# Representing Constraints in Human Bimanual Manipulation for Transfer to Humanoid Robots

Zur Erlangung des akademischen Grades eines

Doktors der Ingenieurwissenschaften

von der KIT-Fakultät für Informatik
des Karlsruher Instituts für Technologie (KIT)

genehmigte
Dissertation

von

Franziska Krebs

aus Heidelberg

Tag der mündlichen Prüfung: 06. Mai 2025

1. Referent: Prof. Dr.-Ing. Tamim Asfour
2. Referent: Prof. Dr. Oliver Brock

# Abstract

Bimanual manipulation is a fundamental skill humans acquire through years of practice, enabling complex coordination for tasks ranging from daily activities to expert-level performances. In robotics, achieving purposeful bimanual coordination remains a challenge due to intricate temporal and spatial constraints. However, mastering this skill is crucial for humanoid robots to enhance task efficiency and enable effective programming by demonstration (PbD) and imitation learning (IL).

This thesis develops methods for learning task models for bimanual manipulation from human demonstrations. The work consists of three parts: The first is the development of a taxonomy for bimanual manipulation based on human motion analysis and findings from neuroscience and robotics. In the second part, methods for recognizing the categories defined by the taxonomy are developed and evaluated using new data sets for bimanual manipulation in a household context. Finally, the implementation of the constraints resulting from the respective categories in control strategies for humanoid robots is shown.

**Bimanual Manipulation Taxonomy** In robotics, especially in robotic grasping, taxonomies are a common technique to cope with the complexity of the problem, in this case the high number of degrees of freedom of the hands. This thesis introduces a novel taxonomy designed to formalize the temporal and spatial constraints governing bimanual coordination. Grounded in insights from neuroscience and robotics, the taxonomy is developed to facilitate both the analysis of human movements and the synthesis of bimanual tasks for humanoid

robots. It explicitly accounts for coordination, interaction, and role differentiation between the hands. Notably, in scenarios with limited demonstrations, the taxonomy serves as a powerful framework for generating functional robot movements.

**Recognition of Bimanual Categories in Human Demonstrations** Methods for recognizing bimanual categories in human demonstrations are developed and evaluated using high-precision marker-based motion capture data from the *KIT Bimanual Manipulation Dataset* and RGB-D data from the *KIT Bimanual Actions Dataset* and its extension. First, a rule-based approach is presented for categorizing bimanual manipulation tasks, relying on spatial relationships between hands and objects. This method provides a clear decision-making process and demonstrates high performance on marker-based motion capture data. Additionally, a learning-based approach utilizing Graph Neural Networks (GNNs) is introduced to improve recognition of bimanual categories, even with less precise RGB-D data typically found in robotics. Spatial relations between hands and objects are encoded in graphs, and relevant features are extracted to optimize classification. The approach not only improves recognition accuracy with motion capture data but also shows robust classification performance with RGB-D data.

**Taxonomy-driven Execution of Bimanual Tasks** To equip humanoid robots with effective bimanual manipulation skills, this work investigates the integration of constraints derived from bimanual categories into a comprehensive task model. A control framework is proposed that incorporates category-specific constraints into actions associated with distinct bimanual categories. The effectiveness of the approach is demonstrated through real robot experiments. Furthermore, the broader implications of these constraints for sequences of bimanual manipulation actions are discussed, highlighting their potential for improving task execution.

# Deutsche Zusammenfassung

Beidhändige Manipulation ist ein komplexer motorischer und kognitiver Prozess, den Menschen in den ersten Lebensjahren erlernen. Diese Fähigkeit entwickelt sich schrittweise, beginnend mit dem Greifen einzelner Gegenstände, über das Übergeben von Gegenständen von der einen Hand in die andere, bis hin zum gleichzeitigen Gebrauch mehrerer Gegenstände. Im Alltag ist beidhändige Manipulation eine unverzichtbare Fähigkeit für eine Vielzahl von Aktivitäten. Dies reicht von Aktivitäten des täglichen Lebens wie der Zubereitung von Mahlzeiten zu handwerklichen Tätigkeiten. Mit viel Übung kann die Koordination zwischen beiden Händen ein beeindruckendes Niveau erreichen, wie Musiker beweisen.

Die Fähigkeit des Menschen, Objekte mit beiden Händen zu manipulieren, ist einzigartig und bleibt für Roboter bislang unerreicht. Die zielgerichtete Koordination der Hände in der Robotik ist eine besondere Herausforderung, da mehrere zeitliche und räumliche Beschränkungen berücksichtigt werden müssen. Dennoch ist dies eine entscheidende Fähigkeit für humanoide Roboter. Sie verbessert die Effizienz der Aufgabenausführung und ermöglicht die effektive Abbildung menschlicher Demonstrationen - die in der Regel beide Hände einbeziehen - im Rahmen der Programmierung durch Demonstration (PbD) oder des Imitationslernens (IL).

Ziel dieser Arbeit ist es, Methoden zum Lernen von Aufgabenmodellen für beidhändige Manipulationen aus menschlichen Demonstrationen zu entwickeln. Zu diesem Zweck werden menschliche Manipulationen analysiert, um die zugrundeliegenden Strategien zu extrahieren, die von Menschen bei der

Ausführung solcher Aufgaben verwendet werden. Basierend auf dieser Analyse und inspiriert von Erkenntnissen aus den Neurowissenschaften und der Rehabilitationswissenschaft wird eine Taxonomie zur Kategorisierung beidhändiger Manipulationen vorgeschlagen, die explizit die damit verbundenen Herausforderungen in der Robotik berücksichtigt. Es werden Methoden für die Erkennung dieser Kategorien basierend auf unterschiedlichen Eingabemodalitäten entwickelt. Schließlich werden kategoriespezifische Robotersteuerungsstrategien verwendet, um sicherzustellen, dass die durch die zeitliche und räumliche Koordination beider Hände auferlegten Beschränkungen bei der Ausführung beidhändiger Manipulationsaufgaben auf einem humanoiden Roboter erfüllt werden.

Die Arbeit besteht aus drei Teilen: Der erste ist die Entwicklung einer Taxonomie für beidhändige Manipulation, die auf der menschlichen Bewegungsanalyse und Erkenntnissen der Neurowissenschaften und der Robotik basiert. Im zweiten Teil werden Methoden zur Erkennung der durch die Taxonomie definierten Kategorien entwickelt und anhand neuer Datensätze für beidhändige Manipulationen in einem Haushaltskontext evaluiert. Abschließend wird die Umsetzung der sich aus den jeweiligen Kategorien ergebenden Randbedingungen in Kontrollstrategien für humanoide Roboter gezeigt.

**Eine Taxonomie der beidhändigen Manipulation** In der Robotik, insbesondere beim robotergestützten Greifen, sind Taxonomien eine gängige Technik, um die Komplexität des Problems zu bewältigen, insbesondere die hohe Anzahl von Bewegungsfreiheitsgraden der Hände. In dieser Arbeit wird eine neue Taxonomie vorgeschlagen, die zur Formalisierung der zeitlichen und räumlichen Beschränkungen zwischen den beiden Händen verwendet wird. Die Taxonomie wird auf der Grundlage des bisherigen Wissens aus den Neurowissenschaften und der Robotik konzipiert, wobei der Schwerpunkt auf der Anwendbarkeit für die Analyse menschlicher Bewegungen und die Synthese von zweihändigen humanoiden Roboteraufgaben liegt. Die Taxonomie berücksichtigt die Aspekte der Koordination, Interaktion und unterschiedlichen Rollen der Hände. Insbesondere für das Lernen aus wenigen Demonstration, kann die Taxonomie als wertvolles Werkzeug für die Generierung funktionaler Roboterbewegungen dienen.

**Erkennung beidhändiger Kategorien in menschlichen Demonstrationen** Die Arbeit entwickelt Methoden zur Erkennung der beidhändigen Kategorien der vorgeschlagenen Taxonomie. Die entwickelten Ansätze werden anhand hochpräziser markerbasierter Motion-Capture-Daten des *KIT Bimanual Manipulation Dataset* sowie RGB-D-Daten des *KIT Bimanual Actions Dataset* und dessen Erweiterung evaluiert. Zunächst wird ein regelbasierter Ansatz zur Kategorisierung beidhändiger Manipulationsaufgaben vorgestellt, der räumliche Beziehungen zwischen Händen und Objekten nutzt. Diese Methode ermöglicht eine nachvollziehbare Analyse des Entscheidungsprozesses und zeigt eine hohe Leistungsfähigkeit bei markerbasierten Motion-Capture-Daten. Darüber hinaus wird ein lernbasierter Ansatz auf der Grundlage von GNNs entwickelt, um die Erkennung von beidhändigen Kategorien auch bei weniger präzisen RGB-D-Daten, wie sie in der Robotik üblich sind, zu verbessern. Dabei werden räumliche Relationen zwischen Händen und Objekten in Graphen kodiert und relevante Merkmale extrahiert, um die Klassifikation zu optimieren. Neben der verbesserten Erkennungsleistung bei Motion-Capture-Daten zeigt dieser Ansatz eine hohe Robustheit bei der Klassifikation basierend auf RGB-D-Daten.

**Nutzung beidhändiger Kategorien für humanoide Roboter** Ziel dieser Arbeit ist die zuvor entwickelten Konzepte für die Ausführung von Aktionen auf humanoiden Robotern nutzbar zu machen. Um dies zu erreichen, werden zuvor definierte Einschränkungen, die aus beidhändigen Kategorien abgeleitet wurden, in ein umfassendes Aufgabenmodell integriert. Auf der Grundlage der beidhändigen Kategorien wird eine Kontrollstrategie vorgeschlagen, die die Einhaltung der zeitlichen und räumlichen Beschränkungen der jeweiligen Kategorie gewährleistet. Der Ansatz wird durch Experimente mit einem realen Roboter validiert. Zudem werden die weitergehenden Implikationen dieser Beschränkungen für Sequenzen beidhändiger Manipulationshandlungen analysiert und ihr Potenzial zur Verbesserung der Aufgabenausführung herausgearbeitet.

# Acknowledgment

This thesis is the result of my work as a research scientist in the High Performance Humanoid Technologies Lab ($H^2T$) of the Institute for Anthropomatics and Robotics (IAR) at the Karlsruhe Institute of Technology (KIT).

First, I would like to extend my sincere gratitude to Prof. Dr.-Ing. Tamim for providing an environment in which the field of humanoid robotics can prosper.

I am also deeply grateful to my colleagues and friends at the $H^2T$. You all have greatly contributed to my learning, hands-on experience, and the motivation to continue my PhD journey.

To my family, friends, and my partner – thank you from the bottom of my heart for your constant support, patience, and belief in me throughout every step of this journey.

I gratefully acknowledge the institutions and funding agencies that supported my research, particularly the Carl Zeiss Foundation, the Baden-Württemberg Ministry of Science, Research and the Arts (MWK), and the German Federal Ministry of Education and Research (BMBF).

Karlsruhe, July 2025 *Franziska Krebs*

# Contents

# CHAPTER 1

# Introduction

Bimanual manipulation involves complex motor and cognitive processes that humans learn over the first years of their lives (Kimmerle et al., 2010; Swinnen, 2002). Bimanual skills develop progressively, beginning with grasping of individual objects, advancing to handing them over, and eventually mastering the simultaneous use of multiple objects. In daily life, bimanual manipulation is an indispensable skill for various activities. With extensive practice, the coordination between both hands can achieve an impressive level of proficiency, as demonstrated by musicians.

The human ability to manipulate objects using both hands remains unprecedented and has not yet been achieved by robots (Billard and Kragić, 2019). Achieving goal-directed coordination of the hands in robotics is particularly challenging due to the necessity of accommodating multiple temporal and spatial constraints. Satisfying these constraints is essential for tasks such as carrying a large object, which is infeasible with a single hand and demands precise inter-hand collaboration. Therefore, bimanual manipulation is a critical capability for humanoid robots. It enhances task execution efficiency and enables the effective mapping of human demonstrations—which typically involve both hands—within the frameworks of programming by demonstration (PbD) or imitation learning (IL). Consequently, bimanuality must be integrated into the development of task models derived from human demonstrations for robotic execution. These models need to incorporate the intricate temporal and spatial constraints and dependencies between the hands.

Achieving successful bimanual manipulation is challenging due to the complexity of the high-dimensional configuration space that must be considered. This challenge is similar to robotic grasping, where taxonomies have been introduced to address the complexity of hand design and grasp synthesis (Kamakura et al., 1980; Cutkosky, 1989; Feix et al., 2015). While previous studies have applied taxonomies to broader manipulation scenarios (Bullock et al., 2012; Borràs and Asfour, 2015), this thesis specifically focuses on bimanual manipulation. Categorizing bimanual coordination patterns can serve various purposes, such as learning task models from human demonstrations, enhancing human-robot collaboration, improving action recognition, and deriving constraints for the coordinated execution of robotic bimanual manipulation tasks.

This chapter introduces the central research questions addressed in this thesis and outlines the contributions made toward answering them. It then provides an overview of the structure of this thesis.

## 1.1. Problem Statement

The aim of the thesis is to explore how robots can learn from and replicate human manipulation capabilities, with a particular focus on utilizing both hands effectively and mastering the spatial and temporal constraints inherent in bimanual manipulation tasks. A humanoid robot should be able to autonomously extract those constraints from human demonstration, incorporate them into a task model, and account for them when executing the respective tasks. In order to achieve these goals, the following questions and challenges are addressed in this thesis:

- **How can the constraints arising in human bimanual manipulation be formalized into a taxonomy?** When humans engage in bimanual activities, there is typically a form of coordination between the hands, implying the presence of constraints governing their interaction. To effectively capture these essential task constraints, they must be first identified and then formalized. Previous research in robotic grasping has shown that taxonomies can be a powerful tool for describing interdependencies of hand joint movements and provide a structured approach to categorizing constraints. In the context of bimanual manipulation, the research question arises: How can a taxonomy for bimanual manipulation be defined?

- **How can the bimanal coordination patterns, so-called bimanal categories, formulated by this taxonomy be recognized in human demonstrations?** In order to integrate the usage of a taxonomy in Learning from Demonstration (LfD) frameworks, such bimanal categories should be automatically detectable in human demonstrations. The required segmentation and classification should be ideally performed using visual data collected by the robot.
- **How can the constraints associated with different bimanual categories defined in the taxonomy be leveraged to develop control strategies for bimanual manipulation?** Control methods for humanoid robots are needed that can adhere to the various constraints specific to each category of the taxonomy and maintain these constraints even in the presence of external perturbations.

## 1.2. Contributions

This thesis aims to advance methods for learning task models for bimanual manipulation based on human demonstrations. To achieve this, human bimanual manipulations are analyzed to uncover the underlying strategies utilized in such tasks. Drawing on insights from neuroscience and rehabilitation science, a taxonomy for categorizing bimanual manipulations is proposed, explicitly addressing the associated challenges in robotics. Learning-based methods are then developed and evaluated for recognizing these categories across various input modalities. Finally, category-specific robot control strategies are implemented to ensure that the temporal and spatial coordination constraints of both hands are satisfied during the execution of bimanual manipulation tasks on a humanoid robot. An overview of the main contribution is depicted in Figure 1.1.

**A Bimanual Manipulation Taxonomy** This thesis combines insights from neuroscience and robotics to propose a taxonomy of bimanual manipulations designed for both analyzing bimanual human manipulations and synthesizing bimanual tasks for humanoid robots. In robotics, particularly within robotic grasping, taxonomies are a prevalent technique for addressing complexity, as they provide reasonable assumptions about task constraints with minimal initial information, even if they may not match the specificity of application-specific

solutions. The proposed taxonomy addresses key aspects such as coordination, physical interactions, and different roles of the hands during bimanual manipulations. This contribution facilitates the development of more efficient and effective bimanual manipulation strategies in humanoid robotics, leveraging a structured approach to understanding and replicating human bimanual capabilities.

**Recognition of Bimanual Categories in Human Demonstration** The thesis develops methods for recognizing bimanual categories in human demonstrations and evaluates them on the *KIT Bimanual Manipulation Dataset*, which consists of high-accuracy marker-based motion capture data, as well as on the *KIT Bimanual Actions Dataset* and its extension for single-view RGB-D data, which is commonly available on robots.

To detect bimanual categories in human demonstrations, the thesis introduces a rule-based approach that leverages spatial relations between hands and objects. This approach provides human-interpretable introspection into the decision-making process and performs well on marker-based motion capture data. Furthermore, recognizing the need for robustness when handling less precise RGB-D data typically encountered in robotics, the thesis develops a learning-based method based on GNNs. This approach encodes spatial relations between hands and objects into graphs, leveraging extracted features to enhance category recognition. In addition to improving performance on motion capture data, the method demonstrates the capability to robustly recognize bimanual categories in RGB-D data.

**Leveraging Bimanual Categories for Robot Control** This thesis contributes to the execution of bimanual tasks by integrating bimanual categories from a predefined taxonomy into a structured task model suitable for humanoid robots. The proposed approach enables task execution with minimal prior knowledge, relying only on a few demonstrations without requiring a dynamic model of the environment or force-torque sensors. A task-space impedance controller ensures compliant behavior while preserving spatial and temporal constraints, allowing the system to adapt reactively to external disturbances. Additionally, transitions between bimanual action categories are designed to respect predefined temporal constraints, improving task execution efficiency. The framework is validated through real-robot experiments, demonstrating its feasibility for bimanual task execution.

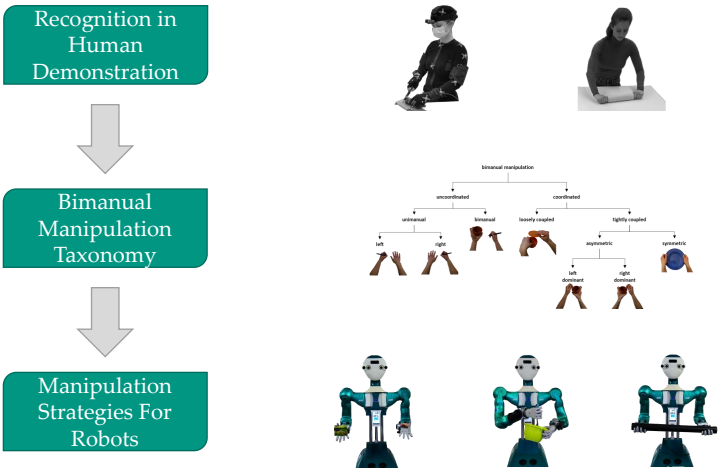



Figure 1.1.: Overview of the three parts of the thesis: A bimanual manipulation taxonomy, recognition of bimanual categories in human demonstration and leveraging those categories for robot control strategies.

