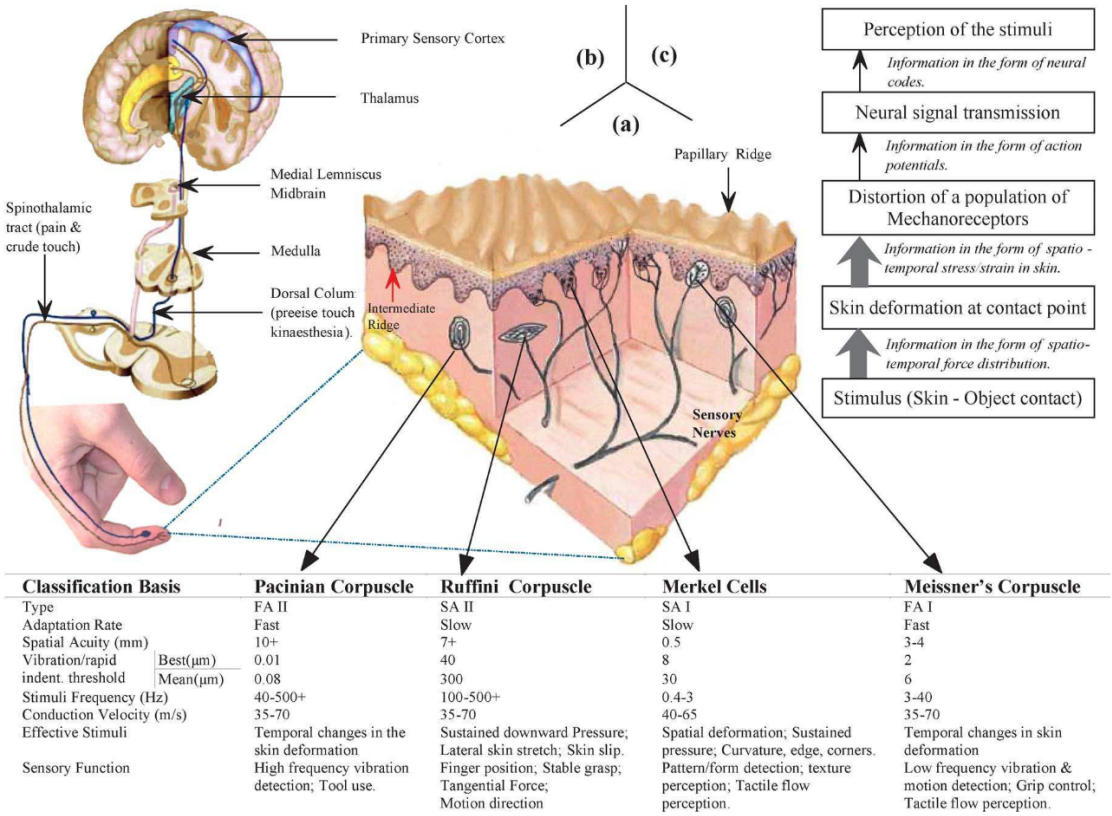
to the Slow Adaptive I types and react to static pressure. These cells have a small receptive field, therefore an accurate spatial acuity, and are sensitive to even minor impacts. They are widely distributed throughout the body, but are particularly dense in the fingertips. The second category of slow adaptive cells (Slow Adaptive II), are Ruffini Corpuscles, also referred to as Bulbous Corpuscles. These receptors are sensitive to static pressure and horizontal stretch. They are distributed in the fingertips and also in joint capsules, where they register the position of the joints and the speed of deflection. Consequently, they can also be classified to the kinesthetic haptic systems.

These categories can also be applied to tactile sensors in robots, where there are sensors with high resolution, sensors that react to vibrations, and sensors that can measure high forces. As with humans, there are no sensors that can measure all types of tactile information.

2.2 Tactile Sensor Technologies

Tactile sensors are technologies that are capable of detecting touch, pressure, or vibrations (see right side of Figure 2.1), among other stimuli, and subsequently converting them into electrical signals. Such sensors are frequently employed in robotics and automation to identify objects, ascertain their characteristics, or enable precise interactions.

As previously discussed, the human sense of touch can be divided into four mechanoreceptors when it comes to perceiving contact. The receptors differ in terms of their response time, spatial acuity, their ability to discern changes in pressure up to vibrations, and the pressure threshold at which they become active. These factors and requirements can also be applied to tactile sensors for the use in robotics. Also the various sensor technologies differ in this context and can be partially attributed to the receptors. An overview of the advantages and disadvantages of these technologies are listed in Table 2.1 from [18]. When integrating tactile sensors on robot platforms, however, there are other requirements besides the selection of sensitivity, precision, and reaction time. Depending on the application, the sensors must be particularly robust or flexible. Furthermore, the sensors should be energy-efficient, particularly in the context of mobile



Classification Basis		Pacinian Corpuscle	Ruffini Corpuscle	Merkel Cells	Meissner's Corpuscle
Type		FA II	SA II	SA I	FA I
Adaptation Rate		Fast	Slow	Slow	Fast
Spatial Acuity (mm)		10+	7+	0.5	3-4
Vibration/rapid indent. threshold	Best(μm)	0.01	40	8	2
	Mean(μm)	0.08	300	30	6
Stimuli Frequency (Hz)		40-500+	100-500+	0.4-3	3-40
Conduction Velocity (m/s)		35-70	35-70	40-65	35-70
Effective Stimuli		Temporal changes in the skin deformation	Sustained downward Pressure; Lateral skin stretch; Skin slip.	Spatial deformation; Sustained pressure; Curvature, edge, corners.	Temporal changes in skin deformation
Sensory Function		High frequency vibration detection; Tool use.	Finger position; Stable grasp; Tangential Force; Motion direction	Pattern/form detection; texture perception; Tactile flow perception.	Low frequency vibration & motion detection; Grip control; Tactile flow perception.

Figure 2.2: This graphic from Dahiya et al. [8] shows the in part (a) the location and classification of mechanoreceptors. In (b) the tactile signal transmission is shown from the fingertip to the brain, and in part (c) the event chain from stimulus to perception. Furthermore, the graphic shows the tactile receptor types in the skin, and their properties like spacial acuity, stimuli frequency, and adaptation rate.

platforms. Since tactile sensors are mostly used to interact with the environment, they are often subject to wear and tear, so ideally the sensors should be cost-efficient, durable, and easy to maintain. Tactile sensors often react to heat and humidity and provide different values in different environments. In order to ensure accurate measurement, it is essential to calibrate the sensors. In their review, Chi et al. [18] provide an overview of the most significant sensor technologies employed in the field of robotics. Some of these sensor principles are employed in our own work, and thus they are presented and summarized here.

Capacitive – In a capacitive sensor, two electrodes form the plates of a capacitor whose capacitance can be measured. This can change, for example, by changing the distance between the electrodes or if a conductive material comes close to them. One example of capacitive sensors are touchscreens on smartphones. [19, 20]

Piezoresistive – The basic principle of piezoresistive sensors is that the acting forces cause a change in the electrical resistance. One example is that of strain gauges, where a thin conductive metal is applied to a surface in a meandering pattern. This cable is then energized and the voltage is measured. If the surface deforms, the resistance of the wire changes, which can be measured. [21, 22, 23]

Piezoelectric – Piezoelectric materials generate voltage when deformed. This effect can be used to build sensors. Polyvinylidene fluoride (PVDF) and its copolymers are the most commonly used materials for these tactile sensors, as they are lightweight and easy to process. However, a significant drawback of these sensors is that they are ineffective at measuring static forces. Such sensors are often used as pickups in stringed instruments, given their aptitude for measuring vibrations. [24, 25]

Optical – This category of tactile sensor employs light intensity measurement, using optical fibers to ascertain pressure strength. One advantage of these sensors is that they exhibit no interference in arrays. Camera-based sensors also belong to this category, where a small camera is typically attached behind a deformable material which measures the deformation during contact. [26, 27]

Inductive – For this type of sensor, a coil is used to generate a magnetic field. Changes in this magnetic field can be measured in order to draw conclusions about contacts or distances. One advantage of this technology is that it is capable of measuring distances without contact. However, a disadvantage is that it only works when used with conductive or ferromagnetic materials. Such sensors are often used for precise detection and distance measurement of metallic machine parts. [28]

Magnetic – Magnetic sensors are capable of measuring magnetic fields. This can be achieved through the use of Hall sensors, which are able to measure the voltage that arises perpendicular to a current flow when a magnetic field is present. By molding a magnet in silicone over such a sensor, a tactile sensor can be constructed. With this sensor, it is possible to measure shear forces, besides normal forces. Otherwise, this type of sensor requires a magnetic field in the contact to be measured, but also has the advantage that it works without contact. [29, 30]

Binary – Binary tactile sensors are very simple and therefore favorable sensors that only measure whether or not contact is present. A simple example of this principle can be observed in the end-stop mechanism of 3D printers, where a mechanical switch is activated upon contact, which then closes a circuit and thus detects the presence of contact. Such sensors are also frequently used in robotic vacuum cleaners.

Table 2.1: An overview about advantages and disadvantages about standard tactile technologies is given in [18]

Transduction Mechanisms	Advantages	Disadvantages
Capacitive	High sensitivity High spatial resolution Large dynamic range Temperature independent	Stray capacitance Complex measurement circuit Cross-talk between elements Susceptible to noise Hysteresis
Piezoresistive	Simple construction High spatial resolution Low cost Compatible with VLSI	Hysteresis High power consumption Lack of reproducibility
Piezoelectric	High frequency response High accuracy High sensitivity High dynamic range	Poor spatial resolution Charge leakages Dynamic sensing only
Optical	Good reliability Wide sensing range High repeatability High spatial resolution	Non-conformable Bulky in size Susceptible to temperature or misalignment
Inductive	Linear output High sensitivity High power output High dynamic range	Low frequency response Poor reliability More power consumption

2.3 ROS

The operation of our robots is based on ROS [31], which is an open-source middleware that is widely employed within the robotic community. ROS provides a communication structure between different software components, which are referred to as nodes and modularize the software development for robots. Components can thus be developed individually and made easily accessible to others. The communication is divided into three fundamental concepts. The first variant is called *topics*, which offer n to n communication. *Messages* are sent asynchronously from a node on topics and other nodes can listen as required. This functionality is employed, for instance, in the context of sensor data. The sensor is not interested in whether and how many are listening, the data is provided in one direction to the whole system. A further variant is that of *services*, which is a synchronous blocking type of communication. A service is provided by a server node, a client node can then send a *request* which is then processed by the server, that answers at some point with a response. In the meantime, the client code is no longer executed. This type of communication is typically used to change configurations or to trigger one-time actions. A third method of communication is via *actions*, which are employed for longer-lasting, interruptible, or feedback-capable tasks. Actions offer extended functionality compared to services, as they not only have a request and a response, but can also provide intermediate statuses during execution. Actions also communicate via a one-to-one connection, whereby the client does not

block but can asynchronously check the status and request a response. Actions are typically used for the navigation of mobile platforms or the motion control of robotic arms. The registration of topics, services, and actions is conducted with the use of designated names, thereby facilitating the clear assignment of these elements. Communication requests are registered with the master, which then establishes the connection.

In addition to the aforementioned communication tools, ROS offers a number of other potentially useful applications, including RViz [W1], Gazebo [W2], and Plotjuggler [W3]. Such tools can be used for the visualization of the entire robot, the plotting of data, and the simulation of tasks. The modularization facilitates the exchange of information with other groups and the integration of code on other robots.

ROS has a few shortcomings, it is not real-time capable and can only run on linux operating systems [32]. The developers of ROS focused these problems by changing the whole architecture and come up with ROS2 [33]. To accomplish the real-time requirement ROS2 uses the Data Distribution Service (DDS) [34] for industry-standard real-time communication systems and end-to-end middleware.

However, given that ROS2 is still a relatively new technology and the robots employed in this research are operated with ROS1, we will always refer to the first version when we talk about ROS.

2.4 Unity

In addition to robots and tactile sensors, XR was employed in some of our experiments. The development of applications for VR and MR devices can be facilitated through the use of the game engine Unity [W4], which was used in this thesis. Unity is a robust cross-platform development environment for the creation of interactive 2D and 3D content, particularly for games, simulations, and VR applications. The user interface is intuitive, the application programming interface (API) is extensive, and the programming languages supported include C# and JavaScript. Unity facilitates the integration of animations, physics, lighting, and audio. The resulting applications can be compiled into standalone programs, enabling their export and execution on external devices, including smartphones and XR devices, without a connection to a computer. During development, these applications can also be executed on the PC and transferred to the corresponding device via cable, which significantly simplifies development with XR devices. ROS cannot be executed directly on XR devices; thus, an interface between these systems is necessary. Initially, scientists have implemented their own interface [35], but around 2018 Bischoff [W5] provided the first open-source interface called ROS# for the community. Furthermore, this enabled the importation of Unified Robot Description Formats (URDFs), messages, and services into Unity, as well as the subscription to and publication on topics. Messages are converted to and sent in JSON format, this has the disadvantage of slowing down the communication for large messages. Furthermore, applications for Universal Windows Platform (UWP) and other

XR environments were not compatible and could only be realized in a circuitous manner. A few years later, Unity itself implemented an interface, the Unity-Robotics-Hub [W6]. This is somewhat more lightweight and requires more steps for integration, but communicates directly with TCP and is therefore significantly faster. Developing with XR devices also works without any detours.

2.5 Virtuality Continuum

In the field of virtual environments, there are a lot of terms that are often blurred or used incorrectly in the literature. An overview of the taxonomy is provided by Milgram and Kishino [36], the authors present the concept of the Virtuality Continuum (VC), which is shown in Figure 2.3. It is often not entirely clear which term is the right one for a specific scenario, which is why some terms are often used analogously to each other. In the following description, the most important terms are clarified to prevent confusion.

Virtuality Continuum (VC) – This term describes the connection between a completely real world and a completely virtual world, irrespective of the display devices employed. All other terms in this list can be situated somewhere on this continuum. [36]

Real Environment – This includes the real world, which consists exclusively of physical objects. But it also includes videos that are played on conventional video displays. [36]

Virtual Environment (VE) – In comparison to the Real Environment, the Virtual Environment is the other extreme of the Virtuality Continuum. This is not bound to the display device but rather describes a world created exclusively through the computer graphics. A simulation presented on a monitor, for instance, also falls within this category. [36]

Virtual Reality (VR) – In Virtual Reality, the user is completely immersed in a synthetic world using immersive Head-Mounted Display (HMD) and only interacts with this, no longer with the real world. [36]

Augmented Reality (AR) – The term Augmented Reality is used when a real environment is augmented with individual computer-generated objects. This technology is independent of the displaying device. The user is able to perceive reality on a screen, for example, in a

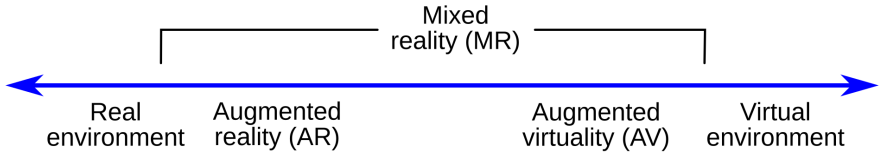


Figure 2.3: The Virtuality Continuum first introduced by Milgram and Kishino [36].

video, by wearing a see-through HMD or even a fully immersive HMD in which the real world is projected. [36]

Augmented Virtuality (AV) – Augmented Virtuality can be defined as a situation in which a predominantly virtual world is extended by the real world, either visually or through interaction. In this case, the user is wearing an immersive HMD. It is notable that this term is rarely used in the literature, with the broader term Mixed Reality being used instead. [36]

Mixed Reality (MR) – The term Mixed Reality is a broad one, encompassing all combinations of real and virtual environments. Especially in the gray area between Augmented Reality and Augmented Virtuality, this term is used when a decision is difficult. This term is completely independent of the displaying device. [36]

Extended Reality (XR) – Extended Reality is often used as an umbrella term for Virtual Reality, Augmented Reality, and Mixed Reality. This includes everything on the scale of the Virtuality Continuum except for the Real Environment, and, unlike Mixed Reality, also includes the Virtual Environment. Independent of the device used. [37]

3 Multimodal Classification

In this section, the first fundamental question (**FQ1**) will be addressed, the objective is to use tactile sensors in order to collect data from the surrounding environment of the robot. In the following chapter, we will investigate the combination of tactile and audio modalities in a classification experiment.

People often rely on their visual sense when orientating oneself in the environment or distinguishing objects. If this sense is impaired, for example in the absence of light or for blind people, the other senses become more important and it is crucial that we can rely on them. It is equally important for robots to utilize different modalities to explore and interact with the environment. This can also be observed in current research, where the gathering of multimodal information is becoming increasingly important for the performance of tasks that are more robust than those performed with just one sense. Nevertheless, the majority of publications on multimodal work have visual information as one modality (e.g. [38, 39, 40, 41, 42]). In their scenario, Güler et al. [43] demonstrate that vision and touch are equally effective for recognizing the contents of milk cartons. This illustrates that there are scenarios in which the robot can rely on its senses in a complementary manner. Furthermore, they show that multimodal use is even more robust, particularly in instances where one modality is not functioning optimally. In most cases, tactile sensors are in direct contact with the surface they are supposed to collect information from. This may be the surface of an object being grasped [44, 45], or in more unusual task, reading braille [46, 47].

In this work we want to focus on what happens when we cannot use the visual sense, for example when identifying the contents of visually indistinguishable cans. Furthermore, in such a scenario, there is no direct contact with the object or objects to be classified. The act of checking whether a container is empty, whether it is a packet of chewing gum or a carton of drinks, is a common human practice. However, to find out what is inside, we open the container. This raises the question of whether it is possible for a robot to perform this task without opening the container. The obvious idea is to use audio data for this purpose. However, since these are often noisy in robot environments, particularly from loud fans, the question now arises, to what extent the multimodal combination together with tactile data helps to enhance classification, or whether it is even possible to achieve similar results with tactile data alone, as in [43]. This brings us to the following research question:

> Can a combination of tactile and audio data help to explore the environment by classifying the content of visually indistinguishable containers?

To answer this question, an experiment was conducted in which our robot grasps and shakes a series of visually indistinguishable objects with different contents. The used sensor is inspired by the human finger and is capable of measuring vibrations at over 1000 Hz [48]. In comparison, the human finger measures vibrations at up to approximately 700 Hz [8]. Furthermore, the cans are shaken directly in front of a microphone, specifically set up for this purpose, in order to obtain data that is as noise-free as possible. The noise level will be added at a later stage to facilitate more effective control. With this setup, training data is collected and neural networks are used to train uni- and multimodal models.

This chapter is structured as follows. After the introduction, the related work with approaches similar to ours is presented in Section 3.1. Section 3.2 describes the setup, including the robot platform, sensors used, and experiment materials. We describe the classification approach in Section 3.3, followed by the evaluation in Section 3.4 including an experiment, the results, and the discussion. Section 3.5 summarizes the work in a conclusion and gives ideas for possible subsequent.

3.1 Related Work

The central objective of this research is to conduct an analysis and classification of objects within a can or capsule that are not directly accessible. This topic has been explored by numerous authors, who have published a multitude of approaches which include unimodal and multimodal combinations. Since the focus of our work is on tactile and acoustic classification, these are also the focal points of our related research. However, there are also approaches that are based exclusively on the visual modality. These typically employ transparent containers to identify the content. A popular scenario is the recognition of the fill level of liquids in jars, for which both conventional computer vision techniques and machine learning approaches are utilized [49, 50, 51].

More often than the contents of containers, the properties of the objects touched are determined and analyzed. For example by analyzing audio signals when hitting objects [52, 53], or with tactile data when stroking surfaces [54, 55]. Other authors have developed multimodal approaches to classify objects by manipulating them with the robot [56].

3.1.1 Tactile Analysis and Classification

The first comparable experiment in which a robot explores the properties of the contents of an object using tactile sensors was carried out by Chitta et al. [57] in 2011. In one experiment, they used a mobile robot, to classify 4 different bottles and determine their internal state, whether they were open or closed and whether they were filled with liquid. The bottles were placed between the fingers and the gripper closed without moving the object. Several characteristics are measured, such as the position of the gripper, the closing speed and the force measured by the tactile sensor arrays of their robot. They tested a decision tree, a Support Vector Machine (SVM),

and Neural Networks to classify the objects, with the decision tree performing best at 93.9 %. It is not clear how important the tactile arrays are for classification. The classification of the internal state depends strongly on the type of container, ranging from 94.8 % to 32.5 %.

In 2016, Chen, Snyder, and Ramadge [58] conducted an experiment in which they attached a contact sensor to four different containers and classified 12 different contents, subsequently approximating the number of these objects. The contact sensor is a contact microphone that was attached to the outside of the containers and is capable of sampling vibrations at a frequency of 5 kHz. A linear Support Vector Machine was employed to classify the classes with an accuracy of 94 %. By attaching the sensor directly to the container, they were able to achieve a high signal resolution with little noise. The selected classes differ significantly in size, weight, material, and shape (e.g., ball bearing, acrylic piece, rubber ball). In [59], the authors also use a contact microphone to estimate the amount and flow of granular material using recurrent neural networks.

Saal, Ting, and Vijayakumar [60] present an approach to determine the viscosity of different liquids in a bottle based purely on tactile data by shaking the bottle back and forth with their robotic hand-arm setup. They vary the shaking frequency and rotation angle to speed up the estimation. The force measurements of a three-finger gripper with 6 tactile sensor arrays and a total of 486 taxels are used in their approach, the readout frequency is not specified. Using Gaussian Processess (GPs), they approximate the non-linear function between sensor data, actions, and viscosity, and adjust the shaking frequency using active learning. The three liquids water (1 cst, 0 ($\log_{10}$)), motor oil (120 cst, 2.07 ($\log_{10}$)), and glycerine (1200 cst, 3.07 ($\log_{10}$)), each 160 g in identical bottles, are used to train the model. In this work, the content is not classified, instead the viscosity is approximated. To test the model, a mixture with a viscosity of 1.47 ($\log_{10}$) is also used and a mean squared error of 0.72 is achieved. Other works that use tactile sensors to estimate liquids dynamics are [61] and [62]. Similar, but with rigid objects, Sundaralingam and Hermans [63] learn the dynamics of rigid objects with modifiable center of mass.

In their study, Guo, Huang, and Yuan [64] present an approach for estimating four properties of enclosed objects in identical containers. The aforementioned properties are content mass, content volume, particle size, and particle shape. While mass and volume are determined using a force torque sensor in the wrist, they use a modified GelSight sensor with a high-speed camera (frame rate of 815 Hz) for the size and shape of the particles. The camera captures the movements of 70 markers on a gel pad that exerts pressure on the container, subsequently averaging over the 30 highest movements as principle vibration signal. The estimation of the parameters is conducted through the utilization of Multi-Layer Perception (MLP). In total, 37 different materials, ranging from fine flour to beans, were used, and the particle size was estimated with a mean absolute error of 1.1 mm, the particle shape with a mean absolute percentage error of 75.6 %, the weight with an accuracy of 1.8 g, and the volume with 6.1 ml.

3.1.2 Acoustic Analysis and Classification

One work in which the contents of capsules are classified, which forms the basis for our own research is the work of Eppe et al. [65]. In their work, they utilize a NICO robot [66] to shake 10 relatively different materials within small capsules and classify them using audio signals, as well as estimating their weight. In contrast to many other works, including ours, they do not rotate the capsules to generate sounds, but instead move the arm back and forth in a jerky motion. The audio data is preprocessed with Mel Frequency Cepstral Coefficients (MFCC), the type is then classified with Gated Recurrent Units (GRUs) and the weight is estimated with Long Short-Term Memorys (LSTMs). They achieved a mean absolute error for the estimated weight of 3.51 g and an accuracy of 91 % for the classification.

In their study, Jin et al. [67] present a method for classifying objects using acoustic signals in containers in open environments. To this end, the authors present and utilize the kernel k-nearest neighbor algorithm in an open environment (OSKKNN). Initially, four standard machine learning methods were evaluated in a closed environment, wherein 20 object classes in three different containers were classified. Subsequently, the most effective algorithm, kernel k-nearest neighbor (KKNN), is employed to identify unknown objects, and to subsequently learn and classify these unknown objects with a second KKNN classifier. In this instance, the authors use the same 20 object classes, split into 10 known and 10 unknown objects. In the closed environment experiment, an accuracy of 85.5 % was achieved, while in the open environment, an accuracy of approximately 83 % was attained. This indicates that the accuracy remains largely unaffected despite the presence of unknown objects. It is notable that only specific classes within the dataset are similar, and the extent to which the classification is enhanced by the use of different containers remains unclear, as they are consistently assigned to a fixed class.

Liquids can also be considered as objects within a container. The estimation of dynamics has already been described in the preceding section. In their study, Liang et al. [68] present an approach to determine the fill level when pouring liquids into different containers using acoustic signals. They recorded a dataset comprising force and torque values, motion trajectories, and images as people poured liquids. With recurrent neural networks, they were able to determine the fill level with an accuracy of less than one millimeter when pouring water. They then successfully transferred the method to their robot.

3.1.3 Multimodal Analysis and Classification

In 2014, Sinapov, Schenck, and Stoytchev [69] used machine learning to learn object relations with their robotic arm. They filled identical types of containers in 3 different colors with 4 different objects. In addition, they filled each material in three weight classes, so that a total of 36 containers could be recognized. The aim was not so much to learn every combination, but rather to learn the relations, for example that container a is heavier than container b or has the same color. They used three modalities: first, images of the object were taken with a stationary

camera, then the object was explored with 10 manipulations of the robot using proprioceptive data from the joints and a microphone to identify further features of the containers. Among other things, they showed which modality can be used to determine which feature and which exploration step is best suited for this.

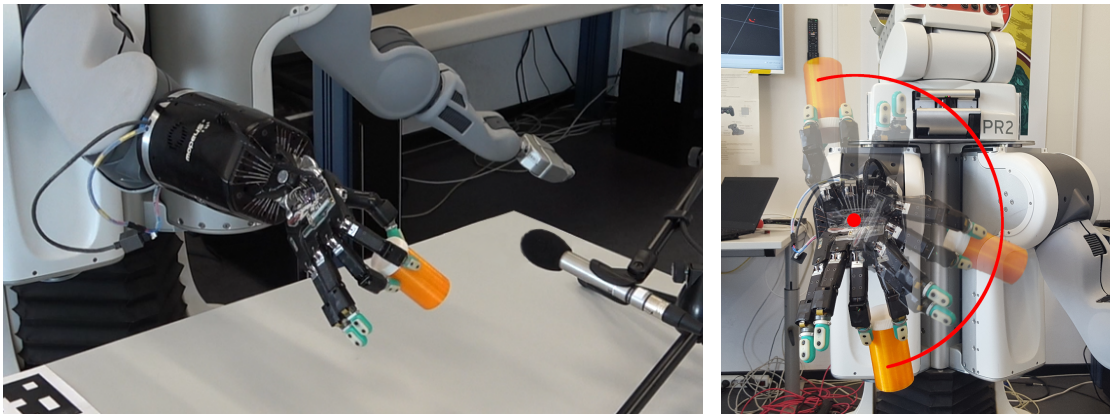
In their work, Spisak, Kerzel, and Wermter [70] utilize three modalities of their robot to classify containers and their contents, and evaluate the efficacy of different fusing strategies. The robot is equipped with a fish camera in the head, tactile fingertips and proprioceptive data from the joints. These modalities are fused in three different approaches, which are evaluated and compared. In the first approach, three individual classifiers, one for each modality, vote on the class. In the second approach, the outputs of the classifiers are combined in a dense layer. In the third approach, all the data are fed together into a network. A Convolutional Neural Network (CNN) is employed for the visual modality, whereas multiple dense layers are used for tactile and proprioception. The evaluation contains seven classes, comprising three different containers that can be either empty or full, with one that can also be half full. With an accuracy of 90.6 %, the multimodal approach is demonstrated to be the most successful. Furthermore, the analysis reveals the reasons behind this success and identifies the strengths and weaknesses of each modality.

Piacenza, Lee, and Isler [71] learn to pour using their robotic arm and a two-finger gripper with BioTac sensors. Their goal is to eliminate the need for a dedicated force-torque sensor and instead use the torque values from the joints and the tactile data from the BioTac sensors. They take the data from 250 pouring motions and train a neural network to determine the amount of liquid poured. To do this, 19 impedance values from the tactile sensors are fed into encoders and then into an MLP along with the proprioceptive sensor data from the robot. They compare the performance of the multimodal approach with tactile and proprioceptive data from each modality. They achieve an average accuracy of 10 ml deviation from the target.

A work that combines visual and tactile data is that of Güler et al. [43] from 2014. They use their robotic arm to grasp and squeeze containers with two fingers of their gripper. They record data from the tactile arrays on the fingers as well as the depth image from a kinect camera. The aim is to classify the contents of identical cardboard containers, which can be water, yogurt, flour, rice, or an empty container. They compare the four algorithms k-means, Quadratic Discriminant Analysis (QDA), k-nearest neighbor (kNN), and SVM, with SVM achieving the highest classification accuracy of 95 %. They also analyze the influence of the two modalities on the result.

3.2 Setup

The setup used in this work is depicted in Figure 3.1. We use a PR2 [72] platform equipped with a Dexterous Shadow Hand [W7] and five BioTac [7] tactile sensors to shake the containers and measure the tactile data. The audio data is recorded with an external microphone, which is placed on the table in front of the robot to be as close as possible to the sound source.



(a) (b)

Figure 3.1: The classification setup. **(a)** The content of an orange medicine container is recognized with the help of sensor data from three tactile BioTac sensors on the Dexterous Shadow Hand robot gripper and the audio signal from a microphone placed in front of it. On the right side **(b)**, the movement that the robot makes to generate rattle sounds and vibrations to recognize the content is shown. The robot hand rotates 180 degrees up and down with a rotation speed of 0.8 which corresponds to one radian per second.

3.2.1 Robot Platform

The robot platform utilized in this experiment is the mobile two-armed PR2 [72]. The robot is capable of omnidirectional movement on a planar surface and of performing complex manipulations with its two arms and grippers. Additionally, the robot is equipped with a series of visual sensors. The robot is designed for the use in service robotics, which encompasses its ability to interact with the domestic environment and assist with everyday tasks. For this purpose, the robot's dimensions are approximately equivalent to those of an adult human in terms of height and workspace. The abilities of the robot are not explicitly necessary for the experiment; however, the ability to analyze objects, which is learned by the robot in this work, could be integrated into larger scenarios. Therefore, it was advantageous for us to have the characteristic noises and vibrations emitted by the robot in our data. In this particular case, only a rotational movement is required, which the PR2 is capable of performing with its forearms. The right forearm, which includes a gripper, has been replaced by a five-finger Dexterous Shadow Hand [W7], allowing the robot to perform a wider range of fine motor tasks than it could with the standard parallel gripper on its left arm. With this anthropomorphic hand, which has 20-Degrees of Freedom (DoF), the robot is able to perform dexterous manipulations and in-hand manipulation [73].

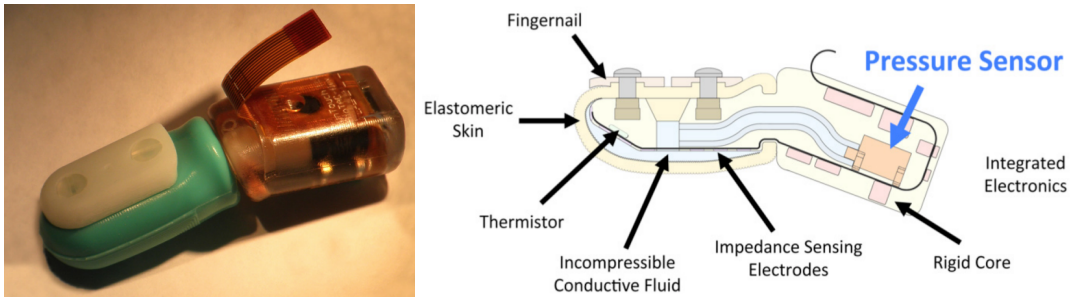


Figure 3.2: A cross-section of the bio-inspired tactile sensor BioTac. The sensor is filled with a conductive liquid between the rigid core marked with the black line and the outer silicone cover. The pressure and vibrations generated by contact and used in our experiment are measured with the sensor marked in blue. [48]

3.2.2 Tactile Sensors

The tactile sensors called BioTac, manufactured by the company SynTouch, can be seen on the fingertips of the robotic hand in Figure 3.1, as well as the cross-section of a fingertip in Figure 3.2. The tactile device is inspired by the human fingertip and was presented by Wettels et al. [7] in 2008. It has been used for various task, like force estimation, slip detection [74], or tumor localization [75]. The sensor has a bone-like core, a soft skin-like silicone cover, and a fingernail-like plate on the back that connects the core and cover. The core is equipped with various sensors and the space in between is filled with a conductive liquid. Different modalities are combined in the sensor, with which forces, vibrations, and temperatures can be measured. For this purpose, the finger has a hydro-acoustic pressure sensor at the end of the bone (marked in blue) to measures the pressure of the liquid. At the tip of the bone there is a thermistor that measures temperature. There are also 19 sensing electrodes distributed over the tip of the finger which measure the emitted current from 4 reference electrodes, so that the deformations of the silicone sheath can be deduced. Since we have no direct contact with the pills to be classified in the cans, we can ignore the temperature, the deformation of the cover, but also the force exerted by the fingers. Measuring vibrations with the sensors is essential for this work because the pills rattle against each other and the container when shaken. In terms of its vibration properties, the liquid inside the sensor is indistinguishable from water [48]. The analog signal of the pressure sensor is amplified with a gain of 10 and low-pass filtered with 1040 Hz to measure the absolute pressure (DC). To extract the vibrations from this signal, it is band-pass filtered between 10 and 1040 Hz with a gain of 99.1, resulting in the dynamic fluid pressure (AC) value. These vibrations are later used for the classification.