## 1.3. Structure of the Thesis

The thesis is organized into six chapters, outlining related work and presenting the primary contributions of this research.

**Chapter 2** introduces relevant related work. First, it explores the neuroscientific foundations of bimanual manipulation and reviews existing taxonomies in neuroscience, rehabilitation science, and robotics. Next, methods for bimanual action and category recognition are detailed, followed by an overview of bimanual control strategies. The chapter concludes with a survey of existing datasets dedicated to bimanual human manipulation.

**Chapter 3** presents the developed *Bimanual Manipulation Taxonomy* based on the design principles of coordination, interaction, and hand roles during bimanual manipulations. Further, the temporal and spatial constraints associated with the different categories of the taxonomy are formalized.

**Chapter 4** describes the datasets collected for this research, aimed at analyzing bimanual manipulation actions and detecting the categories of the taxonomy. This includes the multimodal *KIT Bimanual Manipulation Dataset*, which provides high-precision marker-based motion capture recordings of human and object movements, along with other data modalities. Additionally, the existing *KIT Bimanual Actions Dataset*, which consists of single-view RGB-D data, is ex-

tended with new scenarios to cover a broader range of bimanual manipulation strategies.

**Chapter 5** focuses on the development of recognition methods tailored for the simultaneous segmentation and classification of categories defined in the *Bimanual Manipulation Taxonomy* from human motion recordings. To this end, a rule-based and a learning-based approach are proposed and evaluated on the data presented in Chapter 4.

**Chapter 6** introduces the proposed approach for executing bimanual tasks based on a taxonomy of bimanual categories. It details the task model, control framework, and constraint integration, followed by an experimental evaluation demonstrating the feasibility of the approach.

**Chapter 7** concludes the thesis by summarizing the contributions and the achieved results. It provides a critical review of both the strengths and weaknesses of the proposed methods. Additionally, the chapter discusses potential extensions and future research directions.

# CHAPTER 2

# Related Work

Bimanual manipulation in robotics remains an emerging research area with significant untapped potential (Billard and Kragić, 2019). The increased complexity of bimanual tasks compared to single-arm manipulation is underlined by Smith et al. (2012).

This chapter reviews the current state of the art in several fields relevant to this thesis. It begins with neuroscientific foundations of bimanual manipulation, examining existing taxonomies from neuroscience, rehabilitation science, and robotics. The discussion then transitions to human activity recognition, focusing on the identification and classification of bimanual categories. Next, the chapter explores how constraints inherent to bimanual manipulation are addressed in robotic control frameworks, including both explicit constraint modeling and implicit approaches leveraging deep learning techniques. Finally, an overview of existing datasets for human bimanual manipulation is provided, highlighting their relevance to this area of research.

## 2.1. Bimanual Manipulation Taxonomies

In this thesis, we aim to develop a taxonomy that describes bimanual manipulation and use it to enhance bimanual manipulation capabilities of humanoid robots. To achieve this, we first review and discuss previous re-

search on taxonomies from the fields of neuroscience, rehabilitation, and robotics.

### 2.1.1. Bimanual Manipulation in Neuroscience

Given that this thesis seeks to draw inspiration from human bimanual manipulation, we initially examine the current body of knowledge on human bimanual manipulation, with a particular emphasis on aspects that hold potential relevance for robotics. We consider common behavioral patterns in human temporal and spatial coordination, the underlying cognitive representations, and the involved brain regions.

The study of motor control and learning in bimanual manipulation and coordination has a long-standing history in neuroscience, neuro-rehabilitation, and the clinical assessment of motor impairments affecting daily activities. Bimanual manipulation and coordination in humans represent a complex process that develops during childhood (Barral et al., 2006; Kimmerle et al., 2010) and can be affected by neurodegenerative diseases and brain pathologies (Hung and Zeng, 2020; Roebuck-Spencer et al., 2004).

As stated by Swinnen and Wenderoth (2004), two primary theoretical frameworks dominate the study of bimanual motor control: (i) the dynamic pattern perspective and (ii) the information-processing perspective. The dynamic pattern perspective views biological systems as comprising various subsystems that evolve over time, where behaviors emerge in a self-organized manner. These subsystems can be understood as multiple underlying constraints that influence and shape the resulting bimanual behavior. This perspective aligns with the optimal control theory, a widely used approach in robotics. In contrast, the information-processing perspective conceptualizes bimanual movements as tasks that encounter structural interference due to the finite capacity of neural resources, leading to neural leakage during bimanual activities. By default, both arms are strongly coupled by what is termed *neural cross-talk*. The objective in learning bimanual coordination is to overcome this coupling, thereby enabling distinct behaviors of both hands. This perspective is prevalent in neuroscience and is crucial to consider when evaluating various studies in this domain.

In the realm of bimanual manipulation research, a disproportionate focus exists regarding the types of tasks examined. The majority of studies concentrate on cyclic bimanual coordination, while object-oriented and goal-directed bimanual

tasks have been largely overlooked (Obhi, 2004). Cyclic motions in bimanual coordination can be subdivided into in-phase and out-of-phase movements (Kelso, 1984). Furthermore, Swinnen (2002) introduces anti-phase patterns, which describe cyclic movements with a phase angle difference of 180° between the two end-effectors. The author asserts that anti-phase patterns are less stable than in-phase movements but more stable than out-of-phase movements. Non-cyclic, goal-oriented tasks have received less attention historically. Perrig et al. (1999) analyzed interlimb synchronization and temporal correlation in goal-directed bimanual tasks. For instance, in the task of pulling a drawer with one hand and grasping a peg inside with the other, the study demonstrated not only near-synchronous motion onset but also simultaneous goal achievement. The hypothesis was further reinforced by Kelso et al. (1979), who observed a tendency for both arms to initiate and conclude their movements at nearly the same time, even when the amplitudes of the movements differed. A different hypothesis was presented by Guiard (1987) who stated that in role-differentiated bimanual actions the non-dominant hand usually starts its motion slightly earlier than the other hand.

Ivry et al. (2004) underscore the critical role of action representation in human bimanual coordination, asserting that limitations in bimanual manipulation significantly overlap with cognitive constraints. The authors demonstrate how various task representations influence the extent of *neural cross-talk* as described by the information-processing theory. Specifically, they highlight the importance of action representations in coordinating both arms. For instance, spatial interference (*neural cross-talk*) during response planning differs markedly between congruent and incongruent hand motions when tasks are symbolically encoded, such as using symbols (e. g., letters) to indicate target locations that require interpretation or translation into spatial coordinates (e. g., nearer or farther targets). Conversely, when tasks are directly cued through visual representation of the target location, hand behavior for congruent and incongruent motions becomes more similar. This influence of task conceptualization extends beyond spatial constraints to temporal ones. According to Ivry et al. (2004), discrete motions can be conceptualized using an event structure, where the complexity depends on the number of salient events. This framework aligns with the findings of Swinnen (2002), who demonstrated that in-phase motions, representable by a single salient event, are inherently more stable than anti-phase motions, which necessitate the representation of two salient events. The observed stability and faster execution of simpler encodings suggest that cognitive limitations play a pivotal role in bimanual manipulation. These findings also imply that designing

compact and minimally complex task representations in robotics could better align robotic behavior with human performance.

The differentiation of roles between the hands is a further characteristic feature of human behavior. Anthropological evidence, including studies of tool production techniques, tool usage, and cave art, suggests a long-standing preference for right-handedness (handedness bias) that dates back to at least the time of the Neanderthals (Cashmore et al., 2008). While this bias is often interpreted as reflecting the superiority of the dominant hand, Guiard (1987) proposes an alternative perspective, viewing it as a division of labor where each hand performs distinct, complementary roles. Tasks that exhibit these differentiated roles are described as *asymmetrical* in the context of his work. He identifies three overarching principles governing such bimanual actions: (i) the dominant hand operates within a reference frame established by the non-dominant hand; (ii) the hands exhibit a contrast in spatio-temporal scales, with the non-dominant hand performing macrometric actions and the dominant hand handling micrometric actions; and (iii) the non-dominant hand often precedes the dominant hand in task execution. In Nakamura et al. (2019), the factors influencing the choice between unimanual, self-handover, and symmetric bimanual actions in transport tasks are investigated. The study finds that self-handover is primarily used to transfer an object between the right and left hemispheres. Further, bimanual transport, while offering increased stability compared to unimanual transport, requires greater effort.

The neural implementation of bimanual manipulation involves multiple brain regions rather than a single location. These include the corpus callosum (Swinnen, 2002; Wiesendanger and Serrien, 2001; Ivry et al., 2004), cerebellum (Tracy et al., 2001), primary motor cortex, premotor cortex, and supplementary motor areas (Toyokura et al., 1999, 2002). In many cases, distinct areas are specialized for managing different types of constraints. The cerebellum is particularly critical for representing temporal information, especially in encoding event structures during discontinuous movements (Ivry et al., 2004). In contrast, the corpus callosum, which facilitates communication between the left and right cerebral hemispheres, plays a key role in mediating spatial interactions during bimanual tasks.

The cognitive neuroscience of motor control has provided valuable insights into the constraints governing bimanual movements in humans, including spatial and temporal interference effects. A key finding is that the conceptualization of task goals significantly influences the patterns of interference between the two movements (Ivry et al., 2004). While the constraints observed in human biman-

ual coordination may not directly translate to robotic systems, neuroscientific findings offer a powerful framework for developing efficient action representations in robotics. By considering resource limitations and leveraging these principles, robotic systems can be designed to emulate human-like coordination and adaptability in bimanual tasks.

### 2.1.2. Taxonomies in Neuroscience and Rehabilitation Science

Several classifications of bimanual manipulation have been proposed in the past. Guiard (1987), analyzes bimanual manipulation by considering the roles of the two hands. For symmetrical movements, both hands perform the same role, such as transporting a large box. In contrast, for asymmetric movements, the hands assume different roles; one hand stabilizes an object while the other acts on it. Such movements are also referred to as *role-differentiated bimanual manipulations (RDBMs)* (Kimmerle et al., 2010). Within these manipulations, the dominant and non-dominant hands often have specific roles. According to Guiard, the non-dominant hand provides a spatial frame of reference within which the dominant hand moves. Additionally, there is a contrast in the spatio-temporal scale of the motions of the hands. Sainburg (2002) introduced the dynamic dominance hypothesis, suggesting that the dominant hand is more effective in adapting to novel task dynamics.

Shirota et al. (2016) compare different definitions, assessment methods, and robotic devices for therapy. Their goal is to standardize the terms and methods in robotic and sensor-based assessments and establish a common language for communication and collaboration among clinicians, neuroscientists, and engineers researching interlimb coordination. The authors also present a taxonomy of interlimb activities, which does not have a hierarchical structure but defines terms for describing one limb independently (e. g., periodic) and in relation to the other (e. g., in-phase).

Kantak et al. (2017) present a classification of bimanual tasks (see Figure 2.1) to study bimanual coordination in rehabilitation. According to their proposed classification, bimanual tasks can be characterized by the symmetry of arm movements, the task goal, and the necessity of cooperative interaction.

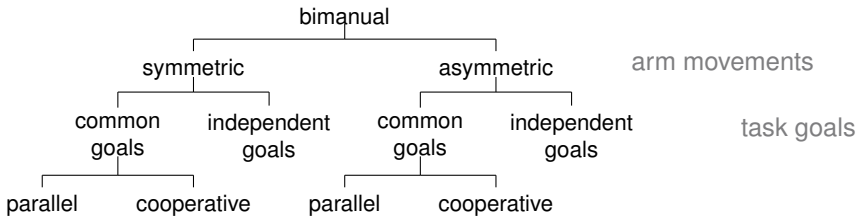



Figure 2.1.: Taxonomy presented in (Kantak et al., 2017).

These insights from human bimanual coordination can play a significant role in the execution of goal-directed bimanual tasks by robots.

### 2.1.3. Taxonomies in Robotics

In robotics, grasp taxonomies are a widely used method to manage the complexity of hand design and grasp synthesis (Kamakura et al., 1980; Cutkosky, 1989; Feix et al., 2015; Arapi et al., 2021). While prior research has utilized taxonomies for broader manipulation contexts (Bullock et al., 2012; Borràs and Asfour, 2015), in this section we specifically focus on bimanuality. Such a classification of bimanual patterns can be utilized for learning task models from human demonstrations, enhancing human-robot collaboration, recognizing actions, and establishing constraints for the coordination and execution of robot bimanual manipulation tasks.

Some studies in the literature focus on specific types of bimanual manipulation rather than offering a comprehensive classification. For instance, Pais Ureche and Billard (2018) explore the extraction of constraints in asymmetric bimanual tasks from human demonstrations, where tasks are either executed autonomously by a robot or in collaboration with a human. This work assumes a master-slave dynamic between the end-effectors and explicitly accounts for the interaction forces. Similarly, Pais and Billard (2014) examines hand dominance by analyzing the force-motion relationship between the hands to identify key task constraints. Laghi et al. (2018) introduce a framework for intuitive bimanual telemanipulation, which includes a dual-arm teleoperation mode and a control strategy for symmetric motions, where both robot arms are controlled by a single human arm. The operator can seamlessly switch between control strategies using gesture commands. Experimental results from box transport tasks demonstrate that the symmetric control strategy leads to improved performance and reduced variance in two out of three tasks.

Paulius et al. (2019) introduce a manipulation motion taxonomy that focuses on features such as *contact type*, *engagement type* (rigid/soft), *trajectory type*, *contact duration* (discontinuous/continuous), and *manual operation* (unimanual/bimanual). However, bimanual manipulation is addressed only as a binary attribute rather than a primary focus. In contrast, Yao et al. (2021) present a hand pose taxonomy specifically designed for high-precision, bimanual fine-manipulation tasks, like those in watchmaking. This taxonomy is based on the analysis of virtual fingers in relation to force and torque demands. Although it is applied to each hand individually, the authors illustrate how the hand pose matrix can describe and visualize functional distributions across both hands. A sparse matrix indicates low variance in hand pose combinations, while concentration in the upper or lower diagonals suggests handedness. This method is more suitable for analyzing hand pose selection throughout an entire task rather than for classifying individual motion segments.

Several studies have categorized bimanual manipulations to facilitate subsequent planning or control processes. For instance, Zöllner et al. (2004) presents a classification of dual-arm manipulation, which is used for task segmentation in robot programming by demonstration. In this work, bimanual tasks are divided into two main categories: *uncoordinated* and *coordinated*. Coordinated tasks are further classified as either *symmetric* or *asymmetric*. In symmetric tasks, both hands grasp the same object, whereas in asymmetric tasks, they manipulate different objects. Although the authors introduce these categories and demonstrate their application for motion segmentation within the three subcategories, their planning approach only incorporates the distinction between coordinated and unimanual tasks. Another example for the application in robot control is the work of Surdilovic et al. (2010). Bimanual manipulation tasks are categorized into *non-coordinated* and *coordinated* tasks. Within coordinated tasks, the authors further distinguish between *goal-coordinated* and *bimanual* operations, which can be either *symmetric/asymmetric* or *congruent/non-congruent*. The resulting taxonomy is illustrated in Figure 2.2.

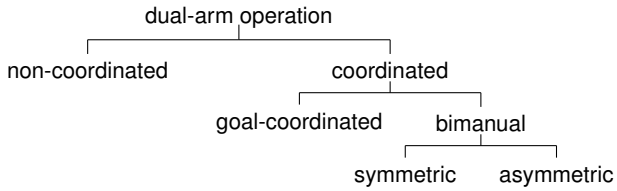



Figure 2.2.: Taxonomy presented in (Surdilovic et al., 2010).

While this taxonomy offers greater detail compared to the one proposed by Zöllner et al. (2004), it is introduced only briefly, with no in-depth definitions or practical applications discussed. The primary focus of the study was to develop an impedance controller for contact-based bimanual operations. Similar to Zöllner et al. (2004), this work also considers the decomposition of tasks into single-arm and bimanual actions.

Another application of bimanual manipulation classification is in the intuitive control of prosthetics, as described in Volkmar et al. (2019). Movements are categorized based on onset and direction into *unimanual*, *bimanual synchronous*, and *bimanual asynchronous*. Once a bimanual category is recognized, the prosthesis automatically controls wrist rotation accordingly. In Rakita et al. (2019), a bimanual action vocabulary is introduced to enhance the performance of dual-arm teleoperation systems. This vocabulary includes categories such as *fixed offset*, *one hand fixed*, *self-handover*, and *one hand seeking*. Additionally, Park and Lee (2016) leverage a taxonomy for subsymbolic motion representation, introducing the Extended-Cooperative-Task Space (ECTS) for coordinated motions between two end-effectors. The ECTS coefficients are used to split motions into absolute and relative components, allowing categorization as uncoordinated or coordinated, with further subcategories of parallel, blended, and serial. The work in Boehm et al. (2021) focus on the online recognition of bimanual coordination modes in teleoperated robots, distinguishing between different symmetry types and relative movement directions.

### 2.1.4. Discussion

The goal is to propose a taxonomy with clearly differentiable categories and special focus on applications in bimanual robotic manipulation. As discussed in Section 2.1.1, clear categorizations of bimanual manipulation exist in neurorehabilitation. However, within this field, less attention is paid to deriving methods and tools for automatic recognition of different bimanual actions but to the assessment of therapy progress in the context of bimanual motor coordination. In particular, Kantak's taxonomy (Kantak et al., 2017) contains important criteria. While symmetry is of great importance in neurology due to the characteristic muscle activation or the way of interhemispheric communication, the focus in robotics is rather on the fulfillment of or dependencies on task goals to be achieved. In contrast, previous approaches in the field of robotics focus less on the precise definition and classification of different categories but on providing a schema that supports the development of control and planning strategies

for dual-arm manipulation tasks. Accordingly, some consider a limited set of bimanual categories (Pais Ureche and Billard, 2018; Volkmar et al., 2019). In other cases the presented taxonomies are very promising but their precise definition is lacking since the focus of the respective publication was purely on the technical implementation (Surdilovic et al., 2010). Taking into account previous work in neuroscience (in particular Kantak et al. (2017)) and robotics, we propose a bimanual manipulation taxonomy that is not primarily tailored to the evaluation of therapy progress in rehabilitation, but rather dedicated to the representation of bimanual robotic manipulation tasks. This includes learning such representations from human motion data and making use of this knowledge to improve the execution of bimanual manipulation tasks in humanoid robotics.