3.2.3 Microphone

A conventional all-purpose condenser microphone from Behringer (Single Diaphragm Condenser Microphone B-5) was utilized to record the audio signals. To minimize the presence of background noise, a cardioid characteristic was selected. This configuration amplifies sound originating from the front more than sound from the sides, which enables more effective isolation of the sound generated by the pills within the containers. The microphone was positioned as illustrated in Figure 3.1a and samples the audio signal at a frequency of 44.1 kHz.

3.2.4 Pill Container

The containers for this experiment should have specific properties. Firstly, they should be visually indistinguishable, as no image data is to be used. They should also not be flexible so that information about pill properties can not be accessed by pressing them. We decided to 3D print the containers, and, to give them a meaningful appearance and stay in the pill context, we made them look like medicine boxes. The base is printed in translucent orange, the lid in white (see Figure 3.3). The boxes have a thickness of 2.5 mm with a 20 mm radius and a height of 80 mm. We used an Fused Deposition Modeling (FDM) printer with translucent orange Polylactic Acid (PLA) plastic for the container and white PLA for the lid.

Figure 3.3: The 3D printed containers used in the experiment, with a thickness of 2.5 mm with a 20 mm radius and a height of 80 mm.

3.2.5 Pill Classes

In order to test the robot's ability to identify objects, an experiment was designed with eight different classes of objects. To ensure the robot would not be able to identify the content too easily, objects with similar properties were selected. Given our decision to utilize containers similar to medical cans in terms of their visual appearance, we also want to stay in this category

when selecting the contents. Tablets often have a distinctive shape, comparable dimensions, and consistent characteristics, as they are typically consumed in their entirety without mastication. In order to avoid the classification of real medications, a selection of food supplements and sweets similar to tablets was made. Table 3.1 illustrates the eight selected classes, with a weight per pill ranging from 0.31 g (*B-Complex*) to 2.2 g (*Calcium*). The tablets are between 8 and 20 mm in length or diameter.

3.3 Approach

We use a deep learning approach for the classification of pills within containers, utilizing both acoustic and tactile data. The classification process can be divided into several steps: Initially, information relevant to the classification is selected or filtered from the raw signals. In the second stage, the data undergoes preprocessing to ensure that the neural network is not overwhelmed by an excess of irrelevant information. Subsequently, Recurrent Neural Networks (RNNs) are employed to determine the contents of the containers, while the robot shakes the sample during the experimental procedure.

3.3.1 Sample Selection

In order to classify the pill type as robustly as possible, the raw audio and tactile signals are filtered automatically before the classification. During the shaking motion, a lot of unwanted data is recorded which provide no added value for the recognition. The recording can be roughly divided into three phases: the first phase involves the abrupt acceleration and deceleration of the rotational movement of the robot hand, resulting in unwanted vibrations and unusable data. During the second phase, the arm rotates, but the pills in the can do not produce any noise or vibrations because they are still too stable. The last phase is the one we want to filter and use for our classification. Here the pills fall around in the container and generate data.

Table 3.1: Pill classes included in the data set and used in the experiment

		 **Magnesium**	 **Calcium**	 **B-Complex**	 **Big Mints**	 **Chew**	 **Small Mints**	 **Vitamin B**	 **Candy**
Weight per pill		1.27*g*	2.2*g*	0.31*g*	1.2*g*	1.13*g*	0.5*g*	0.55*g*	0.6*g*
Sample count	One pill	166	99	104	110	118	137	126	137
	Small amount	228	405	222	212	299	218	260	391
	Half full	239	407	232	413	251	293	290	336
	Full	237	184	242	215	243	296	304	174
	Overall	820	1095	800	950	911	944	980	1008

In Figure 3.4 the raw tactile and audio signals of a shaking movement are plotted. The bottom graph shows the joint position of the forearm rotation. The movement starts with a 180 degree lifting rotation which takes 4.5 seconds followed by the lowering of the object which takes the same amount of time (the motion is also depicted in Figure 3.1b). To extract the relevant data of the third phase from the signals, two filters are applied. In the first filter, the signals of the first phase that are noisy due to the jerk of the abrupt change in speed are extracted. Only the data within the joint position range of -1 to 1.5 radians is passed on. All filtered values are marked red. It is noticeable that mainly the tactile data is affected by the jerk. The arm could also be accelerated and decelerated slowly, but this would only disturb the signal in a different way, so we decided to make an abrupt, clearly recognizable break instead. The second phase states that data in which the pills do not move produce neither sound nor vibrations and are therefore useless for classification. To filter these silent data, we look at short time windows of 0.2 seconds of the audio signal and determine the maximum amplitude. If this is below a threshold, the data is not used further (see yellow regions in Figure 3.4). All remaining values contain useful features for the classification and are colored green.

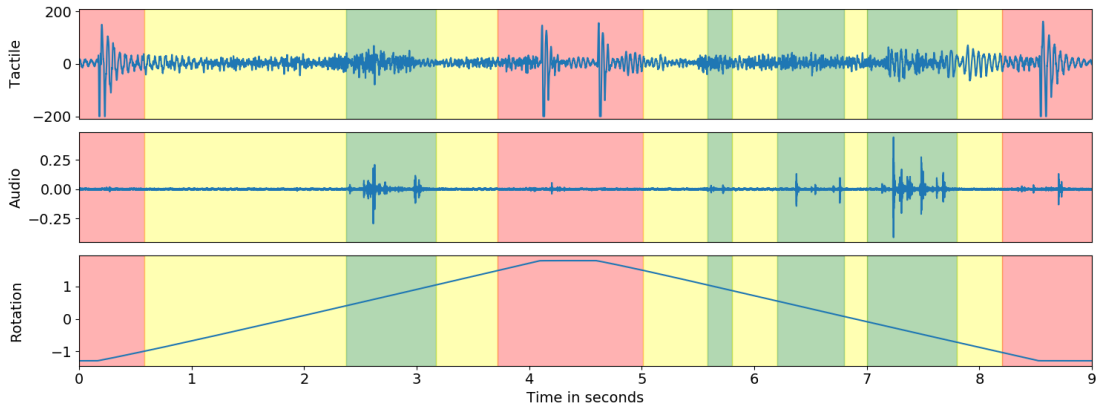


Figure 3.4: The top plot shows the raw tactile signal from 10 *Big Mint* pills shaken at an angular velocity of 0.8 radian per second. In the middle the raw audio data and in the lower plot the current joint position of the forearm, each for an entire shaking sequence. The audio data clearly shows the point in time at which the contents of the container produce sound. In the first filter step, data is filtered during the direction change of the rotation. This data contains only noise and cannot be used for analysis. The removed data is marked in **red**. In the second step, the signal strength is used for filtering. All data below a threshold is filtered as it does not contain any useful information for classification. The filtered areas are marked in **yellow**. The **green** areas show the extracted data after all filters have been applied.

3.3.2 Mel Frequency Cepstral Coefficients

The raw audio signals are not suitable for being fed directly into the network. For a compact representation of the frequency spectrum, MFCC [76] are often used for speech recognition or to analyze music. They could also be used for music synthesis [77] and, as in our case, for acoustic classification [65, 78]. To convert the raw audio signal into meaningful data, multiple steps are applied. In the first step, the signal is divided into equal time windows. A Fourier transformation [79] is then applied to these windows, which describes how often different frequencies are present in the original signal. In the second step, the powers of the resulting spectrum are mapped in overlapping windows to the mel scale [80], which is a perceptual scale of pitches where listeners perceive each interval as being equally spaced. Frequencies (Hz) are mapped to *mels* here. Afterwards, the logs of the powers at each mel frequency is taken. In the last step, a discrete cosine transform is applied to the mel log powers. In the resulting spectrum shown in Figure 3.5, the amplitudes are the MFCC. During these steps there are some parameters that can be adjusted. The most important are the window length, the window step size, the number of resulting mel coefficients, low and high frequency filters, and some more, which were not important in our case and were left at the default values of the python library used. We used MFCC for both the audio and tactile signals and determined the appropriate parameters with Tree-Parzen-based hyperparameter optimization [81].

The audio signal was recorded at a frequency of 44.1 kHz. The optimal results were obtained with a window size of 0.03, a step size of 0.02, and 21 resulting Mel coefficients.

The vibrations were recorded at a frequency of 1000 Hz with the tactile sensor. The high frequency allows for the utilization of audio processing techniques. In preliminary experiments, it was determined that applying MFCC to the data prior to its entry into the network resulted in enhanced classification accuracy in comparison to the raw values and is therefore suitable for our application. The optimal results were obtained with a window size of 0.04, a step size of 0.04, and 9 resulting Mel coefficients. Moreover, the frequency spectrum was reduced to a range of 4 to 440 Hz, as this achieved better results.

3.3.3 Network Architecture

For our approach, three neural networks are needed. One for the unimodal acoustic approach, one for the unimodal tactile approach, and one for the multimodal audio and tactile approach. Since our goal is to investigate whether tactile information adds value to the pure acoustic signal in order to increase the robustness of the classification, and not to optimize the classification accuracy per se, a simple network architecture is chosen that works for both the unimodal and, with slight modifications, the multimodal approach. We have followed the architecture of Eppe et al. [65] and used their approach also for the tactile signal, as the vibration signal from the BioTac sensors behaves similar to an audio signal. As described in the previous section, MFCC are applied to the raw signal streams to extract processable information from the enormous

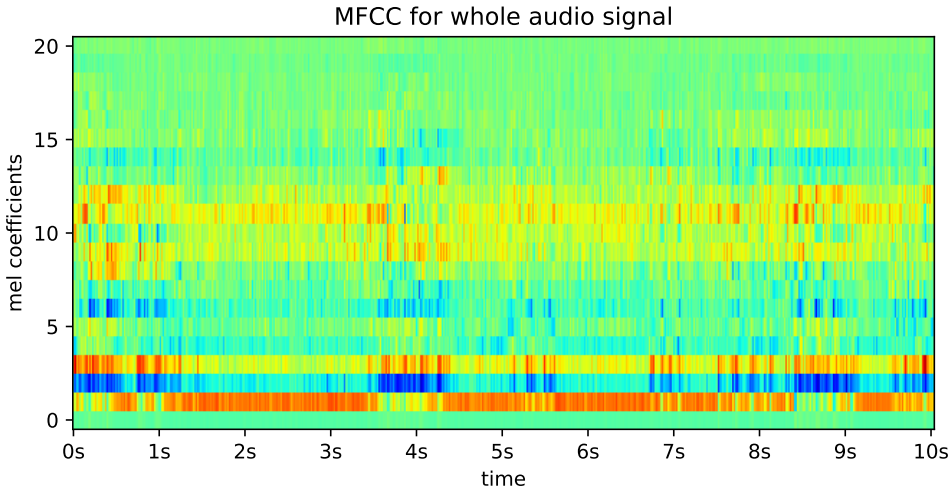
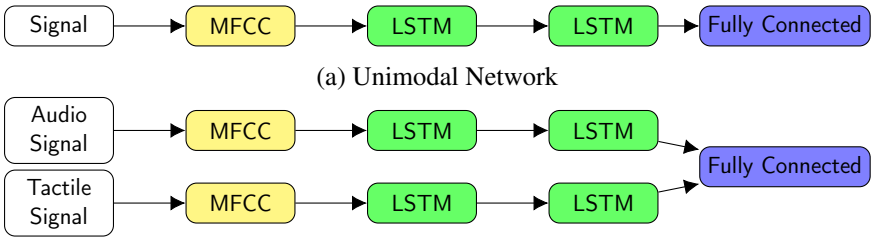


Figure 3.5: The resulting MFCC spectrum of an audio signal of a complete unfiltered shake sequence. A window length of 0.03 s, a step size of 0.02 s and 21 resulting Mel coefficients were used for this spectrum.

amount of data. Networks that are particularly well suited for the classification and regression of audio data are Recurrent Neural Network (RNN), as they can process input sequences or continuous streams [77, 82, 83]. Another advantage of these networks is that the sequences can be of different lengths. This means that one type of pill is recognized with high confidence after a short time, while others are only reliably classified after a few shakes. Both cases with the same network. Eppe et al. [65] have investigated that LSTM [84] and GRU [85] perform better than Simple Recurrent Networks (SRN) on similar data. In preliminary experiments, we found that in our case LSTM achieve better results than GRU and decided on this architecture.

In the first layer of our unimodal networks we use MFCC for preprocessing the data. Subsequently, two consecutive LSTMs form the next layers, as these have achieved the best classification results for all modalities in our experiments. Both layers use a Rectified Linear Unit (ReLU) activation function. For training we also applied a dropout after the LSTM layers. In the last layer we use a fully connected layer and apply a softmax activation to classify the data. The network architecture can be applied to both modalities due to the analogous properties of acoustic and tactile signals (see Figure 3.6a). The unimodal networks diverge only in their respective MFCC, as mentioned before, parameters and the number of nodes in the LSTMs. We have optimized the parameters of the networks with hyperparameter optimization [81], separately for the modalities. In regard to the audio-based classifier, the optimal results were obtained with 400 hidden units in the initial LSTM, 90 hidden units in the subsequent one, and a dropout rate of 0.34. In the tactile version, the optimal values for the first and second LSTM hidden units were found to be 180 and 90, respectively. The highest classification accuracy was achieved with a dropout rate of 0.7.

For the multimodal network, the audio and tactile data must sooner or later be brought together in the network. Given the differences in sample rate, frequency range, and sequence size due to the MFCC parameters, we have decided to initially integrate the data in the fully connected layer. However, we intend to maintain the network structure and number of hidden units consistent with the unimodal architecture (see Figure 3.6b). The fully connected layer for the multimodal case is configured in a manner analogous to that of the unimodal architecture. The models were implemented with Keras [W8] and trained with the adam optimizer [86], a learning rate of 0.001, and the default momentum parameters of β_1=0.9, β_2=0.99 of the Keras deep learning framework.



(a) Unimodal Network

(b) Multimodal Network

Figure 3.6: The network architecture of the unimodal network is shown in **(a)** and that of the multimodal network in **(b)**. The unimodal networks mainly differ from each other in the MFCC parameters, the number of coefficients fed into the network, and the length of the sequences fed into the pipeline. The multimodal network combines the unimodal networks for audio- and tactile-based classification by concatenating the last output of the recurrent layers and feeding it into a fully connected layer. In the last fully connected layer of both architectures, a softmax activation is applied.

3.4 Evaluation

3.4.1 Experiment

To evaluate our method and answer the research questions, a dataset was recorded in which the robot shake 8 containers filled with the pills shown and presented in Table 3.1 and Section 3.2.5. To record the data, the robot arm is straightened out pointing to the microphone as shown in Figure 3.1a. The container is grasped in a tripod grasp with the thumb, index, and ring finger, and is firmly placed by a supervisor, before the shaking movement and the data recording starts. In this way, only the fingertips come into contact with the container, thus reducing the transmission of vibrations generated by the robot. The applied force and finger joint states remain constant during all sample recordings to generate homogeneous data. For one sample, the robot rotates its forearm for 180 degrees, as shown in Figure 3.1b, beginning with lifting the object, pausing for a short time, and lowering the object again. Each container is shaken 12 times, then changing the rotation velocity from 0.8 to 1.0 radian per second, and shaken 12 more times. For each

pill class, 4 different amounts of pills are recorded resulting in 96 samples per pill class. This results in a total of 768 samples for the 8 classes shown in Table 3.1. In addition to the raw tactile and audio data, the joint position of the forearm, the class type, the number of pills, and the rotation speed are stored in each sample. The dataset was not recorded under any specific circumstances, besides the sounds of the robot, there were other environmental noises such as running computers and sounds from outside. Most of the sounds were produced by the robot itself.

3.4.2 Results

To evaluate our classification approach and test the robustness of it, a dataset of 8 pill classes was recorded, each with four different amounts of pills in the container. A total of 7508 samples in 768 shaking movements were recorded. To train our networks, we split the data in 80 % training and 20 % testing samples.

To test the robustness of our approach, the microphone was not placed on the robot but directly in front of the sample to minimize ego and ambient noise. In real applications, more noise is to be expected as the microphone is attached directly to the robot. The noise was added to the audio signal afterwards. This only applies to the audio signal, there was no way to reduce or separate the noise for the tactile part. To determine the effect of the robot's ego noises on classification accuracy, we recorded them separately and added a noise amplification factor to our raw audio signals. This noise gain ratio was increased in 0.05 steps between 0 and 1, and the models were trained and tested separately for each case. The results are shown in Figure 3.7. As the noise has no influence on the tactile model, only one evaluation step was carried out with the tactile architecture. Ten iterations of training and testing were carried out in each evaluation step. In addition to the mean accuracies of the audio (blue), tactile (yellow), and multimodal (green) models, the 25- and 75-percentiles are also shown in the graph. The best classification accuracy in the test split of the data set was 56.06 % for tactile-only data, 89.1 % for acoustic-only data and 91.23 % for multimodal input. As the noise increases, the test accuracy for the audio and multimodal models decreases to the point where the audio model is only guessing, with a noise gain of 0.55. The multimodal model relies only on the tactile data from a noise factor of 0.5.

To evaluate the performance of the algorithms, a confusion matrix is shown in Table 3.2. It shows the classification results for a noise ratio of 0.3. This factor represents a more realistic scenario than without noise and is also a value at which the tactile signal has a significant influence on the classification results. The rows in the matrices indicate the actual pill class, the columns the predicted pill class. The table therefore indicates how often a type was correctly classified and, if not, with which class it was confused. For this noise ratio, the accuracy for the audio network is 58.75 %, for the tactile network it is 51.25 %, and for the multimodal network 71.63 %. In section **(a)**, the audio part, it is clear to see that the distribution of confusions is even, regardless of the pill type. Whereas in the tactile model **(b)**, some pills are more likely to be misrecognized, especially *Candy* and *Small Mints*, and therefore a more accumulated inse-

curity. In contrast, the accuracy of the multimodal-based counterpart has a significantly higher accuracy of 71.63 % (see Table 3.2 **(c)**). The confusion matrices reveal no significant within-pair confusion, suggesting that all models are generally effective at distinguishing between different classes.

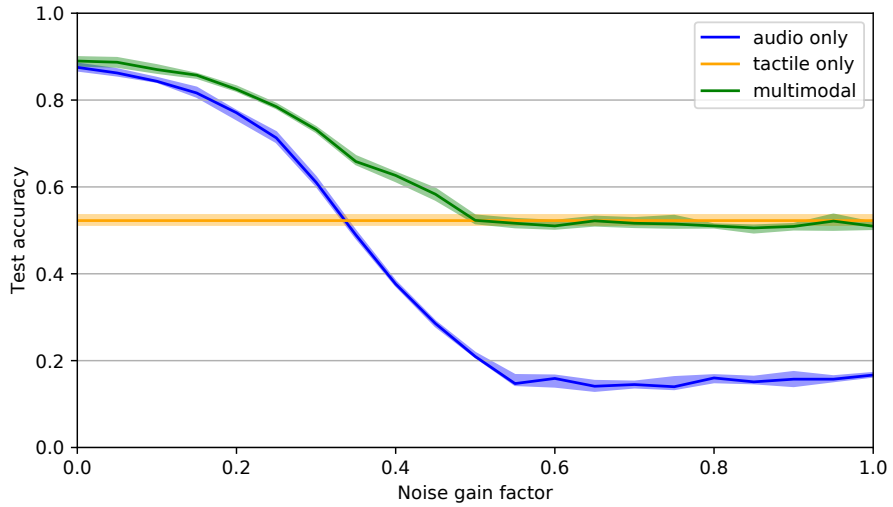


Figure 3.7: The graph shows the classification accuracies of the respective networks. The audio network in blue, the tactile network in yellow and the multimodal network in green. To test the robustness of the models, the added noise on the audio signal is increased on the x-axis.

Table 3.2: Confusion matrices for all three models at a noise ratio of 0.3

	(a) Audio only ∅58.75 %								**(b) Tactile only** ∅51.25 %								**(c) Multimodal** ∅71.63 %							
	Magnesium	Calcium	B-Complex	Big Mints	Chew	Small Mints	Vitamin B	Candy	Magnesium	Calcium	B-Complex	Big Mints	Chew	Small Mints	Vitamin B	Candy	Magnesium	Calcium	B-Complex	Big Mints	Chew	Small Mints	Vitamin B	Candy
Magnesium	0.42	0.21	0.02	0.19	0.03	0.01	0.05	0.07	0.49	0.07	0.13	0.09	0.05	0.14	0.0	0.01	0.7	0.08	0.09	0.07	0.04	0.0	0.01	0.02
Calcium	0.1	0.68	0.0	0.06	0.0	0.01	0.03	0.12	0.12	0.46	0.02	0.15	0.1	0.06	0.01	0.09	0.15	0.72	0.01	0.02	0.0	0.0	0.01	0.09
B-Complex	0.07	0.02	0.61	0.01	0.0	0.18	0.09	0.01	0.03	0.0	0.66	0.03	0.03	0.24	0.02	0.0	0.08	0.01	0.62	0.05	0.03	0.12	0.09	0.0
Big Mints	0.11	0.05	0.1	0.5	0.05	0.03	0.05	0.11	0.02	0.05	0.03	0.44	0.08	0.23	0.09	0.08	0.05	0.06	0.04	0.56	0.12	0.05	0.09	0.04
Chew	0.14	0.05	0.0	0.14	0.51	0.0	0.06	0.09	0.08	0.11	0.04	0.12	0.35	0.18	0.07	0.05	0.07	0.03	0.02	0.1	0.68	0.02	0.02	0.07
Small Mints	0.01	0.02	0.07	0.02	0.01	0.83	0.02	0.03	0.05	0.02	0.12	0.07	0.06	0.63	0.03	0.02	0.0	0.01	0.02	0.01	0.01	0.92	0.03	0.01
Vitamin B	0.08	0.09	0.05	0.06	0.0	0.07	0.47	0.18	0.0	0.0	0.05	0.05	0.05	0.19	0.63	0.01	0.0	0.0	0.04	0.08	0.06	0.01	0.74	0.08
Candy	0.06	0.11	0.02	0.05	0.02	0.03	0.04	0.68	0.02	0.04	0.02	0.19	0.17	0.11	0.01	0.44	0.0	0.03	0.0	0.04	0.03	0.02	0.09	0.79

3.4.3 Discussion

In order to answer the research question of whether tactile information in conjunction with acoustic signals in a multimodal network can provide added value and contribute to robustness in classification, a shaking experiment was conducted with different pills in identical containers. Figure 3.7 illustrates that with minimal noise, a multimodal network exhibited only marginally better than a unimodal audio model, with accuracy rates of 91.23 % and 89.1 %, respectively. However, when noise levels were elevated, the performance of the unimodal network declined to a significantly greater extent than that of the multimodal network. This indicates that the tactile data may offer additional value, which is not present in the audio data.

To verify this assertion, one may consult the convolution matrix presented in Table 3.2. The most compelling evidence in support of this hypothesis is *Vitamin B*. This particular pill is often confused by the audio network, whereas it is consistently identified by the tactile network. The audio network, in particular, identifies the pill as *Candy*, which is not the case in the other network. This discrepancy can be attributed to the material of the pills. While the majority of pills are rather hard, *Vitamin B* has a soft hull, resulting in a quieter noise when shaken. The increasing noise level presents a challenge for acoustic recognition, this does not affect the vibrations resulting in a robust tactile classification. The multimodal network takes advantage of this and uses the information from both modalities for even more robust detection of *Vitamin B*. With *Candy* and *Magnesium*, it is evident that the different unimodal networks confuse the pills with different other pills. The overall recognition of these pills is rather poor in both networks, especially for *Magnesium*, but, on the other hand for the multimodal network comparatively high.

Our research question can therefore be answered by stating that multimodal classification with acoustic and tactile signals in a noise-free space brings a small but not significant improvement compared to unimodal acoustic recognition. However, as soon as the ambient noise becomes stronger, information can be obtained from the tactile data that is not contained in the audio data and thus contributes to the robustness of the multimodal network and improves the classification accuracy by up to 30 % (with a noise gain factor of 0.5 in our experiment).

3.5 Conclusion

This work demonstrates the efficacy of employing tactile sensors to improve the accuracy and robustness of multimodal classification systems to learn more about the environment of the robot and analyze characteristics of grasped objects and thus provides an answer to the first fundamental question (**FQ1**). Moreover, the objective is to highlight the impact of multimodal usage with regard to the audio and tactile modalities. Our findings indicate that such an approach can enhance the precision of classification outcomes under optimal conditions while also improving the adaptability of the multimodal system to noisy signals. This shows, that the tactile data provides the network with supplementary information that is not available from the other modality.

Our work is based on the research conducted by Eppe et al. [65], who employed a deep learning approach to achieve the classification of optically identical capsules based solely on the acoustic modality. The aforementioned work and the work of Chen, Snyder, and Ramadge [58] both achieve slightly higher accuracy in their classification, but they employed a greater variety of materials in their experiments, which are easier to distinguish, and they recorded a markedly larger amount of data. Our findings indicate that our approach could be a valuable addition to both existing methods.

Other properties, such as the weight of the entire contents, the number and size of the individual objects, or even the material and hardness, could be determined using similar approaches to analyze the object. However, this approach is constrained by the capabilities of certain tactile sensors which are able to perceive vibrations. A classic tactile array will not achieve satisfying results in this context, as these typically lack the capacity to measure vibrations. Another possibility for future work would be the use of interactive sensing, which means that the robot would learn movements that would enable it to determine the content more quickly and robustly. The specific movements would likely be contingent upon the pill's characteristics and could potentially be learned through reinforcement learning [87, 88].

4 Robot State Estimation

We have successfully demonstrated that tactile data can be utilized for classification and provide added value in a multimodal setup together with audio data, in which the robot collects information about its environment. In the next step, we aim to investigate the extent to which tactile data can be used to gather information about the intrinsic state of the robot, to investigate the second fundamental question (**FQ2**). In the following work we want to combine the tactile modality together with proprioceptive information in form of motor states and use regression to to estimate the current joint state of the robotic gripper during grasping motions.

For some robotic scenarios, it is sufficient to simply grasp an object and place it roughly at another location. In these cases, the priority is to ensure that the object is securely and stably gripped. For other applications, it is necessary to grasp with precision, cautiously, or at a specific location. For both cases, there are grippers designed specifically to address these needs, but there are also grippers that aim to cover all possible scenarios. Other important design considerations include the cost, robustness, accuracy, and complexity of the gripper.

Robot hands like the Shadow Dexterous Hand [W7] can grasp very precisely, perform in-hand manipulations, and monitor their own state accurately. However, they are expensive, fragile, and complex to control, making them unsuitable for widespread use. In contrast, simple and inexpensive grippers, are often not equipped with enough sensors to measure their own state and perform precise grasps or manipulations. The qb SoftHand, for example, is very robust and adapts to the objects it grasps [89, 90]. However, achieving precise grasping or determining joint angles is very difficult for this gripper, which is generally a challenge in soft robotics.

In this work we will look at a gripper that is built for robust industrial applications and also for research scenarios. The underactuated 3-Finger Adaptive Robot Gripper built by Robotiq [W9] (see Figure 4.1) is, on the one hand, very robust and adapts to objects, is comparatively inexpensive and widely used, but on the other hand has only few sensors to analyze a grasp. In the following work the joint positions of this gripper during grasps are evaluated.

There are various approaches or sensors to determine the position of a motor or rotation. The most common ones are encoders and resolvers. In general, an encoder operates by detecting changes in the distance or position of a moving part and converting these changes into an electrical signal that can then be interpreted by a control system [92]. A resolver is an electrical transformer that measures the inductive coupling between two copper windings with an rotating conductive metal in between. The rotating metal piece changes the amplitude of the voltage, which is used to measure the angle of rotation [93, 94]. Another type of sensor is a potentiometer [95], which outputs a resistance value depending on its absolute position. These sensors are

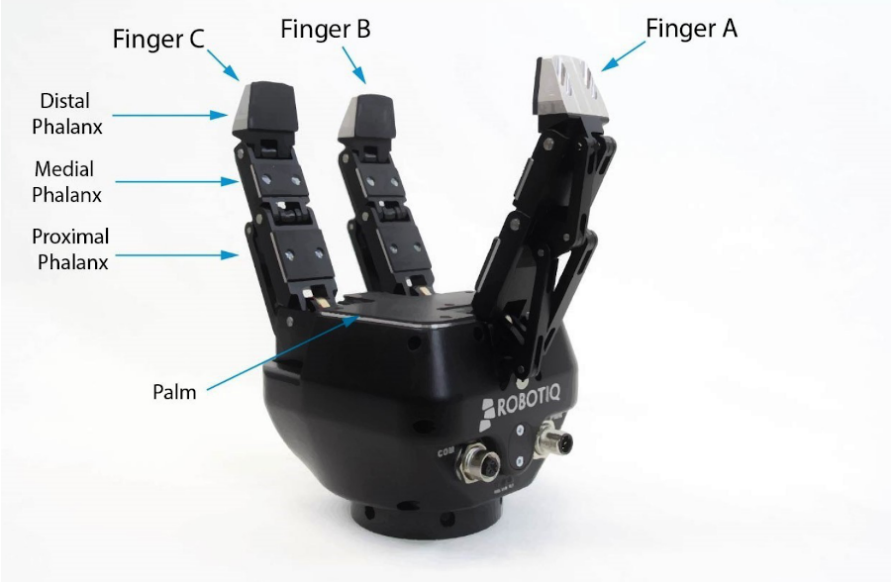


Figure 4.1: The Robotiq 3-Finger Adaptive Robot Gripper with named fingers and phalanxes. [91]

very affordable but cannot measure continuous joints. Hall effect sensors can also be used to measure joint angles. To do this, the magnetic field of a magnet attached to the axis of rotation is measured [96, 97]. The sensors presented so far typically need to be placed directly on or around the rotating axis. In common stepper or servo motors, these sensors are already built-in. Retrofitting them, for example, to measure a non-actuated joint, is often not trivial, especially in the case of robot grippers. Particularly when grippers are designed for human environments, having multiple fingers closely running past each other, there is little space available around the rotation axes. One approach that does not allow the motor position to be read out directly, but does not have the problem of having to be placed directly or close to the axis of rotation is the use of Inertial Measurement Unit (IMU) sensors. Due to their drift, they are not particularly suitable for traditional motor position determination, but they are useful for measuring the angle between two links [98]. Visual tracking systems are also frequently used, but are more suitable for laboratory conditions or under restricted movements like [12] or [73]. Under realistic conditions, visual tracking system have problems with occlusion.

Franchi and Hauser [12] use contact information to determine joint angles of the same gripper used in the work at hand. The approach is based on mathematical analyses and reaches its limits as soon as dynamics occur on the grasped object. Furthermore, it is only implemented in simulation. As we successfully integrated tactile arrays on the fingertips of our gripper [23, 22], we have decided to investigate to what extend it is possible to use this sensor technology to estimate the joint positions. A deep learning approach is applied to learn the behavior of the fingers and estimate the joint positions. The approach is also applied to the real robot hand and

evaluated with different objects.

This chapter is structured as follows: In Section 4.1 the work related to this approach is presented. Section 4.2 presents the used gripper, the custom contact sensors, and the tracking system to determine ground truth values. In the succeeding section we describe our deep learning approaches, the basic function of the analytical baseline method, and conduct an experiment in Section 4.4, in which 20 different objects are grasped. In Section 4.5 the results are presented and discussed in Section 4.6. Finally, Section 4.7 concludes this chapter.

4.1 Related Work

In 2019, Sintov et al. [99] published a paper in which they presented a transition model for their underactuated gripper. The gripper is composed of two fingers, each of which is equipped with two compliant joints with springs. A tendon through the entire length of the finger controls them. Similar to our work, they use fiducial markers to measure the ground truth of the joints. Using Gaussian Processes (GP) and Neural Networks, they learn the state of the gripper based on trajectory sequences. As input, they consider the angle of the actuators and their load, along with the positions of the objects, which are exclusively cylinders. The outcome enables the gripper to move the cylinder between the fingers and learn how to move the object along paths using in-hand manipulation.

Similar to the previously mentioned paper, Van Hoof et al. [100] also learn how to manipulate cylinders with an underactuated gripper with two-jointed fingers. The cylinder is positioned on a plane and contacted from both sides by the fingertips of the gripper, which are then moved back and forth. In contrast to the previous work, no state model of the hand is learned here, but the movements. Reinforcement learning is utilized to learn movement models, which are then used to fulfill the intended task. This is achieved by employing the motor values and tactile data from sensors at the fingertips. Other works that learn in-hand movements instead of state models are [101, 102, 103]

Soft grippers are a cheap and popular way of producing robust grippers. Matsuno, Wang, and Hirai [104] present a soft gripper, with flexible 3D-printed fingers with air chambers on the back. When filled with air pressure, the finger bends in order to adapt to the objects to be grasped. With such a gripper, fragile objects can be gripped robustly without hesitation. However, determining the state of such a finger is not trivial as it has way more Degrees of Freedom then a rigid one. By utilizing electro-conductive yarn on the back of the finger, which alters the resistance when the finger is bent, it is possible to estimate the state of the finger. With the change in resistance, the diameter of the object can be determined, which results in conclusions about the bending state of the finger. Similar works with soft grippers are [105, 106, 107].

Another field of research is the determination of the object pose in the hand, which is not only a problem for underactuated, but also for fully actuated or anthropomorphic robotic hands [108, 109]. An approach using a gripper with a similar design to that employed in previous related

work and the gripper we use is that of Azulay, Ben-David, and Sintov [110]. They use kinesthetic features in the form of actuators angle and tendon torque, together with tactile data, to estimate the pose of the object in an initial model. Subsequently, they learn a transition model to perform in-hand manipulations. Another publication about object pose estimation with underactuated grippers is [111].

To the best of our knowledge, Franchi and Hauser [12] present the only work that estimates the finger states of a Robotiq 3-Finger Gripper, which is the same gripper used in our work. Their approach is based entirely on the mathematical analysis of the behavior of the fingers. Later in this chapter, in Section 4.3.1, the approach is discussed in more detail and compared to our neural network approach. One limitation of this work is that the evaluation was conducted exclusively on static objects, which does not reflect the complexities encountered when working with real-world objects. Additionally, the approach was used to simulate the gripper. An additional example of a mathematical approach to determining the state of a gripper is [112].

4.2 Hardware Setup

4.2.1 3-Finger Adaptive Gripper

The gripper used in this setup is a 3-Finger Adaptive Gripper from Robotiq [W9]. The gripper is designed to be robust and precise for advanced manufacturing and robotic research. With its adaptive three fingers, it enclosures objects more then a simple parallel or two finger gripper and therefore provides a more stable grasp in general.

A picture of the gripper is shown in Figure 4.2. The gripper has three fingers, two of them close in parallel (finger B and C), the other one closes from the opposite direction (finger A). Each finger has three joints, mechanically coupled on the backside to close adaptively. Each finger is controlled by one motor located in the lowest joint. The range of motion and axis of rotation of the motor are depicted in Figure 4.3. When the finger closes, joint 1 closes first, joint 3 opens, and joint 2 remains unchanged. As soon as joint 1 has reached its limit of movement, joint 2 begins to close, while joint 3 remains unchanged as it is at its negative limit in this case. If joint 2 also reaches its limit, only joint 3 closes. This behavior changes when an object is touched on any finger segment. The motor is differentiated with an 8 bit integer value and can therefore move to theoretical position values between 0 and 255. The actual value range is between 6 (fully open) and 240 (closed). A fourth motor between finger B and C controls a scissor movement of these two finger. It can spread the fingers and close them till they touch. With these movement options, the gripper can be used in different modes. In *basic mode*, fingers B and C are parallel and all three fingers close together, the scissor motor remains unchanged. In *wide mode* and *pinch mode*, the gripper behaves in the same way, except that fingers B and C are spread or the tips of the fingers touch, respectively. The last mode is called *scissor mode* where only the forth motor is controlled to open and close finger B and C to each other. It is also possible to control all motors individually in order to control the hand as desired.

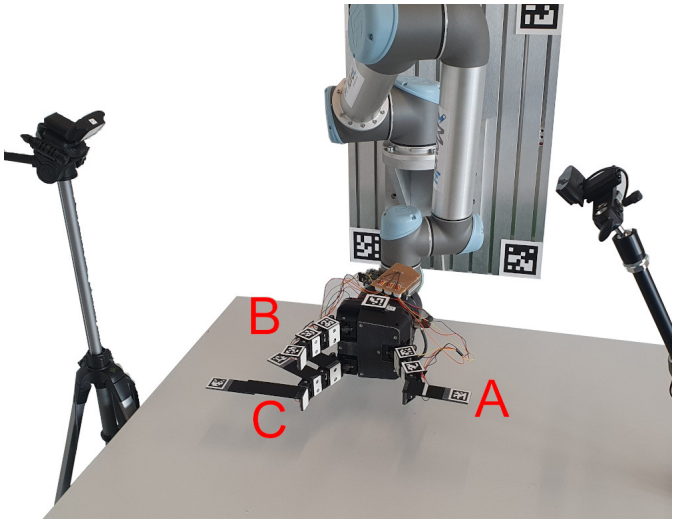


Figure 4.2: The experiment setup to record training and test data. Our gripper is equipped with 9 tactile sensors including AprilTags to measure the ground truth joint states, one at each phalanx. An additional tag is attached to the side of the palm to measure the first joints. Two cameras are positioned at an angle on the side of the gripper to monitor the tags.

Besides the four motor position values, the hand provides several other measurements. The general status of the gripper, if it is in movement, or if one, two, or all fingers have stopped. Additionally, a value representing instantaneous current consumption which only strikes if a lot of force is applied, which usually is only the case when the hand is about to stop, and an object detection value.

4.2.2 Contact Sensor

To solve the task, contact sensors can be used as the approaches are on the one side only interested in binary *contact* or *no contact* information for the baseline, or force values for the deep learning approach. In this work we use our own custom built contact sensor described in [23] and [22]. Instead of building tactile arrays for all links, we decided to build simpler one single taxel per phalanx sensors as only one microcontroller to read all 9 sensors is needed. The sensors are integrated together with the fiducial markers on custom designed finger pads (see Figure 4.4).

The resistance-based contact sensor is build from several layers (see Figure 4.5). The bottom layer of a plastic foil as shown in the picture is left out, as the aluminum foil is glued directly to the adapter which serves as supply. On top of it, the piezoresistive velostat foil is placed followed by the sensing aluminum tape layer. To cover this, a layer of transparent adhesive tape is stuck over it and folded over at the sides so that all layers are attached to the adapter. The last

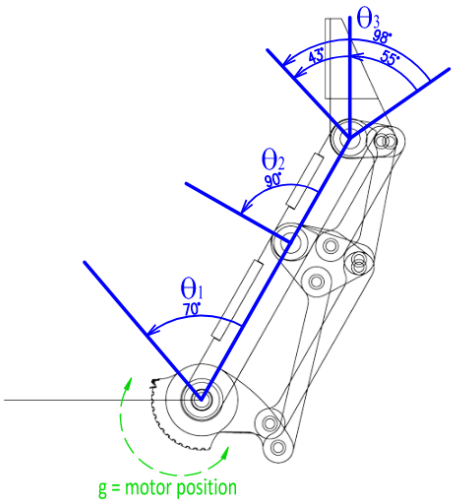


Figure 4.3: The motion angles of the three joints. The displayed pose represents the zero position of the fingers. In green are the direction of motion and axis of rotation of the motor with position g depicted. [12]

layer is a 1 mm thick silicone mat, which serves to protect the sensor, make it more robust, and distribute the applied pressure.

When pressure is applied to the sensor, the resistance of the velostat foil changes which is measured by a microcontroller. The sensor is incorporated into a voltage divider alongside a 10 kΩ resistor in the readout circuit. Consequently, the voltage output from the divider corresponds to the level of pressure. The output is measured with a rate of 50 Hz and a resolution of 10 bit by an analog to digital converter in the microcontroller. At the beginning of each grasp, the readout is zeroed and normalized between 0 and 1. Preliminary tests have shown that a good threshold value for detecting static contacts is 0.7. When only one finger is touching the object and thus moving it, the value is usually lower then 0.4.

4.2.3 Visual Tracking

To measure the ground truth joint angles of each finger, we decided to use a visual tracking system as it can be integrated quickly and easily, and also be removed if required. A fiducial marker was attached to each phalanx and the palm of the gripper as shown in Figure 4.2 and Figure 4.6. Since the original pads and fingertips were replaced and redesigned anyway with the new contact sensors, the tags were integrated directly into the new pads (see Figure 4.4). The tags are observed from the side by two conventional usb cameras with a frame rate of 30 Hz, a resolution of 720p, and a distance of 50 cm. To reduce occlusion, extensions have been added to finger B and the fingertip of finger A. These extensions would make a permanent use of the tracking system not feasible, as they would significantly restrict the range of movement and

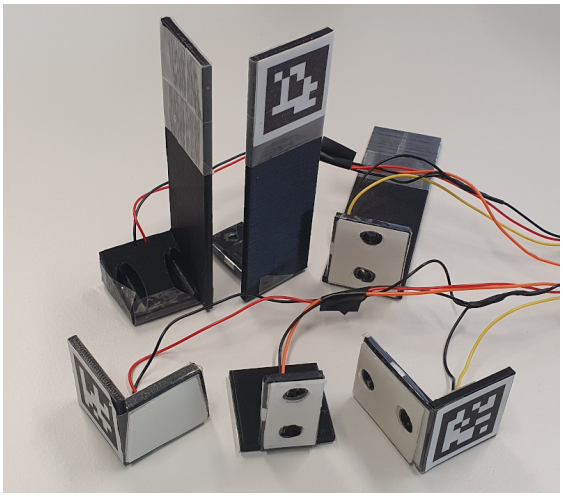


Figure 4.4: Custom designed and built finger pads for the proximal and middle phalanx and a fingertip for the distal one. Each peace is 3D-printed and equipped with custom-built contact sensor and AprilTag to track the position. The tag holders are of varying lengths to minimize occlusions.

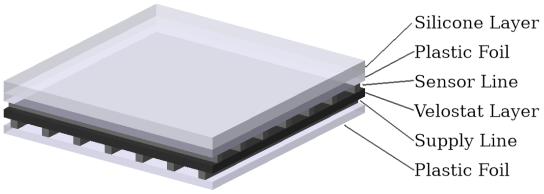


Figure 4.5: Layers of the custom built tactile sensor used for the contact sensor at hand. [22]

make collision-free movements difficult. In addition, the markers are observed from fixed angles and optimal tracking only works in this position when the arm is not moving.

The fiducial marker system used is AprilTag 3 [113] with the tag family 36h11. All tags are placed in a plane with the z-axes pointing out from the markers. With the dot product it is possible to calculate the angle between the x- or alternatively the y-axes between two tags, then the angle of position 0 is subtracted to get the current angle of a joint. With an accuracy of 0.3 degree, the angle between two links can be calculated.

4.3 Approach

Our aim is to be able to determine the joint states of the finger joints. To achieve this goal, it is necessary to integrate additional hardware on the robot. One approach is the use of Hall effect sensors as shown by Kargov et al. [114]. Here, a magnet is mounted on the rotating axis and the

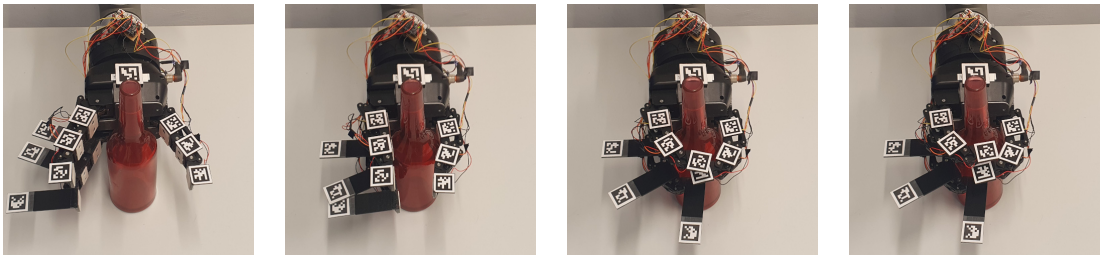


Figure 4.6: An exemplary grasp with the Robotiq gripper. The fingers adapt to the object when closing and enclose it. The pictures show from left to right the fully opened state, the first contact with the object, the adaptive behavior in the middle of the grasp, and the final grasp.

magnetic field is read without contact by a Hall effect sensor and the joint position is determined. Other hardware solutions can also be used, as in the work of Seel, Raisch, and Schauer [115], who use IMUs to determine the angle of a human knee, this can also be transferred to a robot. The disadvantage of these approaches is the considerable hardware outlay. As the space in front of, behind, and next to the individual phalanges is limited, it is not easy to integrate such solutions retrospectively. The advantage is that the joint position can be determined at any time. Visual tracking of the hand is also possible, Andrychowicz et al. [73] used a PhaseSpace system to track the fingertip positions of their shadow hand to estimate the hand state. The disadvantage is, as already mentioned in our visual based ground truth generation, that the visual markers are often obscured when the robot is not at a fixed position.

Another approach, implemented by Franchi and Hauser [12] is the use of contact information, in their case only in simulation. Contact sensors are often already present on grippers or are integrated anyway to obtain more information about the grasp. Depending on the sensor, they are very flat and take up little space. The disadvantage of contact sensors is that the joint position cannot be read out directly, but must be estimated from the context of the grasp.

We decided to use contact sensors as the fingertips were already equipped with tactile sensor arrays in previous work [23, 22]. The objective is to find out how accurately the joints of the hand can be predicted with contact sensors on all phalanges. Franchi and Hauser [12] did this in simulation with an analytical approach in which they set up a mathematical model for the fingers (see Section 4.3.1). In their approach, there is only one model which is applied to all fingers, without considering the other two fingers. In practice, this has the disadvantage that it only calculates the joint states correctly for static objects. As soon as the object moves, the calculation leads to incorrect joint states. To capture this nonlinear behavior of the fingers, a deep learning approach is utilized in the work at hand. On one hand, the fingers are considered separately, similar to the analytical baseline approach. On the other hand, we aim to collectively learn all fingers simultaneously to account for mutual influences. This leads to the following two research questions:

1. Can a newly designed recurrent neural network approach outperform the existing baseline for state estimation and compensate object movements?

2. Does an approach that takes all fingers into account at the same time perform better than not differentiating between the fingers?

4.3.1 Baseline - Mathematical Analytical Approach

Franchi and Hauser [12] show, it is possible to calculate the state of the hand as long as the grasped object is static. The authors divide the closing movement of the hand into four phases to calculate the relative change of the joint states from step to step (see Table 4.1). A phase transition is determined by whether a finger segment is in contact with an object (c_n) or reaches its limit (l_n). The full state tuple has all six limit and contact information ($c_1, c_2, c_3, l_1, l_2, l_3$). In the first phase (1) neither a contact nor a limit at any joints occur. In 1', the fingertip hits its negative limit and does not open any further. For the next phase (2), the first phalanx is in contact or in a limit. Again, 2' determines that the fingertip is in its negative limit, however, this limit can be discontinued again. In phase three (3), the middle phalanx is in contact or limit. If the hand closed this far, it is not possible, that the fingertip will stay in the negative limit except for an immediate contact and therefore change to phase (4), which determines the final sate of the hand as no joint will change anymore. Phases can also be skipped, for example, if the first contact occurs at the fingertip. While it is theoretically possible to return to a previous phase, it is not the case in reality, as the formulas only cover closing motions.

The three joint angle changes are calculated depending on the current phase with the following equations:

$f_1(x,u) = m_1 u$, with $m_1 = \Theta_{1,max}/140$

$f_2(x,u) = m_2 u$, with $m_2 = \Theta_{2,max}/100$

$f_3(x,u) = m_3(g)u$, with $m_3(g) = \Theta_{3,min} + (\Theta_{3,max} - \Theta_{3,min})/(255 - g)$

The changes in g from one time step to the next are described with $u \in [-1,1]$, while 1 is a closing and -1 an opening motor step. A full state machine is shown in their paper [12].

Table 4.1: Analytical joint state calculation [12]

Phase	State tuples	$\Delta\Theta_1$	$\Delta\Theta_2$	$\Delta\Theta_3$	Δg
1	(0,0,0,0,0,0)	$f_1(x,u)$	0	$-f_1(x,u)$	u
1'	(0,0,0,0,0,-1)	$f_1(x,u)$	0	0	u
2	(1,0,0,0,0,0),(0,0,0,1,0,0)	0	$f_2(x,u)$	$-f_2(x,u)$	u
2'	(1,0,0,0,0,-1),(0,0,0,1,0,-1)	0	$f_2(x,u)$	0	u
3	(·,1,0,·,0,0),(·,0,0,·,1,0)	0	0	$f_3(x,u)$	u
4	(·,·,1,·,·,0), (·,·,0,·,·,1)	0	0	0	u

4.3.2 Recurrent Neural Network Approach

Without a contact, the state of the hand is known at every possible point in time, i.e. at every motor position g. However, as soon as contact is made with an object, this linear behavior changes and the state dependents on the previous step, g-1 or g+1. RNNs are designed for such cases and map temporal or state dependencies and learn on sequential information. The most common RNNs are LSTMs [116] and GRUs [85]. LSTM networks performed better in preliminary tests, which is why GRUs ar not considered any further at this point.

Equal Finger Network

As the baseline approach does not differentiate between finger A, B, and C we wanted to build a network using the same information, to see if it outperforms this approach. This network therefore receives sequences of one motor position and three contact values as input and outputs three joint angles. The model is trained with the sequences of all fingers. During data recording we do not distinguish between the fingers which means one grasp produces three sample sequences. Using Optuna [117] for hyperparameter tuning, the parameters of the network are optimized. The parameters to be determined are the number of LSTM layers, LSTM neurons, optional linear layers after the LSTM, activation functions, learning rate, and weight decay. The hyperparameter optimization found the following parameters to be most effective: 148 neurons in one LSTM layer and a hyperbolic tangent (tanh) activation function. It was trained with an adam optimizer for 1000 epochs, a learning rate of 0.002, and a weight decay of 0.00001.

Entire Hand Network

In order to learn influences between the fingers during dynamic object movements, the fingers are now considered individually. This means that three motor positions and nine contact values are entered sequentially into the network with an output of nine joint positions. For the data recording it means, that one grasp produces one sample sequence. Just as for the other network, hyperparameter tuning is used to determine the best parameters, again with Optuna [117]. The parameters to be determined are again the number of LSTM layers, LSTM neurons, optional linear layers after the LSTM, activation functions, learning rate, and weight decay. The hyperparameter optimization found the following parameters to be most effective: 37 neurons in one LSTM layer and a tanh activation function followed by one linear layer with 184 neurons and a ReLU activation. It was trained with an adam optimizer for 1000 epochs, a learning rate of 0.0009, and a weight decay of 0.00002.

4.4 Experiments

To test and evaluate the three approaches against each other, we grabbed different objects and compared the resulting movements and final grasps with the ground truth data from our AprilTag

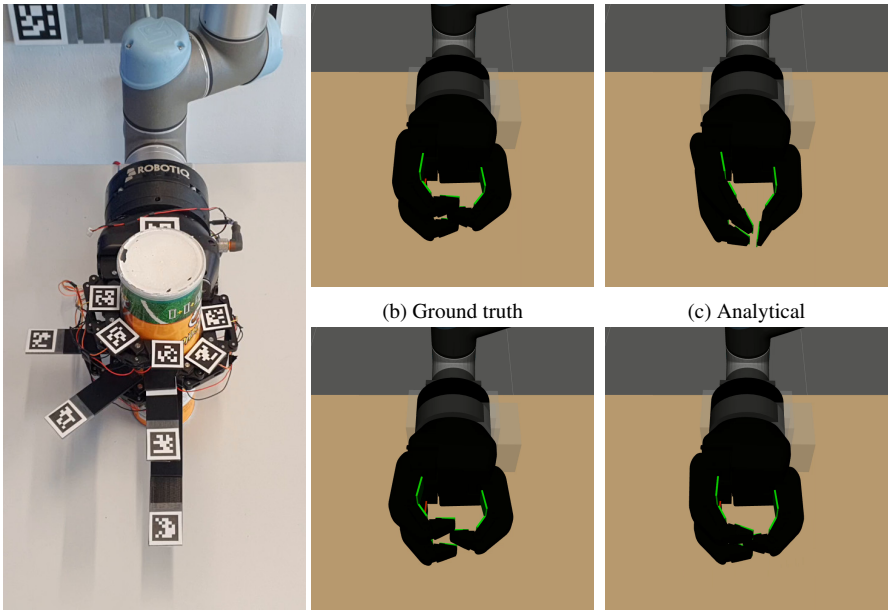
Figure 4.7: 20 objects varying in size, shape, weight, and stiffness were used to record the dataset for the experiment. Some of the objects are from the YCB object set [118].

setup. The used experiment setup is shown in Figure 4.2, the arm is fixed in the shown position approximately 15 cm above the table and will not move during data collection to produce comparable training data. In total, 20 different objects were grasped with the gripper, which are shown in Figure 4.7. Some of the objects are from the YCB object set [118] and differ in shapes, sizes, and stiffness. The objects are placed by hand on the table within the fingers and moved slightly between the grasps. Each object was grasped three times, resulting in a total of 60 grasps, which means 60 sequences for the *entire hand* and 180 for the *equal finger* network.

In Figure 4.6 an exemplary grasp is shown. The hand begins in a fully opened position and closes all fingers with the same speed. As soon as the first contacts are made, the adaptive behavior of the hand begins and the fingers wrap around the object, which is shown in the center pictures. In the right photo, the grasp is finished and neither the fingers nor the object is moving anymore. One grasping sequence contains exactly this movement, from a fully opened gripper till the end of the grasp.

During data recording the fingers are closed slower than usually to increase the tag detection accuracy. As the neural networks recurrent dependency is the state rather than the time in our case, the closing speed does not matter. The recording of a sequence is stopped when the last finger stopped moving, meaning they are in a limit or in contact. When a marker is occluded and not correctly detected anymore, the sample collection for this finger is terminated and will not be continued if the tags are detected again.

The recorded data is used for all three approaches. In the case of the deep learning ones, the dataset is split into training (80 %) and test (20 %) set.



(a) Pringles can (b) Ground truth (c) Analytical (d) Equal Finger RNN (e) Entire Hand RNN

Figure 4.8: Estimation results of the end state of grasping a Pringles can.

4.5 Results

With the recorded dataset described in Section 4.4 the three approaches are evaluated.

Figure 4.8 depicts the visualizations of the state estimations of the individual methods for a Pringles can. Additionally, the figure presents an image of the real robot with ground truth markers attached, along with the visualization of this ground truth data. Visualizations of the grasps of other objects are shown in the appendix in Chapter 8.

For the results in Table 4.2, the average difference between the ground truth and the estimated joint position for all motor positions in all grasps are listed. To get a better understanding of the individual joints, the errors are listed individually and overall. Since the most interesting part of the state estimation is the final state, as this can be used for further applications rather than the complete movement, both results are listed. The average error or accuracy is given in radians. It is noticeable that joint 1 between the palm and the proximal phalanx has the highest accuracy for all approaches, followed by joint 2 between proximal and middle, and joint 3 between middle and distal to be the least accurate, independent of the full movement and end state. The baseline approach reaches an overall accuracy of 0.142 radian (0.248 end state), the *equal finger* approach reaches 0.089 radian (0.133 end state), and the *entire hand* network has the most accurate results with 0.04 radian (0.079 end state).

These results are consistent with the boxplots in Figure 4.9 for the full closing movement and Figure 4.10 for the end state error. The plots show the median value as horizontal red line within

Table 4.2: Joint state estimation accuracies in radians

Whole closing motion	Joint 1	Joint 2	Joint 3	Overall
Analytical	0.084	0.145	0.196	0.142
RNN equal finger	0.057	0.093	0.117	0.089
RNN entire hand	0.026	0.048	0.047	0.040
End state				
Analytical	0.140	0.275	0.329	0.248
RNN equal finger	0.088	0.149	0.163	0.133
RNN entire hand	0.045	0.085	0.107	0.079

the box, the box itself indicates the 25th and 75th percentiles, and the lines show the interquartile range times 1.5. The distribution of errors is significantly higher with the analytical approach than with the other two, both for the entire movement and for the final state. In the most extreme case for joint 3 at over 45 degrees (0.8 radian). The plots also show that the distribution is significantly lower for the entire hand. It is striking that the deviation for joint 3 in the entire hand network is smaller than for joint 2, but only for the full movement.

Figure 4.11 gives more information about when errors occur, the graph shows the amount of deviation according to the motor position g. At the beginning of the movements the error is very small in all three cases and increases more and more. It can be seen, that the deep learning approaches are much more accurate than the analytical approach. Again, the three joints are plotted separately, red shows joint 1, green joint 2, and blue indicates joint 3.

To understand when the errors occur, an exemplary grasp is plotted in Figure 4.12, divided into the three fingers and the three joints, with the same color scheme as before. The joint states are plotted in the upper graphs and the corresponding contact values are plotted in the lower graphs over the motor position g. In addition, the lines in the upper part are divided into ground truth as a continuous line, analytical estimation as a dotted line, and the results of the network for the entire hand as a dashed line. The results of the equal finger network have been omitted to maintain clarity. It can be seen that the lines deviate from the ground truth state as soon as contacts occur.

4.6 Discussion

In the first research question we wanted to find out if a neural network approach can outperform the existing analytical approach and compensate object movements. The results presented in Section 4.5 have shown that both deep learning approaches perform better than the baseline. The reason for this can be seen in Figure 4.12, as long as no contact appears, all estimations perform equally well. When a phalanx contact occurs, the finger position for the analytical way will not change anymore even though, the real joints still move due to object movement.

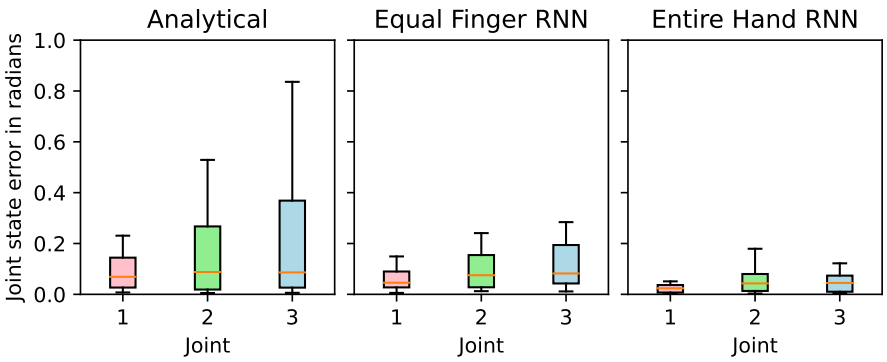


Figure 4.9: Each boxplot show the joint state error distribution in radians for the three joints during the full closing trajectory. The left one shows the results from the analytical method, the middle one from the RNN approach treating all finger the same way, and on the right side the RNN approach using the entire hand.

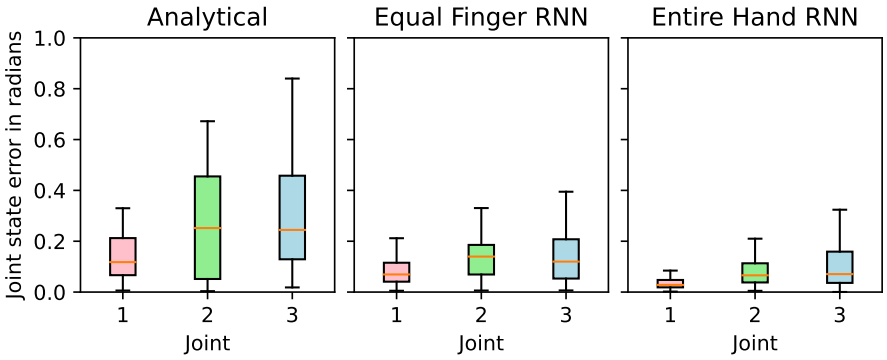


Figure 4.10: Each boxplot show the joint state error distribution in radians for the three joints for the last state when the grasp is complete. The left one shows the results from the analytical method, the middle one from the RNN approach treating all finger the same way, and on the right side the RNN approach using the entire hand.

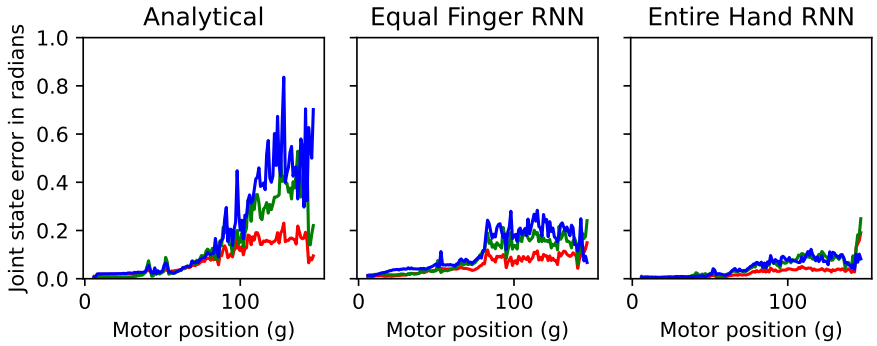


Figure 4.11: The deviation of the joints from the ground truth distributed over the motor position g of the entire dataset. Blue indicates the fingertip, green the middle, and red the proximal joint. The left one shows the results from the analytical method, the middle one from the RNN approach treating all finger the same way, and on the right side the RNN approach using the entire hand.

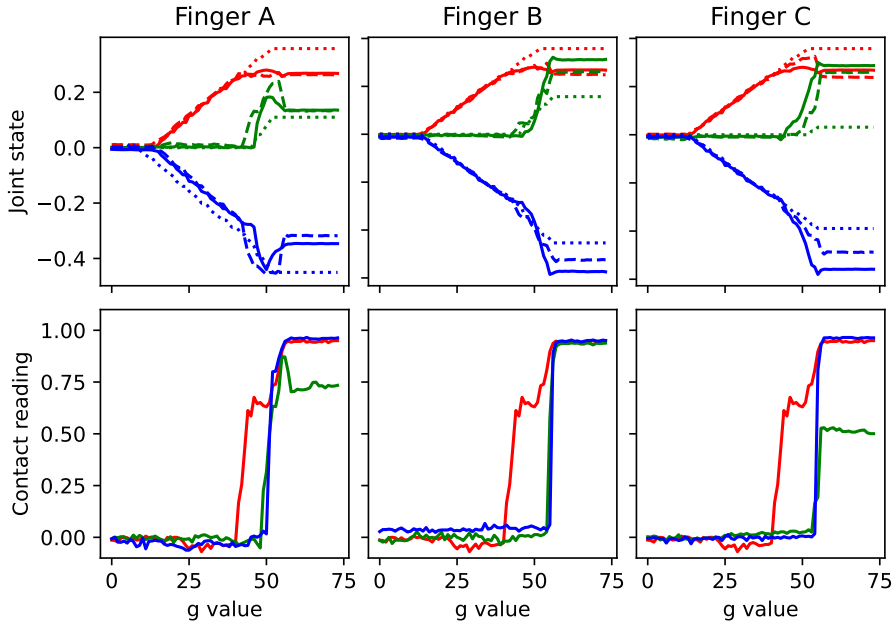


Figure 4.12: In the upper graphs, the states of all 9 finger joints are plotted during one exemplary grasp of the wooden block. The ground truth data is plotted as solid line, the analytical method as dotted line, and the RNN approach for the entire hand with a dashed line. In the lower plots, the normalized contact readings of the respective finger are shown. Blue indicates the fingertip, green the middle, and red the proximal joint.

The other two approaches are more stable there and behave more like the ground truth curve. This is also the reason why the error in Figure 4.11 increases a lot for the analytical approach. Another problem is the different forces acting on the object, as soon as all fingers have contact with the object and exert forces from two sides. The object moves in the hand and the adaptive fingers react to these forces and object movements. This behavior is not covered in the analytical approach, but is learned in the deep learning methods. As expected, the model for the entire hand is even more robust here. In-hand movements occur for almost all objects, which increases the error for the baseline.

Regarding the second research question we wanted to investigate the difference between the two neural networks. Since the networks hardly differ architecturally, but in one case only receive information from one finger, while in the other case they receive all values, it is obvious that this information provides added value in determining the joint states.

For further applications like grasp analysis and evaluation, the end state is of greater interest than the entire movement, especially the accurate starting phase without a contact is negligible. As depicted in Figure 4.11, it becomes apparent that the analytical method exhibits high levels of inaccuracy towards the end. Furthermore, Figure 4.10 illustrates the discrepancy in final states, emphasizing a notable deviation from the true state. In comparison, the recurrent neural network method, which takes the entire hand into account, is more accurate and is therefore suitable for further applications.

The results have shown that the errors increase significantly from joint 1 to joint 3 (see Figure 4.9 and Figure 4.10). To explain the behavior of the individual joint errors, one can examine the range of motion. The fingers behave adaptively due to the coupling mechanism behind them. This adaptive free space is noticeable smaller for joint 1 compared to joint 2 and 3, whereby joint 3 can change the most. This implies that solely from the motor position, one can predict the location of joint 1 more accurately than joints 2 and 3.

With an average accuracy of 0.04 radians (2.29 degrees) and an average end-state accuracy of 0.079 radians (4.53 degrees), the state of the hand can still be determined very accurately.

4.7 Conclusion

On our way to utilizing tactile data in the robotic environment, we have seen that tactile sensors can be used to learn more about the environment or more precisely about characteristics of grasped objects (see Chapter 3). In this chapter, we wanted to use tactile data to learn more about the internal state of the robot and investigate the second fundamental question (**FQ2**). With a new deep learning approach, we have shown that the state of an underactuated gripper can be determined only with the use of contact sensors and one motor position. The baseline for this work was presented by Franchi and Hauser [12]. The authors used mathematical analysis to determine the condition of the hand. However, this method reaches its limits as soon as gripped objects move. With our RNN approach, we were able to robustly determine the state of the hand

even under these circumstances. Furthermore, we have shown that it is advantageous to consider all three fingers together rather than individually. The best results achieved had an accuracy of 0.040 radians or 2.29 degrees with the *entire hand* model for the full motion and 0.079 radians or 4.53 degrees for the end state.

In the next step, the results of this research can be used to analyze and evaluate grasps. Other works have used similar approaches to perform in-hand manipulation of objects [100, 110].

5 Mixed Reality Teleoperation

So far, tactile sensors have been used to collect intrinsic and extrinsic information with the robot independently of the human. In one case tactile data was used together with audio and in the other case together with proprioceptive data to see whether a multimodal advantage can be achieved. In the next step, we want to involve the human and communicate the tactile data measured by the robot with visual and acoustic cues, to address the third fundamental question (**FQ3**) to augment the human perception.

In 1954, Goertz and Thompson [119] presented one of the first approaches to remote control a robot. Since then, the field has evolved with new robots, new interfaces, and new applications. In 1994, the first 6-DoF robot that could be remotely controlled via a desktop application over the internet was introduced by Goldberg et al. [120] allowing to grasp and manipulate objects with a robot at another laboratory, in another country, and even in space. An early overview on the topic is presented in [121].

Mastery of teleoperation typically needs a period of training. This is particularly the case in environments where mistakes cannot be tolerated, such as in surgical tasks. Especially in this context, it is important to implement simple interfaces [122, 123]. But it is also advantageous for other applications to keep the system as simple as possible, especially for non-expert users.

A further challenge is to track the user's movements. This can be achieved through the use of hardware such as IMUs or LEDs, which are attached to key points on the user's body and thus enable to track it [124, 125]. Alternatively, via tracking gloves, some of which are able to provide additional feedback [126]. These approaches have the disadvantage that it takes a long time to prepare the user to put on the devices and that movements are sometimes restricted. An advantage is that movements can often be transferred one-to-one to the robot. XR devices usually come with integrated hand or controller tracking, enabling interaction with the virtual world. This data can also be used for teleoperation, albeit with limitations. For instance, while a 6-DoF pose per hand and the pose of the head can be obtained, information about the entire arm and body are missing. On the other hand, these devices can be set up and ready for use in a matter of seconds.