Based on the bimanual manipulation taxonomy proposed in this thesis, Ziaeetabar et al. (2024) present a modification of the taxonomy by considering *hand spatial relation* (*closed hands*, *crossed hands*, *stacked hands*) and the *precision level* (*low*, *medium*, *high*). Unlike our taxonomy, which assigns a manipulation to a single leaf node, the enriched taxonomy allows a manipulation to be associated with multiple nodes. The authors use the taxonomy to create semantic action descriptions.

## 2.2. Bimanual Action and Category Recognition

In this section, we provide a systematic review of prior research on the recognition of bimanual categories in human demonstrations. We define bimanual categories as cases where a single label represents the coordinated activity of both arms, highlighting their interplay and synchronization. This contrasts with classical action recognition, which typically focuses on task-related subactions performed by individual hands, such as e. g., cut, approach, or open. Although bimanual category recognition differs in focus, it can be situated within the broader framework of Human Activity and Recognition (HAR). HAR aims to automatically detect and classify human activities from video or other sensor data (Aggarwal and Ryoo, 2011). Consequently, we also examine the latest advancements and methodologies in the field of HAR, particularly in the subfield of action recognition.

In this section, we address the tasks of *segmentation* and *classification*. *Segmentation* involves the temporal decomposition of a continuous data stream by identifying the start and end points of distinct, coherent segments. *Classification*,

in contrast, entails assigning labels from a predefined set to these temporally delineated segments.

### 2.2.1. Category Recognition

There are several works addressing the recognition of certain bimanual categories. However, the definition of those categories as well as the available sensor data differs significantly.

Several approaches stem from the medical field. Boehm et al. (2021) employ a rule-based classification approach to recognize bimanual coordination modes in robot-assisted surgery. These modes are characterized by the *direction* (e. g., move together or away) and *symmetry* (e. g., point or mirror) of the movements. The classification relies on the analysis of end-effector positions over time. To improve bimanual interaction with a semi-autonomously controlled prosthetic hand, Volkmar et al. (2019) differentiate between *unimanual*, *bimanual asynchronous*, and *bimanual synchronous* movements. These movements are detected through a rule-based classification of data obtained from two inertial measurement units (IMUs) attached to the prosthesis and the other hand. Another rule-based approach in robotics is presented by Zöllner et al. (2004), who consider factors such as whether both hands are in a grasped state and whether a closed kinematic chain is formed.

Miller and Wade (2021) classify similar categories based on motion symmetry to monitor the rehabilitation progress of post-stroke patients. Artificial neural networks are applied to both raw IMU data and features extracted from this data. Rakita et al. (2019) proposed an approach for teleoperation of dual-arm robots utilizing a defined set of bimanual action categories termed the *bimanual action vocabulary*. This vocabulary includes specific actions such as *one hand seeking*, *self-handover*, *fixed offset*, and *one hand fixed*. To classify these actions, they employed a sequence-to-sequence recurrent neural network (RNN) capable of predicting the most likely bimanual action category based on observed data sequences.

Comparative quantitative analysis of existing approaches is challenging due to variations in defined categories and the use of different sensory modalities. Some methods, such as Boehm et al. (2021); Volkmar et al. (2019), employ rule-based approaches, which offer intuitive classification and straightforward error analysis in cases of misclassification. In contrast, other approaches, like Miller and Wade (2021), leverage neural networks to handle noisy and complex

data, aiming for robust classification performance. Additionally, approaches such as that of Rakita et al. (2019) draw inspiration from neural processes in humans to enhance teleoperation strategies for dual-arm robots. Overall, there are only a few works focusing on the classification of bimanual categories, and to our knowledge, none utilize RGB or RGB-D sensor data for solving this problem.

### 2.2.2. Action Recognition

The problem of recognizing bimanual categories can be viewed more broadly as a Human Activity Recognition (HAR) problem, where the objective is to assign semantic labels to human activities based on time-series data collected from various sensors. In this context, numerous methods have been developed that utilize various sensor modalities, often incorporating vision data (e.g., RGB or RGB-D) commonly available on humanoid robot platforms. Consequently, we will also examine related research in the domain of human action recognition. We begin by elaborating on general methods employed for action recognition, with a particular focus on the concept of Graph Neural Networks, which will be utilized in this thesis. Subsequently, we present a detailed overview of approaches for bimanual action recognition, categorized into two main groups: *skeleton-based methods*, which rely on predefined features such as detected objects and human skeleton representations, and *visual feature-based methods*, which additionally process raw image data.

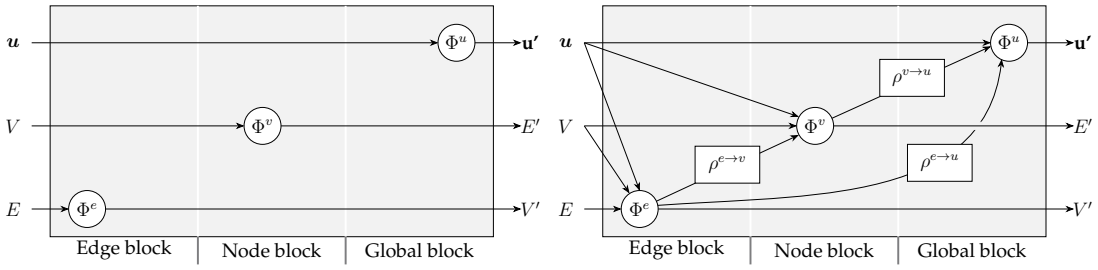
Khaire and Kumar (2022) conducted a recent survey focusing on deep learning approaches and RGB-D-based recognition of human actions, human-human interactions, and human-object interactions. Their work provides insights into relevant datasets and commonly employed techniques. Specifically, for human-object interaction detection, which involves identifying manipulation-related actions, the survey highlights that predominant methods rely on graph-based representations. Another recent review by Sun et al. (2022) explores human action recognition, emphasizing the utilization of various sensor modalities. Skeleton-based action recognition is considered a distinct modality alongside RGB and depth data. Approaches in this domain are categorized into three main types: Recurrent Neural Networks (RNNs), including variants such as Long Short-Term Memory networks (LSTMs) (Liu et al., 2017; Song et al., 2017); Convolutional Neural Networks (CNNs) (Yang et al., 2019; Tang et al., 2022); and GNNs (Shi et al., 2019; Dreher et al., 2020). GNNs are particularly notable because they not only maintain the expressive capabilities of graph structures

but also excel in handling inputs of varying sizes. This adaptability makes GNNs well-suited for capturing intricate relationships within skeleton-based action recognition tasks.

### Graph Neural Networks

While the skeleton modality, as discussed in Sun et al. (2022), focuses solely on describing the human body in manipulation tasks, it is crucial to consider the involvement of objects as well. This aspect is explicitly addressed in Dreher et al. (2020), where GNNs are employed for action recognition and segmentation. In this approach, each hand and detected object in the scene is represented as a node in a graph-based structure. The edges connecting these nodes encode spatial relationships between the hands and objects, facilitating comprehensive understanding of interactions in bimanual tasks.

*Graph Networks (GN)* (Battaglia et al., 2018) is a framework operating on graph-structured representation. In this context a graph is defined as a 3-tuple $G = (u, V, E)$, with $u$ being the global attribute of the graph, $V$ the set of nodes in the graph and $E$ the set of edges. The set of nodes $V$ consists of the node attributes $v_a \in V$ and the edges $E$ of 3-tuples $e = (e_a, s, r) \in E$. Within the edges, $e_a$ represents the edges attributes and $s$ and $r$ are the sender and receiver nodes in $V$. GN blocks are the main computation unit of the GN framework. Each GN block contains the *update* functions $\Phi^e, \Phi^v, \Phi^u$ and the *aggregation* functions $\rho^{e\to v}$, $\rho^{e\to u}$, $\rho^{v\to u}$. A full GN block is depicted in Figure 2.3b. Computations proceed from edge to node, to global level.




(a) Independent GN block (b) Full GN block

Figure 2.3.: Graph Network (GN) blocks as described in (Battaglia et al., 2018).

When employing GN frameworks the input graphs determine how representations interact instead of this being determined by the model architecture. This way the expressive power of grasps can leveraged for the respective task. At the same time entities, in this case nodes of the input graph and their relations

are treated as sets and are therefore invariant to permutations. Further, since per-edge and per-node functions are reused across all edges and nodes, GNs support a form of combinatorial generalization meaning that graphs can have different input sizes and shapes.

GN blocks can be combined to create various architectures. A common architecture design is the *encode-process-decode* configuration (see Figure 2.4). $G_{in}$ is transformed into a latent representation by an encoder $G_{enc}$, then a shared core block $GN_{core}$ is applied M times and finally the result is decoded by $G_{dec}$.

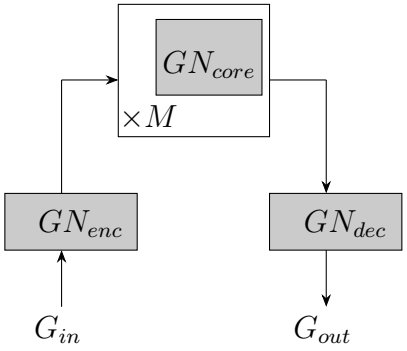



Figure 2.4.: Encode-process-decode architecture. By concatenating GN blocks different network architectures can be created. $\times M$ indicated that the block is concatenated $M$ times with itself. This figure is taken from (Battaglia et al., 2018).

### Skeleton-Based Approaches

This section focuses on work that leverages previously extracted features, such as human skeleton representations and detected objects. Graph Networks (GNs) in the *encode-process-decode* configuration (Figure 2.4) were utilized for hand-wise action classification in Dreher et al. (2020). The encoding and decoding stages consist of independent graph network blocks, while the core processing block is a fully connected network. Multi-layer perceptrons (MLPs) serve as the update functions. This method operates not on raw image data but on extracted 3D bounding boxes of hands and objects, along with their spatial relationships. In the fully connected input graphs, nodes represent hands and objects within the scene, with spatial relationships encoded as edge attributes. To incorporate temporal dynamics, these scene graphs from multiple frames are linked by temporal edges that connect the same entities across frames. The Graph Neural Network (GNN) is applied to classify short sequences. By employing

a sliding window approach, simultaneous classification and segmentation are achieved. Ziaeetabar et al. (2024) introduces an approach called the *Bimanual Graph Neural Network (BiGNN)*, which, similar to Dreher et al. (2020), utilizes scene graphs and GNNs. The key distinction lies in its more sophisticated architecture, which incorporates a hierarchical Graph Attention Network (GAT). The multi-head attention mechanism in BiGNN facilitates the capture of both local and global contextual information. The extracted features are subsequently processed by a Temporal Convolutional Network (TCN) to model temporal dependencies. Additionally, unlike the one-hot encoding used in previous works, BiGNN employs numerical encoding for attributes, offering a more nuanced representation.

Pyramid Graph Convolutional Network (PGCN) (Xing and Burschka, 2022) follow the basic idea of downsampling the data to distill essential spatial information and then upsample it back to the temporal scale of the input. It consists of three central elements: an attention-based graph convolutional encoder, a temporal pyramid upsampling decoder and a convolution-based predictor. The proposed temporal pyramid pooling (TPP) layer is capable of extracting spatial information with various temporal scales and improves the segmentation performance significantly. The authors subsequently extended their work by proposing the Uncertainty Quantified Temporal Fusion Graph Convolution Network (UQ-TFGCN) (Xing and Burschka, 2024). This approach incorporates an attention-based graph convolutional encoder and a novel temporal fusion decoder comprising multiple parallel temporal-pyramid pooling blocks. UQ-TFGCN addresses the challenge of over-segmentation in human action recognition while also providing confidence values for predictions. However, the increased number of parameters results in significantly higher computational demands. Compared to PCGN, UQ-TFGCN demonstrates a modest improvement in performance.

Interactive Spatiotemporal Token Attention Networks (ISTA-Net) (Wen et al., 2023) utilize skeleton data to learn relationships between entities without requiring initial adjacency information. ISTA-Net employs interactive spatiotemporal tokens (ISTs) and integrates multi-head self-attention blocks with 3D convolutions to capture inter-token correlations. Unlike other approaches that rely on graph-based methods, ISTA-Net ensures entity permutation invariance through entity rearrangement. Notably, ISTA-Net is originally designed for classification tasks rather than segmentation.

### Visual Feature-Based Approaches

Several recent graph-based models in HAR utilize raw visual features instead of only previously extracted features. One example are Asynchronous-Sparse Interaction Graph Networks (ASSIGN) (Morais et al., 2021) which take framewise visual features extracted from the 2D bounding boxes of humans and objects from a complete recording as input. ASSIGN considers the segmentation and the classification task explicitly in their network architecture, thereby preventing over-segmentation. The framework (see Figure 2.5) consists of two layers with different functionalities. In the *frame layer*, outputs of Bidirectional Recurrent Neural Network (BiRNN) units send messages to other entities. Entities correspond to hands and objects in the scene. Based on the current recurrent state and the messages from neighboring nodes which are weighted by an attention mechanism, the detector decides whether this frame is considered a segmentation node or not. In case of a detected segmentation point a change signal is sent to the corresponding node in the *segment layer*, triggering a segmentation point. Afterwards the previous segment is classified based on information of the frame- and segment level.

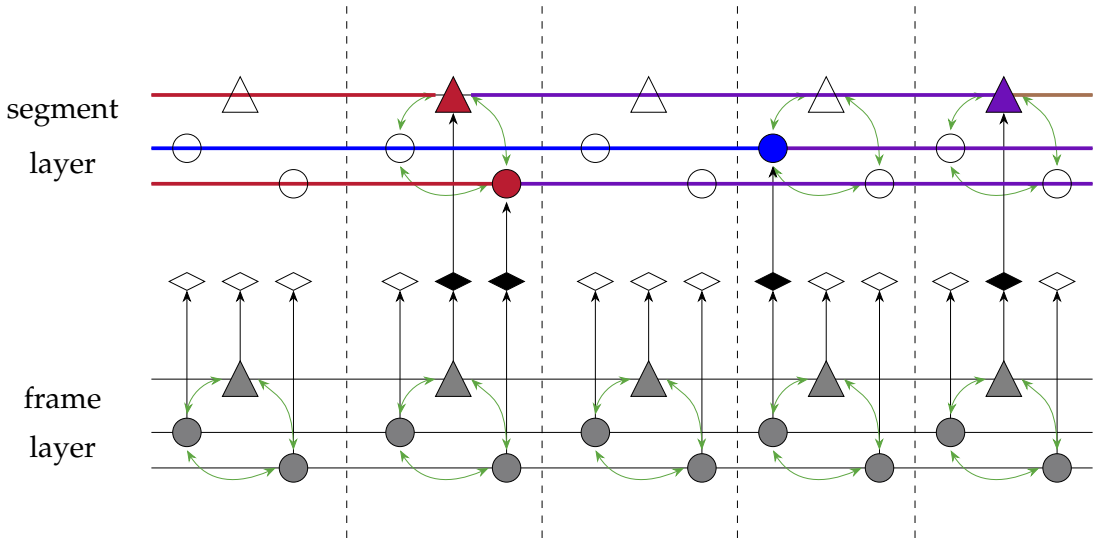



Figure 2.5.: The ASSIGN architecture (Morais et al., 2021) consists of two layers. Nodes represent human (triangle) and object (circle) entities. Temporal edges (black arrows) are modeled with BiRNNs, spatial edges (green arrows) with message passing. At each time step, the frame level is updated, and for each entity, it is determined whether the corresponding segment-level entity changes its state (indicated by solid diamonds) or skips the update (indicated by hollow diamonds). This sparse change signal results in asynchronous and sparse updates at the segment level.

Another approach considering visual input data is the Spatio-Temporal Interaction Graph Parsing Network (STIGPN) (Wang et al., 2021). The two-stream architecture consists of a visual stream including a feature extraction network and bounding boxes of the entities, and a semantic stream that uses bounding boxes and object knowledge. Each stream has the same network architecture and its own predictor. A self-attention mechanism is used to reset edges in the dense graph and infer a sparser one. Like ASSIGN, STIGPN uses a BiRNN. The approach in its original form is for classification and not for segmentation.

### 2.2.3. Discussion

Since the methods presented in the previous section are partially evaluated on the same datasets, we provide a quantitative comparison in Table 2.1. Apart from the GNN (Dreher et al., 2020) most approaches perform in a similar range, with the PGCN (Xing and Burschka, 2022) performing the best on the Bimacs dataset (Dreher et al., 2020).

Table 2.1.: Comparison of action recognition approaches on HOI datasets. The best-achieved scores are highlighted respectively for each dataset.

| Dataset | Model | Classification | | Segmentation | |
|---|---|---|---|---|---|
| | | macro F1 | micro F1 | F1@10 | F1@25 |
| Bimacs (Dreher et al., 2020) | GNN (Dreher et al., 2020) | 0.630 | 0.640 | 0.41 | 0.35 |
| | ASSIGN (Morais et al., 2021) | 0.798 | 0.826 | 0.84 | 0.81 |
| | PGCN (Xing and Burschka, 2022) | **0.815** | **0.869** | **0.89** | **0.86** |
| CAD120 (Koppula et al., 2013) | ASSIGN (Morais et al., 2021) | 0.878 | **0.899** | **0.88** | **0.85** |
| | STIGPN (Wang et al., 2021) | **0.919** | 0.880 | - | - |
| H2O (Kwon et al., 2021) | ISTA-Net (Wen et al., 2023) | - | **0.891** | - | - |

However, not only the quantitative performance is relevant but also general aspects, which are compared in Table 2.2. This includes whether the approaches use previously extracted higher-level features or rely on raw *visual features*. Additionally, some methods explicitly handle *segmentation*, while others focus on classification but can be adapted for segmentation with a sliding window

approach. This aspect often affects whether the method is suitable for *real-time* applications. For example, methods such as ASSIGN (Morais et al., 2021) and PGCN (Xing and Burschka, 2022) require the entire demonstration for inference, making them unsuitable for real-time use. Another key consideration is object knowledge, which involves the model's ability to recognize and track the same symbolic object across multiple data points from different demonstrations. This means the model is able to identify that the same object (e. g., a rolling pin) is held during symmetric motions and retain this information during inference.

Table 2.2.: Comparison of the different approaches based on general aspects.

| Model | Visual Features | Segmentation | Object Knowledge |
|---|---|---|---|
| GNN (Dreher et al., 2020) | | | ✓ |
| STIGPN (Wang et al., 2021) | ✓ | | |
| ASSIGN (Morais et al., 2021) | ✓ | ✓ | |
| PGCN (Xing and Burschka, 2022) | | ✓ | ✓ |
| ISTA-Net (Wen et al., 2023) | | | |

In this thesis, the focus is on detecting different categories of bimanual manipulation. While, as shown in Section 2.2.1, for this purpose rule-based approaches are commonly used in the field of category recognition, recent action recognition methods show the enormous potential of learning-based approaches. This thesis will design and evaluate both rule-based and learning-based methods for the simultaneous classification and segmentation of bimanual categories in human demonstrations.

## 2.3. Constraints in Robotic Bimanual Manipulation

To implement efficient and robust bimanual control policies, it is essential to consider the various constraints involved. In Section 2.3.1, we analyze approaches that explicitly address these constraints. Although this necessitates a comprehensive understanding of the underlying constraints to develop the respective approach, the resulting controllers are human-understandable. Recently, learning-based approaches have gained popularity, which we will discuss in Section 2.3.2. In these approaches, bimanual coordination is not explicitly encoded but implicitly emerges from the learned policy.

### 2.3.1. Explicit Consideration of Constraints

Bimanual manipulation is influenced by various types of constraints that govern the coordination between hands. Pek et al. (2023) classify these constraints into two primary categories: spatial and temporal, each of which can be further subdivided. Temporal constraints arise in scenarios where specific sequences of actions are required, such as opening a bottle before pouring, or when a continuous spatial-temporal relationship must be maintained, such as ensuring a stable relative pose between hands while jointly carrying a large object. Spatial constraints, on the other hand, define dependencies in the spatial domain. These constraints may exist at a symbolic level, such as placing an object on top of another, or at the trajectory level, as seen in tasks like stirring a cup held by the other hand. In robotics, force-based constraints are also frequently considered. These involve the transfer of forces between the hands, either directly, as in clamping an object between the hands, or indirectly through tools, such as holding a vegetable in one hand while using a peeler with the other.

#### Temporal Constraints

Several approaches exist for describing temporal constraints. Allen's Interval Algebra (Allen, 1983) is a well-established formalism for describing temporal relations using 13 distinct relations, such as *before* and *during*. Building on this, Dreher and Asfour (2022) propose a softened formulation of temporal relations that is applicable to real-world data, particularly for bimanual manipulation tasks. They further represent temporal task models using graphs and infer subtasks from these representations. In a more recent work, Dreher and Asfour

(2024) use Gaussian mixture Models for the representation of temporal differences between semantic temporal waypoints. They further show how fuzzy Allan relations can be inferred from those.

Another well-established formalism is Linear Temporal Logic (LTL) (Pnueli, 1977) which allows a mathematically precise and unambiguous formulation of temporal dependencies. In Roşu and Bensalem (2006) it is shown that a discrete variant of Allen's temporal logic can be translated into an equivalent LTL formula. Puranic et al. (2021) use an extension of LTL, signal temporal logic (STL) to rank demonstrations based on their temporal specifications and to infer rewards for model-free reinforcement learning.

Other methods for describing temporal constraints include task precedence graphs (Pardowitz et al., 2007) to describe sequential dependencies between different subtasks. Zöllner et al. (2004) propose one Petri net to model the concurrency in bimanual manipulation. In this net, each hand can assume one of two conditions (*ready* or *active*). Events are triggered either when the task of one specific hand is triggered or if there are new tasks. Compared to alternative methods like behavior trees, Petri nets inherently provide a robust framework for modeling concurrent and parallel processes. This makes them particularly advantageous for representing shared resources, such as the coordination of two hands in the context of bimanual tasks. These constraints can also be directly implemented into control strategies. For instance, Mirrazavi Salehian et al. (2018) present an approach for coordinating multiple agents to reach a moving object while avoiding self-collision, using a centralized inverse kinematic solver formulated as a quadratic program.

### Spatial Constraints

Spatial constraints can be described at a symbolic level using spatial relations such as *right of*, *behind*, and *close* (Ziaeetabar et al., 2018; Kartmann et al., 2020) or at a trajectory level. Trajectory-level coordination often employs a leader-follower approach (Luh and Zheng, 1987; Zhou et al., 2016) or the cooperative-task space (CTS) method (Uchiyama and Dauchez, 1992), both of which focus on physical interaction tasks.

Aksoy et al. (2011) introduced the extraction of symbolic relations, such as *touching* and *overlapping*, from segmented images. They analyzed the temporal evolution of spatial relations during tasks by employing transition matrices referred to as Semantic Event Chains (SECs). Building on this foundation,

Ziaeetabar et al. (2018) extended SECs to include a broader spectrum of static and dynamic spatial relations. Static relations, such as *contact*, *right*, and *front*, can be derived from individual frames, whereas dynamic relations, such as *moving together* and *getting close*, require temporal information from preceding frames. The evaluation of these spatial relations was facilitated by approximating objects using 3D bounding boxes, which were generated from RGB-D images. Similar symbolic spatial relations are considered by (Kartmann and Asfour, 2023), which focus on incremental learning of a generative, sub-symbolic representation of spatial relations based on cylindrical probability distributions (Kartmann et al., 2020, 2021) by having the robot request more information from the human. To this end, symbolic representations commonly used in human language are grounded in sub-symbolic representations to bridge the gap between abstract concepts and sensory data.

Other approaches focus on spatial relations at the trajectory level. Traditionally, bimanual motions have been represented in two primary ways: the Cooperative Task Space (CTS) representation (Uchiyama and Dauchez, 1992; Chiacchio et al., 1996), which models the manipulation of a rigid object by a dual-arm system, and leader-follower approaches, where the motion of one hand serves as the reference frame for the other (Smith et al., 2012). Park and Lee (2015, 2016) introduced the Extended Cooperative Task Space (ECTS), a unified framework that integrates the CTS representation with the leader-follower approach, providing a more comprehensive representation of bimanual motion.

The velocity vector of the first ($\dot{\boldsymbol{x}}_1$) and the second manipulator ($\dot{\boldsymbol{x}}_2$) are described as

$$\begin{bmatrix} \dot{\boldsymbol{x}}_1 \\ \dot{\boldsymbol{x}}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{I}_6 & -(1-\alpha)\boldsymbol{I}_6 \\ \beta \boldsymbol{I}_6 & \alpha \boldsymbol{I}_6 \end{bmatrix} \begin{bmatrix} \dot{\boldsymbol{x}}_a \\ \dot{\boldsymbol{x}}_r \end{bmatrix} \tag{2.1}$$

dependent on the absolute velocity ($\dot{\boldsymbol{x}}_a$) and the relative velocity ($\dot{\boldsymbol{x}}_r$) and parameterized by the balance coefficient $\alpha \in [0, 1]$ and the coordination coefficient $\beta \in 0, 1$. Based on the parameterization of $\alpha$ and $\beta$ four modi can be described: orthogonal (uncoordinated), serial (analogous to leader-follower), parallel (where hand motions are relative to an absolute trajectory between them) and blended. Recently, ECTS combined with quadratic programming was used for grasping and tossing an object in motion (Bombile Bosongo and Billard, 2022).

Several studies leverage the concepts of *global* and *relative* task definitions in the context of bimanual manipulation. These definitions pertain to the reference

frame in which a trajectory is *relative*, which can be a *global* reference frame within the environment or relative to other frames in the scene, such as objects or frames on the robot. One approach describes tasks as sequences of relative, global, and local targets, where local means with respect to the robot frame (Stavridis et al., 2021). Another approach represents tasks as absolute and relative skills encoded as Dynamical Movement Primitives (DMPs) (Pairet et al., 2019). Liu et al. (2023) propose a relative parameterization method for bimanual manipulation based on Gaussian Mixture Models (GMMs) and motion generation methods based on optimal control for leader-follower and synergistic motions. Furthermore, some researchers introduce a hierarchy prioritizing relative goals over global ones (Tarbouriech et al., 2022), while others use deep learning methods to predict when relative actions are required (Kim et al., 2024). Gao et al. (2024) present a framework for visual imitation learning of bimanual manipulation actions, based on hybrid master-slave relationships between the hands and extracted key points on objects. This framework allows for a flexible combination of action definitions, both relative to the other hand and the environment, which can be automatically learned.

### Force-based Constraints

Force-based constraints between the hands (directly or via tools and objects) play an essential role in tightly-coupled bimanual tasks. The cooperative transport of a single object has been extensively investigated. In these tasks, a certain "squeezing" force must be applied to maintain the object's stability between the hands. The methods employed for such tasks can be categorized into two main approaches: object-level impedance control and a combination of Projected Inverse Dynamic Control (PIDC) with the grasp matrix. Object-level impedance control has been explored by researchers such as Wimböck et al. (2012) and Shahbazi et al. (2017). The combination of PIDC and the grasp matrix has been discussed in works by Walker et al. (1991), Dehio et al. (2018), and Gao et al. (2018a). Interaction forces also serve as significant constraints in asymmetric tasks, as described in Pais Ureche and Billard (2018).

In summary, various methods have been developed to address temporal, spatial, and force constraints in bimanual manipulation. These methods are typically designed for specific application scenarios and are not easily adaptable to different types of actions. In this work, we aim to establish general representations for spatial and temporal constraints between the hands, drawing from established concepts. Furthermore, we will develop template versions of

these constraints tailored to the different categories in the Bimanual Manipulation Taxonomy. By representing a task as a sequence of bimanual categories, we can derive the appropriate temporal and spatial constraints for the entire task.

### 2.3.2. Implicit Consideration of Constraints

In addition to the methods described in the previous subsection, several approaches implicitly encode the constraints between the hands within the task model. Early work predominantly employed Movement Primitives. Later, those approaches were supplemented with deep learning-based methods which can be split into methods using imitation learning and reinforcement learning.

#### Movement Primitives for Bimanual Manipulation

Common representations of movement primitives for general manipulation tasks, not specifically bimanual cases, include Dynamic Movement Primitives (DMPs) (Ijspeert et al., 2002a), Probabilistic Movement Primitives (ProMPs) (Paraschos et al., 2013), Task-Parameterized Gaussian Mixture Models (TP-GMMs), and Via-Point Movement Primitives (VMPs) (Zhou et al., 2019). Methods of task-specific generalization of DMPs are presented in (Ude et al., 2010). To extend DMPs for interaction with the environment, Coupling Movement Primitives (CMPs) (Gams et al., 2014) integrate force/torque feedback. This framework has been demonstrated to facilitate bimanual tasks in scenarios where control of the two arms is distributed rather than centralized. *Distributed*, in contrast to *centralized*, refers to the absence of a central coordinating entity, with each controlled arm governed by its own independent controller (Umlauft et al., 2014). Similarly, another approach to cooperative manipulation under distributed control, based on DMPs, is proposed by Umlauft et al. (2014). Their method implements formation control by incorporating a cooperation feedback term derived from artificial potential fields.

For centralized bimanual manipulation, Coordinate Change Dynamic Movement Primitives (CCDMPs) (Zhou et al., 2016) offer an alternative by defining the DMPs governing the follower's motion within the reference frame of the leader. This leader-follower configuration enables coordinated control. Batinica et al. (2017) propose a control strategy for bimanual manipulation that achieves

low trajectory errors and compliant control without relying on explicit task models, utilizing Compliant Movement Primitives (CMPs) (Deniša et al., 2015). The end effectors are compliant in their absolute task but stiff in their relative task. This is accomplished by formulating tasks with separately adjustable stiffness and damping for the relative and absolute tasks, and by applying additional torque based on virtual force translation. However, the CMPs and corresponding torque profiles are task-specific, limiting generalizability. Additionally, since CMPs are defined in joint space, the approach is effective only when the robot's joint configuration matches that of the demonstration.

In some cases, movement primitives are used to reduce the action space in imitation learning. Xie et al. (2020) represent movement primitives as graph recurrent neural networks (RNNs with a graph attention layer) and combine a high-level planner for sequencing 13 pre-learned primitives and a low-level controller to combine primitive dynamics. The learning model is specifically designed to capture relational information within the scene, thereby facilitating bimanual manipulations. The sequence of primitives is inferred based on the observed sequence of states. The safe, interactive movement primitive learning (SIMPLe) algorithm (Franzese et al., 2023) offers a method for encoding policies using graphs based on Gaussian process regression. This approach enables the interactive adaptation and synchronization of single-arm demonstrations through kinesthetic human feedback, providing a foundation for versatile and responsive bimanual coordination.

Building on the concept of combining primitive actions with more learning-based control techniques, Avigal et al. (2022) propose a garment folding framework. Their method integrates a neural network with predefined bimanual action primitives, allowing the prediction of gripper pose pairs and the parameterization of action sequences. This hybrid approach leverages the strengths of both data-driven predictions and structured primitives. Extending to reinforcement learning applications, Amadio et al. (2019) present a symmetrization method for probabilistic movement primitives (ProMPs). By mapping a single-arm MP to the second arm, they enable efficient learning of coordinated bimanual tasks, demonstrating accelerated convergence in reinforcement learning scenarios. Stepputtis et al. (2022) introduce an imitation framework for contact-rich bimanual manipulation tasks. Their approach employs Bayesian Interaction Primitives (BIP) for temporal and spatial reasoning, highlighting the critical role of force/torque data in phase estimation and task success. Applied to insertion tasks, their study also assesses manipulation robustness, user mental demand, and the effectiveness of teaching methods such as kinesthetic guid-

ance and teleoperation. The orchestration of predefined, hand-crafted skills in bimanual scenarios has been explored using a Large Language Model (LLM) to generate action plans that include both coordinated and uncoordinated sections (Chu et al., 2024). Results show that integrating bimanual category labels, as outlined in taxonomies, into the LLM prompts improves the effectiveness of this approach.

### Imitation Learning

Imitation learning involves acquiring a desired behavior by replicating the actions of a demonstrator (Osa et al., 2018). In this thesis, the term specifically refers to supervised, deep learning-based methods for Learning from Demonstrations (LfD). Recently, transformer architectures have emerged as the most commonly used models in imitation learning, demonstrating strong performance across various applications. In some instances, these architectures are integrated with classifiers and leverage movement or action primitives.

An early approach leveraging transformers for bimanual manipulation is presented in Kim et al. (2021). Self-attention is used to mask out sensory input which is not relevant for the current task. Building on the potential of transformer-based approaches, ALOHA (Zhao et al., 2023) has recently emerged as a notable framework. A low-cost bimanual robot setup including a teleoperation interface is presented as well as the imitation learning approach Action Chunking with Transformers (ACT). This end-to-end learning policy is using action chunking and temporal assembling to help with non-markovian behavior (depending on past states) and mitigating the effect of compounding errors. The approach is evaluated based on about 10 minutes of training data for each task. The tasks are selected to be precise, contact-rich and dynamic. The performance of ACT is compared to multiple baseline approaches. The work is extended to mobile manipulation in Fu et al. (2024b). The hardware setup is adapted to allow for mobile manipulation, and the policy learning method is applied for mobile manipulation tasks. The authors show that co-training with the data from the static ALOHA dataset can be efficient and robust for different data mixtures. This work is extended in Lee et al. (2024) through the integration of hierarchical attention mechanisms, designed to capture dependencies between the joint states of the arms and visual input. By employing a hierarchical attention encoder along with a multi-arm decoder, the authors demonstrate, through both simulation and real-world experiments, that this approach achieves superior performance compared to ACT. Their results indicate that incorporating mechanisms to capture

inter-arm coordination within the network architecture can significantly enhance the effectiveness of the imitation learning framework.

Some works include different hand roles in the design of their approach. Grannen et al. (2023) differentiate between a *stabilizing* and an *acting* role for the hands. Policies are learned for both hands individually. The policy for the stabilizing hand is conditioned on a visual key point which is extracted from the environment and the policy for the acting hand. Similarly, in Liu et al. (2024a) the arms are assigned either an *acting* or a *stabilizing* role. They present VoxAct-B, which is a voxel-based and language-conditioned method that leverages Vision Language Models (LVMs). Additionally, they extend the RLBench benchmark (James et al., 2020) with asymmetric bimanual tasks (open drawer, open jar, put item in drawer). In a parallel work, Grotz et al. (2024) also propose a benchmark for bimanual manipulation tasks by extending the RLBench benchmark. A new network architecture called *PerAct*$^2$ is presented which extends to existing *PerAct* framework (Shridhar et al., 2023) to bimanual cases. *PerAct*$^2$ takes a 3D voxel grid, proprioception data and a language goal as input and generates a discretized action (6 Degree of Freedom (DoF) end-effector pose, gripper state, flag for collision-free motion planning) for each arm. Unlike e. g., Zhao et al. (2023) action goals are generated in task-space instead of in joint space. To enable coordination between both arms a new self-attention module is suggested.

The Bimanual Keypose-Conditioned Consistency Policy (BiKC) (Yu et al., 2024) is a hierarchical imitation learning framework that incorporates a high-level keypose predictor and a low-level trajectory generator using consistency models (CMs) as the visuomotor policy. This design addresses the typical inference latency found in diffusion policies by enabling CMs to produce outputs in a single step, whereas standard diffusion models require multiple denoising steps. Evaluated within the ALOHA framework, this approach shows improvements in both success rate and operational efficiency.

Drolet et al. (2024) provide a comparison of imitation learning algorithms for bimanual insertion tasks in a MuJoCo simulation. They compare Generative Adversarial Imitation Learning (GAIL), Implicit Behavioral Cloning (IBC) (Florence et al., 2022), Dataset Aggregation (DAgger) (Ross et al., 2011), Behavioral Cloning (BC) (Pomerleau, 1988), Action Chunking Transformer (ACT) (Zhao et al., 2023), and Diffusion policy (Chi et al., 2023) trained on 200 expert demonstrations. They found that ACT and Diffusion performed best in particular in regard to hyperparameter and noise tolerance, compute efficiency and training stability.