A non-trivial modality for teleoperation, especially in combination with XR devices, is haptic. The transmission of vibration or small deformations via small wearables is possible for haptic feedback; however, these often reduces the immersion and make tracking with XR devices more difficult [127]. Another form of contact less haptic feedback is via ultrasound, but this can also restrict the range of movement and requires an extra device [128].

In the following chapter we present our MR teleoperation approach. The objective is to imple-

ment real-time teleoperation for one or two robot arms simultaneously. The system should be as free as possible from hardware dependencies in order to facilitate the integration of new robots and to enable the replacement of the XR device, for example to utilize only the tracking, to provide virtual augmentations in AR or to switch completely to VR. Furthermore, it is our intention to present tactile information from the robot to the user, without any additional hardware other than the XR display, in form of visual and acoustic cues through the HMD. The contribution of this work is as follows:

- Development of an MR-based real-time teleoperation system combining XR devices with arbitrary robot arms
- Visual and auditory feedback of tactile readings during MR teleoperation
- Evaluation of the teleoperation system regarding the user experience and usability

The rest of this chapter is organized as follows. Section 5.1 provides an overview of related work on teleoperation and feedback methods in XR. The teleoperation approach is presented in Section 5.2 including the robot setup, tracking, and jogging of the robot. We then evaluate our system in a user study in Section 5.3, describe the results in Section 5.4, followed by the discussion. Section 5.5 offers a summary of this part and suggestions for subsequent work.

5.1 Related Work

The work related to our approach can be divided into two parts, on the one hand implementations of XR teleoperation approaches, on the other hand feedback in Virtual Environments.

5.1.1 Teleoperation with Extended Reality

The benefit of teleoperation in virtual environments is, that the devices typically provide input options, such as controllers or hand tracking. The poses and interaction possibilities provided can be optimally employed for teleoperation without the necessity to wear an elaborate tracking suit or similar beforehand. There is no consensus among scientists who integrate XR devices with robots regarding the use of the terms VR, AR, and MR. This often makes it difficult to compare the approaches they take. In this related work, we use the taxonomy as presented in Section 2.5. Consequently, some authors refer to their work as VR teleoperation, but it is described here as MR.

In 2024, Cheng et al. [129] presented a teleoperation approach with active visual feedback. Meaning that the user in MR is able to control the robot's head through their own head movement, thereby enabling active visual exploration of the robot's environment. The robot is equipped with a stereo camera whose two camera perspectives are displayed directly in the VR glasses, allowing the user to see through the eyes of the robot. A closed-loop inverse kinematics algorithm,

based on Pinocchio, calculates the joint angles for the robot's arms from the user's hand pose. The authors utilize the system for imitation learning. A few years earlier, Zhang et al. [130] presented a comparable setup for complex manipulation tasks with imitation learning. In their MR teleoperation application, they present the robot's environment to the user with point clouds from a depth camera instead of a stereo camera.

In their study, Penco et al. [131] present a system through which they teleoperate a humanoid robot with assistive autonomy in MR. This implies that, on the one hand, they can plan movements with a virtual robot before they are optimized and executed, and on the other hand, they can also perform direct teleoperation with probabilistic movement primitives. With their humanoids they can open doors and box against a punching bag. Other works that teleoperate humanoids with MR are [132, 133, 134, 135].

Not only the direct teleoperation of robots is being researched, but also alternative interactive interfaces in AR to control the robot. This can be achieved, for example, by setting a target position, visualizing the planned trajectory, and then executing the movement [35, 136, 137].

5.1.2 Feedback Methods in Extended Environments

In their work, Chan et al. [138] present a combination of direct and pre-planned teleoperation. Users are able to specify a path for the robot to follow. Subsequently, the planned trajectory can be traversed manually via a gesture. The user is provided with feedback on the force applied at the end effector during the movement. Two distinct feedback methodologies for this force are evaluated in a study. One is a visual arrow above the gripper that changes in size, and the other is haptic vibrations on the forearm. They achieve superior outcomes with the haptic feedback and attribute this to a cognitive overload with visual feedback. Other studies have reached a different conclusion, with better results achieved through the use of multimodal feedback [139, 140, 141, 142].

An earlier attempt to provide feedback in a desktop VR experiment was conducted by Herbst and Stark [140] in 2005. The researchers employed a data glove that provided vibration feedback at the fingertips, in conjunction with a hand tracking device for control, to enable participants to manipulate virtual boxes within the study. Using different modalities (visual, acoustic, and haptic), they provide feedback on force magnitude so that the candidates can sort the blocks by weight or push the block with the least friction off a stack. The combination of two modalities has been found to result in an increase in performance, in terms of both execution time and the number of transitions. However, the combination of all three modalities has been found to provide no further increase.

One use case of VR and haptic feedback that has attracted considerable attention from the research community is surgical procedures. In their work, Zhou et al. [143] show the impact of haptic feedback to surgeons and how important it is to be able to handle this feedback when training new surgeons. They present this concept through a study in which they conduct surgical operations in a simulator, with and without haptic feedback. Other works in the field of feedback

in surgery are [144, 145, 146].

One effective yet complex method of providing proprioceptive feedback is through the use of exoskeletons. A full-body exoskeleton for teleoperation is proposed by Ishiguro et al. [135], with a particular emphasis on locomotion and arm movements. Electromotors with encoders in the joints are used not only to measure and transmit joint positions, but also to provide force feedback from the remote-controlled robot. Other works in which exoskeletons are used in virtual environments for feedback are [147, 148, 149, 150].

5.2 Teleoperation Approach

This chapter describes our teleoperation approach as shown in the process graph in Figure 5.1. The user's arm and hand movements are tracked, the target poses for the arms and grippers are determined, converted into movements, and then sent to the robot's corresponding controllers. If the user is in a extended environment, the robot is visualized accordingly and, if available, tactile information is provided. Figure 5.2 shows a user teleoperating the PR2's left arm wearing a HoloLens 2 head mounted display which tracks the hand and visualizes the tactile data of the robot. In the following sections, we take a closer look at the individual parts and describe the connections between them in order to successfully implement the teleoperation.

5.2.1 Pose Tracking

In our approach, we control the 6-DoF end effector pose of robot manipulators. In order to transfer the user's movement to the robot, a tracking system that can both capture the movement of the arms and provide a way to interact with the grippers is needed. Possible commercial solutions include the use of data gloves, external visual tracking systems such as PhaseSpace Motion Capture [W10] or OptiTrack tracking system [W11]. Some HMDs, such as the HoloLens 2 [W12] or Meta Quest 2 [W13], also offer solutions for tracking the user's hands to interact with the virtual environment. In the case of the Quest 2, controllers can also be used as an alternative to hand tracking. The presented approach offers both, controller control and hand tracking for teleoperation.

Hand Tracking

For both, the HoloLens 2 and the Meta Quest 2, there are toolkits available that provide hand tracking. In the case of the HoloLens, the relevant toolkit is the Mixed Reality Toolkit [W14]. For the Quest it is the Meta XR All-in-One SDK (UPM) [W15], whereby in our experiments we use the predecessor Oculus Integration SDK [W16]. Both SDKs provide between 24 and 26 joints per hand, which are largely similar. The joints of the HoloLens tracking can be observed in Figure 5.3a. To guarantee smooth and natural teleoperation, none of the frames proved to be suitable. The finger joints move and rotate too much to use them for controlling the arm.

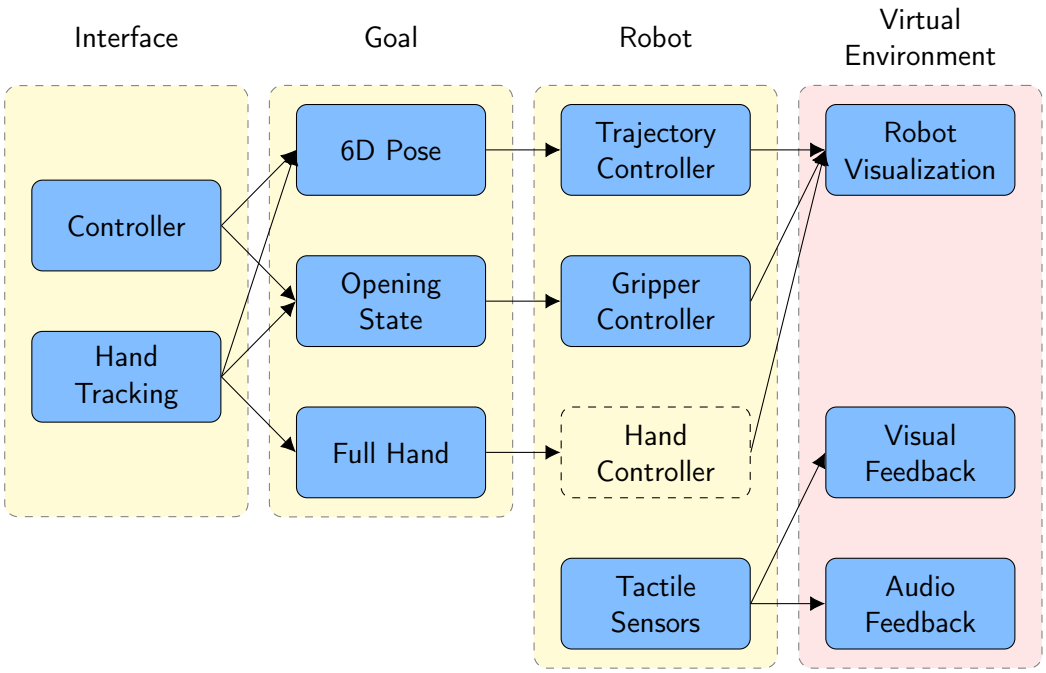
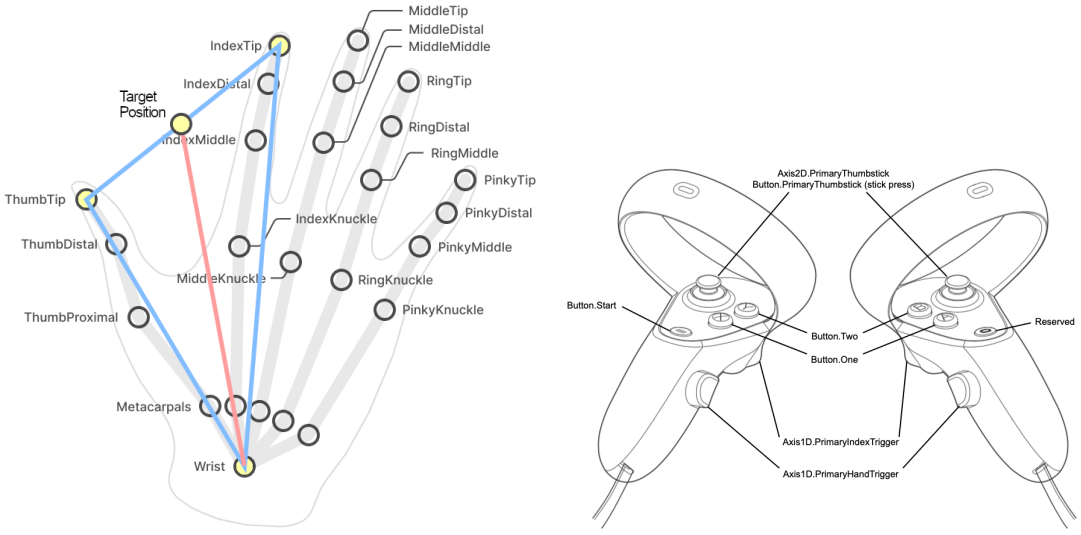


Figure 5.1: The teleoperation process graph from hand tracking to arm movements and control of simple parallel grippers. An interface for possible control of a 5-finger robotic hand is also implemented.

Figure 5.2: On the left side of the graphic, the user remotely controls the robot's left arm and performs a stacking task. The position of the right hand and the distance between the thumb and index finger are transmitted to the gripper. The right side of the image shows the user's view through the HoloLens 2, with the ball providing information on the target position and the arrows providing feedback on the tactile values of the fingertips.

The wrist frame offers greater stability, although it also feels unnatural during teleoperation, independent of the SDK. In order to enable stable teleoperation, it is necessary to use multiple frames of the hand. As shown in Figure 5.3a, a triangle is constructed between the fingertips of the index finger and thumb and the wrist. The position that is transmitted to the tool frame of the end effector is centered between the fingertips, while the orientation results from the straight line between this center point and the wrist for the forward direction and the line between the tips for the rotation. The distance between the fingers is also utilized to open and close the gripper. Additionally, the system provides the transfer of the full hand's state to 5-finger hands. However, this is not implemented in the present work. To start the teleoperation, the remaining three fingers (middle, ring, and little) must be closed. Conversely, to terminate the teleoperation, these fingers are opened again.



(a) (b)

Figure 5.3: On the left (a) the 25 joints tracked by HoloLens 2 are shown [W17]. The triangle between *IndexTip*, *ThumbTip* and *Wrist* is used to determine the position and orientation of the target pose for teleoperation. The right side (b) shows the Quest controller for the teleoperation controller mode [W18]. The teleoperation frame is located in the center of the button area.

Controller

The advantages of controller control are that the tracking of position and orientation are more accurate for fine manipulation tasks. Furthermore, the tracked pose is much more reliable and stable and does not require further stabilization as with hand tracking. In the case of the Meta Quest 2 [W18], infrared LEDs in the controller ring are recognized by the four cameras in the

HMD and the position and orientation are calculated. As long as these are not occluded, the tracking works quite accurately. To activate the arm teleoperation, the button designated as A, or button one, is employed, whereby the arm only moves when the button is pressed. Furthermore, the primary index trigger is used to to open and close the parallel gripper (see Figure 5.3b). A further feature, which was only introduced after the user study, is a reduction in the speed of the arm when the primary hand trigger is pressed. In this mode, it is easier to perform fine manipulations, as the robot arm only moves and rotates at half the original speed. More sophisticated robot hands can only be used to a limited extent in controller mode. In the case of a five-finger hand, only a pinch grasp with index finger and thumb or power grasps are used.

5.2.2 Robot

In this work, we focus on the teleoperation of robotic arms, regardless of whether they are mounted on a mobile platform or stationary ones. It is recommended that the robotic arms have at least 6-DoF, as this allows the independent control of position and orientation of the end effector in Euclidean space and can therefore map all dimensions. Our approach has been integrated and tested in both simulation and on real robots. Figure 5.4 depicts the robots employed in this work: a UR5 equipped with a 3-Finger Robotiq gripper on top and a two-armed PR2. Both robots were tested in simulation (on the left) and on the real hardware (on the right). As all interfaces are independent of the actual robot, it is straightforward to exchange the platform. The prerequisites are that the robot is operated with ROS, that there is a trajectory controller that moves the arm, and a gripper controller that manages the gripper (see Figure 5.1 Robot). More sophisticated grippers, such as the Shadow Hand, can be abstracted as parallel grippers. In a two arm setup, both arms can be controlled at the same time, or individually (see Figure 5.5).

5.2.3 Extended Environment

In our experimental setup, the extended environment serves as the counterpart to the robot. In the course of our experiment and user study, MR is used with the HoloLens 2 [W12]. However, the system is also integrated and tested on the Meta Quest 2 [W13]. With regard to the development with HMDs, the game engine Unity3D [W4] is used, which is compatible with both devices. Furthermore, there are also SDKs for hand tracking. It is noteworthy that the software is capable of supporting a multitude of Virtual and Augmented Reality glasses, and can therefore be easily integrated. Since this part is important for our experiment, but not essential for the teleoperation approach, it is colored red in Figure 5.1, which should indicate that it is optional. As long as there is another interface for hand or controller tracking.

Bridge between virtual environment and robot

As robot and VR device operate in different environments, a bridge between ROS and Unity is necessary. Two prominent systems exist for this purpose: ROS# [W5], which provides a number

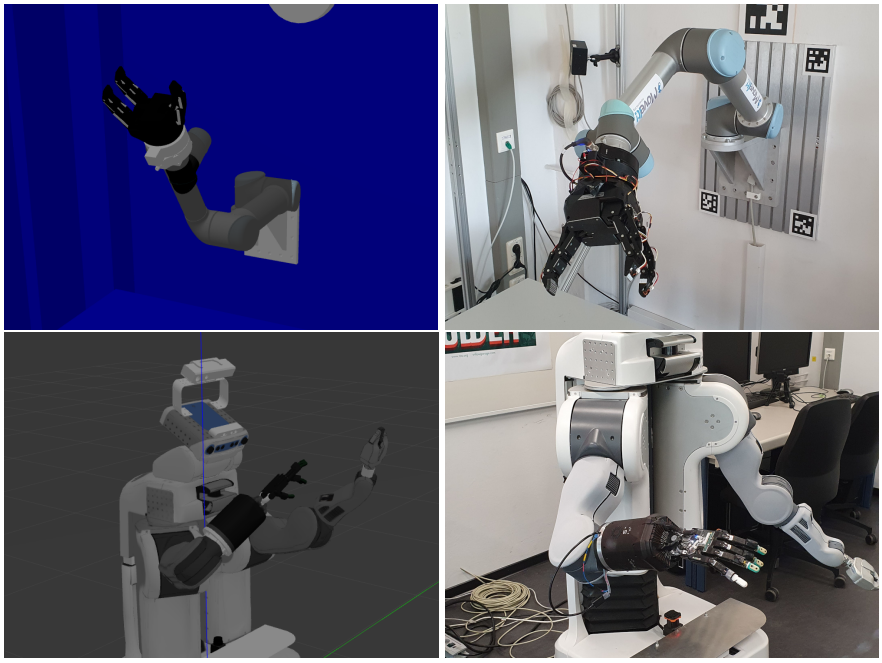


Figure 5.4: The presented teleoperation system is platform independent and has been tested, among others, on a UR5 with a Robotiq gripper (top row) and both arms of a PR2 (bottom row). The system is suitable for both simulation (left column) and real robot systems (right column).

of libraries and tools for the communication between ROS and .NET applications. On the other hand, Unity itself offers a few more slimmer packages for communication [W6]. Furthermore, both packages also work with HMDs. ROS# was developed slightly earlier and thus employed in our user study. After Unity introduced their software, we tested it and switched because it is not only slimmer, but also have a better performance due to the direct TCP communication. Both versions provide the option of importing robot URDFs, ROS messages, services, and actions, thereby making the development with both environments easier.

5.2.4 Registration

In the context of Virtual Reality, where there is no interaction with the physical elements of the real world, it is not necessary to know the location of the user or the HMD within the room. But, in the case of MR, in which virtual objects are blended into the real world, the device must perceive the real world in order to display these objects realistically. To achieve this, the HoloLens uses depth sensors to perceive the surrounding environment and determine its position within it. One step further, it is necessary to know where objects in the real world are located in order to integrate them into the virtual world, consequently, both worlds need to be aligned. In our case, we want to display visual markers on the robot, therefore, it is essential to know

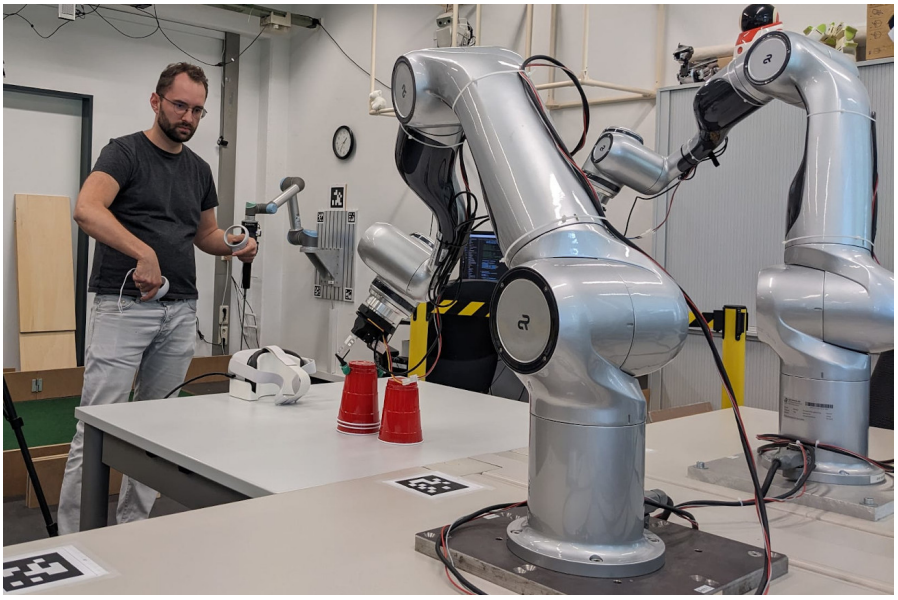


Figure 5.5: The teleoperation approach in a dual arm scenario with the Meta Quest 2.

the robot's location. To address the issue of registration, one potential solution is to place an anchor in the environment that can be detected by both the HMD and the robot. This anchor can then be used to determine the spatial relationship between the two worlds. An AprilTag [113] is employed as the anchor, which is fixed within the room and is identified by both the glasses and the robot. AprilTags are fiducial markers whose position and orientation can be determined by conventional computer vision. This enables the calculation of the transformation from robot to marker and from marker to glasses, thus also determining the transformation from robot to glasses. The accuracy of the tags at the size and distance we use was determined by Kalaitzakis et al. [151] to be 2 cm, which is sufficient for our scenario. However, since the calculation of the transform on the HoloLens significantly limited the performance of the glasses at the time of the experiment, the two worlds are only synchronized with each other at the beginning and not continuously.

5.2.5 Jogging

Jogging a robot means that the robot moves a short distance and rotation. By repeating this movement, the robot can be guided to specific positions and orientations. Most industrial robots offer this movement option on their control panels to move the individual joints of the robot either in joint space or the end effector position in Euclidean space. The jog_control [W19] package was taken as the basis for our jogging approach and extended so that absolute poses could be approached in space, rather than jogging or rotating the end effector in specific directions. The robot is given a 6-DoF pose, which should be reached with the end effector. When

using a VR device for the teleoperation, the target position is indicated with a small semitransparent sphere (see Figure 5.2). A linear movement of the end effector is calculated by means of interpolation between the current pose of the robot and the target pose. The inverse kinematic of these points is then calculated with the help of BioIK [152]. As this type of teleoperation only controls the pose of the end effector, it is possible to specify constraints such as the position of the elbow. For example, in scenarios involving two arms, the elbow should, if possible, behave in such a way that it does not point in the direction of the other arm. The position of the hand is mapped in a one-to-one relationship with that of the robot, such that a movement of 10 cm by the hand is reflected in a corresponding movement of 10 cm by the robot. In order to prevent the robot from making any unforeseen movements and to allow the user to readjust their hands, the position of the hand at which the teleoperation was started is taken as the reference point. If the user interrupts the control, this reference point is reset as soon as the teleoperation is started again.

5.2.6 Tactile sensing

To communicate tactile information to the user, tactile sensors on the robot fingers are necessary. In the following user study, a PR2 with tactile arrays on each of its parallel grippers is used. With 22 taxels, 15 of which are on the inner contact surface, the applied force is measured. Therefor, the work of Romano et al. [153] is used to estimate the normal forces. These forces can be visualized via a HMD, or acoustically via loudspeakers.

5.3 User Study

The objective of the user study is to evaluate the usability and performance of the system, as well as the use of visual and acoustic substitutions of tactile data, both uni- and multimodal. Given the circumstances of the ongoing covid19 pandemic, the study was conducted as a pilot with a limited number of participants, due to the restrictions on access to the laboratory facilities. The participants were asked to perform a simple manipulation task under different conditions. With this experiment, following questions should be answered:

1. Does the communication of tactile feedback through visual and auditory cues to the user improve the quantitative performance of a teleoperation task?
2. Is the new MR robot setup applicable for teleoperation tasks regarding the usability for non-expert users?

5.3.1 Experiment Setup

The teleoperation approach presented is independent of the input device and also of the remote-controlled robot. The setup configuration used in this study is pictured in Figure 5.2. In the user

study, participants were prompted to remotely control the left arm of the PR2, as this component is both robust and not too complex. The mobile platform and height-adjustable torso enabled the robot to be placed in the center of the room, thereby facilitating greater flexibility for users in moving around the robot to optimize their positioning for teleoperation. Furthermore, the robot is equipped with tactile arrays on the fingertips, the functionality of which has been analyzed in other works. The HoloLens 2 was selected as the tracking device, as it enables users to receive immediate feedback from the actual robot through MR, allowing them to observe its behavior. Additionally, there are pre-existing solutions for hand tracking for this device, as detailed in Section 5.2.1. A table is set up in front of the robot, which is also used for registration by markers on top of the table (see Section 5.2.4). The cups are placed at the edge of the table to facilitate gripping from the side without collision.

5.3.2 Tasks

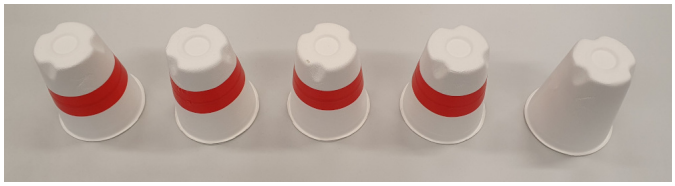
To evaluate the system, the candidates are given the task of stacking five cups with the remote-controlled robot. In Figure 5.6a, these five cups are lined up next to each other; they are to be picked up one after the other in the red area by the robot's left gripper and placed on the right unmarked cup. As little force as possible should be used to avoid destroying the cups. Furthermore, the participants were also instructed to be as fast and precise as possible. In order not to make the task too challenging, the cups were always placed in a row next to each other so that they could be easily grasped from the side. With this scenario we can measure whether different amounts of force are used with the different conditions and how fast the participants are. A cup is considered as lost as soon as it slips out of the gripper and falls over. If the cup subsequently lands in the same initial position, it can be grasped once more. Once all four marked cups have either been stacked or fallen over, the task is considered complete, and the next condition is set up. The three cups on the right side of Figure 5.6b are used in the tasks. They have different degrees of stability and size, in the experiment, we want to test whether these properties have an influence on the performance. The hardness of the cups decreases from left to right.

5.3.3 Conditions

In all conditions, the position of the jogging target is indicated by a semi-transparent sphere on the wrist. Furthermore, the tactile data from the gripper sensors is visualized or displayed in different ways:

Color (C)

The ball used to indicate the target position changes its color depending on the force applied. Prior to contact, the sphere is white (0 N). As soon as a contact is made, the color changes proportionally to green until a force of 5 N is reached. At this point, the color transitions to



(a)

(b)

Figure 5.6: The top picture (a) shows the user study task, where participants are asked to pick up the cups in the red area and place them one by one on the unmarked cup on the right. They should do this as quickly as possible without destroying the cups. The different cups are shown at the bottom (b). The left cup is only used in the warm-up phase, the other three have different strengths to measure their influence on performance.

yellow with a full saturation at 7.5 N and subsequently changes to red at 10 N. The color does not change beyond this point. See Figure 5.7a.

Size (S)

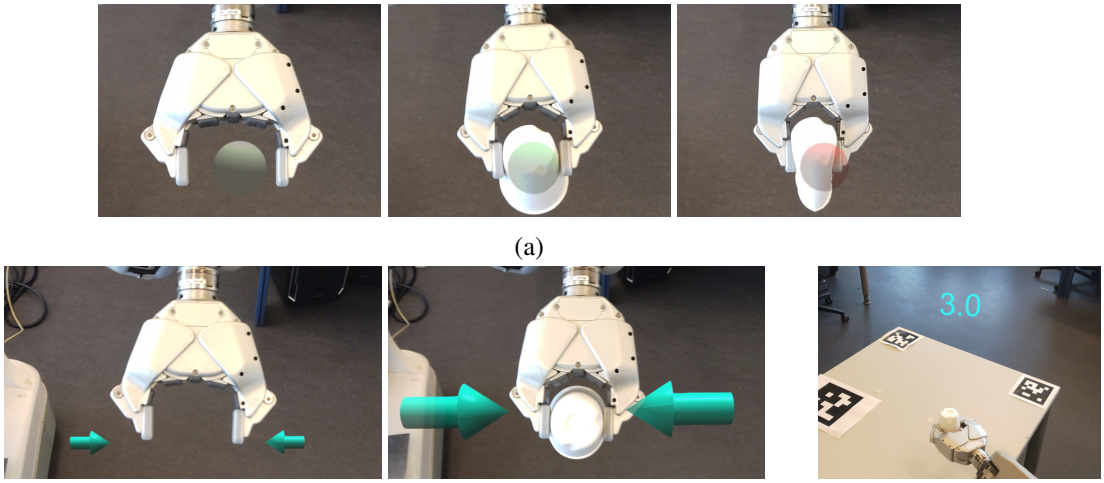
Arrows are visualized next to the fingers to represent the measured force with a size-changing indicator (see Figure 5.7b). In the absence of contact, these arrows have a small size, as illustrated in the left image of the graphic. As soon as force is applied, these arrows expand in proportion to the magnitude of the force, reaching their maximum size at 10 N. To ensure comparability across conditions, both arrows are presented with the same size, which corresponds to the average force value.

Text (T)

The measured force is visualized in the form of a text above the gripper (see Figure 5.7c). This text corresponds to the applied force in Newtons, again with a maximum value of 10 N.

Audio (A)

In order to provide the user with tactile feedback in the form of audio signals, a sine tone is played via the internal loudspeakers of the HMD as soon as contact is made. This sine tone begins at a frequency of 200 Hz (>0 N) and is increased to 600 Hz in proportion to the maximum



(a) (b) (c)

Figure 5.7: The three virtual replacements of tactile feedback. (a) shows the measured force as different colors of the position ball, (b) changes the size of the arrows, and (c) shows the measured force as text above the gripper.

force of 10 N. As soon as more force is applied, the 600 Hz is played as a beep, similar to a variometer.

5.3.4 Measurements

As the system mainly aims at a good user experience, the subjective perception of the participants as well as the objective performance of the candidates during the teleoperation is measured. To evaluate the objective factors, the following variables were measured: execution time, average applied force, and success rate. The time taken for the execution time was recorded, beginning with the first active movement and concluding with either the stacking of the last cup or the cup fallen over. The average force was calculated from all tactile measurements that detected contact on both tactile arrays. The success rate was determined by the number of cups that were successfully stacked, with a maximum value of 4.

In order to evaluate users' subjective perceptions, they were asked to complete various questionnaires. One is the NASA Task Load Index (NASA TLX) [154], which can be used to assess the perceived workload. To evaluate the usability of the approach, the System Usability Scale (SUS) [155] was employed, which is an established method for quantitatively analyzing usability. Simulator sickness is a common issue in XR experiments, the Simulator Sickness Questionnaire (SSQ) [156] was utilized to assess this problem. Additionally, we employ the AttrackDiff2 [157] Questionnaire to measure the perceived Hedonic and Pragmatic Quality. A custom questionnaire to enquire about the various substitutions was used. These are the questions and statements with the possible answers in brackets:

- Which condition did you like best? (8 conditions)
- Which condition was the easiest for you? (8 conditions)
- The arrows were distracting. (5-point Likert scale)
- The sphere was distracting. (5-point Likert scale)
- The number was distracting. (5-point Likert scale)
- The sound was distracting. (5-point Likert scale)
- Which visualization was more helpful? (5-point Likert scale, three questions with the combinations of the visual substitutions)
- The sound was helpful. (5-point Likert scale)
- The information from the tactile sensors helped me a lot. (5-point Likert scale)
- Which kind of cup was the easiest for you? (Three cups and no difference)

5.3.5 Conduct

Participants

The study was conducted during the ongoing corona pandemic, which resulted in restricted access to offices and laboratories. Consequently, only a pilot study with five participants was feasible. The participants ranged in age from 23 to 57 years (M = 34.4, SD = 13.18) and were employees of the working group. All participants had normal or corrected-to-normal eyesight and none had equilibrium issues. Of the five participants, three had prior experience with AR headsets and three had experience with hand tracking.

Procedure

Before the participants took part in the study, all devices were disinfected and cleaned, the specified safety distance was maintained, and the participants and the supervisor wore masks throughout the experiment. At the beginning of the experiment, the participants were informed of the experiment's structure and provided with an overview of the experimental setup and the tasks they would be performing. Subsequently, the participants completed the initial questionnaire, comprising a demographic section and the first part of the SSQ, on a computer provided for this purpose. Prior to completing the actual task, participants were given a brief period of familiarization. This allowed them to get used to the Head-Mounted Display and familiarize themselves with the robot's behavior during teleoperation. They were informed about the interaction possibilities. Using a cup that was not provided in the experiment (Red cup in Figure 5.6b), the participants were able to practice grasping and lifting objects without receiving

any feedback on the tactile data. Following the warm-up phase, the participants were presented with the actual experiment, which was divided into three phases, with a short break allowed between each phase. The combination of eight conditions and three cups meant that the participants were required to stack the cups eight times in each phase. The sequence was randomized to ensure that neither the cups nor the conditions remained the same from one trial to the next. Following the completion of the three phases, participants were asked to complete the remaining questionnaires. In total, the participants wore the HMD for approximately 45 minutes; including questionnaires and preparation, one session lasted around 60 minutes.

5.4 Results

5.4.1 Quantitative Results

The average execution time and success rate are presented in Table 5.1. The values obtained from all participants are then aggregated and averaged across the conditions. The results demonstrate that the candidates were notably slower when presented with the colored ball or the arrows, both in uni- and multimodal in conjunction with audio feedback. The recorded times ranged from 120.2 seconds to 127.2 seconds. The fastest times were 104.6 seconds in the absence of any feedback and 102.7 seconds when the audio were combined with the text over the gripper. With only text and only audio, the test subjects completed the task in 112.2 and 111.2 seconds, respectively. The mean execution time for all conditions and all cups is 118.2 seconds, with cup 1 taking 118.2 seconds, cup 2 taking 115.2 seconds, and cup 3 taking 112.9 seconds.

With regard to the success rate, there were no significant differences; the participants performed best when they were presented with the arrows in combination with audio, with a success rate of 91.6 %. Conversely, the lowest success rate was observed with the colored ball, at 83.3 %. On average, participants successfully stacked 87.1 % of the cups. The different cups had no influence on the success rate.

Figure 5.8 illustrates the average force applied across all cups. It is evident that a greater force is applied without any feedback on the tactile readings. The remaining conditions have a minimal variation, regardless of the type of feedback employed, whether visual, acoustic or multimodal combinations. Figure 5.9 shows the average force applied to the three cups. It can be observed that an average of 4.5 N was applied to cup 1, 5.1 N to cup 2 and 3.0 N to cup 3, showing a notable difference.

Table 5.1: Average Execution Time & Success Rate

	No	C	S	T	A	C+A	S+A	T+A
time	104.6 s	120.2 s	127.2 s	112.2 s	111.2 s	120.8 s	125.7 s	**102.7 s**
rate	88.3 %	83.3 %	85 %	88.3 %	86.6 %	85 %	**91.6** %	88.3 %

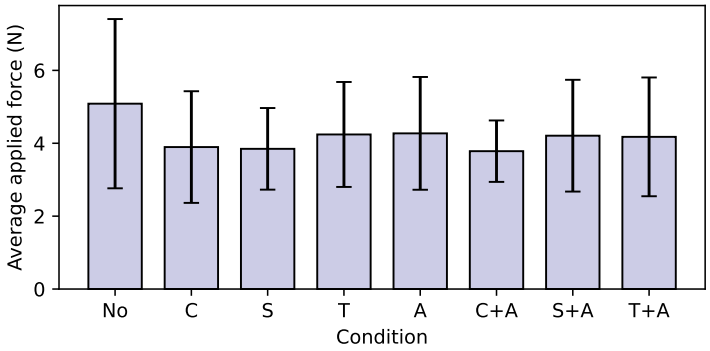


Figure 5.8: The bars show the average force applied by all users during the experiment. Where *No* is no feedback on the tactile data, *C* is the colored sphere, *S* is the resized arrows, *T* is the text above the gripper, and *A* is the audio feedback, plus the results of the three multimodal combinations.

5.4.2 Qualitative Results

Table 5.2 presents the findings of the NASA TLX questionnaire. The score is expressed on a scale of 0 to 100, with 100 representing the highest level of load. The result is divided into six distinct scales, as illustrated in the table. Notably, the values for overall performance and frustration level are particularly low. The overall workload is 56.1. The simulator sickness questionnaire showed that the value before the experiment was 153.38 ($SD = 85.8$) and 228.17 ($SD = 171.62$) after the experiment. This increase was to be expected. Furthermore, the SUS mean score was 77 ($SD = 13.04$), which is above the average score of 69.5 [158]. Figure 5.10 provides a summary of the results from the AttrakDiff2 questionnaire. This questionnaire is based on a model that divides attractiveness into two main components: Pragmatic Quality (PQ) and Hedonic Quality (HQ). The PQ dimension indicates how well a product supports the user in accomplishing tasks, while the HQ dimension captures the extent to which a product appeals to or resonates with the user on an identity level. Participants gave positive ratings for the system's attractiveness across both dimensions.

The evaluation of the custom questionnaire indicate that the participants prefer multimodal feedback to individual substitutions. The combination of arrows and audio (*S+A*) was the most preferred option, with three participants indicating a preference for this over the color-changing ball and audio (*C+A*), which was favored by two participants. This was also the outcome of the comparison between the visual feedback methods. These were compared individually on a 5-point Likert scale. The responses were transformed into a scoring system, whereby the visualizations were assigned no points if the vote was neutral and 1 or 2 points depending on how strongly the tendency was towards this method. In total, the color (*C*) feedback received 14 points, the arrows (*S*) 10, and the text (*T*) 1 point. This confirms the preference for color and

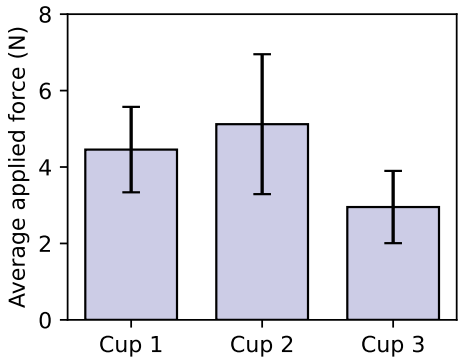


Figure 5.9: The bars show the average force applied by all users to the different cups during the experiment. Where Cup 1 is the white, most rigid cup, Cup 2 the brown, and Cup 3 is the softest blue one.

size. In response to the question of which feedback was most helpful, two individuals selected *S+A*, two selected *C+A*, and one selected *S*. Four people indicated that the substitutions were not particularly distracting, with a mean value of 1.2 on a scale of 0 to 4. One subject indicated a high value, which suggests the presence of personal preferences. The result to the question of whether the tactile feedback was particularly helpful in the task was mixed with a mean value of 2.4 on a 5-point Likert scale. With the last question we asked about the different cups, three people said that cup 1 was the easiest to handle, two said that it was cup 2.

5.5 Discussion

The objective of the first research question (see Section 5.3) is to answer whether providing information regarding the tactile data of the robot gripper results in quantifiable performance enhancements of the teleoperation. The performance of the system is evaluated based on three

Table 5.2: NASA TLX scores.

	M	*SD*
Mental Demand	56.7	25.28
Physical Demand	53.4	13.94
Temporal Demand	53.4	18.26
Overall Performance	26.67	9.13
Effort	60.0	19.00
Frustration Level	40.0	22.36
Overall Workload	56.1	11.35

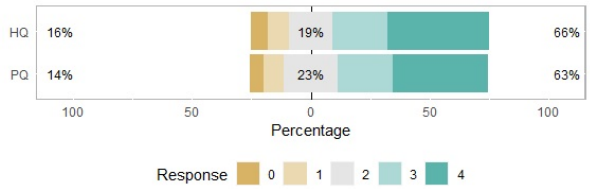


Figure 5.10: The graph shows the result of the AttrakDiff2 questionnaire, with the top graph showing the percentages of Hedonic Quality (HQ) and the bottom graph showing Pragmatic Quality (PQ) on a 5-point Likert scale.

key metrics: execution time, force applied to the object, and success rate. The time taken by the candidates to complete the task was comparatively longer when they were required to use either the arrows (*S*) or the colored ball (*C*) as a condition. This can be attributed to the phenomenon of perceptual overload, whereby too much information can cause an overload and therefore a reduction in the execution time [159], as both conditions were visualized nearby the robotic gripper and therefore in the visual region of interest. Users indicated that the text was not a favorite condition and that they probably paid less attention to it. Consequently, they were better able to concentrate on speed under this condition. It can be inferred that users were also less distracted by audio feedback and were therefore better able to concentrate on the task. A comparison of the force exerted in each condition reveals that less force was applied to the objects when the users had feedback on the tactile data, indicating that the tactile feedback was beneficial. However, which feedback condition did not have a significant impact on the results. One potential reason for this is that the task was too challenging, and the force exerted was not the primary focus. Additionally, no notable differences were observed in the success rate. With the different cups, the influence of the hardness of objects on the teleoperation was measured. Although the participants indicated that the softest cup was the most difficult to handle, it required the least force. This can be attributed to the fact that the participants had to be more careful, otherwise the cup would be crushed. The cups had no significant influence on the other measurements relating to success rate and execution time.

Regarding the qualitative evaluation of our study, the questionnaires showed that our system is both user-friendly and accepted by the users and is considered suitable for teleoperation tasks. This is supported by the above average SUS and AttrackDiff2 results. Users indicated that they preferred the visual feedback close to the gripper, even though the qualitative results showed that performance was worse with this feedback. If the focus of teleoperation is purely on performance, such distracting visualizations should be avoided in favor of an unused modality, such as audio in this case. If the focus is purely on user experience and usability, visual feedback is a good option, and, if possible in a (redundant) multi- over unimodal way. Overall, this pilot study does not provide precise results, but it does show first important trends, as all users were very satisfied with the system and stated that the teleoperation was easy and intuitive to use,

especially as all participants were inexperienced in this area and had little or no experience with teleoperation or virtual environments. These results also answer our second research question that our MR setup is suitable for teleoperation even if the users are inexperienced.

5.6 Conclusion

In this work, we have shown that tactile data can be used to support humans in teleoperation tasks, improve performance, and enhance user experience and usability. Therefore the third fundamental question (**FQ3**) could be positively answered, as the system show a positive influence using tactile feedback for augmented perception. The presented system is independent of both the input device and the robot to be controlled and let users remotely control robots in real-time. The hardware can be exchanged rapidly and effortlessly, as long as it meets the criteria for supporting ROS on the robot side and Unity on the XR display. The pose of the robot end effector is controlled by the pose of the user's hand or a controller. The same applies to the gripper of the robot.

In a pilot user study, we demonstrate, using the example of a 7-DoF robot arm on the PR2 platform, that the system can be operated by non-expert users to perform precise manipulation tasks, as evidenced by a success rate of 87.1 %. The objective was to demonstrate that the performance of the system and its usability could be enhanced by the presentation of tactile data. The results indicate that there is an improvement in the quantitative performance in certain areas, such as the reduction of the average force applied. However, there is also a deterioration in performance due to the provision of too much information, which primarily affects the execution time. Nevertheless, both unimodal and multimodal feedback have been shown to enhance the usability and user experience, whereby the users preferred multimodal feedback.

To enhance the functionality of our system, the control of 5-finger hands can be integrated in future work. The majority of current XR devices offer robust 5-finger tracking capabilities. With the appropriate mapping, the state of users hands can then be transferred to more sophisticated robotic hands. Furthermore, the system presents the potential for autonomous, side-by-side collaboration between humans and machines, including temporary teleoperation and autonomous behavior of the robot.

6 Haptic Feedback

In the final experiment in this thesis, we intend to use the robot to provide haptic feedback and to answer the last fundamental question (**FQ4**). Up to this point, the attention has been primarily directed towards the tactile sensors and the robot itself. However, in this chapter, the focus will shift towards the human aspect, with the aim of creating a comprehensive and engaging tactile and haptic experience. We already dealt with Extended Reality (XR) in this thesis, and this chapter will also deal with the combination of robots and, in this case, Virtual Reality (VR). The objective is to enhance the sense of immersion in VR with a device designed and constructed for this purpose, thereby creating an even more realistic impression of the virtual environment.

An example of haptic interaction in VR and AR scenarios, which is also frequently used in academic studies, is the act of pressing a virtual button [160]. For commercially available XR devices like the Meta Quest 2, when a button is pressed, the user receives audio feedback in the form of a clicking sound or, in the case of controller usage, in the form of additional vibrations, which are less reminiscent of pressing a button. Additionally, the button also changes visually, so that the user's virtual hand presses the button down. There is a lot of research into improving this type of feedback in human-computer interaction scenarios, with the aim of improving the quality of the VR experience (further presented in Section 6.1). Approaches that provide the user with haptic feedback can be broadly classified under the term Encountered-Type Haptic Display (ETHD). The device presents a surface to the user to encounter within a specific workspace. A tracking system is utilized to capture human movements, enabling the precise positioning of haptic surfaces at appropriate locations. By the use of virtual environments, the presence of the haptic device can be obscured, enabling the illusion of realistically touching surfaces or objects in VR. Mercado, Marchal, and Lécuyer [13, p.2] defines ETHDs as follows:

> *"In the context of human interaction with a virtual or remote environment, an Encountered-Type Haptic Display is a device capable of placing a part of itself or in its entirety in an encountered location that allows the user to have the sensation of voluntarily eliciting haptic feedback with that environment at a proper time and location."*

This definition also includes wearable devices, in which the to encountered surface is not in constant contact with the user. The limitations of these devices, with the exception of wearable technology, are that they only provide feedback in form of a single point of contact [161]. Consequently, more complex shapes can only be explored to a limited extent or not at all.

We want to address this issue by presenting the user with physical 3D models of the virtual objects. Therefore we design and construct an inconspicuous table, on which objects can be

moved and rotated with a plotter-like mechanism beneath the tabletop. The characteristics of the haptic display will be evaluated through a series of experiments. Furthermore, the usability and user experience will be investigated in a user study.

The main contribution of this work can be summarized as follows:

- Design and construction of an Encountered-Type Haptic Display for human-computer interaction scenarios in Virtual and Augmented Reality with a three-dimensional movement (x, y, and rotation) of arbitrary objects on a tabletop
- Utilization of Hall effect sensors for wireless position and interaction detection of objects on the table surface

The remainder of this chapter is structured as follows. Section 6.1 provides an overview of existing related work on ETHDs and magnetic control. The subsequent chapter outlines the design of our table. Section 6.3 evaluates the technical possibilities and Section 6.4 the subjective quality of the device in a user study. Finally, a summary of the results is presented and suggestions for potential ideas for future research.

6.1 Related Work

This section will summarize two related topics. The first will investigate approaches to integrate haptic feedback in the form of ETHDs. The second will analyze the literature that uses comparable hardware approaches to ours, namely magnetic control.

6.1.1 Encountered-Type Haptic Displays

ETHDs can be categorized in different ways, depending on their underlying technology in *grounded* and *ungrounded* devices. *Grounded* displays have a fixed workspace as the robotic devices are firmly positioned on the ground, like robotic arms or fixed platforms, as the device presented in this chapter.

Mercado, Marchai, and Lécuyer [162] employ a UR5 robot equipped with various end effectors in conjunction with an HTC Vive HMD. They use the system to create the illusion of a large flat surface. Usually a large surface is achieved by interrupting the contact and displacing the end effector. The authors investigate the extent to which a constant contact with the surface is achievable during interaction. To prevent users from reaching into empty space and to provide cues for interaction, a circle with a diameter of 5 cm was visualized indicating the contact area. Its position and color indicated when and where a contact with the surface was possible. During an experiment, five different interaction techniques to move the interaction area where tested in which three had intermitted contacts and two continuous contacts. Furthermore, the participants were asked to color three shapes on the surface with their index fingertip. The study demonstrated that both, user experience and performance for interaction techniques involving

interruptions were better. Reasons for this included easier control and inadvertent interactions during the coloring task. More examples of robotic arm approaches can be found in [163, 164, 165, 166, 167, 168, 169, 170, 171].

Furumoto et al. [172] presented the concept of a midair balloon interface, which is an example for a fixed platform. They create acoustic radiation forces using Airborne Ultrasound Phased Arrays (AUPA) [173] to exert forces on solid objects. For their setup, they positioned 11 AUPA devices spherically with a diameter of 528 mm, allowing interaction access to the balloon while still enabling efficient three-dimensional movements. In [174] they showed, that the device can be used in virtual environments with a prototype scenario. More devices are presented in [175, 176, 177, 178, 179, 180, 181].

In contrast, *ungrounded* devices are not fixed to a specific location but can move within space. Examples of these include mobile platforms, Unmanned Aerial Vehicles (UAV), or wearable devices.

A common approach for haptic feedback with mobile devices is by attaching adjustable haptic surfaces to mobile robots [182, 183]. This provides the opportunity to utilize multiple robots to offer feedback at more than one contact point and even use swarm intelligence. Some research has also led to the development of displays that can alter their shape, such as pin arrays, in order to exhibit different shapes [184, 185]. Suzuki et al. [186] presents an approach with tabletop-size mobile robots with height and orientation changeable surfaces. The system is easily scaleable to provide multi point feedback or larger areas. Furthermore, objects can be attached to the surface to present perfectly shaped feedback in virtual environments. Kim and Follmer [187] use a swarm of small wheeled robots to indicate haptic patterns to the users hands and arms. With their robots, they generate both tactile and kinesthetic feedback depending on the applied force. Possible scenarios for the swarm robots are indicating notifications, directional cues, or remote social touch. To track arm motions, they used a wristband in their experiments.

An example for UAVs is presented by Abtahi et al. [188], for their setup, they encased a quadcopter with aluminum mesh to enable safe interaction with the device. The drone can land at any location to provide dynamic passive haptic feedback in the form of a box on-demand. Furthermore, they attached different textures and physical objects to the drone to be encountered by the human at arbitrary positions mid-air. Such devices can also be used to simulate weight, surface stiffness, and lateral force [189]. More examples of Unmanned Aerial Vehicles are [190, 191, 192].

Another category of ungrounded ETHDs are wearable devices such as the device developed by Ariza Nunez et al. [160]. This device comprises a mechanical thimble that can provide haptic feedback through tapping at the fingertip and vibration. In addition, electrodes on the forearm are used to provide proprioceptive feedback through electrical tendon stimulation in order to obtain a sensation of stiffness, contact and activation. The setup is evaluated in a study wherein participants are instructed to press virtual buttons. Other works with wearable devices includes, for example [193, 194, 195].

6.1.2 Magnetic Control

Magnetic control is more commonly used in a different dimension than in our work, namely for nanoparticle manipulation [196, 197]. For an overview of this topic, see the review by Abedini-Nassab, Pouryosef Miandoab, and Şaşmaz [198]. In particular, three methods are presented with which the particles are manipulated. The first is with electromagnets or permanent magnets, which are often used to mix or separate liquids, as the hardware is too large for precise manipulation of nanoparticles [199, 200]. To precisely manipulate particles, embedded microwires or micro-coils are often used [201], which is the second method, or the third with magnetic film [202]. However, the hardware for the last two methods is significantly more complex.

Another example where magnetic control is used is in arc welding. Nomura, Morisaki, and Hirata [203] have presented an approach in which four permanent magnets are used to turn the typically round cross-section of the arc plasma into an elliptical shape which significantly improves the welding quality. This is just one example of many, as magnetic control is very popular in welding, as this review shows [204].

Mahoney and Abbott [205] attached a permanent magnet to their 6-DoF robotic arm, enabling the control of a capsule endoscope within a liquid-filled stomach. They developed a method to use this setup to control the position of the capsule in three dimensions and the orientation in two. They have distances of around 25 cm between the magnet and capsule to keep them in equilibrium. Other research that uses magnetic control for endoscopy are [206, 207, 208]

The hardware that is probably closest to ours is the commercially available Atari Pong table [W20]. It is a mechanical reproduction of the 1970s arcade game. A two-axis system is employed to move a magnet underneath a glass plate, thereby moving the ball on the top in two dimensions. Four magnets are attached to the square ball in a 2x2 matrix, the magnets are polarized crosswise to make the ball more stable. Furthermore, a layer of Polytetrafluoroethylene (PTFE) tape is applied beneath the ball. With this configuration, the ball is capable of reaching speeds of up to 1.58 m/s. In contrast to our approach, the system has two fewer Degrees of Freedom and lacks the ability to detach the object. Additionally, the system does not gather any information about the object, as there is no interaction intended.

6.2 Encountered-Type Haptic Display

6.2.1 Requirements

In a typical scenario the user is situated within a Virtual or Augmented Reality environment and interacts with the virtual objects that are present in their immediate vicinity. It should be possible for the user to make physical contact with these objects and perceive their dimensions to be accurate in relation to the virtual object, and furthermore, to explore them in a manner that is both intuitive and realistic. These physical feedback objects have to be placed by the haptic display within reach of the user and at the right time, so that the user does not reach into the void.

Furthermore, it should also be possible to present different physical objects, in the best case arbitrary of size, shape, and stiffness. The device should be designed unobtrusive in a way that it can be used for AR, ensuring the maintenance of the illusion of a realistic environment. For instance, the presence of a robot arm in front of the user would immediately disrupt the illusion. This also means that the user should not be provided with a feedback glove or exoskeleton.

Figure 6.1: The Encountered-Type Haptic Display during the *Whac-A-Mole* user study. The user is wearing a Meta Quest VR Headset and gets haptic feedback with the ETHD at the intended position. The participants were supposed to perceive the table as a normal table, which is why any motors visible on the sides were inconspicuously concealed.

6.2.2 Design

Given that a typical robotic arm, drone, exoskeleton, glove, or similar robots and devices are unsuitable due to their direct exposure to the user, we decided to modify an everyday object to a haptic display, but keeping appearance. The selected item is a table. This attracts little attention at first, but still offers the possibility of discreetly attaching a mechanism beneath the table top. As illustrated in Figure 6.1, the table appears to be a conventional table, yet it is equipped with a 3-axes system under the tabletop shown in Figure 6.2. The end effector, which is equipped with permanent magnets, is capable of moving along two axes, enabling the manipulation of attached objects along the table surface. An additional axis enables the magnets to be lowered from the tabletop, thus allowing the manipulation of different objects by detaching and attaching. Furthermore, this axis is capable of being rotated, thus enabling the object to be rotated on the other side of the surface. With this configuration, manipulations in four dimensions beneath the table and three dimensions above are achieved. As the table surface is too thick, a hole is sawn out in the center and a thinner plate is placed over it. The size of the hole also determines the

size of the workspace in which the objects can be placed. This workspace can be scaled by using a larger table and longer axes. Furthermore, the entire structure should minimally extend downwards allowing users to sit at the table without restriction. We have raised the table slightly for people with long legs. The used hardware is described in more detail in the following section, followed by the software used.

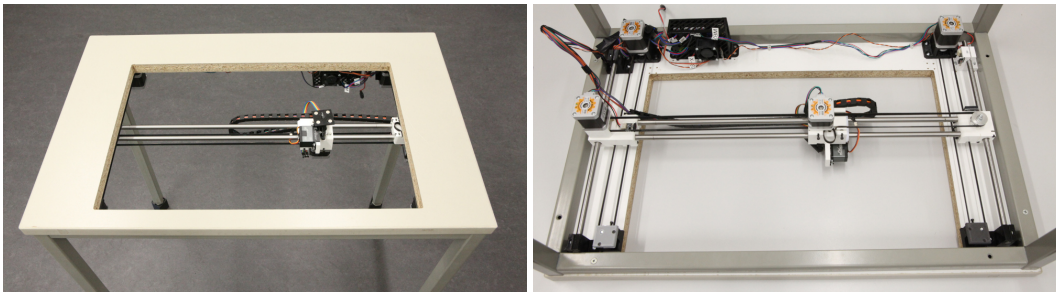


Figure 6.2: Hardware design of the haptic display. The left picture shows the top of the table without an attached surface. The hole has the maximum size possible with this table, the remaining edge is necessary to attach the plotter mechanism. In the right picture, the table is shown from below. Two motors in the top corners move the y-axis, while the motor for the x-axis moves on the y-guides. In the end effector, a fourth servo and a stepper motor are attached to control z- and yaw-axis. The controller board is attached in the upper left corner.

Hardware

Figure 6.2 illustrates the fully assembled Encountered-Type Haptic Display (ETHD) from both, an overhead and a lateral perspective, in each case without the thin surface. This table was chosen as it has no struts under the tabletop, but instead a frame along the edges. Overall its dimensions are 80 cm in length, 50 cm in width, and 74 cm in height. The table top is made of pressed wood and is 2.5 cm thick, thereby elements can simply be screwed in. For a tabletop with this thickness, the use of an electromagnet would be impractical as a magnet strong enough to create a magnetic field that overcomes this distance would be too large. Instead, a 56.3 x 36 cm hole was cut into the plate and a 5 mm thin sheet of PTFE was laid over the entire surface of the table. This is the maximum hole size we could cut in order to be able to attach all the components. The aforementioned hole size yields to a workspace of 47.2 x 26 cm, within which objects can be placed. Furthermore, this approach allows for the additional saving of 2.5 cm of downward expansion. For the y-axis, two 40 cm long metal shafts with a diameter of 1 cm were attached as far out as possible beneath the table top. The mount for these guides was custom designed and 3D printed and is shown in Figure 6.4a. For the x-axis, two mounts were printed (see Figure 6.4d and Figure 6.4e), ball bearings are attached to them to slide over the y-axis, one on each side. Between these two mountings there are two further metal guides with a diameter

of 1 cm and a length of 60 cm parallel to each other with a center distance of 3 cm between them. The custom designed end effector, shown in Figure 6.3 on the left side, runs on these two guides, once again using two ball bearings. The y-axis is driven by two motors that are fixed to the wooden tabletop. Each motor drives a belt that is attached to the x-axis mounts, which can be tensioned via a screw on the opposite pulleys. With regard to the x-axis, a motor is attached to one mount and a pulley to the other. This can be moved using a belt that runs underneath the parallel rails and is fixed to the end effector. The electronic cables are guided in a channel along this axis so that they do not hang down and do not contact the user. The custom end effector combines two axes of movement. On the one hand, a servo motor moves the axis which is orthogonal to the surface up and down, and on the other hand, this axis can rotate by a stepper motor. The designed models are shown in Figure 6.4. The missing part in this kinematic chain, where the magnets are attached, is shown in Figure 6.3 on the right side. These magnets must be lowered from the table's surface for the detachment of objects, yet they offer the potential for the rotation of objects on the table through the application of four magnetic fields. We have obtained the most effective results with the use of 8x3 mm cylindrical magnets attached to the tool, as illustrated in Figure 6.3 on the right side. Four Hall effect sensors [96], positioned at 90-degree angles to one another, have been integrated into the center of the tool. These sensors are used to measure the magnetic field of a reference magnet, which is positioned centrally in the object above the table. The data gathered from this measurement is then used to get information about the object. The three stepper motors, the servo motor, and two end stops, one for the x- and y-axis to calibrate the system, are driven by a Bigtreetech SKR mini E3 V2.0 board, with integrated motor drivers. The data from the Hall effect sensors is read out with another Arduino Nano board at 400 Hz per sensor. The hardware specification including the accuracy findings from our experiment in Section 6.3 is shown in Table 6.1.

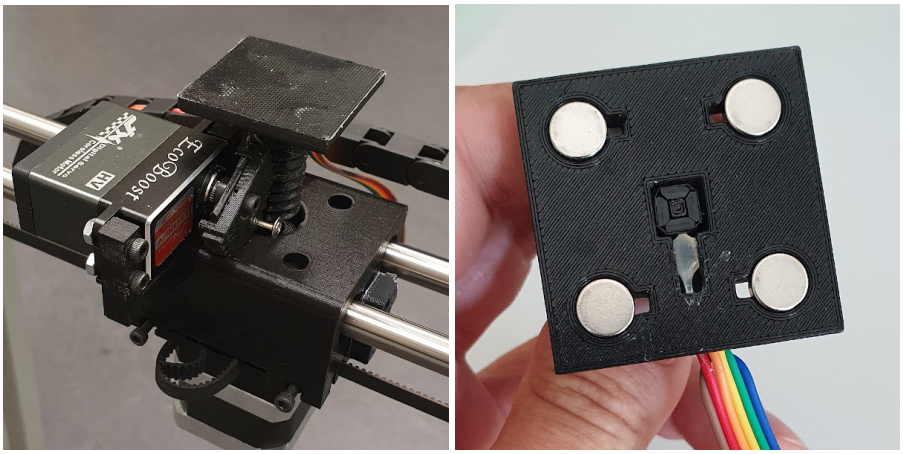
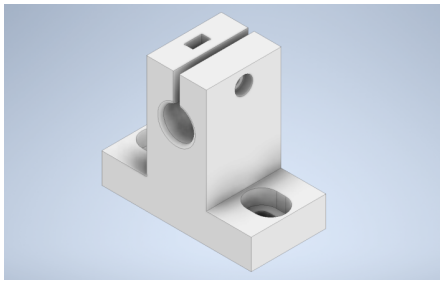


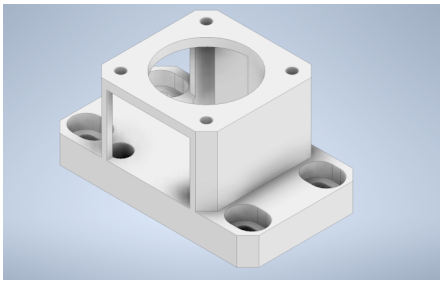
Figure 6.3: The custom build end effector of the ETHD. The left picture shows the combination of a servo motor for the yaw-rotation and a stepper motor for up and down z-movements, to attach and detach objects. To estimate the position of a reference magnet within the object on top of the table, four Hall effect sensors are placed in the center of the tool (right picture).

Table 6.1: Hardware Specification

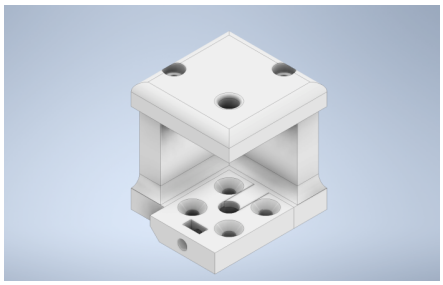
Parameter	**Specification**
Table Size (H×W×D)	74 cm×50 cm×80 cm
Workspace	47.2 cm×26 cm
Payload	210 g
Speed	66 cm/s
Object Estimation Accuracy	0.5 mm
Power Supply	12 V
Motors	4 Stepper, 1 Servo Motor
Control Type	Belt-driven
Surface Thickness	5 mm



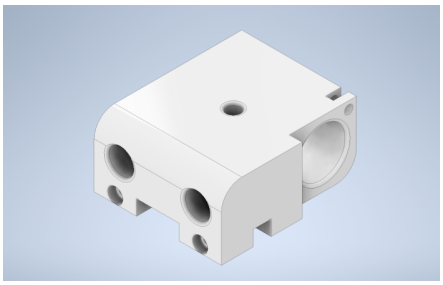
(a) Guide holder



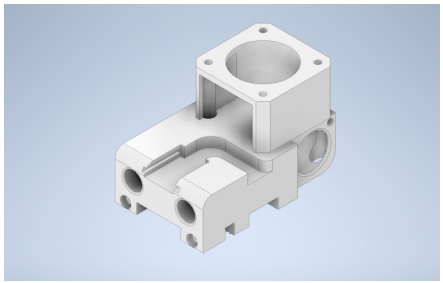
(b) Motor holder



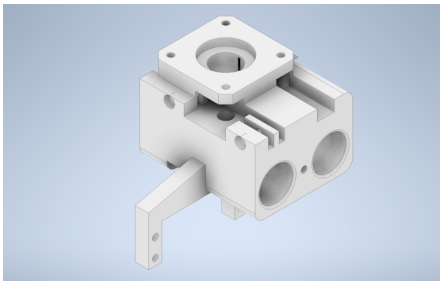
(c) Pulley holder



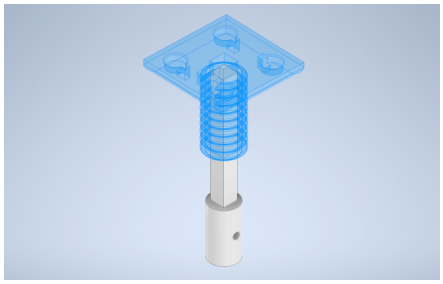
(d) x-axis holder



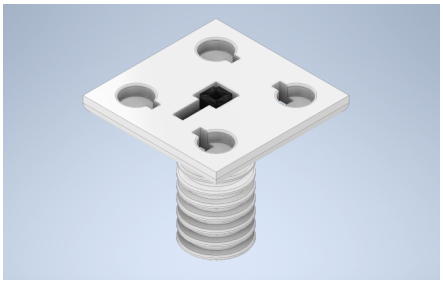
(e) x-axis holder with motor connection



(f) End effector



(g) Motor shaft extension for the end effector



(h) Tool

Figure 6.4: The models are produced via 3D printing and subsequently assembled on a tabletop for the construction of the haptic display. (a) The guide holder is required for the mounting of the y-axis guides. The motor holder (b) and pulley holder (c) serve to tension the belt, which is responsible for enabling the y-axis movement. The x-axis is formed by the two holders (d) and (e), which feature a motor connection, and two metal guides. The end effector (f), a shaft extension (g) for the rotary motor, and a tool (h) in which the magnets are located.

Software

Given the analogous architectural configuration of our system and that of a 3D printer, it is reasonable to check whether the same software can be used. The mainboard mentioned in Section 6.2.2 has been designed for 3D printer and also supports the use of the open-source Marlin [W21] firmware. This firmware takes over the complete control of 3D printer, encompassing, for example, the movement of the nozzle, extrusion of the filament, temperature control of the print bed, or fans. As the software is open-source, it is also possible to integrate custom components. Communication with the firmware or with the printer then takes place via G-codes. These G-codes can be used to control the individual components of the system, such as the aforementioned movements of the end effector. These codes are typically generated by a slicing program that creates print paths from a 3D model. These slicers then generate a list of contiguous G-code commands, which are then processed by the printer in a sequential manner. One disadvantage for our system is that it is not designed to react in real-time and interactively to the current movement, i.e. a movement command across the table cannot be interrupted. However, since the table is supposed to react to spontaneous movements of the user, we have to generate this behavior. One potential solution is the command M410 Quickstop, which can interrupt such movements. However, after the command is called, the motors lose their position due to the abrupt stop and require recalibration, which is not a viable option. To enable interactive reaction to changes, movements are interpolated and the corresponding G-code commands are sent individually. In this way, movements can be interrupted and updated. In order to ensure compatibility with our robots, a ROS interface for the generation of the G-code used for the setup has been implemented. Furthermore, as already described in the previous chapters, the Unity-Robotics-Hub enables the communication with Unity in order to be able to communicate with the table not only from a computer, but also directly from the virtual environment. Consequently, the positions can be controlled directly with a pose message via a topic. The generated code is transmitted to the ETHD at a rate of 100 Hz.

Simulation

The entire system is simulated within the Gazebo environment, providing a platform for testing and validating the setup in the absence of physical hardware (see Figure 6.5). The simulation environment was mainly used to test the VR application. The interfaces are identical to those of the actual hardware, with simple pose messages, the end effector can be moved and, consequently, also the object on the table. However, the simulation has limitations, as it was only employed for testing purposes. The behavior of the object does not precisely mirror that of the real-world counterpart, without the dragging effect due to friction. Additionally, the processes of attaching and detaching are not provided. To change the object, the simulation environment must be restarted.