Table 2.3 provides a summary of relevant studies published after 2020, focusing on the neural network architectures employed. The table explicitly identifies the use of diffusion policies, transformer-based models and classifiers, while other deep learning architectures are not specified. Additionally, it highlights cases where movement primitives or primitive actions are integrated into the proposed approaches and whether distinct roles of the hands/arms are assumed. As shown in Table 2.3, relatively few studies focus on diffusion models for bimanual manipulation. This might be due to the significant inference latency inherent to diffusion models due to iterative denoising. However, diffusion models are frequently used as a baseline for comparison (Fu et al., 2024b; Grotz et al., 2024; Drolet et al., 2024). While primitive-based approaches were more prevalent in earlier work, they remain in use, often in combination with modern techniques such as transformers. Further, it is interesting to note that several works do not rely on pure end-to-end learning but include some prior knowledge in their model. This can mean using classifiers e. g., to select suitable, predefined primitives (Grannen et al., 2022; Avigal et al., 2022) or assuming the hands taking over distinct roles such as the *stabilizing* and *acting* hand in Grannen et al. (2023).

Table 2.3.: Publications after 2020 employing deep learning-based approaches for robotic bimanual manipulation.

| Publication | Diffusion | Transformer | Classifier | Primitives | Hand Roles |
|---|---|---|---|---|---|
| (Kim et al., 2021) | | ✓ | | | |
| (Liu et al., 2022) | | ✓ | | ✓ | ✓ |
| (Grannen et al., 2022) | | | ✓ | ✓ | ✓ |
| (Stepputtis et al., 2022) | | | | ✓ | |
| (Avigal et al., 2022) | | | ✓ | ✓ | |
| (Liu et al., 2023) | | | | ✓ | ✓ |
| (Franzese et al., 2023) | | | | ✓ | |
| (Zhao et al., 2023) | | ✓ | | | |
| (Grannen et al., 2023) | | | ✓ | | ✓ |
| (Fu et al., 2024b) | | ✓ | | | |
| (Lee et al., 2024) | | ✓ | | | |
| (Kim et al., 2024) | | ✓ | ✓ | ✓ | |
| (Yu et al., 2024) | ✓ | | | | |
| (Grotz et al., 2024) | | ✓ | | | |
| (Liu et al., 2024a) | | ✓ | | | ✓ |

### Reinforcement Learning

While imitation learning is limited to the available demonstrations, Reinforcement Learning (RL) allows the robot to discover new policies by exploring the state-action space (Billard et al., 2016). As mentioned in (Oh et al., 2024) deep RL holds specific challenges for bimanual manipulation, namely 1. *credit assignment* (each arms contribution to the reward function), 2. *vanishing memory* (restricts handling multiple action sequences) and 3. the *exploration-exloitation trade-off* which is more severe in the bimanual case.

Several studies explore RL with sparse rewards in the context of bimanual manipulation. This assumption is particularly relevant in robotic manipulation, where success is typically defined by the complete and accurate execution of the task. For instance, in a pick-and-place scenario, the task is considered successful only if the object is safely and correctly placed in the target location. Chitnis et al. (2020) focus on improving the sample efficiency of model-free RL for sparse-reward tasks by learning to compose skills. Their approach involves dividing the problem into two parts: learning a state-independent task schema, which is a sequence of skills, and a state-dependent policy for selecting appropriate parameters. The method assumes the availability of a library of generic skills. Tasks are bimanual, but the bimanuality is not explicitly considered in the model design. Similarly, Zhang et al. (2021) also address sparse rewards by decomposing a task into subtasks and using curriculum learning. The approach ensures that each robot focuses on distinct sub-tasks, minimizing the risk of one arm dominating the other or causing operational conflicts. The framework leverages an attention mechanism and a multilayer perceptron (MLP) to learn the desired position shifts and gripper states for both arms. In another work, Li et al. (2023) propose a symmetry-aware actor-critic framework that captures the interchangeable roles of the two arms in bimanual handover and rearrangement tasks. To address sparse rewards, they utilize curriculum learning and introduce an object-centric relabeling technique based on Hindsight Experience Replay (HER) (Andrychowicz et al., 2017). As an alternative to curriculum learning, Schwarke et al. (2023) leverage intrinsic motivation to guide agent exploration, utilizing Random Network Distillation (RND) as a mechanism for exploration incentives. This approach is applied to tasks such as door opening and package manipulation, demonstrating its effectiveness in mobile bimanual manipulation with a wheeled-legged robot.

A significant challenge in RL is the high cost and potential safety risks associated with training models on real-world data. As a result, RL models are

frequently trained in simulation environments like MuJoCo or Isaac Gym. However, deploying these learned policies on real robots involves the often difficult task of sim-to-real transfer, where the discrepancies between the simulated and real-world environments must be effectively addressed. Kataoka et al. (2022) present a reinforcement learning approach for a bimanual connect task and describe in detail how to modify the RL framework, particularly the simulator modifications, to facilitate sim-to-real transfer. Lin et al. (2023) instead focus on RL and the corresponding sim-to-real transfer for bimanual tasks using tactile sensing.

In contrast to previously discussed approaches where robot arms do not have hands at all (Lin et al., 2023; Schwarke et al., 2023) or only two-jaw grippers (Kataoka et al., 2022; Chitnis et al., 2020; Zhang et al., 2021), in other works two dexterous, multi-fingered hands are controlled. (Yang et al., 2024) present the Asymmetric Dexterity (AsymDex) framework which leverages the inherent symmetry between the hands to reduce the dimensional complexity in Rl of bimanual manipulation tasks, thereby enhancing sample efficiency. The framework comprises two main components. First, following an approach similar to Grannen et al. (2022), AsymDex assigns specialized roles to each hand: the *facilitating hand* repositions and reorients the object, while the *dominant hand* performs intricate, dexterous manipulation. Second, AsymDex learns the *relative* motion between the two hands, focusing on the coordinated interaction between them.

The BiDexHD framework (Zhou et al., 2024) standardizes task construction from existing bimanual datasets and utilizes a teacher-student policy learning approach. The teacher learns state-based policies across related tasks using a unified reward function, followed by policy distillation. The student consolidates these multi-task policies into a vision-based policy. Evaluation is conducted on tasks spanning six categories from the TACO dataset (Liu et al., 2024b).

In addition to standard single-agent RL methods, cooperative Multi-Agent Reinforcement Learning (MARL) can be utilized for bimanual manipulation tasks. (Chen et al., 2022) present the benchmark Bi-DexHands for bimanual manipulation based on two floating dexterous hands and tasks with varying complexity associated with different levels of human motor skills. They compare single-agent RL algorithms such as Proximal Policy Optimization (PPO) (Schulman et al., 2017) or Soft-Actor Critic (SAC) (Haarnoja et al., 2018) with multi-agent approaches such as Heterogenous-Agent Proximal Policy Optimization (HAPPO) (Kuba et al., 2021) and Multi-Agent Proximal Policy Optimization

(MAPPO) (Yu et al., 2022) where each agent represents a hand. Results indicate that single-agent PPO generally outperforms multi-agent methods, although the performance gap diminishes for more complex tasks. SAC consistently shows the lowest performance. Gu and Demiris (2024) combine MAPPO with a variable impedance action space for the contact-rich task of bathing bed-bound people using two robot arms. They show the superior performance of MAPPO compared to PPO for their application. Focusing on safe MARL, (Zhan and Chin, 2024) propose Multi-Agent Constrained Proximal Advantage Optimization (MACPAO) which ensures adherence to safety constraints in each iteration by sequentially updating the agents. Their approach is evaluated based on multiple bimanual object manipulation tasks with two floating dexterous hands.

Recent progress in RL includes the introduction of Transformer Reinforcement Learning (TRL). A notable development in this domain is the Multi-Agent Transformer (MAT) (Wen et al., 2022), designed to integrate cooperative MARL with general sequence models (SMs). MAT employs an encoder-decoder structure and utilizes the multi-agent advantage decomposition theorem to recast the joint policy optimization task as a sequential decision-making problem. Evaluated against various benchmarks, including Bi-DexHands (Chen et al., 2022), which addresses bimanual dexterous manipulation, MAT demonstrates superior performance compared to state-of-the-art MARL methods such as MAPPO and HAPPO. Temporal Context Transformer Reinforcement Learning (TC-TRL) (Oh et al., 2024) extends these advancements by addressing long-horizon tasks through a hybrid offline-online learning approach. Offline learning with behavior cloning (BC) reduces excessive exploration, while online learning employs a temporal-context transformer module integrating an observation encoder, a multi-agent critic network, and an action decoder to generate optimal action sequences. For long-horizon tasks, TC-TRL achieves superior performance compared to MAT, MAPPO, and HAPPO.

The approaches discussed in this section are summarized and compared in Table 2.4. Among them, PPO-based methods encompass both single-agent and multi-agent reinforcement learning techniques leveraging PPO. The comparison highlights that the application of RL for bimanual, dexterous manipulation on real robots remains a largely underexplored area.

Table 2.4.: Comparison of publications in the field of RL considering the usage of dexterous (multi-fingered) hands, the execution of real-world experiments and the method used for RL: Soft-Actor Critic (SAC), approaches based on Proximal Policy Optimization (PPO) or using Transformers.

| Publication | Dexterous Hands | Real-Robot Experiments | SAC | PPO-based | Transformer RL |
|---|---|---|---|---|---|
| (Chitnis et al., 2020) | | ✓ | | ✓ | |
| (Zhang et al., 2021) | | | ✓ | | |
| (Kataoka et al., 2022) | | ✓ | | ✓ | |
| (Chen et al., 2022) | ✓[1] | | ✓[2] | ✓[2] | |
| (Wen et al., 2022) | ✓[1] | | | | ✓ |
| (Li et al., 2023) | | ✓ | ✓ | | |
| (Lin et al., 2023) | | ✓ | | ✓ | |
| (Schwarke et al., 2023) | | ✓ | | ✓ | |
| (Gu and Demiris, 2024) | | ✓ | | ✓ | |
| (Zhan and Chin, 2024) | ✓[1] | | | ✓ | |
| (Yang et al., 2024) | ✓[1] | | | ✓ | |
| (Zhou et al., 2024) | ✓ | | | ✓ | |
| (Oh et al., 2024) | ✓ | | | | ✓ |

[1] only hands considered (no robotic arm) [2] benchmarking several methods

### 2.3.3. Discussion

Bimanual coordination constraints can be addressed through various approaches. One strategy involved explicit modelling of temporal and spatial constraints, as discussed in Section 2.3.1. This approach requires a detailed understanding of the task, providing human-interpretable insights into the control framework, which facilitates error detection and correction.

Alternatively, implicit learning-based methods (see Section 2.3.2) autonomously infer bimanual constraints and embed them directly into their models. While these methods eliminate the need for explicit task modeling, imitation learning approaches often require extensive training data and face challenges in generalizing beyond the provided training distribution. Reinforcement Learning (RL) methods, in contrast, do not rely on pre-recorded training data and can independently explore solutions of the task. However, the design of effective reward functions can be tedious and task-specific. Furthermore, since most RL approaches are trained in simulation, transferring learned

policies from simulation to real-world applications remains a persistent challenge.

This thesis primarily utilizes explicit constraint formulations that integrate temporal and spatial constraints for bimanual manipulation. These constraints are grounded in the proposed taxonomy, which defines general coordination patterns in bimanual manipulation tasks rather than being tailored to specific tasks. In contrast, recent trends and advances in robotic bimanual manipulation often rely on learning-based methods. While this thesis does not expand upon such approaches, an analysis of the related literature highlights that general knowledge of bimanual coordination patterns, as outlined in our taxonomy, can inform model architectures (e. g., Grannen et al. (2023); Liu et al. (2024a); Yang et al. (2024)) or provide insights that facilitate subaction scheduling (Chu et al., 2024). This suggests that taxonomy-supported approaches hold significant potential in areas such as imitation learning and reinforcement learning, where they could enhance model design and improve task execution strategies.

## 2.4. Datasets for Bimanual Manipulation

This section reviews relevant human motion datasets, with a particular focus on multi-modal bimanual recordings of daily household and kitchen activities. The datasets are categorized based on the sensor modalities used. An overview of the most relevant related works is presented in Table 2.5. Our comparison considers several key factors including the provided action annotations, the sensor modalities, the accuracy of whole-body pose and object interaction, and the variations captured within each action type. Many datasets offer unconstrained recordings of various subjects performing naturally in unstructured environments, aiming to capture a broad variance in data across all dimensions. Nevertheless, the explicit introduction of single variations in object types and relations, as well as bimanual execution of actions, can be advantageous for research focused on generalizing bimanual task models. This approach allows for a more precise analysis and comparison of how different task parameters influence execution, thus facilitating a deeper understanding of coordination in bimanual manipulation tasks

### 2.4.1. Single-View Video Datasets

Several large-scale datasets capturing video recordings of humans performing various actions in different daily scenarios are available. Head-mounted video cameras are frequently utilized to record a human subject's field of view, creating datasets of daily activities in natural environments (e. g., Damen et al. (2022); Fathi et al. (2011); Cai et al. (2017)). These cameras are convenient to attach to the human body and can continuously capture the subject's workspace, even during mobile manipulation. The *EPIC Kitchen Dataset-100* (Damen et al., 2022), which includes 100 hours of long-term unscripted kitchen activities, is the largest annotated egocentric action dataset. Similarly, the *20BN-Something-Something dataset* Goyal et al. (2017) offers a large collection of very short video clips containing both first- and third-person human-object interactions. However, it is important to note that both datasets lack recordings of the human whole-body motion and have been collected within an unknown or changing camera coordinate system.

In Das et al. (2013), the *YouCook* dataset was created by collecting and annotating 88 publicly available third-person cooking videos from YouTube. By contrast, the *MPII Cooking Activities Dataset* (Rohrbach et al., 2016) consists of 27 hours of annotated videos captured using a static camera setup, depicting subjects preparing real dishes in a kitchen environment. These datasets typically consist of a large volume of videos captured in diverse, unstructured settings, benefiting from the relatively low effort required for data collection. Recordings are either sourced from open video platforms or captured in real-world scenarios, as they do not necessitate extensive sensor setups or specialized equipment. The substantial size of these datasets makes them particularly valuable for training and evaluating machine learning algorithms, especially in the domains of action recognition, detection, anticipation, and retrieval.

Various computer vision techniques for extracting 2D (Cao et al., 2019) and 3D (Mehta et al., 2017) human poses, grasp types (Cai et al., 2017), as well as object bounding boxes (Redmon et al., 2016) and 3D poses (Pauwels and Kragic, 2015) from RGB videos, can be applied to derive various types of information from video data. However, achieving accurate information retrieval across diverse scenarios and conditions remains challenging. Moreover, the development of robot manipulation concepts based on video datasets, as outlined in Shao et al. (2020), could greatly benefit from datasets that provide comprehensive insights into human demonstrations.

### 2.4.2. Multi-View and/or Multi-Modal Video Datasets

Collecting multi-view video datasets or employing additional sensor modalities to capture human demonstrations enhances the extraction of comprehensive knowledge from these demonstrations.

The *TUM Kitchen Data Set* (Tenorth et al., 2009) integrates multiple RGB camera streams to reconstruct three-dimensional human poses. Additionally, RFID tags and magnetic sensors are utilized to detect actions such as opening a door or drawer while setting a table. The *50 Salads dataset* (Stein and McKenna, 2013) includes task recordings of different salad preparations, incorporating rough object trajectories obtained from accelerometers. In Pieropan et al. (2014), subjects are observed preparing cereal from multiple viewpoints, with audio signals recorded as an additional sensor modality. Another approach involves using an egocentric RGB-D camera to collect data for classifying grasp types and predicting contact points and forces (Rogez et al., 2015).

*CAD-120* (Koppula et al., 2013) collect RGB-D data from a single view for 10 different high-level, long-term activities in a kitchen context. The *ETRI-Activity3D* dataset (Jang et al., 2020) is designed to capture motions for recognizing daily activities of elderly individuals. RGB-D videos of a large group of both young and elderly subjects are recorded from eight different viewpoints. In the *LEMMA* dataset (Jia et al., 2020), two static RGB-D cameras and two egocentric RGB cameras are used to record two agents cooperating to perform tasks in various kitchen and living room environments. The *KIT Bimanual Actions Dataset* (Dreher et al., 2020) include RGB-D videos from a single camera and focuses on human action recognition in bimanual household tasks emphasizing spatial relations. Action labels are assigned to each hand individually, increasing the granularity of the provided annotations. Similarly, the *H2O* (Kwon et al., 2021) dataset provides multi-view RGB-D (including egocentric) recordings associated with respective object and hand poses and interaction labels. Recent datasets such as *GigaHands* (Fu et al., 2024a) focus on the combination of motion recordings and text annotations. In a studio setting over 50 subjects interact with more than 400 real-world objects are recorded. Based on an instruct-to-annotate pipeline instruction scripts are generated that result in recorded motion sequences with a high variety.

Similar to larger single video datasets, most of the datasets discussed offer primarily unconstrained motion recordings, which are essential for the training and evaluation of action recognition and prediction methods. Additionally, these datasets play a crucial role in tasks that involve learning from human

observations. Examples include identifying changes in 3D semantic relations during bimanual manipulation (Dreher et al., 2020), learning simple motion primitives (Dometios et al., 2018), and understanding object affordances (Koppula et al., 2013). Nevertheless, complex bimanual manipulation scenarios require more sophisticated motion tracking approaches. These scenarios often involve multiple small or featureless objects and occlusions, making it challenging to accurately extract motion trajectories of both subjects and objects.

### 2.4.3. Motion Capture Datasets

Several large-scale human motion databases have been developed (Ionescu et al., 2014; Mandery et al., 2015; Mahmood et al., 2019), yet most do not explicitly focus on object manipulation. The *KIT Whole-Body Human Motion Database* (Mandery et al., 2015) represents a notable exception, as it comprises extensive marker-based motion recordings of whole-body actions while also tracking interactions with objects. This database is continuously expanded with additional motion recordings and augmented with complementary sensor modalities, such as force plates and data gloves. This thesis contributes to the ongoing expansion of the database by incorporating additional motion recordings.

Other multimodal datasets have been introduced to capture specific human activities. The *Carnegie Mellon University Multimodal Activity* (CMU-MMAC) database (De la Torre et al., 2009) records subjects engaged in cooking and food preparation tasks using various sensor modalities. Meanwhile, the *AnDyDataset* (Maurice et al., 2019) captures industrial activities such as screwing and load manipulation under diverse conditions, employing a multi-modal sensor setup for classifying, predicting, or assessing human motions in industrial environments. However, both datasets primarily focus on capturing object motion within video data. By contrast, the *OPPORTUNITY Activity Recognition Data Set* (Roggen et al., 2010) involves subjects performing daily tasks, with their poses tracked using inertial measurement units (IMUs), and interactions with objects and the environment captured using various sensor modalities. In comparison, the *Daily Interactive Manipulation (DIM) Dataset* (Huang and Sun, 2019) is specifically designed for interactive manipulation, focusing on fine motions where objects or tools are manipulated to interact with other objects. The dataset features a custom-built handle equipped with a force-torque sensor attached to various

tools. It includes a large collection of short actions, particularly pouring actions with variations in objects and contents. Alongside force data, object poses are captured using RGB-D sensors. However, this dataset primarily captures subsymbolic information such as position and force trajectories of unimanual motions performed with the sensorized tool. This setup limits natural grasping and manipulation behaviors and does not include human motion during task execution. In Pais Ureche and Billard (2018), recordings involve subjects performing a "fruit scooping" task. Similar to the DIM Dataset, a sensorized tool is used, but in this case, the human hand and forearm are tracked using multiple sensors, including motion capture. The task involves bimanual manipulation where one hand guides a robotic arm to hold the fruit while the other performs the scooping action. The *GRAB: GRasping Actions with Bodies* dataset (Taheri et al., 2020) offers extensive marker-based motion capture recordings, encompassing full 3D human shape and pose sequences, including hand and face motions, as subjects interact with 51 different 3D-printed objects. This dataset emphasizes whole-body grasping and the accurate estimation of grasp contact surfaces. However, it primarily focuses on interactions with individual objects during bimanual grasping actions. Motion capture datasets exhibit various focuses. The *MoGaze* dataset (Kratzer et al., 2020) focuses on gaze tracking in unimanual pick-and-place tasks. It considers full body motions including eye-gaze and the workspace geometry. Other recent datasets focus on hand interactions with single articulated objects (Fan et al., 2023), or specific actions like flipping food during grilling (Pereira et al., 2022).

Recently, Liu et al. (2024b) introduced a large-scale dataset for hand-object manipulation, encompassing a wide variety of tool-action-object compositions in daily activities. Objects are tracked using markers that are subsequently removed from the data, while hand poses are detected through vision methods. Based on this dataset, three benchmark tasks are proposed: compositional action recognition, generalizable hand-object motion forecasting, and cooperative grasp synthesis. OAKINK 2 (Zhan et al., 2024) is another dataset focusing on hand-object interactions. This paper emphasizes hand mesh reconstruction, task-aware motion fulfillment (TaMF), and complex task completion (CTS). Additionally, Zakour et al. (2024) presents a dataset of hand-object interactions (HOIs), recording up to two subjects simultaneously and including longer sequences. Key contributions of this dataset are multi-view, multi-hand 3D pose annotations, and the tasks of Hand Mesh Recovery (HMR) and Hand Action Segmentation (HAS). A recent large dataset using marker-based motion tracking of the upper body and additionally recording with a stereocamera system

is BiCap (Carmona and Yu, 2024). The focus is on spatial, object and participant variations for the actions pour, open and pass. Only a task plan but no segmentation is provided for each recording.

### 2.4.4. Discussion

Recently, robot motion datasets have gained prominence in which demonstrations are generated not by recording human demonstrators but by directly executing actions on specific robotic platforms, such as through teleoperation.

This includes datasets where data from a single robot arm is collected such as BridgeData V2 (Walke et al., 2023), the Open X-Embodiment dataset (O'Neill et al., 2024) and (Khazatsky et al., 2024) but also datasets where two robot arms are considered such as ALOHA (Zhao et al., 2023), mobile ALOHA (Fu et al., 2024b) and BRMData (Zhang et al., 2024). The Open X-Embodiment dataset (O'Neill et al., 2024) is a collection of multiple existing robot training sets and seeks to train "generalist" policies, inspired by the success of general-purpose models in computer vision and NLP. They demonstrate that co-training on diverse, large-scale datasets enhances model performance, particularly when the original dataset is small. Most models, however, are specialized for specific robots, tasks, and environments, restricting their generalization. Further, most datasets only collect data with a single arm not capturing the coordination between both arms which is essential for humanoid robots. When using a different robot, embodiment transfer challenges similar to those encountered when learning from human data persist. Human motion recordings remain valuable due to their cost-effectiveness, accessibility, and the direct encoding of the demonstrator's motion intelligence, independent of specific teleoperation interfaces or robotic platforms.

An overview of the most relevant related works is provided in Table 2.5. Datasets are listed chronologically based on the year of their publication. Our comparison is based on the provided action annotations, sensor modalities, particularly the accuracy of whole-body pose and object interaction, and the captured variations within an action type. Only datasets that include bimanual manipulation actions are considered. At the time of writing this thesis, several of the more recent datasets were not yet publicly accessible, rendering it impossible to conclusively verify certain aspects.

Our *KIT Bimanual Manipulation Dataset* (Krebs et al., 2021) (see Chapter 4.1) exceeds prior datasets in key aspects, particularly in multi-modality, action

variability, and precise tracking of whole-body, hand, and object motions (see Table 2.5). The dataset extends the *Whole-Body Human Motion Database* (Mandery et al., 2015). The dataset's design is guided by the objective of enabling comprehensive learning of task models from human demonstrations, with an emphasis on capturing variations in object types and their relationships in bimanual tasks. Furthermore, it includes manually annotated actions for each hand, enhancing the dataset's granularity and utility for learning and analytical purposes. Considering these aspects our dataset was also not yet matched by more recently published datasets.

Table 2.5.: Overview of human motion datasets for bimanual object manipulation.

| General | | | | Label | | Action Variations | | | | Sensors | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | Marker-based Motion Capture | | | | | Vision | | | | Glove | | | Attached | | | | | | | | | | |
| | | | | | | | | | | | | | | | | | | | | | | | Type | | | | | | | Location | | | |
| Reference | Household actions | Bimanual execution | 3D object models | Fine-grained actions | Separate per hand | Unconstrained | Spatial variations | Tracked objects | Bimanual execution | Other[3] | Human whole-body | Hand configuration | Face mimics | Primary object | Secondary objects | RGB | Depth | Multi-view | Egocentric | Pressure | Kinematics | Both hands | Gyroscope | Magnetometer | Accelerometer | Torque sensor | Force sensor | Pose sensor | Other[4] | Human whole-body | Human arm | Other body parts | Objects | Environment | Audio |
| (De la Torre et al., 2009) | ✓ | ✓ | | | | ✓ | | | | | ✓ | | | | | ✓ | | ✓ | ✓ | | | | ✓ | ✓ | ✓ | | | | | | ✓ | | | ✓ |
| (Tenorth et al., 2009) | ✓ | ✓ | | ✓ | ✓ | ✓ | | | | | | | | | | ✓ | | ✓ | | | | | | | | | | | ✓ | | | | ✓ | |
| (Pham and Olivier, 2009) | ✓ | ✓ | | ✓ | | ✓ | | | | | | | | | | | | | | | | | | | ✓ | | | | | | | ✓ | | |
| (Roggen et al., 2010) | ✓ | ✓ | | ✓ | ✓ | ✓ | | | | | | | | | | | | | | | | | ✓ | ✓ | ✓ | | | | ✓ | ✓ | | | ✓ | ✓ | ✓ |
| (Stein and McKenna, 2013) | ✓ | ✓ | | ✓ | | ✓ | | | | | | | | | | ✓ | ✓ | | | | | | | | ✓ | | | | | | | | | |
| (Koppula et al., 2013) | ✓ | ✓ | | ✓ | | ✓ | | ✓ | | | | | | | | ✓ | ✓ | | | | | | | | | | | | | | | | | |
| (Pieropan et al., 2014) | ✓ | ✓ | ✓ | | | ✓ | | | | | | | | | | ✓ | ✓ | | | | | | | | | | | | | | | | | |
| (Bullock et al., 2014) | ✓ | ✓ | | | | ✓ | | | | | | | | | | | | | ✓ | | | | | | | | | | | | | | | |
| (Mandery et al., 2015) | ✓ | ✓[1] | ✓ | | | | ✓ | ✓ | | ✓ | ✓ | | | ✓ | ✓ | ✓ | | | | | | | | | | | | | | | | | | |
| (Goyal et al., 2017) | ✓[1] | ✓[1] | | | | ✓ | | | | ✓ | | | | | | | | | ✓[1] | | | | | | | | | | | | | | | |
| (Maurice et al., 2019) | | ✓ | | ✓ | | | | | | ✓ | ✓ | | | | | ✓ | | | | ✓ | ✓[1] | | ✓ | ✓ | ✓ | | | | ✓ | | | | | |
| (Huang and Sun, 2019) | ✓ | | | | | | ✓ | ✓ | | ✓ | | | | ✓[2] | ✓[2] | ✓[1] | ✓[1] | | | | ✓ | | | | | ✓ | ✓ | ✓[2] | | | | | ✓ | |
| (Dreher et al., 2020) | ✓ | ✓ | | ✓ | ✓ | ✓ | | | | | | | | | | ✓ | ✓ | | | | | | | | | | | | | | | | | |
| (Damen et al., 2022) | ✓ | ✓ | | ✓ | | ✓ | | | | | | | | | | ✓ | | | ✓ | | | | | | | | | | | | | | | |
| (Jia et al., 2020) | ✓ | ✓ | | ✓ | | ✓ | | | | | | | | | | ✓ | ✓ | ✓ | ✓ | | | | | | | | | | | | | | | |
| (Taheri et al., 2020) | | ✓[1] | ✓ | | | ✓ | | | | | ✓ | ✓ | ✓ | ✓ | | | | | | | | | | | | | | | | | | | | |
| (Jang et al., 2020) | ✓ | ✓ | | ✓ | | ✓ | | | | | | | | | | ✓ | ✓ | ✓ | ✓ | | | | | | | | | | | | | | | |
| (Nicora et al., 2020) | ✓ | ✓ | | ✓ | | | | | | | ✓[5] | | | | | ✓ | | ✓ | ✓ | | | | | | | | | | | | | | | |
| (Krebs and Meixner et al.) | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | | ✓ | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ | | | | | ✓ | ✓ | | | | |
| (Kwon et al., 2021) | ✓ | ✓ | ✓ | ✓ | | | | | | | | | | | | ✓ | ✓ | ✓ | ✓ | | | | | | | | | | | | | | | |
| (Fan et al., 2023) | ✓ | ✓ | ✓ | | | | | ✓ | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | | | | | | | | | | | | | | |
| (Liu et al., 2024b) | ✓ | ✓ | ✓ | ✓ | | | | ✓ | | | | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | | | | | | | | | | | | | | |
| (Zakour et al., 2024) | ✓ | ✓ | ✓ | ✓ | | | | | | | | | | ✓ | ✓ | ✓ | ✓ | ✓ | | | | | | | | | | | | | | | | |
| (Zhan et al., 2024) | ✓ | ✓ | ✓ | ✓ | | | | ✓ | | | ✓ | | | ✓ | ✓ | ✓ | | ✓ | ✓ | | | | | | | | | | | | | | | |
| (Carmona and Yu, 2024) | ✓ | ✓ | ✓ | | | | ✓ | ✓ | | | ✓[6] | | | | | ✓ | ✓ | ✓ | | | | | | | | | | | | | | | | |
| (Fu et al., 2024a) | ✓ | ✓ | ✓ | ✓ | | | | | | | | | | | | ✓ | | ✓ | | | | | | | | | | | | | | | | |

[1] only partially [2] switched from motion capture to pose sensors [3] mass, speed, etc. [4] RFID, power, pressure sensor, etc. [5] only right arm [6] only upper body

# CHAPTER 3

# Bimanual Manipulation Taxonomy

Manipulation remains a significant challenge in robotics. Compared to unimanual manipulation, bimanual tasks introduce additional complexities, including arm coordination strategies, redundancy resolution in closed kinematic chains, self-collision avoidance, and advanced force-based control (Smith et al., 2012). To address this complexity inherent to highly redundant systems, grasping taxonomies (Kamakura et al., 1980; Cutkosky, 1989; Feix et al., 2015) have been widely employed in robotics.

Building upon prior work in neuroscience (Kantak et al., 2017) and robotics (see Section 2.1), we propose a bimanual manipulation taxonomy specifically designed for robotic applications which was originally published in Krebs and Asfour (2022). Unlike taxonomies developed primarily for evaluating therapy progress in rehabilitation, this taxonomy focuses on representing bimanual robotic manipulation tasks. It facilitates the formation of task models for bimanual manipulation and explicitly encodes the spatial and temporal constraints arising between the hands.

## 3.1. Design Principles

In the following, we elaborate on the critical aspects of bimanual manipulation that must be considered when developing a novel taxonomy for bimanual

manipulation tasks in robotics. These key aspects include: (i) coordination between both hands, (ii) physical interaction between both hands and (iii) the role of each hand.

**Coordination** This criterion addresses whether there is any form of spatial or temporal coordination between the hands as defined by task-specific constraints. Uncoordinated movements are, in essence, simultaneous unimanual actions. An action is considered uncoordinated if the same outcome can be achieved by performing the action consecutively in any order with a single arm. In such cases, both arms operate without spatial or temporal coordination and do not pursue directly connected goals. For instance, one hand may hold a coffee cup while the other hand takes notes. Here, each hand adheres to its task-specific constraints, but spatial coordination is limited to avoiding collisions, and there is no temporal coupling.

In humans, uncoordinated actions performed by individual hands cannot be excessively complex. In the example mentioned, one hand can effectively write because holding the cup is a highly automated action requiring minimal cognitive resources. Practically, distinguishing between coordinated and uncoordinated movements can be challenging, as the relationship often emerges over time rather than at a single point. For example, if one hand closes a chest lid while the other hand holds an object, the connection may be unclear. However, if it is observed that one hand initially opened the chest to allow the other hand to retrieve something, then the actions are likely coordinated. Another challenging category involves actions that include support poses. In Borràs and Asfour (2015), the authors present an approach for distinguishing between support poses and manipulation by analyzing transitions within a whole-body support pose taxonomy. For instance, when an individual leans on a table with one hand to grasp a distant object with the other, there is an interdependence despite the seemingly different activities of the hands. Without the support, the person would be unable to reach the distant object. In contrast, there are situations within the context of support poses where both hands operate independently. For example, one hand may hold a cup of coffee while the other hand grips a handrail while climbing a staircase. It is evident that some cases are difficult to recognize from pure observation, even for humans, and are only apparent if the action is repeated in a different manner.

**Interaction** In bimanual manipulation, physical interaction between hands may or may not occur; regradless of high-level coordination imposed by spa-

tial or temporal constraints. This interaction entails the transmission of forces between hands, directly or via common objects. Examples include holding a large object with both hands or using one hand to hold an object while the other applies force to it using a tool. Understanding these interactions is critical for bimanual dexterous manipulation tasks, where interactions with objects and potentially between hands are pivotal for successful task completion.

**Hand Roles** In bimanual manipulation tasks, the hands often assume distinct roles, as discussed in (Guiard, 1987), where symmetry refers not to the movement, but to the abstract roles of the hands. Symmetric manipulation tasks entail both hands assuming identical roles, exemplified by jointly holding and transporting a large box, forming a closed kinematic chain. Conversely, asymmetric movements entail the hands undertaking different roles, such as one hand stabilizing an object while the other performs an action on it, as in stirring a cup while holding it. In humans, hand roles in asymmetric tasks are closely tied to hand dominance, with the non-dominant hand typically serving as the stabilizing hand, providing a reference frame for the dominant hand. Additionally, temporal-spatial differences exist, with the dominant hand usually exhibiting a higher frequency of movement. However, these roles are not rigidly assigned to the right or left hand and may be altered to optimize task completion. For instance, finer manipulation tasks, like closing a bottle lid, may be performed with the left hand if the right hand already holds the bottle from drinking.

## 3.2. Taxonomy for Bimanual Manipulation

Building on the discussion of key aspects in bimanual manipulation in Section 3.1, we propose a taxonomy specifically tailored to bimanual manipulation tasks from a robotics perspective, as illustrated in Figure 3.1. The objective is to create a taxonomy that facilitates the learning of task models for bimanual manipulation from human observations, enabling their implementation on bimanual robotic systems, such as humanoid robots.

At the top level, we distinguish between *coordinated* and *uncoordinated* bimanual actions based on the presence of spatial and/or temporal constraints needed for task execution. Uncoordinated actions resemble simultaneously executed unimanual actions e. g., holding a cup of coffee in one hand while writing with

the other, This encompasses instances where one hand serves no explicit task, so-called *unimanual* actions. These actions do not require complex coordination between the hands on motor level, as each hand operates independently. In contrast, coordinated actions involve varying degrees of interdependence between hands. We define *loose* coupling as coordination characterized by constraints on hand actions that primarily involve common via points (spatial) or synchronization points (temporal), without a persistent trajectory dependency. The concept of *coordinated*, *loosely coupled* actions can be exemplified in the context of self-handovers, where one hand passes an object to the other. Before physical interaction between the hands occurs, spatial and temporal constraints must be satisfied to ensure a smooth transition into the handover process. This requires the hands to be appropriately positioned and synchronized but without a strong reliance on the specific motion trajectory.

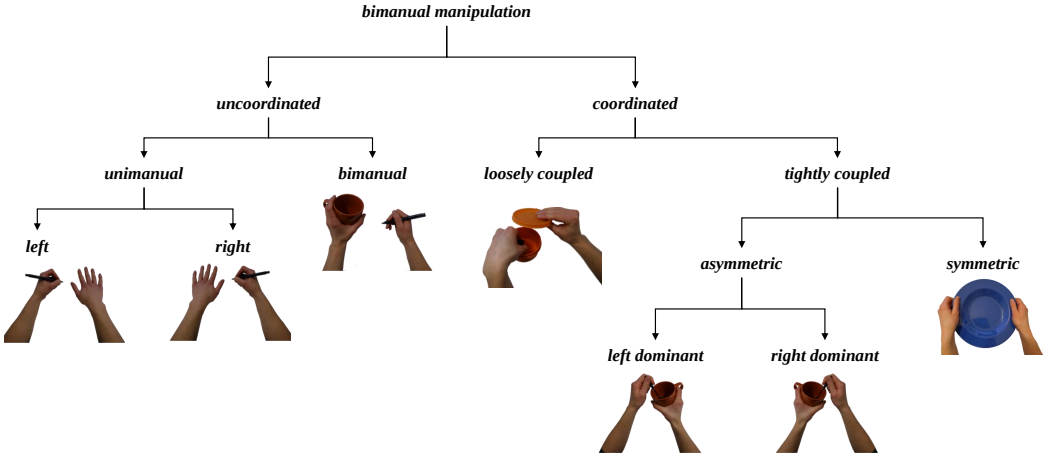



Figure 3.1.: Bimanual Manipulation Taxonomy. Tasks are classified based on the aspects *coordination*, *interaction*, *hand role* and *symmetry*.
Source: Krebs and Asfour (2022) © 2022 IEEE

Moreover, we introduce the category of *tightly coupled* coordinated actions. These are characterized not only by spatial and temporal constraints but also by force-based constraints resulting from contact-rich interactions between the hands and trajectory-level dependency. Within this category, we refer to the work of Guiard (1987) regarding hand roles in *asymmetrical* activities, where the non-dominant hand provides a reference frame for the dominant hand, guiding its motion and influencing trajectory formulation. Determining how the roles of the hands are assigned cannot be solved only based on the handedness of the demonstrator but can differ across tasks. These relations between the hands in such cases form a functional coupling where their roles and motions are highly interdependent. Activities where both hands assume identical roles

are denoted as *symmetric*. Although primarily addressing hand roles rather than motion trajectories, this often coincides with symmetrical motion patterns, especially when both hands manipulate the same object. Here, the dependency between hands is notably stronger due to a fixed transformation between the hands.