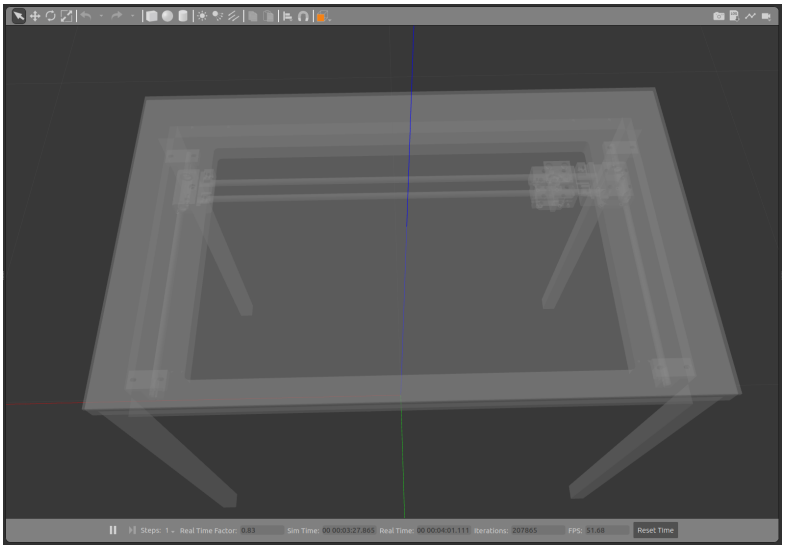


Figure 6.5: Gazebo simulation of the Encountered-Type Haptic Display (ETHD).

6.3 Technical Evaluation

6.3.1 Payload, Velocity, and Magnet Size

In order to find the maximum weight that the end effector is capable of moving and the maximum speed at which it can travel without losing the object, an experiment was conducted. In this experiment, as illustrated in Figure 6.6, a square container was 3D printed with a mass of 20 g when empty. The container is placed on a small square plate in which the magnets are located. The plate can be replaced, allowing the magnets to be exchanged in order to identify the optimal magnet size. Below the table, the tool can be changed, again, to find the ideal magnet size. No Hall sensors are used in this experiment. The experiment consists of two steps. In the first step, the container is moved across the table in a square trajectory three times, with the speed increasing after each round. If the container successfully completes all three rounds without being separated from the end effector, the weight in the container is increased. Conversely, as soon as the container fails to complete one of the three rounds, the run is no longer considered successful and the maximum payload is reached. In the next step, the speed is slowly increased for the last successful weight in order to find the maximum velocity. These two steps are carried out for different magnet combinations. The magnets that we tested were of the following dimensions: 6x3 mm (diameter x height), 8x3 mm, 8x4 mm, and 10x3 mm. The speeds at which they were tested were 10 m/s, 20 m/s and 40 m/s.

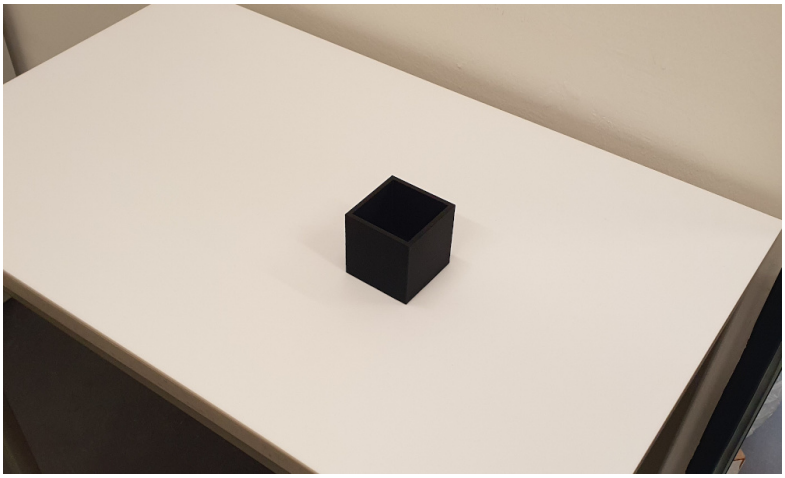


Figure 6.6: The 3D printed container for the payload and velocity experiment.

6.3.2 Friction Compensation

In a world without friction, the object would always sit perfectly above the tool, the mechanism could carry more weight, and move at a faster pace. Unfortunately, since this is not the case, we have to take friction into account. In the preceding section, an analysis was conducted to determine the maximum weight that can be moved and the fastest rate of movement that the end effector can achieve. Furthermore, the object is dragged by the friction during movement, which means that it is not directly above the tool, but a few millimeters behind it. In this context, two distinct types of friction are distinguished. The first is static friction, which occurs when the object is at rest. Only when this is overcome the object begin to move and slide over the contact surface. Kinetic friction, on the other hand, occurs during movement. The extent to which the object lags behind the tool depends on the magnitude of the kinetic friction. As soon as the device stops, the object also stops, and the static friction returns. In a first experiment, we want to find out how large the displacement between the object and the tool is, after the movement has stopped. Therefore, different combinations of magnets are tested to find the best one, also taking into account the results from the payload and velocity experiment. To enable more precise positioning, the four Hall effect sensors will be used to track the position of the object and to correct the approach to achieve a higher positioning accuracy. This will be investigated in a second experiment.

Experiment 1

The objective of this experiment is to investigate the impact of different magnet dimensions on the friction and repetition accuracy. Magnets of the size of 6x3 mm (diameter x height), 8x3 mm, 8x4 mm, and 10x3 mm are tested in different combinations between the object and the tool. A small squared object with an AprilTag [113] on the surface is used as the object, which is tracked

by a calibrated camera. With the tag, the current 6D pose of the object can be determined. This object is used to randomly approach 20 equally distributed points within the workspace in order to determine the repetition accuracy. Each point is approached 20 times, resulting in a total of 400 samples. When a position is approached, a short pause is made to ensure that the recognized tag is stable in the image.

Experiment 2

In order to enhance the precision of the ETHD's positioning, Hall Effect sensors are used to track the reference magnet within the object. If these sensors are precisely assembled, it is theoretically possible to use a simple mathematical analysis to determine the position of a reference magnet. However, since four sensors were constructed by hand into a 3D-printed object, this is no longer a trivial task. Nevertheless, it is a task that can be accomplished with a neural network. The four raw sensor values are therefore fed into the neural network shown in Figure 6.7. Three fully connected layers, the first one with 10 neurons, the second with 20, and the third again with 10 are used, all with ReLU activation. With hyper-parameter optimization the following parameters for the network were determined: Mean Squared Error as loss function together with the adam optimizer, a learning rate of 0.0001, weight decay of 0.07, and a batch size of 5.

In order to create the training data set, a sample object was fixed on the table surface and the tool was positioned exactly underneath it, as the starting reference position. The magnets in the tool and in the object have dimensions of 8x3 mm, the same applies for the reference magnet. The decision in favor of these magnets is explained in more detail in Section 6.3.6. The end effector then scanned a square area of 10 mm around this point in a grid pattern and recorded a total of 24000 training sample tuples of ground truth position and sensor values.

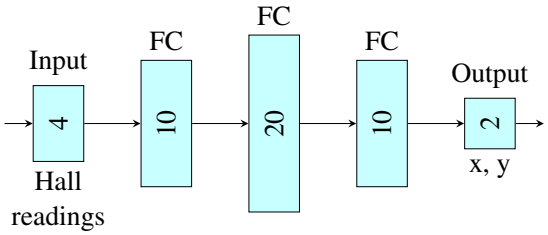


Figure 6.7: The neural network architecture used for the position estimation of a reference magnet on top of the table. The readings of four Hall sensors in the center of the tool below the table were used as input, as output the x-y-coordinates of the magnet relative to the tool. The three fully connected hidden layers have 10, 20, and 10 neurons all with a ReLU activation function.

6.3.3 Object Recovery

The display has been designed with the intention of interactions within virtual environments, whereby the user is able to interact with and manipulate the objects present on the table. Such interactions may, for instance, consist of briefly pressing the object, or the user may pick up the object and place it in a different position. It can also happen that the object is accidentally moved during the interaction. Furthermore, due to too much friction, the object may stuck without any external impact. However, this is only a minor problem, as long as the maximum speed is maintained. A recovery mode has been implemented to prevent the experiment from having to be interrupted or the immersive illusion from being destroyed. The recovery process is based on conventional computer vision techniques. A camera is positioned to observe the entire surface of the table. If required, it can be placed inconspicuously in the room. Initially, the tabletop is filtered out of the image, as it is white and stands out well against the background, simple edge detection can be used here. It is assumed that the four edges forming the largest trapezoid in the image represent the table surface (green lines in Figure 6.8). The intersections (red dots) indicate the corners. The table surface is now converted into a grayscale image in which clusters are searched for. The center of each cluster serves as a reference point for the objects, which are then mapped together with the table surface onto a two-dimensional plane (blue dots as corners, green dot as object). If an object's position does not align with the known positions, the object is considered to have moved and can be returned to its original location. Even if the recognized position is not very precise, it is sufficient to be correctly attached by the magnets in the end effector. This type of recovery was sufficient for our scenario. However, for experiments of greater complexity, it may be more appropriate and robust to train a neural network.

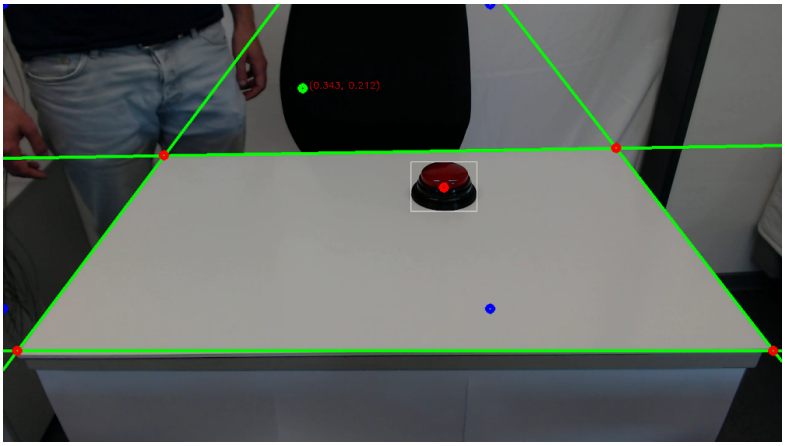


Figure 6.8: The position of objects is monitored continuously. If a user relocates an object, this is identified and the position can be corrected. Conventional image processing techniques are employed to recognize the table and object.

6.3.4 Interaction Detection

In human-robot interaction scenarios, whether in VR, AR, or without a virtual environment, the detection of interactions is crucial. This includes detecting if the object has been taken away and if the object has been touched. For the first part, we can look at the Hall effect values without any deep learning approach. The four sensor values are normalized between 0 and 1. As the noise from the sensors is low enough, it can be assumed that the object is above the tool even if there is a minimal deflection. Using a threshold of 0.03 proves to be a suitable value. Below this value, it can be assumed that there is no magnet above the tool. Regarding the second part, when a person touches the object, the friction compensated position can be used. If the estimated position doesn't match the measured one, the object has been touched.

6.3.5 Results

Payload, Velocity, and Magnet Size

Table 6.2 and Table 6.3 show the results of the first experiment to find the maximum payload and velocity, as well as the optimal magnet size combination. The results indicate that the smaller magnets, with dimensions of 6x3 mm, are insufficient in providing enough force to maintain the connection, regardless of whether they are used in the tool or the object. Even at low weights, friction prevails and the object is quickly lost. Also the combination with larger magnets proved to be ineffective. The maximum weight that can be achieved with these magnets in the tool is 55 g; when they are used in the object, a maximum of 115 g is achieved. With magnets that are too large, the force is so high that the motors do not manage to overcome the static friction and the end effector no longer moves at all. This happens with the combination of 8x4 mm with 10x3 mm magnets. The best result in terms of payload is 235 g overall. This result is achieved with different combinations of magnets, including 8x3 mm with 8x4 mm, 8x4 mm with 8x4 mm, and 10x3 mm with 8x3 mm. The highest speed at which the object is reliably transported is 66 m/s with the magnet combination of 8x3 mm with 8x3 mm. However, these speeds are also almost reached with the other combinations of 8x3 mm, 8x4 mm, and 10x3 mm.

Table 6.2: Maximum payload of the haptic display

		Object			
		6×3 mm	8×3 mm	8×4 mm	10×3 mm
Tool	6×3 mm	20 g	35 g	55 g	35 g
	8×3 mm	105 g	210 g	**235 g**	205 g
	8×4 mm	115 g	225 g	**235 g**	stuck
	10×3 mm	115 g	**235 g**	stuck	stuck

Table 6.3: Maximum velocity of the haptic display

Tool \ Object	6×3 mm	8×3 mm	8×4 mm	10×3 mm
6×3 mm	40 m/s	45 m/s	50 m/s	50 m/s
8×3 mm	50 m/s	**66 m/s**	61 m/s	65 m/s
8×4 mm	53 m/s	60 m/s	64 m/s	stuck
10×3 mm	45 m/s	63 m/s	stuck	stuck

Table 6.4: Repetition accuracy of the table

Tool \ Object	6×3 mm	8×3 mm	8×4 mm	10×3 mm
6×3 mm	1.21	**1.1**	1.36	1.24
8×3 mm	1.18	1.17	1.39	1.39
8×4 mm	1.3	1.44	1.96	2.11
10×3 mm	1.36	1.59	1.89	2.01

Friction Compensation

To compensate the friction, we conducted two experiments. The results of the first experiment are shown in Table 6.4. It is evident that the friction increases with larger magnets, resulting in a corresponding increase in the inaccuracy of the repetition accuracy. The combination of 6x3 mm in the tool and 8x3 mm in the object yields the most accurate positioning, with a standard deviation of 1.1 mm. In contrast, the combination of 8x4 mm in the tool and 10x3 mm in the object exhibits the least accurate positioning, with a deviation of 2.11 mm.

In the second experiment, the objective is to use Hall effect sensors to enhance the precision of the object's positioning by estimating its location. In order to achieve this, a training set was recorded to train the neural network. The values of the data set are shown in Figure 6.9. The color of a data point indicates the degree of deflection of the sensor at that position. For training, the data is split into 70 % training data and 30 % test data, the network was trained in 100 epochs. This approach resulted in a test accuracy of 0.5 mm, which is less than half the initial positioning accuracy of 1.17 mm.

6.3.6 Discussion

The experiments were conducted in order to explore the possibilities and technical limits of our ETHD. It is crucial to select an appropriate combination of magnets in the tool and the object in order to be able to carry objects quickly, precisely, and as heavy as possible. Although the

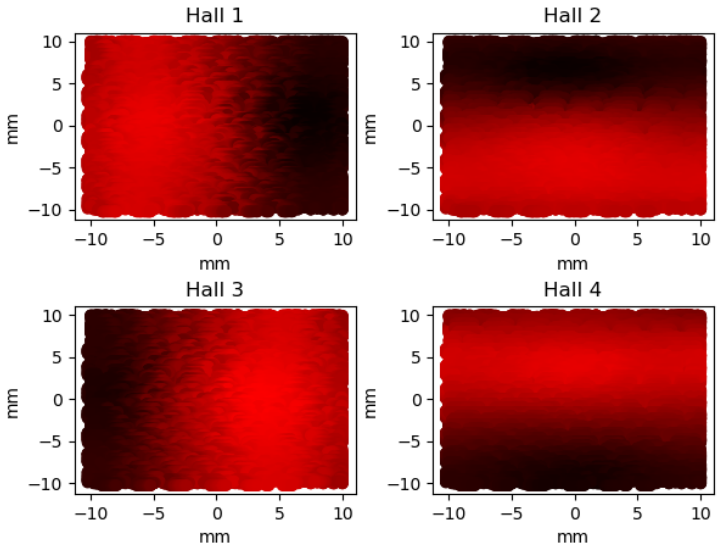


Figure 6.9: The graphic shows training data for the object positioning network. Each diagram shows the recorded training samples of one Hall sensor. One sample contains the x-y position of the reference magnet, and the Hall readout normalized between 0 and 1 mapped to the color space from black to red.

smallest magnets in our tests with 6x3 mm can be positioned quite precisely, they can hardly be transported as the connection quickly loosens both at low weights and at low speeds. With magnets that are too large, the attraction between the object and the tool is so high that the static friction is challenging or impossible to overcome and also leads to inaccurate positioning. As it is not possible to achieve the maximum value in all areas with any combination, a compromise is necessary. It was determined that the optimal choice for both the tool and the object are the 8x3 mm magnets, as this configuration allows for the highest speed, an acceptable level of accuracy, and a relatively high weight. Furthermore, the positioning accuracy has been enhanced through friction compensation. The position of the object is determined by the neural network model, and the target position of the end effector can be adjusted accordingly to achieve a positioning accuracy of 0.5 mm. Furthermore, we can achieve a speed of 66 m/s at which objects are reliably transported without the connection loosening. If the connection get lost, it can be quickly restored with the presented recovery process. A user study was conducted to assess the usability and user experience, as detailed in the subsequent chapter.

6.4 User Study

The haptic display is designed to be used in Virtual and Augmented Reality scenarios. To evaluate the usability, the user experience, and the haptic experience, a user study was conducted.

Furthermore, maintaining the illusion of interacting in a realistic environment where everything behaves naturally is crucial for a good experience. It is therefore important that the haptic feedback is in alignment with the visual representation, otherwise this could cause a break in presence [209, 210]. An object placed incorrectly or with a wrong timing would lead to a discrepancy between the real and virtual world when the user reaches into the void. To prevent this and to evaluate the table, we came up with the following two research questions:

1. How much time does the display have to position an object on the table before the user encounters it?

2. Does the haptic display offer a satisfying user experience and usability?

The objective of the initial research question is to find the velocity at which a user interacts with the table, the amount of time required for the table to position an object at a target location, and the available time for an object placement before the user encounters it. Furthermore, the distance that the object is capable of moving within the workspace is to be determined in order check the coverage.

In the following sections further details on the setup, task, implementation of the study will be described, as well as the results to answer the research questions.

6.4.1 Setup

For this VR user study, the Meta Quest 2 is integrated with the haptic display. The software for the Meta Quest was developed using Unity [W4]. As previously outlined in Section 6.2.2, the use of ROS to generate and transmit G-code commands to the table enables communication via the Unity-Robotics-Hub [W6], as in the previous XR experiments. The ETHD is placed within a standard office setting, with a conventional office chair positioned in front of the table, on which the user is seated during the experiment (as illustrated in Figure 6.1). The setup does not indicate that this is not a conventional desk. In VR, the user observes a simplified environment, rather than the actual office. However, the table is displayed in front of the user, whose position still needs to be registered at the start of the experiment (see Section 6.4.1). Behind the table is a large panel on which the user is shown instructions. Additionally, the user sees virtual representations of their hands. The setup also includes a physical button as shown in Figure 6.10, this button is connected to Wi-Fi and ROS and publishes on a topic as soon as it is pressed or released. This button is equipped with magnets and can be positioned on the table.

Registration

In other experiments, AprilTags have been used to overlay the real and virtual worlds. In order to maintain the table's inconspicuousness in this scenario, the user have to register the physical table in the virtual world. A supervisor could do this before the experiment, but by putting the

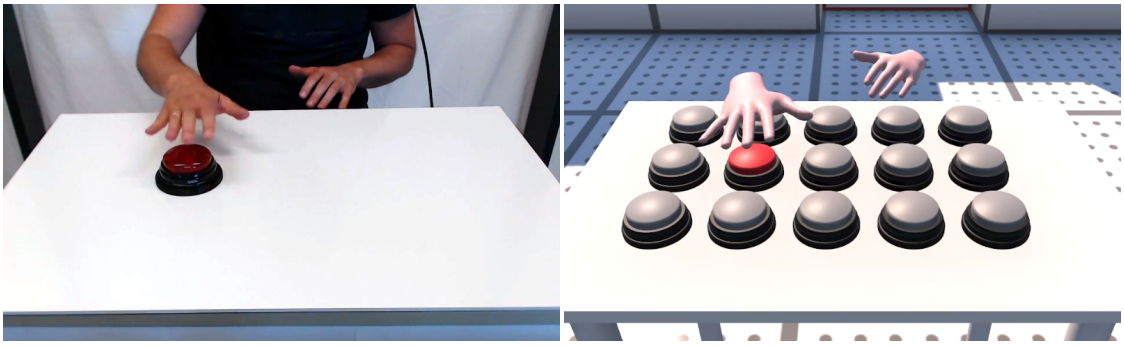


Figure 6.10: *Whac-A-Mole* user study setup in the real world (left) and in VR (right). The user is presented with 15 virtual buttons on a virtual table. To get haptic feedback for all of these buttons, the table substitutes each of them by moving one physical button at the desired position.

VR display up and down, the positioning could get lost. The procedure is as follows: The users explore the table in front of them and feel its position and dimensions. They see the virtual table, but it does not match the one they feel. The user needs to place the tip of the left index finger on the front left corner and perform a pinch gesture with the other hand. In this gesture, the tips of the index finger and thumb touch. Subsequently, the subject touches the right front corner with the tip of the right index finger and performs a pinch gesture with the left hand. Upon completion of this action, the virtual table is transformed, with the front corners situated at the positions of the saved index fingertips. It should be noted that the virtual table may exhibit a slight discrepancy in dimensions, either narrower or wider than the actual table. However, this discrepancy is not perceptible. The virtual table is now aligned with the real table.

6.4.2 Task

During the experiment the participants play two rounds of the arcade game *Whac-A-Mole*. The task in this funfair game is to hit mechanical moles popping out of a box with holes with a hammer and put them back into their hole. The challenge of the game is not knowing which hole the next mole will come out of and hitting it as quickly as possible. A variant of this game replaces the holes and mechanical animals with buttons that are pressed with the hand instead of a hammer. This variant has been recreated in the context of the experiment. The participants are presented with 15 virtual buttons arranged in three rows of five buttons each, located on the virtual table in front of them. All buttons are gray. If they become illuminated in red, the user is prompted to press the button. Only one button is illuminated at a time and remains so for a period of five seconds or until it is pressed. In the course of the experiment, the calculated time available for placing the button was not subjected to live testing. Instead, the next virtual button was only activated once the physical button had reached the target position and when the user returned their hands to a blue area between the table and his body. The game ends when the user

pressed the button 100 times. The experiment is conducted in two conditions: one in which the user presses the button with the flat hand, and the other in which the user presses the button with the index finger only. After the user has pressed 50 buttons, a short pause is initiated, and the user continues with another 50 presses with the other condition. Which condition comes first is decided at random. The participants were not asked to be particularly fast or to use only one particular hand in order to be able to test the influence of these factors on the given research question later on without consciously influencing them.

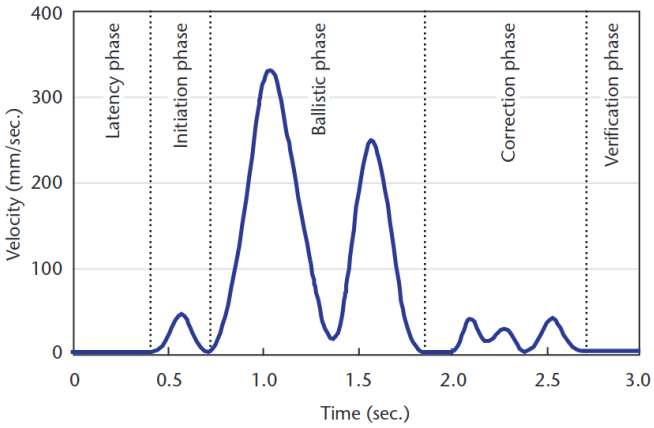


Figure 6.11: The five phases of a movement according to Nieuwenhuizen et al. [211] (graphic taken from the paper). The first latency phase indicates the reaction time of the test subject. First short movements are detected in a short initiation phase. The main part of the distance is covered in the ballistic phase, which is the fastest phase and ends imprecisely near the object before being refined in the subsequent correction phase until the target is reached. In the final verification phase, there is no further movement, it simply lasts until the task is completed.

6.4.3 Measurements

With this experiment we want to measure how fast the users move there arms in order to determine the optimal time for placing an object in a designated location without the user becoming aware that the object was absent from the location seconds before. Secondly, the aim is to gain insight into the participants' subjective perception of the user experience and the usability of the table. In order to get a deeper understanding of arm movements, a closer look at the work of Nieuwenhuizen et al. [211] is taken. Their publication is about phases of movements, which can be observed in Figure 6.11. The process of an arm movement begins with the latency phase, which reflects the user's reaction, followed by an initiation phase, which contains small movements before the actual main phase, the ballistic phase. The ballistic phase represents the largest movement and is characterized by a rapid yet imprecise approach towards the target. This is fol-

lowed by the correction phase, during which slower but more precise movements are employed to compensate for the initial inaccuracy. In the subsequent verification phase, no further movement takes place, this phase ends as soon as the task is recognized as complete. The objective of this study is to investigate the presence of these phases in our scenario and how much time each phase takes.

Furthermore, subjective factors should be analyzed with a series of questionnaires that the users fill out before and after the experiment. A variety of standardized questionnaires for VR will be employed, in addition to customized one. A demographic questionnaire is used to ascertain information about the participants. To find out whether the system is prone to simulator sickness, the Simulator Sickness Questionnaire (SSQ) [156] is used. The NASA TLX [154] is used to determine the level of workload required to operate the system. To quantify the sense of presence we use the Igroup Presence Questionnaire (IPQ) [212], which enables us to measure the extent to which the participants really had the feeling of being in the virtual environment without the illusion being destroyed. The User Experience Questionnaire (UEQ-S) [213] and SUS[155] questionnaires will be used to evaluate the usability and user experience of the haptic display. With a custom questionnaire we asked about the haptic experience specifically about the table. At the latest, the participants will become aware that they have not physically interacted with the virtual objects, but rather with physical objects on the table. The following statements will be rated on a 5-point Likert scale:

- The haptic pressing behavior of the physical buttons matched the virtual ones
- The shape of the physical buttons and the virtual ones matched
- The position of the physical buttons and the virtual ones matched

6.4.4 Conduct

Participants

For the user study, 12 participants could be recruited overall, 10 with an age between 25 and 34 years, one person between 35 and 44, and one between 55 and 64. In terms of gender, 9 people said they are male, two are female, and one person did not want to disclose their gender. The majority of participants (11 out of 12) indicated infrequent VR use, occurring once every quarter or less, while only one participant reported using VR on a weekly basis. In order to participate in the study, no visual impairment must be present.

Procedure

When participants enter the laboratory, the table is positioned directly in front of them. Due to its appearance as a regular table, it doesn't particularly stand out. Initially, participants are required to sign a consent form and receive a brief introduction to the experiment's procedure. Before

commencing, they must fill out the demographic questionnaire and the SSQ questionnaire at a designated computer. Subsequently, they are seated on the long side of the table and put on the VR glasses and headphones, in order to avoid being distracted by the noise emanating from the table. On the panel, users see instructions to register the position of the table and the VR display. Following this, they are presented with a sample button for exploration, followed by placing their hands in the home area. The actual experiment begins afterward. Participants play two rounds of *Whac-A-Mole*. Half of the participants start with 50 flat presses followed by 50 fingertip presses, while the others follow the opposite sequence. After the experiment, participants complete the remaining questionnaires and can choose to be briefed on the objectives and research questions if interested.

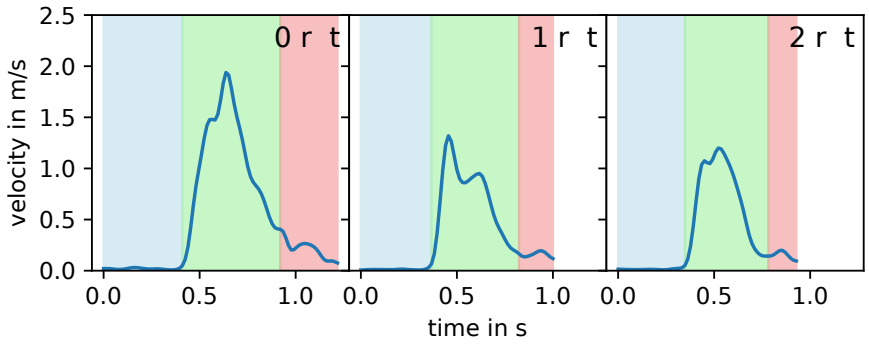


Figure 6.12: Three exemplary velocity plots with colored latency (blue), ballistic (green), and correction (red) phase. An initiation or verification phase can not be observed in the data. All plots are right handed movements (r) in the fingertip (t) condition.

6.4.5 Results

Quantitative Results

A total of 996 usable movements were recorded from the 12 participants, with 759 performed with the right hand and 237 with the left. For the unusable movements, the starting point was not recognized correctly or the button was not pressed. This data was omitted as it falsifies the result. Of these movements, 503 were performed with the fingertip and 493 with the flat hand. The movements have been divided into phases similar to that employed by Nieuwenhuizen et al. [211]. Figure 6.12 illustrates the velocities of three sample movements, all of which were recorded with the right hand and pressed with the fingertip. Only three instead of five phases are recognizable in our case. It is not possible to identify the initiation or verification phases in the data set. Table 6.5 provides a more detailed account of these three phases, it presents the mean durations of the individual phases. As expected, the user's reaction is comparable in both conditions with an average of 385 ms. For precise movements, the average ballistic phase is

440 ms, while for less precise movements, as by the flat hand, it is 417 ms. The correction phase is also significantly shorter, at 42 ms, compared to 98 ms for precise movements. In total, users require an average of 884 ms to complete their movements. However, as it is the fastest movement that should be taken into account, they are listed in Table 6.6. The minimum reaction time was 210 ms, for the ballistic phase it was 111 ms, and for the correction phase it was 14 ms. The fastest movement in total took only 477 ms, which corresponds to almost half a second. A more detailed illustration of the shortest movements can be found in Figure 6.13, which depicts the fastest durations for the 15 buttons. As expected, the nearer buttons were reached with considerably greater speed than the more distant ones. The fastest movements at the front were between 477 ms and 550 ms, whereas the ones at the back took up to 700 ms. Given the maximum speed of the table of 66 m/s, this equates to a distance of approximately 31 cm in the front section and 46 cm in the rear section, which is sufficient to cover the majority of the workspace.

Table 6.5: Average Duration of Movement Phases

	Tip	Flat	Overall
Latency phase	381 ms	389 ms	385 ms
Ballistic phase	440 ms	417 ms	429 ms
Correction phase	98 ms	42 ms	70 ms
All phases	920 ms	848 ms	884 ms

Table 6.6: Minimum Duration of Movement Phases

	Tip	Flat
Latency phase	213 ms	210 ms
Ballistic phase	113 ms	111 ms
Correction phase	13 ms	14 ms
All phases	488 ms	477 ms

Qualitative Results

The SSQ was completed by the participants both before and after the experiment to monitor the change in 16 symptoms. The before values were subtracted from the after values, resulting in the following summarized values for *Disorientation* ($M = 2.320, SD = 17.641$), *Oculomotor Disturbance* ($M = 0.000, SD = 9.142$), *Nausea* ($M = -3.975, SD = 6.378$), and a *Total Simulator Sickness* score ($M = -0.935, SD = 5.319$). These results demonstrate that our system does not lead to simulator sickness.

The results of the NASA TLX indicate that the task was neither *Mentally Demanding* ($M = 12.083, SD = 15.733$) nor *Frustrating* ($M = 7.917, SD = 7.217$). The participants felt that their *Performance* ($M = 9.167, SD = 8.747$) was very good. Due to the fast arm movements, the

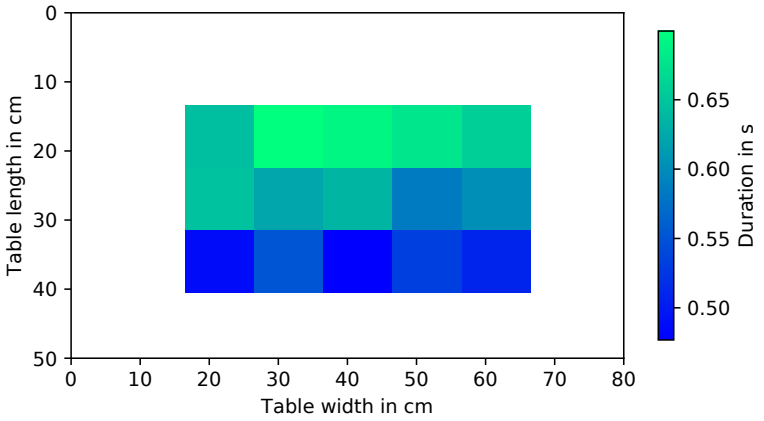


Figure 6.13: The movement duration distribution over the workspace, which is colored.

values for *Physical Demand* ($M = 20.417, SD = 18.885$), temporal demand and required *Effort* ($M = 22.917, SD = 16.714$) were slightly higher.

The results of the IPQ evaluation showed that *Spatial Presence* achieved a high average score of 4.617 ($SD = 0.936$). Additionally, the *Sense of Being There* was also rated highly ($4.833, SD = 1.115$). Users therefore had the feeling of being present in the virtual world. The ratings for *Involvement* ($M = 3.792, SD = 1.044$) and *Experienced Realism* ($M = 2.833, SD = 1.002$) were slightly lower.

The Pragmatic and Hedonic Quality of the system were evaluated using the UEQ-S. Pragmatic Quality refers to the extent to which the system enables the user to accomplish the desired task. Our system achieved a very good result in this assessment ($M = 1.917, SD = 0.779$). The Hedonic Quality, which indicates to the user's enjoyment of the system, was rated as good ($M = 1.354, SD = 1.105$). This places our system in the top 10 % for Pragmatic Quality and in the top 25 % for Hedonic Quality, according to a benchmark data set [214]. With the SUS questionnaire, we reached a mean value of 84.167 ($SD = 9.673$) for the overall usability.

The objective of the custom questionnaire was to ascertain the realism of the behavior between the virtual and physical buttons. A 5-point Likert scale was used for each question. The participants rated the question about the pressing behavior with a 4.417 ($SD = 0.900$), the shapes of the two buttons fitting together with a 4.833 ($SD = 0.389$), and the position with a 4.167 ($SD = 0.937$). Overall, the rating for the behavior of the button was therefore very good.

6.4.6 Discussion

The objective of the user study was to address two research questions. With the first question, we wanted to find out how much time is available to place an object and how large the resulting coverage in our workspace is. The findings indicated that more than half of the table can be covered without the user noticing a mismatch between the two environments. For tasks requiring

greater precision, more time is available for object placement. Furthermore, the movement phases were examined. In contrast to the findings of Nieuwenhuizen et al. [211], only three of the five suggested phases were identified. The simple explanation for the missing validation phase is that it is very short in this context, as the button press immediately completing the task. In such instances, there is quasi-instantaneous validation. Regarding the initiation phase, the respective movement might be so minimal it is not discernible in the data, becoming obscured by the surrounding noise.

The second research question focused on the subjective quality of the system. This was analyzed on the basis of the questionnaires collected. The participants rated the system an excellent Pragmatic Quality and a good Hedonic Quality. In addition, the system obtained a good score on the SUS, which indicates that the usability and user experience of the system is very good. However, the scores for involvement and experienced realism were slightly lower, which can be attributed to the simplified environment. It is likely that these scores would have been higher in a replica of the complete office. The custom questionnaire also yielded positive results, indicating that our haptic feedback approach is perceived as highly realistic. Users reported that they felt able to interact with the virtual objects.

6.5 Conclusion

In the last step of this thesis, we examined the potential of robot-mediated haptic feedback and how this can be used meaningful in a way, that humans can benefit from it, and thus answering the last fundamental question (**FQ4**) on how a robot can be used to provide haptic feedback. To this end, an ETHD has been designed and constructed that enables the provision of haptic feedback to users in Virtual and Augmented Reality environments. This is achieved through the manipulation of physically movable objects on the table surface. As an illustration, the substitution of 15 virtual buttons with only one physical one, as demonstrated in the user study, allows users to perceive the ability to interact with all the buttons. This method enables to improve the virtual experience and make it appear even more realistic. The table is capable of placing objects within a workspace of 47.2 x 26 cm, irrespective of their shape or size. The mechanism comprises a four-dimensional axis system situated beneath the table, which is capable of moving objects across the surface of the table via a magnetic connection. It is important that the table is not recognized as an ETHD by the user, thus enabling the usage in potential AR scenarios. If the workspace is not sufficient, it is possible to build a scaled version under a larger table. With Hall sensors in the end effector we are able to recognize interactions and place objects with a precision of 0.5 mm by compensating for friction. It may also be possible to make virtual avatars appear more realistic using this technology. For example, by moving a cup of coffee on the table in an AR scenario. Furthermore, it may be possible to artificially enlarge the workspace of the table by using hand redirection [215, 216, 217]. The extent to which this can be achieved is a topic for future research.

7 Conclusion and Perspectives

In this work, a wide range of scenarios have been covered to demonstrate the integration of tactile sensors for the collection of information about robots and their multimodal environments. In order to provide a compact overview of the work, we can respond to the fundamental questions posed in Section 1.3 and summarize the main contributions as follows:

Environmental Sensing

The study to answer the first fundamental question (**FQ1**), about the possibilities of improving environmental sensing through tactile sensors demonstrates that the combination of tactile and audio data enables the classification of the content of visually indistinguishable containers with greater accuracy than a unimodal audio network. Moreover, our findings indicate that the multimodal network classification is significantly more robust when tactile data is integrated. Conversely, the classification accuracy of a purely tactile network in a realistic robot scenario with vibration noise is just over 55 percent (for 8 classes). To achieve more precise classification, it is necessary to either reduce the noise of the robot setup or use specialized tactile sensors attached to the container itself. In conclusion, the integration of tactile data enhances the ability to explore the environment, particularly when combined with other modalities.

Intrinsic State Analysis

The objective of this experiment was to investigate to what extent it is possible to make statements about the internal state of the robot with tactile data (**FQ2**). An underactuated gripper, for which the state of the fingers is unknown, was equipped with tactile contact sensors on the individual phalanges. These sensors enable the estimation of the current joint states of the fingers by feeding sequences of data together with the motor state of a grasping movement through a previously trained neural network. The results of this experiment demonstrate that no special sensors are required on the phalanges to measure the joint states; instead, the tactile data can be employed for this purpose.

Augmented Human Perception

The goal of this research is to find out whether humans can be supported by the provision of tactile data measured by the robot (**FQ3**). To this end, a teleoperation experiment has been conducted in which the user is equipped with a Mixed Reality Head-Mounted Display and is

able to teleoperate the robot without the need for additional equipment. The user's hands are tracked by the HMD and transferred to the robot. In a user study, the candidates were presented with the tactile readings of the robots gripper in the form of various visual and acoustic cues on the MR device. The results of different manipulation tasks demonstrated that the performance was enhanced in some areas. However, it was observed that in certain instances, the human operator may become overwhelmed by an excess of stimuli, which could potentially lead to a reduction in performance, particularly in terms of execution time.

Haptic Feedback

In the final fundamental question (**FQ4**), we aim to take a further step by providing the human sense of touch with haptic feedback from the robot. To this end, an ETHD was constructed that enables to move objects on a table in a mechanical manner and place them on demand within the workspace on the tabletop. The haptic display was evaluated in a user study in which participants play a VR version of *Whac-A-Mole* with buttons. By moving objects on the table, the display gave users the impression of interacting with 15 virtual buttons, despite only one physical button being present on the table. This enhanced the realism of the user experience in VR.

The fundamental questions were successfully processed and answered. In all areas, there was added value in using tactile data, and in most cases it was advantageous to use tactile data in combination with other modalities. With robots, this was to be expected, as we assume that more modalities generally mean more information and therefore better results. In instances where human participants were involved, the results did not always improved with the use of more modalities. As Sigrist et al. [159] has already demonstrated, an excess of stimuli can overwhelm individuals, thereby reducing performance. This phenomenon was also observed in our studies.

7.1 Future Work

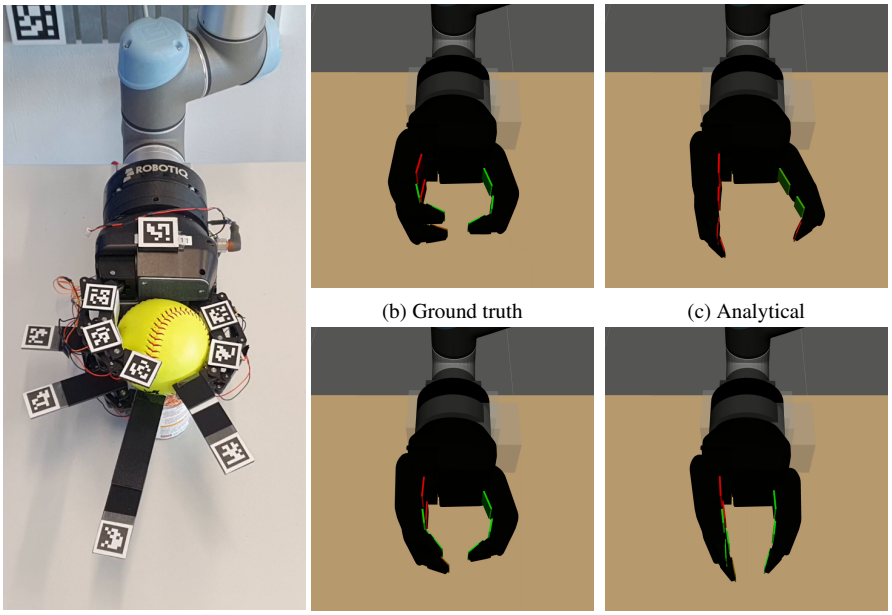
Looking at the application categories from Section 1.2, three of the five categories were covered. One interesting area is that of quality control, which has not been addressed in this thesis. However, it is a topic of great relevance in industrial contexts. As example, there is a growing interest in the implementation of quality control in manufacturing lines using tactile sensors. The increasing availability of affordable robotic manipulators is making it increasingly feasible to automate manufacturing lines. Tactile sensors have the potential to play a major role in this automation process, this could be a topic for future work.

A potential next step of the research would be to combine several categories. In the first experiment, described in Chapter 3, the use of tactile data in object analysis was investigated. This experiment, along with the subsequent one in Chapter 4, demonstrated that the integration

of tactile data with other modalities brings an advantage. In the next step, this knowledge can be used to record data with the teleoperation setup in order to teach the robot to autonomously grasp objects with learning by demonstration. At the same time, the tactile data can be employed to apply a stable grasp and collect information about the grasped object for subsequent analysis. Furthermore, the efficiency of providing tactile information to the user during the learning process can be evaluated.

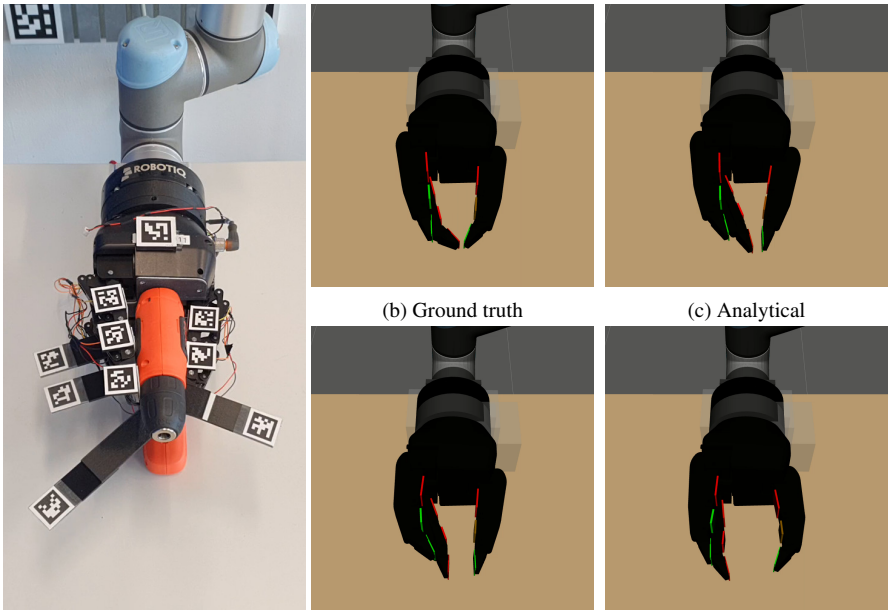
A significant area of research that has emerged in recent years is that of large language model (LLM) first mentioned in [218]. These artificial intelligence systems have been trained on large amounts of text data in order to generate natural language. The applications of these LLMs are diverse, encompassing fields such as chatbots, programming, and content generation. With vision language models (VLMs), these models have been extended to process and generate images and videos [219]. These models can now also be used in robotics, firstly to communicate with the robot, and secondly to enable the robot to perceive and understand its environment without being focused on a specific context. This is possible because these models understand far more than just the experiment for which they are being used. The subsequent stage is the development of multimodal language models, which integrate data from a range of modalities beyond text and vision, including audio and tactile information [220]. Further research could investigate this field, exploring both the possibilities and limits of such technology.

8 Appendix



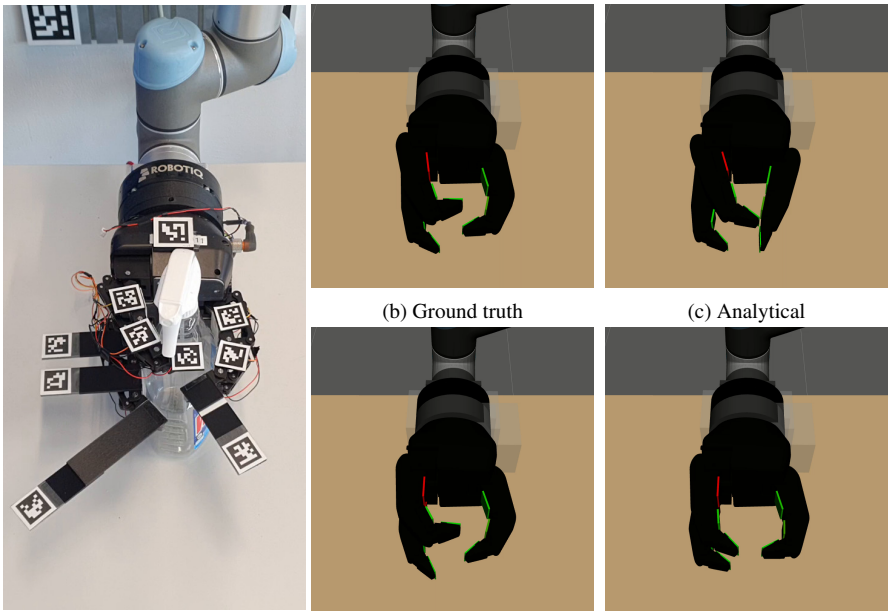
(a) Ball (b) Ground truth (c) Analytical (d) Equal Finger RNN (e) Entire Hand RNN

Figure 8.1: Estimation results of the end state of grasping a ball.



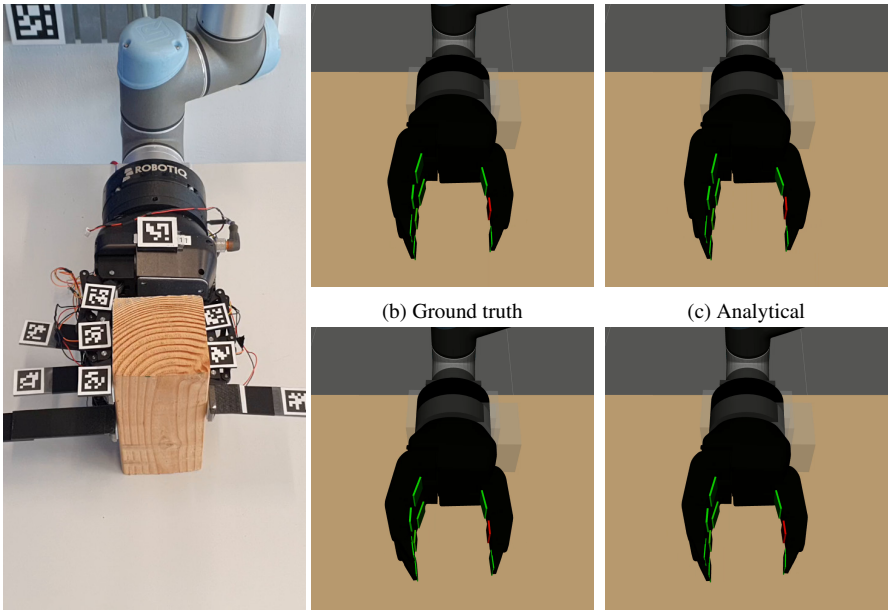
(a) Drill (b) Ground truth (c) Analytical (d) Equal Finger RNN (e) Entire Hand RNN

Figure 8.2: Estimation results of the end state of grasping a drill.



(a) Spray (b) Ground truth (c) Analytical (d) Equal Finger RNN (e) Entire Hand RNN

Figure 8.3: Estimation results of the end state of grasping a spray.



(a) Wooden block (b) Ground truth (c) Analytical (d) Equal Finger RNN (e) Entire Hand RNN

Figure 8.4: Estimation results of the end state of grasping a wooden block.

Bibliography

[1] Roberta L Klatzky and Susan J Lederman. “Touch”. In: *Handbook of Psychology* (2013).

[2] Goran Westling and Roland S Johansson. “Factors Influencing the Force Control During Precision Grip”. In: *Experimental brain research* 53 (1984), pp. 277–284.

[3] Mohsin I Tiwana, Stephen J Redmond, and Nigel H Lovell. “A review of tactile sensing technologies with applications in biomedical engineering”. In: *Sensors and Actuators A: physical* 179 (2012), pp. 17–31.

[4] Leon D Harmon. “Touch-sensing technology- A review”. In: *Society of Manufacturing Engineers, 1980. 58* (1980).

[5] Leon D Harmon. “Automated tactile sensing”. In: *The International Journal of Robotics Research* 1.2 (1982), pp. 3–32.

[6] Leon D Harmon. “Tactile sensing for robots”. In: *Robotics and Artificial Intelligence* (1984), pp. 109–157.

[7] Nicholas Wettels et al. “Biomimetic Tactile Sensor Array”. In: *Advanced Robotics* 22.8 (2008), pp. 829–849.

[8] Ravinder S Dahiya et al. “Tactile Sensing—From Humans to Humanoids”. In: *IEEE transactions on robotics* 26.1 (2009), pp. 1–20.

[9] Zhen Deng et al. “Grasping force control of multi-fingered robotic hands through tactile sensing for object stabilization”. In: *Sensors* 20.4 (2020), p. 1050.

[10] Niklas Fiedler, Yannick Jonetzko, and Jianwei Zhang. “A Multimodal Pipeline for Grasping Fabrics from Flat Surfaces with Tactile Slip and Fall Detection”. In: *2023 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. IEEE. 2023, pp. 1–6.

[11] Shan Luo et al. “Robotic tactile perception of object properties: A review”. In: *Mechatronics* 48 (2017), pp. 54–67.

[12] Giulia Franchi and Kris Hauser. “Technical report: use of hybrid systems to model the RobotiQ adaptive gripper”. In: *Bloomington, IN* (2014).

[13] Victor Mercado, Maud Marchal, and Anatole Lécuyer. ““Haptics On-Demand”: A Survey on Encountered-Type Haptic Displays”. In: *IEEE Transactions on Haptics* 14.3 (2021), pp. 449–464.

[14] Nathan F Lepora and Benjamin Ward-Cherrier. “Tactile quality control with biomimetic active touch”. In: *IEEE Robotics and Automation Letters* 1.2 (2016), pp. 646–652.

[15] Tanel Kossas et al. “Whisker-based tactile navigation algorithm for underground robots”. In: *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2024, pp. 13164–13170.

[16] Jack M Loomis and Susan J Lederman. “Tactual Perception”. In: *Handbook of perception and human performances* 2.2 (1986), p. 2.

[17] Ravinder S Dahiya and Maurizio Valle. *Robotic tactile sensing: technologies and system*. Vol. 1. Springer, 2013.

[18] Cheng Chi et al. “Recent Progress in Technologies for Tactile Sensors”. In: *Sensors* 18.4 (2018), p. 948.

[19] F Zhu and JW Spronck. “A Capacitive Tactile Sensor for Shear and Normal Force Measurements”. In: *Sensors and Actuators A: Physical* 31.1-3 (1992), pp. 115–120.

[20] Alexis Maslyczyk, Jean-Philippe Roberge, Vincent Duchaine, et al. “A Highly Sensitive Multimodal Capacitive Tactile Sensor”. In: *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2017, pp. 407–412.

[21] Karl Hoffmann. “An introduction to stress analysis and transducer design using strain gauges”. In: *The definitive work on strain gauge measurement*. 2012.

[22] Niklas Fiedler et al. “Low-Cost Fabrication of Flexible Tactile Sensor Arrays”. In: *HardwareX* 12 (2022), e00372.

[23] Niklas Fiedler et al. “A Low-Cost Modular System of Customizable, Versatile, and Flexible Tactile Sensor Arrays”. In: *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2021, pp. 1771–1777.

[24] J Dargahi, M Parameswaran, and S Payandeh. “A Micromachined Piezoelectric Tactile Sensor for an Endoscopic Grasper-Theory, Fabrication and Experiments”. In: *Journal of microelectromechanical systems* 9.3 (2000), pp. 329–335.

[25] Weikang Lin et al. “Skin-Inspired Piezoelectric Tactile Sensor Array with Crosstalk-Free Row+Column Electrodes for Spatiotemporally Distinguishing Diverse Stimuli”. In: *Advanced Science* 8.3 (2021), p. 2002817.

[26] Wenzhen Yuan, Siyuan Dong, and Edward H Adelson. “Gelsight: High-resolution robot tactile sensors for estimating geometry and force”. In: *Sensors* 17.12 (2017), p. 2762.

[27] Mike Lambeta et al. “Digit: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation”. In: *IEEE Robotics and Automation Letters* 5.3 (2020), pp. 3838–3845.

[28] Hongbo Wang et al. “Robust and High-Performance Soft Inductive Tactile Sensors Based on the Eddy-Current Effect”. In: *Sensors and Actuators A: Physical* 271 (2018), pp. 44–52.

[29] Yannick Jonetzko et al. “Encountered-Type Tabletop Haptic Display for Objects On-Demand in Virtual Environments”. In: *2023 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. IEEE. 2023, pp. 1–7.

[30] Dominic Jones et al. “Design and evaluation of magnetic hall effect tactile sensors for use in sensorized splints”. In: *Sensors* 20.4 (2020), p. 1123.

[31] Morgan Quigley et al. “ROS: an open-source Robot Operating System”. In: *ICRA workshop on open source software*. Vol. 3. 3.2. Kobe, Japan. 2009, p. 5.

[32] Yuya Maruyama, Shinpei Kato, and Takuya Azumi. “Exploring the performance of ROS2”. In: *Proceedings of the 13th International Conference on Embedded Software*. 2016, pp. 1–10.

[33] Steven Macenski et al. “Robot Operating System 2: Design, architecture, and uses in the wild”. In: *Science Robotics* 7.66 (2022), eabm6074.

[34] Gerardo Pardo-Castellote. “Omg data-distribution service: Architectural overview”. In: *23rd International Conference on Distributed Computing Systems Workshops, 2003. Proceedings.* IEEE. 2003, pp. 200–206.

[35] Dennis Krupke et al. “Comparison of Multimodal Heading and Pointing Gestures for Co-Located Mixed Reality Human-Robot Interaction”. In: *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. Madrid, Spain: IEEE, 2018, pp. 5003–5009.

[36] Paul Milgram and Fumio Kishino. “A taxonomy of mixed reality visual displays”. In: *IEICE TRANSACTIONS on Information and Systems* 77.12 (1994), pp. 1321–1329.

[37] Peter Le Noury et al. “A narrative review of the current state of extended reality technology and how it can be utilised in sport”. In: *Sports Medicine* 52.7 (2022), pp. 1473–1489.

[38] Tomoaki Nakamura, Takayuki Nagai, and Naoto Iwahashi. “Multimodal Object Categorization by a Robot”. In: *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 2007, pp. 2415–2420.

[39] A. Pieropan et al. “Audio-Visual Classification and Detection of Human Manipulation Actions”. In: *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 2014, pp. 3045–3052.

[40] Fuchun Sun et al. “Object Classification and Grasp Planning using Visual and Tactile Sensing”. In: *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 46.7 (2016), pp. 969–979.

[41] Manfred Eppe et al. “Combining Deep Learning for Visuo-motor Coordination with Object Detection and Tracking to Realize a High-level Interface for Robot Object-picking”. In: *IEEE RAS International Conference on Humanoid Robots (Humanoids)*. 2017, pp. 612–617.

[42] Matthias Kerzel et al. “Neurocognitive Shared Visuomotor Network for End-to-end Learning of Object Identification, Localization and Grasping on a Humanoid”. In: *IEEE Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*. 2019, pp. 19–24.

[43] Püren Güler et al. “What’s in the container? classifying object contents from vision and touch”. In: *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE. 2014, pp. 3961–3968.

[44] Di Guo et al. “Robotic grasping using visual and tactile sensing”. In: *Information Sciences* 417 (2017), pp. 274–286.

[45] Hao Dang, Jonathan Weisz, and Peter K Allen. “Blind grasping: Stable robotic grasping using tactile feedback and hand kinematics”. In: *2011 ieee international conference on robotics and automation*. IEEE. 2011, pp. 5917–5922.

[46] A. Alfadhel et al. “Magnetic Tactile Sensor for Braille Reading”. In: *IEEE Sensors Journal* 16.24 (2016), pp. 8700–8705.

[47] Focko L Higgen et al. “Crossmodal pattern discrimination in humans and robots: A visuo-tactile case study”. In: *Frontiers in Robotics and AI* 7 (2020), p. 540565.

[48] J. A. Fishel and G. E. Loeb. “Sensing Tactile Microvibrations with the BioTac — Comparison with Human Sensitivity”. In: *2012 4th IEEE RAS EMBS International Conference on Biomedical Robotics and Biomechatronics (BioRob)*. 2012, pp. 1122–1127.

[49] Kunal J Pithadiya, Chintan K Modi, and Jayesh D Chauhan. “Selecting the most favourable edge detection technique for liquid level inspection in bottles”. In: *International Journal of Computer Information Systems and Industrial Management Applications* 3 (2011), pp. 11–11.

[50] Chau Do, Tobias Schubert, and Wolfram Burgard. “A probabilistic approach to liquid level detection in cups using an RGB-D camera”. In: *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2016, pp. 2075–2080.

[51] Danilo Pau et al. “Dataset of sodium chloride sterile liquid in bottles for intravenous administration and fill level monitoring”. In: *Data in Brief* 33 (2020), p. 106472.

[52] R. S. Durst and E. P. Krotkov. “Object Classification from Analysis of Impact Acoustics”. In: *Proceedings 1995 IEEE/RSJ International Conference on Intelligent Robots and Systems. Human Robot Interaction and Cooperative Robots*. Vol. 1. 1995, pp. 90–95.

[53] Shan Luo et al. “Knock-Knock: Acoustic Object Recognition by Using Stacked Denoising Autoencoders”. In: *Neurocomputing* 267 (2017), pp. 18–24.

[54] Danfei Xu, Gerald E Loeb, and Jeremy A Fishel. “Tactile Identification of Objects using Bayesian Exploration”. In: *2013 IEEE International Conference on Robotics and Automation*. IEEE. 2013, pp. 3056–3061.

[55] Matthias Kerzel et al. “Haptic Material Classification with a Multi-Channel Neural Network”. In: *International Joint Conference on Neural Networks (IJCNN)*. 2017, pp. 439–446.

[56] Ahalya Prabhakar et al. “Multimodal Sensory Learning for Real-time, Adaptive Manipulation”. In: *arXiv preprint arXiv:2110.04634* (2021).

[57] Sachin Chitta et al. “Tactile sensing for mobile manipulation”. In: *IEEE Transactions on Robotics* 27.3 (2011), pp. 558–568.

[58] C. L. Chen, J. O. Snyder, and P. J. Ramadge. “Learning to Identify Container Contents through Tactile Vibration Signatures”. In: *2016 IEEE International Conference on Simulation, Modeling, and Programming for Autonomous Robots (SIMPAR)*. IEEE. 2016, pp. 43–48.

[59] Samuel Clarke et al. “Learning audio feedback for estimating amount and flow of granular material”. In: *Proceedings of Machine Learning Research* 87 (2018).

[60] Hannes P Saal, Jo-Anne Ting, and Sethu Vijayakumar. “Active estimation of object dynamics parameters with tactile sensors”. In: *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE. 2010, pp. 916–921.

[61] Carolyn Matl, Robert Matthew, and Ruzena Bajcsy. “Haptic perception of liquids enclosed in containers”. In: *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2019, pp. 7142–7149.

[62] Hung-Jui Huang, Xiaofeng Guo, and Wenzhen Yuan. “Understanding dynamic tactile sensing for liquid property estimation”. In: *arXiv preprint arXiv:2205.08771* (2022).

[63] Balakumar Sundaralingam and Tucker Hermans. “In-hand object-dynamics inference using tactile fingertips”. In: *IEEE Transactions on Robotics* 37.4 (2021), pp. 1115–1126.

[64] Xiaofeng Guo, Hung-Jui Huang, and Wenzhen Yuan. “Estimating properties of solid particles inside container using touch sensing”. In: *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2023, pp. 8985–8992.

[65] Manfred Eppe et al. “Deep Neural Object Analysis by Interactive Auditory Exploration with a Humanoid Robot”. In: *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2018, pp. 284–289.

[66] Matthias Kerzel et al. “NICO—Neuro-inspired companion: A developmental humanoid robot platform for multimodal interaction”. In: *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE. 2017, pp. 113–120.

[67] Shaowei Jin et al. “Open-environment robotic acoustic perception for object recognition”. In: *Frontiers in neurorobotics* 13 (2019), p. 96.

[68] Hongzhuo Liang et al. “Making sense of audio vibration for liquid height estimation in robotic pouring”. In: *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2019, pp. 5333–5339.

[69] Jivko Sinapov, Connor Schenck, and Alexander Stoytchev. “Learning Relational Object Categories using Behavioral Exploration and Multimodal Perception”. In: *2014 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2014, pp. 5691–5698.

[70] Josua Spisak, Matthias Kerzel, and Stefan Wermter. “Clarifying the Half Full or Half Empty Question: Multimodal Container Classification”. In: *International Conference on Artificial Neural Networks*. Springer. 2023, pp. 444–456.

[71] Pedro Piacenza, Daewon Lee, and Volkan Isler. “Pouring by feel: An analysis of tactile and proprioceptive sensing for accurate pouring”. In: *2022 International Conference on Robotics and Automation (ICRA)*. IEEE. 2022, pp. 10248–10254.

[72] Wim Meeussen et al. “Autonomous Door Opening and Plugging in with a Personal Robot”. In: *2010 IEEE International Conference on Robotics and Automation*. IEEE. 2010, pp. 729–736.

[73] OpenAI: Marcin Andrychowicz et al. “Learning dexterous in-hand manipulation”. In: *The International Journal of Robotics Research* 39.1 (2020), pp. 3–20.

[74] Z. Su et al. “Force Estimation and Slip Detection/Classification for Grip Control using a Biomimetic Tactile Sensor”. In: *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*. 2015, pp. 297–303.

[75] Morelle S Arian et al. “Using the BioTac as a Tumor Localization Tool”. In: *2014 IEEE Haptics Symposium (HAPTICS)*. IEEE. 2014, pp. 443–448.

[76] Steven Davis and Paul Mermelstein. “Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences”. In: *IEEE transactions on acoustics, speech, and signal processing* 28.4 (1980), pp. 357–366.

[77] Manfred Eppe, Tayfun Alpay, and Stefan Wermter. “Towards End-to-End Raw Audio Music Synthesis”. In: *International Conference on Artificial Neural Networks (ICANN)*. 2018, pp. 137–146.

[78] Erik Strahl et al. “Hear the Egg - Demonstrating Robotic Interactive Auditory Perception”. In: *International Conference on Intelligent Robots and Systems (IROS)*. 2018, p. 5041.

[79] Ronald Newbold Bracewell and Ronald N Bracewell. *The Fourier transform and its applications*. Vol. 31999. McGraw-Hill New York, 1986.

[80] Stanley Smith Stevens, John Volkmann, and Edwin Broomell Newman. “A scale for the measurement of the psychological magnitude pitch”. In: *The journal of the acoustical society of america* 8.3 (1937), pp. 185–190.

[81] James Bergstra, Daniel Yamins, and David Cox. “Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures”. In: *International conference on machine learning*. PMLR. 2013, pp. 115–123.

[82] Tayfun Alpay, Stefan Heinrich, and Stefan Wermter. “Learning multiple timescales in recurrent neural networks”. In: *Artificial Neural Networks and Machine Learning–ICANN 2016: 25th International Conference on Artificial Neural Networks, Barcelona, Spain, September 6-9, 2016, Proceedings, Part I 25*. Springer. 2016, pp. 132–139.

[83] Xavier Hinaut et al. “A recurrent neural network for multiple language acquisition: Starting with english and french”. In: *Proceedings of the NIPS Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches (CoCo 2015)*. 2015.

[84] Sepp Hochreiter and Jürgen Schmidhuber. "Long Short-Term Memory". In: *Neural computation* 9.8 (1997), pp. 1735–1780.

[85] Kyunghyun Cho et al. "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation". In: *arXiv preprint arXiv:1406.1078* (2014).

[86] Diederik P. Kingma and Jimmy Lei Ba. "Adam: a Method for Stochastic Optimization". In: *International Conference on Learning Representations (ICLR)*. 2015.

[87] Manfred Eppe, Sven Magg, and Stefan Wermter. "Curriculum Goal Masking for Continuous Deep Reinforcement Learning". In: *International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*. 2019, pp. 183–188.

[88] Manfred Eppe, Phuong D. H. Nguyen, and Stefan Wermter. "From Semantics to Execution: Integrating Action Planning with Reinforcement Learning for Robotic Causal Problem-solving". In: *Frontiers in Robotics and AI* 6 (2019).

[89] Lael U Odhner and Aaron M Dollar. "Dexterous manipulation with underactuated elastic hands". In: *2011 IEEE International Conference on Robotics and Automation*. IEEE. 2011, pp. 5254–5260.

[90] Lael U Odhner and Aaron M Dollar. "Stable, open-loop precision manipulation with underactuated hands". In: *The International Journal of Robotics Research* 34.11 (2015), pp. 1347–1360.

[91] Robotiq Inc. *Robotiq 3-Finger Adaptive Robot Gripper Instruction Manual*. English. Version evision 2021-05-19. Robotiq Inc. 101 pp. May 19, 2021.

[92] C. Farell Winder. "Shaft Angle Encoders Afford High Accuracy". In: *Electronic Infustries* 18.10 (1959), pp. 76–80.

[93] GA Woolvet. "Digital transducers". In: *Journal of Physics E: Scientific Instruments* 15.12 (1982), p. 1271.

[94] Geoffrey Boyes. "Synchro and Resolver Conversion". In: (1980).

[95] Azizul Othman et al. "Design and development of an adjustable angle sensor based on rotary potentiometer for measuring finger flexion". In: *2016 6th IEEE International Conference on Control System, Computing and Engineering (ICCSCE)*. IEEE. 2016, pp. 569–574.

[96] RS Popovic. "The vertical Hall-effect device". In: *IEEE Electron Device Letters* 5.9 (1984), pp. 357–358.

[97] Sidwell Nkosi, Lloyd Patsika, and Lesedi Masisi. "Development of a low cost rotor position sensor". In: *2020 International SAUPEC/RobMech/PRASA Conference*. IEEE. 2020, pp. 1–5.

[98] Nuno Ferrete Ribeiro and Cristina P Santos. "Inertial measurement units: A brief state of the art on gait analysis". In: *2017 IEEE 5th Portuguese Meeting on Bioengineering (ENBENG)*. IEEE. 2017, pp. 1–4.