Although it holds significant importance in neuroscience (Kantak et al., 2017), we argue that strict geometric motion symmetry is less critical in the context of robotics. For loosely coupled actions, motion symmetry assumes a functional role only in exceptional cases e. g., when conducting an orchestra. While not imperative for bimanual reaching, it may hold relevance in gestures like conducting an orchestra. Consequently, in our taxonomy, we exclusively address symmetry within tightly coupled actions, emphasizing the symmetry of hand role distribution rather than strict geometric symmetry.

In the remainder of this thesis the abbreviation introduced in Table 3.1 will be used for the different categories of the taxonomy.

Table 3.1.: Abbreviations for the categories of the Bimanual Manipulation Taxonomy.

| Bimanual Category | Abbreviation |
|---|---|
| No action | *no_action* |
| Unimanual left | *uni_l* |
| Unimanual right | *uni_r* |
| Uncoordinated bimanual | *uncoord_bi* |
| Loosely coupled | *loosely* |
| Tightly coupled asymmetrical left dominant | *asym_l* |
| Tightly coupled asymmetrical right dominant | *asym_r* |
| Tightly coupled symmetrical | *sym* |

## 3.3. Formalization of Constraints Imposed by the Taxonomy

To provide a structured approach to using the taxonomy in bimanual robot manipulation tasks, a formalization of the constraints imposed by the taxonomy is needed. This will allow the implementation of appropriate coordination strategies. In the following we present an explicit representation of temporal and spatial constraints for each category of the taxonomy, with a particular emphasis on their integration into robot controllers for bimanual task execution. These

constraints are not merely descriptive but functionally essential as they must be maintained even when motion execution is adapted or subjected to external disturbances. This is especially critical for compliant robots, such as humanoid robots designed for human collaboration, where controllers must ensure both functional and safe task execution, even in the presence of external perturbations. The formalization of these spatial and temporal constraints within the taxonomy has been published in Krebs and Asfour (2024).

### 3.3.1. Spatial Constraints

Our objective is to develop a methodology for the formalization of bimanual constraints that aligns with established techniques documented in the literature, as elaborated in Section 2.3.1. We achieve this by formalizing the spatial constraints applied to each hand through the adoption of specific states:

- *Unspecified:* The hand is not actively engaged in a task of interest. Besides avoiding collisions and adhering to soft criteria, such as enhancing human-like appearance, the hand's pose is considered irrelevant.
- *Global:* The hand's target is defined independently of the other hand, either within a fixed world frame or relative to the robot's root frame.
- *Relative:* The hand's target is defined relative to the pose of the other hand.

This aligns with the global, local, and relative constraints as defined by (Stavridis et al., 2021). In their framework, global constraints refer to a world frame, while local constraints pertain to the robot's root frame. In our approach, we focus on the constraints between the hands, which leads us to integrate both frames and include object-centric task descriptions within the *global* constraints.

Table 3.2 delineates the primary constraints pertinent to each category. The primary task space goals are denoted by an $\times$, while secondary task space goals are indicated with an $(\times)$. For asymmetric categories, this configuration essentially adopts a leader-follower paradigm, where the non-dominant hand functions as the leader and the dominant hand as the follower. For *uni_r*/*uni_l*, the movement of the right/left hand is governed by *global* constraints, whereas the other hand remains *unspecified*. The three states of spatial constraints - *unspecified*, *global*, and *relative* - can be applied either to the start and end points of a segment or to the entire trajectory.

Table 3.2.: Spatial constraints for different categories.
Source: Krebs and Asfour (2024) © 2024 IEEE

| Bimanual Category | Right Hand | | Left Hand | |
|---|---|---|---|---|
| | Global | Relative | Global | Relative |
| *uncoord_bi* | × | | × | |
| *uni_l* | | | × | |
| *uni_r* | × | | | |
| *asym_l* | × | | | × |
| *asym_r* | | × | × | |
| *sym* | (×) | × | (×) | × |

We formalize the spatial specification as a directed graph, where nodes represent spatial states and edges denote the trajectories connecting them. The types of edges ($E$) and nodes ($V$) are defined in Figure 3.2.

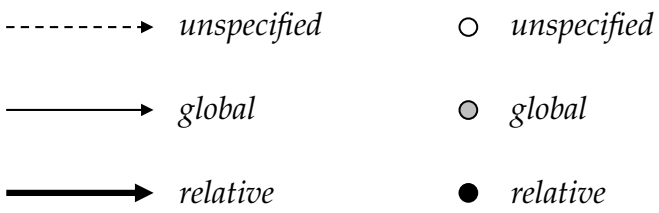



Figure 3.2.: Edges $E$ (left) and nodes $V$ (right) of the spatial graph.
Source: Krebs and Asfour (2024) © 2024 IEEE

We consider all possible tuples $(e, v)$ of trajectories $e \in E$ and goal states $v \in V$ as shown in Figure 3.3. The permutations highlighted in grey are logically impossible due to inherent contradictions: if the entire trajectory is defined globally, the goal state cannot be unspecified; similarly, if the trajectory is defined relatively, the goal state must also be defined relatively. The feasible permutations are grouped based on the goal pose, labeled as A–D. The combination of a relative edge and node (group D) is treated separately, as this scenario is particularly pertinent for actions involving physical interaction between the hands (*tightly coupled* categories). Different combinations of these groups for the left and right hand correspond to various categories within the Bimanual Manipulation Taxonomy (see Table 3.3).

If a task is described as a sequence of categories, the spatial constraints for the entire sequence can be represented using the elements shown in Figure 3.3. The differentiation between *unspecified* and *global* trajectories within the blue (B) and orange group (C) requires further investigation based on the precise motion performed by the respective hand.

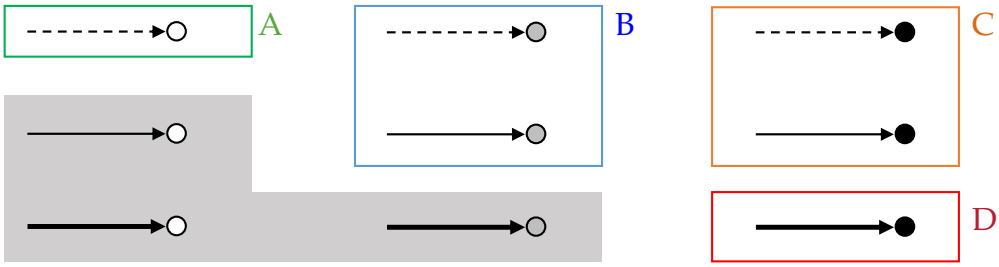



Figure 3.3.: Spatial permutations considering elements defined in Figure 3.2.
Source: Krebs and Asfour (2024) © 2024 IEEE

In the presented application in Chapter 6, we employ the more restrictive formulation with the *global* trajectory definition. The combinations at the bottom right of the Table 3.3, indicated with $\times$, do not exist as they would be underspecified with both goal states being relative to the other. Combinations labeled *asym_r* or *asym_l* could also fall into the *loosely* category if there is no physical interaction (no force constraints) between the hands. The category *sym* is only denoted in brackets in Table 3.3. As shown in Table 3.2 for *sym*, relative constraints are most important for both hands. However, since this leaves the system underspecified, some global grounding needs to be introduced. Conceptually, this can be achieved by defining both hands relative to the same global trajectory. Another option to maintain the relative pose in case of perturbations is to use the asymmetric formulation, ensuring that the perturbed arm is considered the leader (non-dominant hand).

Table 3.3.: Assignment of different spatial combinations to bimanual categories, x indicates that this combination does not exist.
Source: Krebs and Asfour (2024) © 2024 IEEE

| | | Left Hand | | | |
|---|---|---|---|---|---|
| | | A | B | C | D |
| Right Hand | A | *no_action* | *uni_l* | *loosely* | *asym_l* / *loosely* |
| | B | *uni_r* | *uncoord_bi* | *loosely* | *asym_l* / *loosely* |
| | C | *loosely* | *loosely* | $\times$ | $\times$ |
| | D | *asym_r* / *loosely* | *asym_r* / *loosely* | $\times$ | $\times$ / (*sym*) |

### 3.3.2. Temporal Constraints

For the coordinated execution of dual-arm tasks, Zöllner et al. (2004) use Petri nets to describe how task executions of the right/left arm or both arms can be triggered, facilitating the transition between an *active* and a *ready* state for both hands. In this work, we extend the concept described in Zöllner et al. (2004) to define the category-specific constraints and behaviors in response to perturbations. This leads to different Petri net templates shown in Figures 3.4-3.7. Specifically, we add the concept of phase stopping and goal synchronization and formulate those dependent on the bimanual categories. Phase stopping refers to pausing the variable governing the temporal evolution of motion when deviations from the desired trajectory become too large, similar to the concept of phase stopping for Movement Primitives (MPs) presented in Ijspeert et al. (2002b).

Petri nets are described by the 4-tuple $N = (P, T, A, m_0)$, with $P$ being the set of places, $T$ the set of transitions, $A$ the set of arcs and $m_0$ the initial marking of the net. In the following Petri net representations, transitions are represented within rectangles, and places are indicated next to circles.

Petri net templates are defined to depict individual bimanual categories and can be sequenced to outline the entire task. Initial markings are positioned on the left, indicating their initiation from a preceding category, and their transition to subsequent categories is shown on the right.

The category-specific template Petri nets delineate, firstly, the behavior specific to each category in response to perturbations (such as phase stopping), and secondly, the behavior during transitions (e. g., transitioning independently or only when both hands have completed their respective tasks).

The terms *ErrL* and *ErrR* within the transitions $T$ indicate the error of the left and right hand, respectively, computed based on the current pose compared to the pose computed in the previous timestep. The transitions are triggered when the error exceeds a specified threshold.

Figure 3.4 illustrates the Petri net representation for *uni_r*. The temporal evolution of the hands exhibits independence, as reflected in the disconnected paths within the Petri net structure. The *inactive* hand, in this case, the left hand, can directly transition to the next category upon entering the current one. The right hand follows a trajectory with active *phase stopping*, meaning that the progression of the MP is halted if an error exceeding a predefined

threshold is detected. The transition to the completed state occurs once the entire motion primitive has been executed and the target pose has been reached. For *uni_l*, the Petri net structure remains identical, with the roles of the hands reversed.

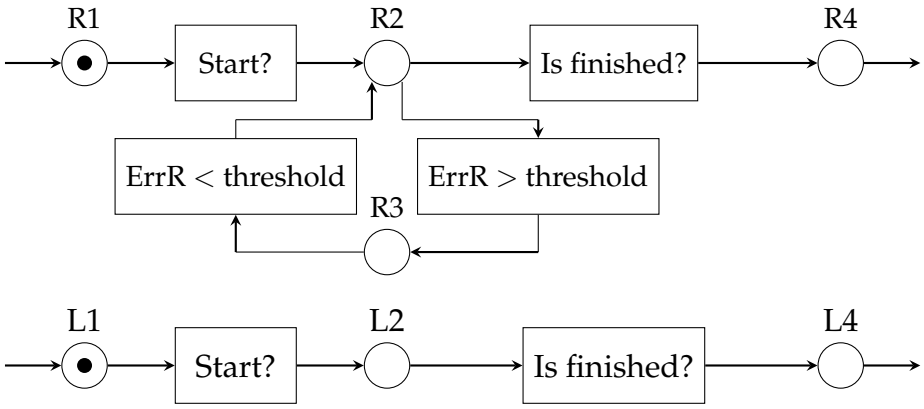



Figure 3.4.: Petri net template for *uni_r* with places $P$ for right/left hand respectively: R1/L1: ready, R2/L2: active, R3/L3: paused, R4/L4: completed. Based on Krebs and Asfour (2024).

In the case of the *uncoord_bi* category (see fig. 3.5), the hands also evolve independently. However, unlike in the *uni_r* case, where only the right hand follows a MP with *phase stopping*, here both hands exhibit this behavior.

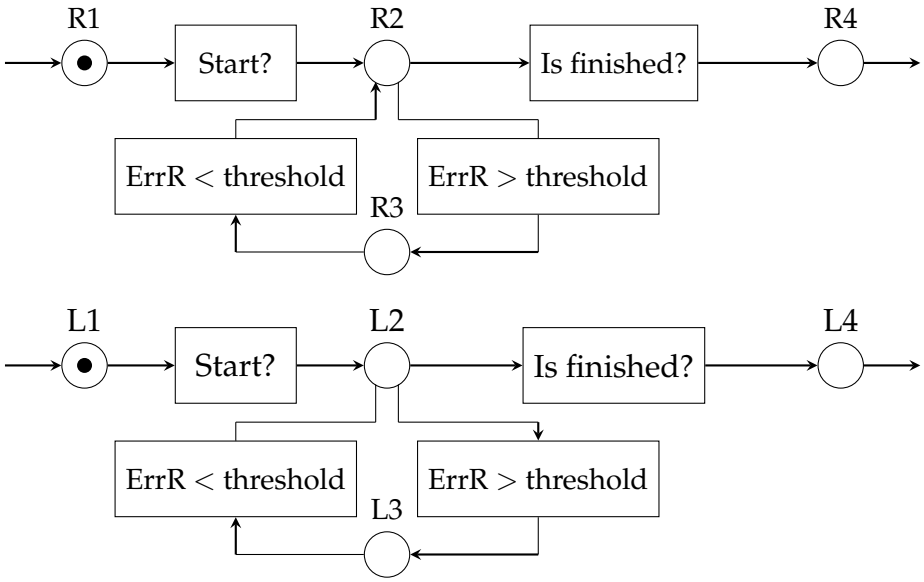



Figure 3.5.: Petri net template for *uncoord_bi* with places $P$ for right/left hand respectively: R1/L1: ready, R2/L2: active, R3/L3: paused, R4/L4: completed. Based on Krebs and Asfour (2024).

The category *asym_r* represents the most complex case (see Figure 3.6). Here, the right hand assumes the role of the *follower* and behaves similarly to the individual hands in the *uncoord_bi* category. Specifically, it follows a predefined MP and is subject to *phase stopping* when its tracking error exceeds a predefined threshold. The temporal execution of the right hand remains independent of the left hand. Conversely, the left hand (or more generally, the *leader* hand) also transitions into *phase stopping* if the right hand's tracking error surpasses the threshold. This constraint ensures that spatial dependencies, such as the relative pose between the hands, are maintained. Initially, both hands are synchronized upon entering a category section; however, the right hand may complete its motion and transition to the next category independently. The category *asym_l* follows the same structure but with the roles of the hands reversed.

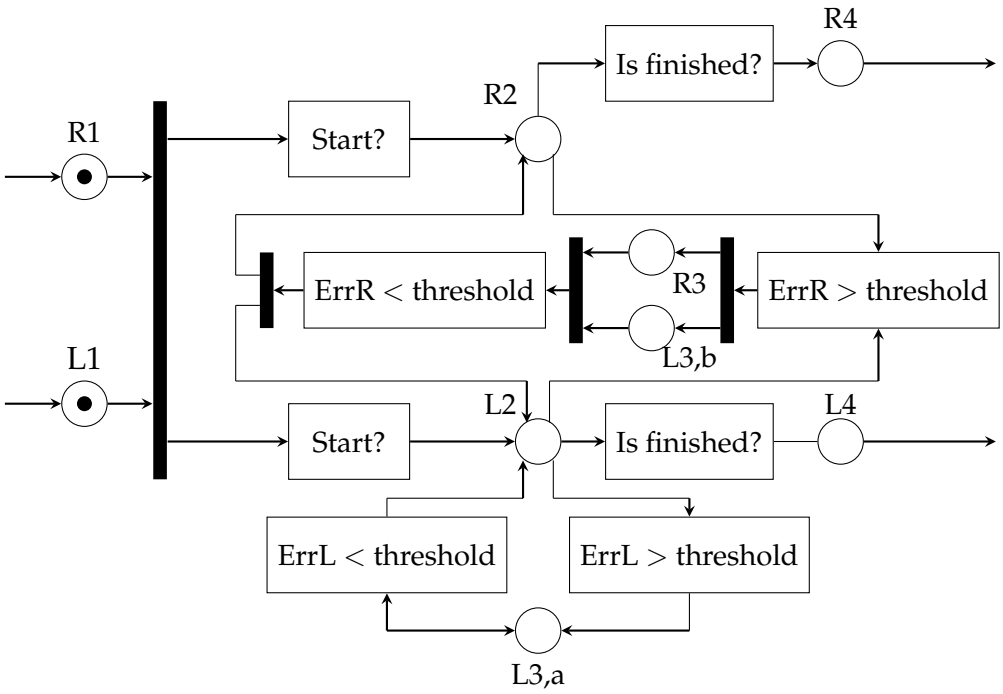



Figure 3.6.: Petri net template for *asym_r* with places $P$ for right/left hand respectively: R1/L1: ready, R2/L2: active, R3/L3: paused, R4/L4: completed. Based on Krebs and Asfour (2024).

Finally, Figure 3.7 depicts the *sym* category. In this case, both hands are fully temporally synchronized, meaning they enter and exit each segment simultaneously. If the tracking error of either hand exceeds the predefined threshold, both hands halt execution. Movement resumes only when the errors of both hands fall within the acceptable range.