[99] Avishai Sintov et al. "Learning a state transition model of an underactuated adaptive hand". In: *IEEE Robotics and Automation Letters* 4.2 (2019), pp. 1287–1294.

[100] Herke Van Hoof et al. “Learning robot in-hand manipulation with tactile features”. In: *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*. IEEE. 2015, pp. 121–127.

[101] Wael M Mohammed et al. “Training an Under-actuated Gripper for Grasping Shallow Objects Using Reinforcement Learning”. In: *2020 IEEE Conference on Industrial Cyberphysical Systems (ICPS)*. Vol. 1. IEEE. 2020, pp. 493–498.

[102] Mingfang Liu et al. “Reinforcement learning control of a humanoid robotic hand actuated by shape memory alloy”. In: *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science* 235.21 (2021), pp. 5736–5744.

[103] Gabriele Maria Achilli, Silvia Logozzo, and Maria Cristina Valigi. “An educational test rig for kinesthetic learning of mechanisms for underactuated robotic hands”. In: *Robotics* 11.5 (2022), p. 115.

[104] Takahiro Matsuno, Zhongkui Wang, and Shinichi Hirai. “Grasping state estimation of printable soft gripper using electro-conductive yarn”. In: *Robotics and Biomimetics* 4.1 (2017), pp. 1–11.

[105] Tianliang Li, Liang Qiu, and Hongliang Ren. “Distributed curvature sensing and shape reconstruction for soft manipulators with irregular cross sections based on parallel dual-FBG arrays”. In: *IEEE/ASME Transactions on Mechatronics* 25.1 (2019), pp. 406–417.

[106] Shinichi Hirai et al. “Measuring McKibben actuator shrinkage using fiber sensor”. In: *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE. 2015, pp. 628–633.

[107] Lin Yan et al. “Curvature Estimation of Soft Grippers Based on a Novel High-Stretchable Strain Sensor with Worm-Surface-like Microstructures”. In: *IEEE Sensors Journal* (2023).

[108] Snehal Dikhale et al. “Visuotactile 6d pose estimation of an in-hand object using vision and tactile sensor data”. In: *IEEE Robotics and Automation Letters* 7.2 (2022), pp. 2148–2155.

[109] Joao Bimbo et al. “In-hand object pose estimation using covariance-based tactile to geometry matching”. In: *IEEE Robotics and Automation Letters* 1.1 (2016), pp. 570–577.

[110] Osher Azulay, Inbar Ben-David, and Avishai Sintov. “Learning Haptic-Based Object Pose Estimation for In-Hand Manipulation Control With Underactuated Robotic Hands”. In: *IEEE Transactions on Haptics* 16.1 (2022), pp. 73–85.

[111] Bowen Wen et al. “Robust, occlusion-aware pose estimation for objects grasped by adaptive hands”. In: *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2020, pp. 6210–6217.

[112] Hailiang Meng et al. “Optimal Design of Linkage-Driven Underactuated Hand for Precise Pinching and Powerful Grasping”. In: *IEEE Robotics and Automation Letters* 9.4 (2024), pp. 3475–3482.

[113] John Wang and Edwin Olson. "AprilTag 2: Efficient and robust fiducial detection". In: *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2016, pp. 4193–4198.

[114] Artem Kargov et al. "Modularly designed lightweight anthropomorphic robot hand". In: *2006 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*. IEEE. 2006, pp. 155–159.

[115] Thomas Seel, Jorg Raisch, and Thomas Schauer. "IMU-based joint angle measurement for gait analysis". In: *Sensors* 14.4 (2014), pp. 6891–6909.

[116] Hasim Sak, Andrew W Senior, and Françoise Beaufays. "Long short-term memory recurrent neural network architectures for large scale acoustic modeling". In: *Interspeech 2014* (2014).

[117] Takuya Akiba et al. "Optuna: A next-generation hyperparameter optimization framework". In: *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 2019, pp. 2623–2631.

[118] Berk Calli et al. "Benchmarking in manipulation research: Using the Yale-CMU-Berkeley object and model set". In: *IEEE Robotics & Automation Magazine* 22.3 (2015), pp. 36–52.

[119] R. C. Goertz and William M. Thompson. "Electronically Controlled Manipulator". In: 1954.

[120] Ken Goldberg et al. "Desktop teleoperation via the world wide web". In: *Proceedings of 1995 IEEE International Conference on Robotics and Automation*. Vol. 1. IEEE. 1995, pp. 654–659.

[121] Craig Sayers and C Sayers. *Remote control robotics*. Springer, 1999.

[122] Mahdi Tavakoli et al. *Haptics for Teleoperated Surgical Robotic Systems*. Vol. 1. Singapur: World Scientific, 2008.

[123] Chenguang Yang et al. "Personalized Variable Gain Control With Tremor Attenuation for Robot Teleoperation". In: *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 48.10 (2017), pp. 1759–1770.

[124] Shuang Li et al. "A dexterous hand-arm teleoperation system based on hand pose estimation and active vision". In: *IEEE Transactions on Cybernetics* 54.3 (2022), pp. 1417–1428.

[125] Shuang Li et al. "A mobile robot hand-arm teleoperation system by vision and imu". In: *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2020, pp. 10900–10906.

[126] Bin Fang et al. "A robotic hand-arm teleoperation system using human arm/hand with a novel data glove". In: *2015 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. IEEE. 2015, pp. 2483–2488.

[127] Samuel B. Schorr and Allison M. Okamura. “Fingertip Tactile Devices for Virtual Object Manipulation and Exploration”. In: *CHI ’17: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. Denver, CO, USA: Association for Computing Machinery, 2017, pp. 3115–3119.

[128] Tom Carter et al. “UltraHaptics: multi-point mid-air haptic feedback for touch surfaces”. In: *Proceedings of the 26th annual ACM symposium on User interface software and technology*. 2013, pp. 505–514.

[129] Xuxin Cheng et al. “Open-television: Teleoperation with immersive active visual feedback”. In: *arXiv preprint arXiv:2407.01512* (2024).

[130] Tianhao Zhang et al. “Deep imitation learning for complex manipulation tasks from virtual reality teleoperation”. In: *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE. 2018, pp. 5628–5635.

[131] Luigi Penco et al. “Mixed reality teleoperation assistance for direct control of humanoids”. In: *IEEE Robotics and Automation Letters* (2024).

[132] Jean Chagas Vaz, Dylan Wallace, and Paul Y Oh. “Humanoid loco-manipulation of pushed carts utilizing virtual reality teleoperation”. In: *ASME International Mechanical Engineering Congress and Exposition*. Vol. 85628. American Society of Mechanical Engineers. 2021, V07BT07A027.

[133] Joao Ramos and Sangbae Kim. “Humanoid dynamic synchronization through whole-body bilateral feedback teleoperation”. In: *IEEE Transactions on Robotics* 34.4 (2018), pp. 953–965.

[134] Kourosh Darvish et al. “Whole-body geometric retargeting for humanoid robots”. In: *2019 IEEE-RAS 19th International Conference on Humanoid Robots (Humanoids)*. IEEE. 2019, pp. 679–686.

[135] Yasuhiro Ishiguro et al. “Bilateral humanoid teleoperation system using whole-body exoskeleton cockpit TABLIS”. In: *IEEE Robotics and Automation Letters* 5.4 (2020), pp. 6419–6426.

[136] A.W.W. Yew, S.K. Ong, and A.Y.C. Nee. “Immersive Augmented Reality Environment for the Teleoperation of Maintenance Robots”. In: *Procedia Cirp* 61 (2017), pp. 305–310.

[137] Michael F. Zaeh and Wolfgang Vogl. “Interactive Laser-Projection for Programming Industrial Robots”. In: *2006 IEEE/ACM International Symposium on Mixed and Augmented Reality*. Santa Barbara, CA, USA: IEEE, 2006, pp. 125–128.

[138] Wesley P. Chan et al. “A Multimodal System using Augmented Reality, Gestures, and Tactile Feedback for Robot Trajectory Programming and Execution”. In: *Proceedings of the ICRA Workshop on Robotics in Virtual Reality*. Brisbane, Australia: IEEE, 2018, pp. 21–25.

[139] Jennifer L. Burke et al. “Comparing the Effects of Visual-Auditory and Visual-Tactile Feedback on User Performance: A Meta-analysis”. In: *ICMI ’06: Proceedings of the 8th international conference on Multimodal interfaces*. Banff, Alberta, Canada: Association for Computing Machinery, 2006, pp. 108–117.

[140] Iris Herbst and Jörg Stark. “Comparing Force Magnitudes by Means of Vibro-tactile, Auditory, and Visual Feedback”. In: *IEEE International Workshop on Haptic Audio Visual Environments and their Applications*. Ottawa, ON, Canada: IEEE, 2005, 5–pp.

[141] Minghui Sun, Xiangshi Ren, and Xiang Cao. “Effects of Multimodal Error Feedback on Human Performance in Steering Tasks”. In: *Journal of Information Processing* 18 (2010), pp. 284–292.

[142] Matthew S. Prewett et al. “The Benefits of Multimodal Information: A Meta-Analysis Comparing Visual and Visual-Tactile Feedback”. In: *ICMI ’06: Proceedings of the 8th international conference on Multimodal interfaces*. Banff, Alberta, Canada: Association for Computing Machinery, 2006, pp. 333–338.

[143] M Zhou et al. “Effect of haptic feedback in laparoscopic surgery skill acquisition”. In: *Surgical endoscopy* 26 (2012), pp. 1128–1134.

[144] Issam El Rassi and Jean-Michel El Rassi. “A review of haptic feedback in tele-operated robotic surgery”. In: *Journal of medical engineering & technology* 44.5 (2020), pp. 247–254.

[145] Angelica I. Aviles-Rivero et al. “Sensory Substitution for Force Feedback Recovery: A Perception Experimental Study”. In: *ACM Transactions on Applied Perception* 15.3 (Aug. 2018), pp. 1–19.

[146] Rajni V Patel, S Farokh Atashzar, and Mahdi Tavakoli. “Haptic feedback and force-based teleoperation in surgical robotics”. In: *Proceedings of the IEEE* 110.7 (2022), pp. 1012–1027.

[147] Anna Karvouniari et al. “An approach for exoskeleton integration in manufacturing lines using Virtual Reality techniques”. In: *Procedia CIRP* 78 (2018), pp. 103–108.

[148] Antonio Frisoli et al. “A force-feedback exoskeleton for upper-limb rehabilitation in virtual reality”. In: *Applied Bionics and Biomechanics* 6.2 (2009), pp. 115–126.

[149] Omar Mubin et al. “Exoskeletons with virtual reality, augmented reality, and gamification for stroke patients’ rehabilitation: systematic review”. In: *JMIR rehabilitation and assistive technologies* 6.2 (2019), e12010.

[150] Alberto Topini et al. “Variable admittance control of a hand exoskeleton for virtual reality-based rehabilitation tasks”. In: *Frontiers in neurorobotics* 15 (2022), p. 789743.

[151] Michail Kalaitzakis et al. “Experimental comparison of fiducial markers for pose estimation”. In: *2020 International Conference on Unmanned Aircraft Systems (ICUAS)*. IEEE. 2020, pp. 781–789.

[152] Sebastian Starke, Norman Hendrich, and Jianwei Zhang. “A memetic evolutionary algorithm for real-time articulated kinematic motion”. In: *2017 IEEE Congress on Evolutionary Computation (CEC)*. IEEE. 2017, pp. 2473–2479.

[153] Joseph M Romano et al. “Human-inspired robotic grasp control with tactile sensing”. In: *IEEE Transactions on Robotics* 27.6 (2011), pp. 1067–1079.

[154] Sandra G. Hart and Lowell E. Staveland. “Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research”. In: *Advances in Psychology*. Vol. 52. Elsevier, 1988, pp. 139–183.

[155] John Brooke. “SUS-A quick and dirty usability scale”. In: *Usability Evaluation In Industry* 189.194 (1996), pp. 4–7.

[156] Robert S. Kennedy et al. “Simulator Sickness Questionnaire: An Enhanced Method for Quantifying Simulator Sickness”. In: *The international journal of aviation psychology* 3.3 (1993), pp. 203–220.

[157] Marc Hassenzahl. “The Interplay of Beauty, Goodness, and Usability in Interactive Products”. In: *Human–Computer Interaction* 19.4 (2004), pp. 319–349.

[158] Aaron Bangor, Philip Kortum, and James Miller. “Determining what individual SUS scores mean: Adding an adjective rating scale”. In: *Journal of usability studies* 4.3 (2009), pp. 114–123.

[159] Roland Sigrist et al. “Augmented visual, auditory, haptic, and multimodal feedback in motor learning: A review”. In: *Psychonomic Bulletin & Review* 20.1 (2013), pp. 21–53.

[160] Oscar Javier Ariza Nunez et al. “Holitouch: Conveying holistic touch illusions by combining pseudo-haptics with tactile and proprioceptive feedback during virtual interaction with 3duis”. In: *Frontiers in Virtual Reality* 3 (2022), p. 879845.

[161] Kotaro Yamaguchi et al. “A Non-grounded and Encountered-type Haptic Display Using a Drone”. In: *Proceedings of the 2016 Symposium on Spatial User Interaction*. 2016, pp. 43–46.

[162] Victor Mercado, Maud Marchai, and Anatole Lécuyer. “Design and evaluation of interaction techniques dedicated to integrate encountered-type haptic displays in virtual environments”. In: *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE. 2020, pp. 230–238.

[163] Bruno Araujo et al. “Snake charmer: Physically enabling virtual objects”. In: *Proceedings of the TEI’16: Tenth International Conference on Tangible, Embedded, and Embodied Interaction*. 2016, pp. 218–226.

[164] Yasuyoshi Yokokohji, Ralph L Hollis, and Takeo Kanade. “What you can see is what you can feel-development of a visual/haptic interface to virtual environment”. In: *Proceedings of the IEEE 1996 Virtual Reality Annual International Symposium*. IEEE. 1996, pp. 46–53.

[165] Scott Devine, Karen Rafferty, and Robin Ferguson. “A Point Contact Encounter Haptic Solution with the HTC VIVE and Baxter Robot”. In: *25th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision*. 2017.

[166] Karen Rafferty, Scott Devine, and Daniel Brice. “A Novel Force Feedback Haptics System with Applications in Phobia Treatment”. In: *25th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision*. 2017.

[167] Victor Mercado, Maud Marchal, and Anatole Lécuyer. "Entropia: towards infinite surface haptic displays in virtual reality using encountered-type rotating props". In: *IEEE transactions on visualization and computer graphics* 27.3 (2019), pp. 2237–2243.

[168] Yaesol Kim, Hyun Jung Kim, and Young J Kim. "Encountered-type haptic display for large VR environment using per-plane reachability maps". In: *Computer Animation and Virtual Worlds* 29.3-4 (2018), e1814.

[169] Yaesol Kim et al. "Synthesizing the roughness of textured surfaces for an encountered-type haptic display using spatiotemporal encoding". In: *IEEE Transactions on Haptics* 14.1 (2020), pp. 32–43.

[170] Victor Rodrigo Mercado et al. "Watch out for the Robot! Designing Visual Feedback Safety Techniques When Interacting With Encountered-Type Haptic Displays". In: *Frontiers in Virtual Reality* 3 (2022), p. 928517.

[171] Sergio Portolés Díez et al. "A Novel Method for Surface Exploration by 6-DOF Encountered-Type Haptic Display Towards Virtual Palpation". In: *IEEE Transactions on Haptics* 14.3 (2021), pp. 577–590.

[172] Takuro Furumoto et al. "Midair balloon interface: A soft and lightweight midair object for proximate interactions". In: *The 34th Annual ACM Symposium on User Interface Software and Technology*. 2021, pp. 783–795.

[173] Shun Suzuki et al. "AUTD3: Scalable airborne ultrasound tactile display". In: *IEEE Transactions on Haptics* 14.4 (2021), pp. 740–749.

[174] Takuro Furumoto et al. "Encounter-type haptic feedback system using an acoustically manipulated floating object". In: *Haptic Interaction: Perception, Devices and Algorithms 3*. Springer. 2019, pp. 183–186.

[175] Hsin-Yu Huang et al. "Haptic-go-round: A surrounding platform for encounter-type haptics in virtual reality experiences". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 2020, pp. 1–10.

[176] Shun Yamaguchi et al. "An encounter type VR system aimed at exhibiting wall material samples for show house". In: *Proceedings of the 2018 ACM International Conference on Interactive Surfaces and Spaces*. 2018, pp. 321–326.

[177] Teppei Tsujita et al. "Development of a surgical simulator for training retraction of tissue with an encountered-type haptic interface using mr fluid". In: *2018 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. IEEE. 2018, pp. 898–903.

[178] Teppei Tsujita et al. "Feedback control of an encountered-type haptic interface using MR fluid and servomotors for displaying cutting and restoring force of soft tissue". In: *Advanced Robotics* 36.24 (2022), pp. 1327–1338.

[179] Teppei Tsujita et al. "Design and evaluation of an encountered-type haptic interface using MR fluid for surgical simulators". In: *Advanced Robotics* 27.7 (2013), pp. 525–540.

[180] Mehmet Murat Aygün et al. "Visuo-haptic mixed reality simulation using unbound handheld tools". In: *Applied Sciences* 10.15 (2020), p. 5344.

[181] Daniel Schneider et al. “Reconviguration: Reconfiguring physical keyboards in virtual reality”. In: *IEEE transactions on visualization and computer graphics* 25.11 (2019), pp. 3190–3201.

[182] Eric J Gonzalez, Parastoo Abtahi, and Sean Follmer. “Reach+ extending the reachability of encountered-type haptics devices through dynamic redirection in vr”. In: *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. 2020, pp. 236–248.

[183] Hiroo Iwata et al. “Circulafloor [locomotion interface]”. In: *IEEE Computer Graphics and Applications* 25.1 (2005), pp. 64–67.

[184] Alexa F Siu et al. “shapeShift: A mobile tabletop shape display for tangible and haptic interaction”. In: *Adjunct Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*. 2017, pp. 77–79.

[185] Daniel Fitzgerald and Hiroshi Ishii. “Mediate: A spatial tangible interface for mixed reality”. In: *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. 2018, pp. 1–6.

[186] Ryo Suzuki et al. “Hapticbots: Distributed encountered-type haptics for vr with multiple shape-changing mobile robots”. In: *The 34th Annual ACM Symposium on User Interface Software and Technology*. 2021, pp. 1269–1281.

[187] Lawrence H Kim and Sean Follmer. “Swarmhaptics: Haptic display with swarm robots”. In: *Proceedings of the 2019 CHI conference on human factors in computing systems*. 2019, pp. 1–13.

[188] Parastoo Abtahi et al. “Beyond the force: Using quadcopters to appropriate objects and the environment for haptics in virtual reality”. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 2019, pp. 1–13.

[189] Muhammad Abdullah et al. “HapticDrone: An encountered-type kinesthetic haptic interface with controllable force feedback: Example of stiffness and weight rendering”. In: *2018 IEEE Haptics Symposium (HAPTICS)*. IEEE. 2018, pp. 334–339.

[190] Mudassir Ibrahim Awan, Ahsan Raza, and Seokhee Jeon. “DroneHaptics: Encountered-Type Haptic Interface Using Dome-Shaped Drone for 3-DoF Force Feedback”. In: *2023 20th International Conference on Ubiquitous Robots (UR)*. IEEE. 2023, pp. 195–200.

[191] Antonio Gomes et al. “Bitdrones: Towards using 3d nanocopter displays as interactive self-levitating programmable matter”. In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 2016, pp. 770–780.

[192] Photchara Ratsamee et al. “Uhd: Unconstrained haptic display using a self-localized quadrotor”. In: *Proceedings of AROB 24th 2019 International Symposium on Artificial Life and Robotics*. 2019.

[193] Jinhyeok Oh et al. “A liquid metal based multimodal sensor and haptic feedback device for thermal and tactile sensation generation in virtual reality”. In: *Advanced Functional Materials* 31.39 (2021), p. 2007772.

[194] Mohammed Al-Sada et al. “HapticSnakes: multi-haptic feedback wearable robots for immersive virtual reality”. In: *virtual reality* 24 (2020), pp. 191–209.

[195] Arata Horie et al. “Encounteredlimbs: A room-scale encountered-type haptic presentation using wearable robotic arms”. In: *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*. IEEE. 2021, pp. 260–269.

[196] Jon Dobson. “Remote control of cellular behaviour with magnetic nanoparticles”. In: *Nature nanotechnology* 3.3 (2008), pp. 139–143.

[197] CS Lee, H Lee, and RM Westervelt. “Microelectromagnets for the control of magnetic nanoparticles”. In: *Applied physics letters* 79.20 (2001), pp. 3308–3310.

[198] Roozbeh Abedini-Nassab, Mahrad Pouryosef Miandoab, and Merivan Şaşmaz. “Microfluidic Synthesis, Control, and Sensing of Magnetic Nanoparticles: A Review”. In: *Micromachines* 12.7 (2021), p. 768.

[199] Jie Wu et al. “Size-selective separation of magnetic nanospheres in a microfluidic channel”. In: *Microfluidics and Nanofluidics* 21.3 (2017), pp. 1–12.

[200] Eriola-Sophia Shanko et al. “Microfluidic Magnetic Mixing at Low Reynolds Numbers and in Stagnant Fluids”. In: *Micromachines* 10.11 (2019), p. 731.

[201] Olivier Lefebvre et al. “Reusable embedded microcoils for magnetic nano-beads trapping in microfluidics: magnetic simulation and experiments”. In: *Micromachines* 11.3 (2020), p. 257.

[202] Vania Silverio et al. “Manipulation of magnetic beads with thin film microelectromagnet traps”. In: *Micromachines* 10.9 (2019), p. 607.

[203] Kazufumi Nomura, Kazuyuki Morisaki, and Yoshinori Hirata. “Magnetic control of arc plasma and its modelling”. In: *Welding in the World* 53 (2009), R181–R187.

[204] Juanyan Miao et al. “Magnetic controlled arc welding technology: a review”. In: *Rapid Prototyping Journal* 30.9 (2024), pp. 1929–1955.

[205] Arthur W Mahoney and Jake J Abbott. “Five-degree-of-freedom manipulation of an untethered magnetic device in fluid using a single permanent magnet with application in stomach capsule endoscopy”. In: *The International Journal of Robotics Research* 35.1-3 (2016), pp. 129–147.

[206] Alberto Arezzo et al. “Experimental assessment of a novel robotically-driven endoscopic capsule compared to traditional colonoscopy”. In: *Digestive and Liver Disease* 45.8 (2013), pp. 657–662.

[207] James W Martin et al. “Robotic autonomy for magnetic endoscope biopsy”. In: *IEEE transactions on medical robotics and bionics* 4.3 (2022), pp. 599–607.

[208] Weiyuan Chen, Jianbo Sui, and Chengyong Wang. “Magnetically actuated capsule robots: A review”. In: *IEEE Access* 10 (2022), pp. 88398–88420.

[209] Anatole Lécuyer. “Simulating Haptic Feedback Using Vision: A Survey of Research and Applications of Pseudo-Haptic Feedback”. In: *Presence: Teleoperators and Virtual Environments* 18.1 (2009), pp. 39–53.

[210] Frank Biocca, Jin Kim, and Yung Choi. "Visual Touch in Virtual Environments: An Exploratory Study of Presence, Multimodal Interfaces, and Cross-Modal Sensory Illusions". In: *Presence* 10.3 (2001), pp. 247–265.

[211] Karin Nieuwenhuizen et al. "Insights from Dividing 3D Goal-Directed Movements into Meaningful Phases". In: *IEEE Computer Graphics and Applications* 29.6 (2009), pp. 44–53.

[212] Thomas W Schubert. "The sense of presence in virtual environments: A three-component scale measuring spatial presence, involvement, and realness." In: *Z. für Medienpsychologie* 15.2 (2003), pp. 69–71.

[213] Martin Schrepp, Andreas Hinderks, and Jörg Thomaschewski. "Design and Evaluation of a Short Version of the User Experience Questionnaire (UEQ-S)". In: *International Journal of Interactive Multimedia and Artificial Intelligence, 4 (6), 103-108.* (2017).

[214] Andreas Hinderks, Martin Schrepp, and Jörg Thomaschewski. "A Benchmark for the Short Version of the User Experience Questionnaire". In: *WEBIST*. 2018, pp. 373–377.

[215] Mahdi Azmandian et al. "Haptic retargeting: Dynamic repurposing of passive haptics for enhanced virtual reality experiences". In: *Proceedings of the 2016 chi conference on human factors in computing systems*. 2016, pp. 1968–1979.

[216] Eric J Gonzalez et al. "A model predictive control approach for reach redirection in virtual reality". In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 2022, pp. 1–15.

[217] André Zenner, Kristin Ullmann, and Antonio Krüger. "Combining Dynamic Passive Haptics and Haptic Retargeting for Enhanced Haptic Feedback in Virtual Reality". In: *IEEE Transactions on Visualization and Computer Graphics* 27.5 (2021), pp. 2627–2637.

[218] A Vaswani. "Attention is all you need". In: *Advances in Neural Information Processing Systems* (2017).

[219] Jingyi Zhang et al. "Vision-language models for vision tasks: A survey". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).

[220] Danny Driess et al. "Palm-e: An embodied multimodal language model". In: *arXiv preprint arXiv:2303.03378* (2023).

Web References

[W1] Dave Hershberger, David Gossow, Josh Faust, William Woodall. *rviz*. Accessed: December 2024. URL: http://wiki.ros.org/rviz.

[W2] Nate Koenig, Andrew Howard. *gazebo*. Accessed: December 2024. URL: https://wiki.ros.org/gazebo.

[W3] Davide Faconti. *plotjuggler*. Accessed: December 2024. URL: https://wiki.ros.org/plotjuggler.

[W4] Unity Technologies. *Unity Real-Time Development Platform | 3D, 2D VR & AR Engine*. Accessed: October 2024. URL: https://unity.com/.

[W5] Martin Bischoff. *ROS#*. Accessed: October 2024. URL: https://github.com/siemens/ros-sharp/.

[W6] Unity-Technologies. *Unity-Robotics-Hub*. Accessed: October 2024. URL: https://github.com/Unity-Technologies/Unity-Robotics-Hub.

[W7] The Shadow Robot Company. *The Shadow Dexterous Hand*. Accessed: September 2024. URL: https://www.shadowrobot.com/dexterous-hand-series/.

[W8] François Chollet et al. *Keras*. Accessed: September 2024. 2015. URL: https://keras.io.

[W9] Robotiq. *3-Finger Adaptive Robot Gripper*. Accessed: November 2024. URL: https://robotiq.com/products/3-finger-adaptive-robot-gripper.

[W10] PhaseSpace Inc. *The Impulse X2E Motion Capture System*. Accessed: October 2024. URL: https://www.phasespace.com/x2e-motion-capture/.

[W11] NaturalPoint, Inc. *OptiTrack*. Accessed: October 2024. URL: https://www.optitrack.com/.

[W12] Microsoft. *HoloLens 2—Overview, Features, and Specs | Microsoft HoloLens*. Accessed: October 2024. URL: https://www.microsoft.com/en-us/hololens/hardware.

[W13] Meta. *Get started with Meta Quest 2*. Accessed: October 2024. URL: https://www.meta.com/de/quest/products/quest-2/.

[W14] Mixed Reality Toolkit Organization. *MixedRealityToolkit-Unity*. Accessed: October 2024. URL: https://github.com/MixedRealityToolkit/MixedRealityToolkit-Unity.

[W15] Meta Horizon. *Meta XR All-in-One SDK (UPM)*. Accessed: October 2024. URL: https://developers.meta.com/horizon/downloads/package/meta-xr-sdk-all-in-one-upm/.

[W16] Meta Horizon. *Oculus Integration SDK*. Accessed: October 2024. URL: https://developers.meta.com/horizon/downloads/package/unity-integration/.

[W17] Microsoft. *Hand tracking*. Accessed: October 2024. URL: https://microsoft.github.io/MixedRealityToolkit-Unity/Documentation/Input/HandTracking.html.

[W18] Meta. *Controller Input and Tracking Overview*. Accessed: October 2024. URL: https://developers.meta.com/horizon/documentation/unity/unity-ovrinput.

[W19] Tokyo Opensource Robotics Kyokai Association. *jog_control*. Accessed: October 2024. URL: https://github.com/tork-a/jog_control.

[W20] UNIS Technology. *The Atari Pong® Coffee Table*. Accessed: December 2024. URL: https://www.kickstarter.com/projects/1461917284/play-atari-pong-in-your-coffee-table.

[W21] *Marlin Firmware Open Source RepRap Driver*. Accessed: November 2024. URL: https://marlinfw.org.

List of Figures

2.1 Haptic-Tactile-Process 10
2.2 Location and classification of mechanoreceptors 11
2.3 Virtuality Continuum 15

3.1 Setup for classification experiment 22
3.2 The BioTac sensor 23
3.3 3D printed containers 24
3.4 Filter steps for audio and tactile data 26
3.5 MFCC spectrum of audio signal 28
3.6 Architecture of classification networks 29
3.7 Classification accuracy plot 31

4.1 The Robotiq 3-Finger Adaptive Robot Gripper 36
4.2 Setup of state estimation experiment 39
4.3 The motion angles of Robotiq 3-Finger Gripper 40
4.4 Tactile finger sensors with AprilTags 41
4.5 Layers of custom built tactile sensor array 41
4.6 An exemplary grasp with the Robotiq gripper 42
4.7 Object set for state estimation experiment 45
4.8 Estimation results of the end state of grasping a Pringles can 46
4.9 Joint state error distribution of closing motion 48
4.10 Joint state error distribution of end state 48
4.11 Joint state deviation to ground truth 49
4.12 Joint states of grasping motion and tactile signals 49

5.1 Teleoperation process graph 57
5.2 Setup of teleoperation experiment 57
5.3 Hand tracking and controller button assignment 58
5.4 Teleoperation system for different robot platforms 60
5.5 The teleoperation approach in a dual arm scenario 61
5.6 Cups used in the teleoperation user study 64
5.7 The three virtual replacements of tactile feedback 65
5.8 Average force applied during the teleoperation experiment 68
5.9 Average force applied to the cups during the teleoperation experiment 69
5.10 Hedonic and Pragmatic Quality of teleoperation system 70

6.1 The Encountered-Type Haptic Display during the *Whac-A-Mole* user study . . 77
6.2 Hardware design of the haptic display 78

6.3 The custom build end effector of the ETHD 80
6.4 Model designs of the ETHD 81
6.5 Gazebo simulation of the ETHD 83
6.6 The 3D printed container for the payload and velocity experiment 84
6.7 The neural network architecture used for position estimation 85
6.8 Recovery visualization of ETHD 86
6.9 The graphic shows training data for the object positioning network 89
6.10 *Whac-A-Mole* user study setup 91
6.11 The five phases of a movement according to Nieuwenhuizen 92
6.12 Three exemplary velocity plots 94
6.13 The movement duration distribution over the workspace 96

8.1 Estimation results of the end state of grasping a ball 103
8.2 Estimation results of the end state of grasping a drill 104
8.3 Estimation results of the end state of grasping a spray 104
8.4 Estimation results of the end state of grasping a wooden block 105

List of Tables

2.1 Tactile technologies 13

3.1 Pill classes included in the data set and used in the experiment 25
3.2 Confusion matrices of classification 31

4.1 Analytical joint state calculation 43
4.2 Joint state estimation accuracies in radians 47

5.1 Average Execution Time & Success Rate 67
5.2 NASA TLX scores. 69

6.1 Hardware Specification 80
6.2 Maximum payload of the haptic display 87
6.3 Maximum velocity of the haptic display 88
6.4 Repetition accuracy of the table 88
6.5 Average Duration of Movement Phases 95
6.6 Minimum Duration of Movement Phases 95

Acronyms

AR Augmented Reality

AUPA Airborne Ultrasound Phased Arrays

AV Augmented Virtuality

CNN Convolutional Neural Network

DDS Data Distribution Service

DoF Degrees of Freedom

ETHD Encountered-Type Haptic Display

FDM Fused Deposition Modeling

GP Gaussian Processes

GRU Gated Recurrent Unit

HMD Head-Mounted Display

HQ Hedonic Quality

IMU Inertial Measurement Unit

IPQ Igroup Presence Questionnaire

KKNN kernel k-nearest neighbor

kNN k-nearest neighbor

LLM large language model

LSTM Long Short-Term Memory

MFCC Mel Frequency Cepstral Coefficients

MLP Multi-Layer Perception

MR Mixed Reality

NASA TLX NASA Task Load Index

OSKKNN kernel k-nearest neighbor algorithm in an open environment

PLA Polylactic Acid

PQ Pragmatic Quality

PTFE Polytetrafluoroethylene

QDA Quadratic Discriminant Analysis

ReLU Rectified Linear Unit

RNN Recurrent Neural Network

RMSE Root Mean Squared Error

ROS Robot Operating System

SRN Simple Recurrent Networks

SSQ Simulator Sickness Questionnaire

SUS System Usability Scale

SVM Support Vector Machine

tanh hyperbolic tangent

UAV Unmanned Aerial Vehicles

UEQ-S User Experience Questionnaire

URDF Unified Robot Description Format

UWP Universal Windows Platform

VC Virtuality Continuum

VE Virtual Environment

VLM vision language model

VR Virtual Reality

XR Extended Reality

Eidesstattliche Erklärung

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Dissertationsschrift selbst verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Sofern im Zuge der Erstellung der vorliegenden Dissertationsschrift generative Künstliche Intelligenz (gKI) basierte elektronische Hilfsmittel verwendet wurden, versichere ich, dass meine eigene Leistung im Vordergrund stand und dass eine vollständige Dokumentation aller verwendeten Hilfsmittel gemäß der Guten wissenschaftlichen Praxis vorliegt. Ich trage die Verantwortung für eventuell durch die gKI generierte fehlerhafte oder verzerrte Inhalte, fehlerhafte Referenzen, Verstöße gegen das Datenschutz- und Urheberrecht oder Plagiate

Hamburg, den 19.12.2024 Yannick Jonetzko