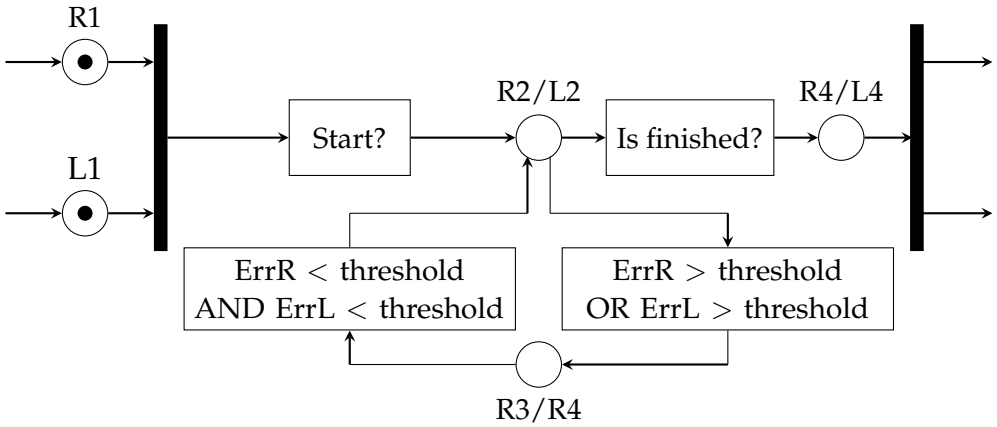



Figure 3.7.: Petri net template for *sym* with places $P$ for right/left hand respectively: R1/L1: ready, R2/L2: active, R3/L3: paused, R4/L4: completed. Based on Krebs and Asfour (2024).

Regarding the category *loosely*, which encompasses all categories with spatial and temporal constraints but without physical interaction, we do not present a specific temporal Petri net template because it encompasses various coordination patterns. Further investigation may be needed to define subcategories within *loosely*.

Spatial and temporal constraints are largely modeled concurrently in this framework, but there are interdependencies. Each segment (category) is formalized with spatial and temporal constraints. Spatial constraints determine whether a *relative* or *global* trajectory is followed within the Petri nets. Temporal constraints can halt the progression of these reference trajectories, thereby influencing spatial outcomes.

### 3.3.3. Transitions Between Categories

In the previous Subsection, we described how spatial and temporal constraints can be formalized within a bimanual category. However, a complete task consists of a sequence of these categories and the associated trajectories.

The temporal conditions for transitions between categories are implicitly derived from their representation as Petri nets. This includes synchronization at the start of a category, such as for *asym_r*/*asym_l* and *sym*. In the case of uncoordinated categories, the hands evolve independently.

In order to model the spatial constraints over time, the representations for different categories can be concatenated to form a common graph. An example

for the case of rolling dough is shown in Figure 3.8. This task consists of the following categories: both hands start from a global resting position, the left hand moves an object out of the way (*uni_l*), both hands move to the rolling pin (*uncoord_bi*), both hands hold the rolling pin and roll out the dough (*sym*), and finally, the hands release the rolling pin and move back to their resting position (*uncoord_bi*). The symmetric segment is represented as a combination of *relative* and *global* constraints since both are needed for a complete description.

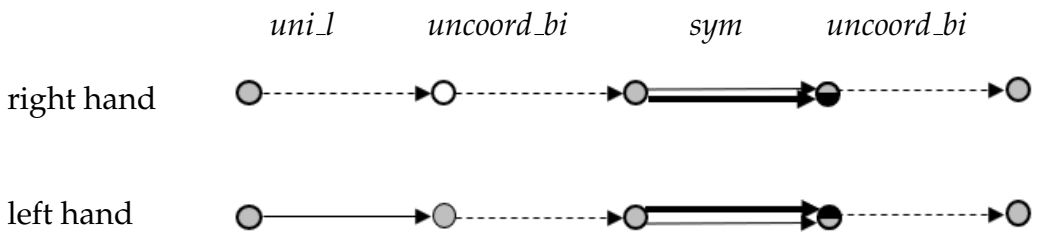



Figure 3.8.: Sequencing of spatial constraints for rolling dough.
Source: Krebs and Asfour (2024) © 2024 IEEE

If no specific trajectory is defined, traditional planners can be employed. Given trajectories can be easily adapted to new start or goal poses by using MPs such as Dynamic Movement Primitives (DMPs) (Ijspeert et al., 2002a), Via-Point Movement Primitives (VMPs) (Zhou et al., 2019), or Probabilistic Movement Primitives (ProMPs) (Paraschos et al., 2013).

## 3.4. Summary

This thesis introduces a taxonomy for bimanual manipulation, drawing inspiration from previous research in robotics, neuroscience, and rehabilitation science. The taxonomy differentiates various coordination patterns in bimanual manipulation tasks by considering key aspects such as coordination and interaction between the hands, the roles of the hands in the task, and the symmetry in task execution. Additionally, we define and formalize the temporal and spatial constraints applicable to actions within the different categories outlined in the Bimanual Manipulation Taxonomy.

This taxonomy contributes to a deeper understanding of bimanual coordination in humans and facilitates the transfer of human strategies for executing bimanual tasks to robotic systems. Unlike existing taxonomies, it is specifically structured

for robotic applications and provides precise definitions for individual categories. Using the taxonomy, bimanual manipulation tasks can be described as a sequence of bimanual categories, encoding the relevant spatial and temporal constraints necessary for execution. This representation enables the selection of appropriate controllers for different phases of a bimanual task and the switching between different control strategies, such as transitioning from a leader-follower control scheme for tightly coupled actions to independent controllers for uncoordinated actions. This approach will be discussed in Chapter 6. Furthermore, the proposed taxonomy can enhance action and intention recognition in human-robot interaction tasks and improve bimanual interactions with prostheses, as demonstrated in (Volkmar et al., 2019).

# CHAPTER 4

# KIT Bimanual Datasets

As highlighted in the state-of-the-art discussion (Section 2.4), there remains a deficiency in human motion datasets that encompass a diverse range of bimanual manipulation actions and the corresponding objects involved. To address this gap, this thesis contributes a novel multi-modal dataset, namely the *KIT Bimanual Manipulation Dataset* (*KIT BMD*), and extends an existing RGB-D dataset, namely the *KIT Bimanual Actions Dataset* (*Bimacs*) (Dreher et al., 2020). Datasets with such a variety of bimanual manipulation actions are essential to validate the proposed taxonomy presented in the previous chapter (see chapter 3). These datasets are not only utilized in this work but are also made publicly available to the broader research community to foster and advance research in bimanual manipulation.

- **KIT Bimanual Manipulation Dataset:** multi-modal recordings using five different sensor modalities
  - **Actions:** 588 recordings, 73 minutes, 2 subjects
  - **Sequences:** 90 recordings, 104 minutes, 6 subjects
- **Extension KIT Bimanual Actions Dataset:** single-view RGB-D data, 120 recordings, 60 minutes, 6 subjects

## 4.1. KIT Bimanual Manipulation Dataset

Our objective is to develop a novel multi-modal dataset of whole-body motions to facilitate the learning of task models for bimanual manipulations. The design of this dataset is driven by the need to provide comprehensive information required for learning such task models from human demonstrations, with a particular emphasis on capturing variations in object types and object relationships within bimanual tasks.

The original *KIT BMD* as presented in Krebs and Meixner et al. (2021)[1] comprises of multi-modal recordings of rather isolated household actions. This dataset was extended in Meixner et al. (2023) with more complex, longer sequences consisting of the actions recorded in the original dataset. The complete dataset is publicly available as part of the KIT Whole-Body Human Motion Database[2].

In the following, we refer to the subparts of this dataset as *KIT BMD Actions* and *KIT BMD Sequences*. This chapter discusses the multi-modal sensor setup used for the recordings and the objects and actions covered. Further, the recording procedure as well as the post-processing of the data is detailed. The provided information holds for both subsets of the dataset if not indicated otherwise.


Figure 4.1.: Subject and multi-modal sensor setup during a recording of *Cut*.

[1]The KIT BImanal Manipulation Dataset resulted from joint work with André Meixner, who contributed equally to the creation of the dataset and publication.
[2]https://motion-database.humanoids.kit.edu/list/datasets/

### 4.1.1. Sensor Setup

The sensor setup comprises five distinct sensor modalities (see Figure 4.1):

- A **marker-based VICON motion capture system** is employed to precisely capture the trajectories of body segments and objects at a frequency of 100 Hz. Figure 4.3 provides an overview of the camera setup in our motion lab, which features nine static motion capture cameras (MX T10) mounted on the walls around the capture area at a height of approximately three meters. Additionally, mobile cameras (Vero[3]) are positioned on tripods around the subject. As depicted in Figure 4.1, the subject wears a full-body suit fitted with optical markers of 14 mm diameter, which are tracked by the infrared cameras. Simultaneously, the experiments are recorded with a synchronized **digital video camera** for documentation purposes.
- For capturing hand grasping movements in bimanual tasks, participants wear **Cyber Glove III data gloves**[4] on each hand. These gloves measure finger joint angles, palm curvature, and wrist angles. In our laboratory setup, we used a right-hand data glove with 22 degrees of freedom (DoF) and a left-hand data glove with 18 DoF, which excludes the distal finger joints. The gloves are calibrated following the procedure detailed in a previous study (Starke et al., 2018), and they record finger joint positions at 90 Hz. Data gloves are the preferred method for tracking finger movements in bimanual tasks, as using the marker-based motion capture system alone would require numerous additional markers on each hand, leading to occlusions and incorrect marker associations during close hand-object interactions.
- Additionally, three **9-DoF inertial measurement units (IMUs)** (Blue Trident by Vicon Motion Systems[5]) are affixed to the body (see Figure 4.2) to measure linear accelerations and angular velocities at a frequency of 225 Hz. The data is up-sampled to 300 Hz to achieve an integer number of sub-samples per frame of the VICON system. The sensors are positioned on anatomical landmarks: one on each forearm near the wrist (dorsal side of the antebrachium above carpals) and one on the back between the shoulder blades (approximately at thoracic vertebra T4).

[3] https://www.vicon.com/hardware/cameras/vero/
[4] https://www.cyberglovesystems.com/cyberglove-iii
[5] https://www.vicon.com/hardware/blue-trident/


Figure 4.2.: The IMU sensors are positioned on both wrists and on the upper back of the subject.

- Furthermore, **three RGB-D cameras** (Azure Kinect[6]) are placed on fixed tripods (see Figure 4.1). These cameras are strategically positioned to provide various viewpoints of the scene and to simulate potential positions of a robot learning from human demonstrations. The video recordings are captured at 30 FPS with 1080p RGB resolution and 640×576 pixel depth resolution.
- To capture **egocentric RGB images** from the subject's perspective, Full HD video recordings are made using a head-mounted action camera (GoPro Hero 8[7]) at 60 FPS in *SuperView* mode with *HyperSmooth* enabled for automatic image stabilization. The RGB images are cropped to focus on key areas of interest.

To accurately determine and monitor the position of the RGB-D and action cameras within the scene, optical markers are affixed to these cameras. These markers enable the motion capture system to record their precise poses. For privacy protection, subjects are anonymized in all video recordings.

For KIT BMD *Sequences* an additional table has been incorporated into the scene (see Figure 4.3 right). This modification not only offers extra space for manipulation but also creates a kitchen nook. The placements of the digital video, RGB-D, and motion capture cameras were subtly altered to adapt to this new setup. Each recording starts with the subject positioned in a T-pose, approximately one to two meters away from the tables.

[6] https://learn.microsoft.com/de-de/azure/kinect-dk/hardware-specification
[7] https://gopro.com/content/dam/help/hero8-black/manuals/HERO8Black_UM_ENG_REVB.pdf

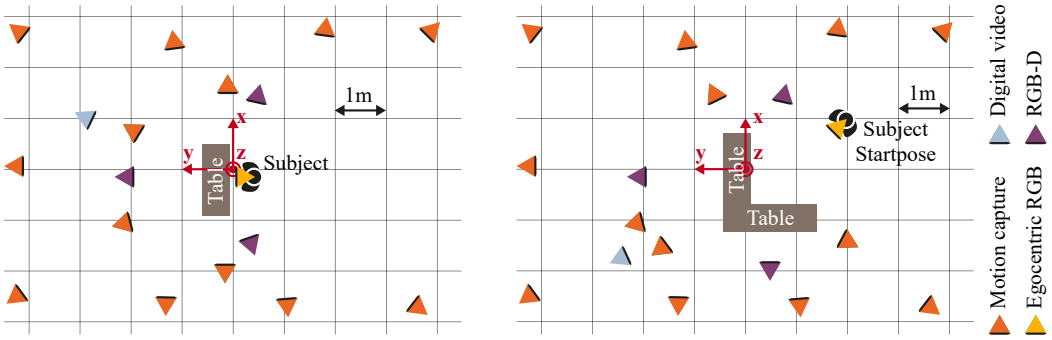



Figure 4.3.: The arrangement of cameras utilized in the KIT BMD *Actions* (left) and the KIT BMD *Sequences* (right) is illustrated. The bold line indicates the orientation of each camera. The red coordinate system designates the origin of the motion capture system.
Sources: Krebs and Meixner et al. (2021) © 2021 (left), Krebs and Leven et al. (2023) © 2023 IEEE

### 4.1.2. Actions and Objects

The dataset was recorded using a total of 21 real household objects and food items (e. g., cucumber), as illustrated in Figure 4.4. All actions took place on or behind a table with a height of 88 cm, matching the height of standard kitchen counters. Each object had at least four markers attached to it to track its pose using the motion capture system. The markers used varied in size, with options of 6 mm, 9.5 mm, or 14 mm depending on the object. The dataset includes 3D models of all objects, created either with a 3D scanner or CAD software for objects with simple geometry.

#### KIT BMD Actions

Twelve manipulation actions (see Table 4.1) commonly used in household activities were selected for the dataset, emphasizing the variety of objects used and the bimanual nature of the actions.

Although some actions, such as pouring, can be executed with a single hand, this dataset specifically focuses on performing these tasks with both hands. Many of these actions involve asymmetric movements (Guiard, 1987), where one hand stabilizes an object while the other manipulates it. For instance, the left hand might hold a cup steady on a table or in midair while the right hand pours water from a bottle. Additionally, some tasks require both hands to hold the same object, such as sweeping or rolling out dough, or involve self-handovers. Special cases include walking while manipulating an object or holding an object


Figure 4.4.: Objects used in the KIT Bimanual Manipulation Datasest.
Source: Krebs and Meixner et al. (2021) © 2021 IEEE

by enclosing it with the entire arm. All actions were recorded as they would be naturally performed by right-handed individuals. To capture variations in natural human action execution, each variation was repeated three times by each subject. Additionally, semantic variations within each action type were included, such as different object locations. For example, a bottle might be positioned on the left or right side of a cup, and placed either close to or farther from the subject. These variations are crucial for studying how motion trajectories adapt to new scenes and situations. The dataset also incorporates different objects, focusing on changes in specific task parameters such as object height or diameter (e. g., small vs. large cup or bowl). Various tool-based actions are included as well, such as scooping with a spoon or ladle, noting that the handling of tools differs even when used for the same purpose. Different methods of executing bimanual manipulation actions are considered, for example, the left hand might hold, tilt, or lift a bowl while the right hand stirs. In some cases, the object held during manipulation can vary. There is an imbalance in the number of recorded variations per action type. This discrepancy arises because certain actions offer more opportunities for interesting parameter variations (e. g., position, direction, height), resulting in more diverse executions.

### KIT BMD Sequences

The *KIT BMD Sequences* consists of three different scenarios, each consisting of multiple of the individual actions recorded previously. The scenarios are

*preparing a meal*, *preparing a pie*, and *cleaning up*. Table 4.1 shows which actions are part of which scenario. Additional details about the scenarios are available in the recording descriptions in our motion database[8].

Table 4.1.: Assignment of actions to different scenarios.

| | | **Scenarios** | | |
|---|---|---|---|---|
| | | *preparing a meal* | *preparing a cake* | *cleaning up* |
| **Actions** | Close | ✓[1] | ✓[1] | |
| | Cut | ✓ | | |
| | Mix | ✓ | | |
| | Open | ✓ | ✓ | |
| | Peel | ✓ | | |
| | Pour | ✓ | ✓ | |
| | Rollout | | ✓ | |
| | Scoop | ✓ | ✓ | |
| | Stir | | ✓ | |
| | Sweep | | | ✓ |
| | Transfer | ✓ | | |
| | Wipe | | | ✓ |

[1] optional

### 4.1.3. Recordings Procedure

This study received approval from the ethics committee of the Karlsruhe Institute of Technology, Karlsruhe, Germany. Participants provided written informed consent before the experiments, agreeing that the data could be publicly available for research purposes in the *KIT Whole-Body Human Motion Database*. For privacy reasons, the faces of subjects are blurred in all publicly available visual recordings.

#### KIT BMD Actions

Two healthy young adults (1 male, 1 female) participated in the experiments. Both individuals are right-handed, have normal vision, and do not have any upper limb orthopedic impairments. Anthropometric data including body height, hand segment lengths, and weight were recorded for each subject. It should be noted that only these two subjects are included in the dataset. This decision

[8] https://motion-database.humanoids.kit.edu/details/datasets/3521/?listpage=1

was made due to the extensive effort required to record each subject and the emphasis on capturing a high number of action variations and repetitions. The dataset focuses on maintaining comparability by not introducing variations between different subjects performing various actions or tasks. While the subjects were familiar with the tasks performed, they were instructed to execute the actions in the manner they would typically perform them in their own home environments.

During each recording session, the subjects begin and end standing behind a table with their hands placed flat on its surface. Specific configurations at the start and end of the scene, including hand and body posture, are standardized. Task instructions, such as "cut off three slices of the cucumber", are provided to the subjects. The exact execution details, such as the temporal synchronization of hand movements and types of grasps used, are left to the subject. Each action was recorded with three repetitions per subject, resulting in a total of 588 demonstrations. Across different variations of an action, the order of recordings was consistent within subjects, but the sequence of actions (e. g., scoop, pour) varied. The duration of each action recording ranged from 5 to 15 seconds.

#### KIT BMD Sequences

Six healthy, right-handed individuals (three male and three female) carried out sequences of manipulation tasks within three everyday household scenarios. The two subjects of the *KIT BMD Actions* also were part of the subjects recorded for longer sequences. Before each recording subjects were briefly informed about the overall objectives of the respective scenario. However, they could freely decide the sequence and manner of performing the individual actions. After each iteration, the instructor randomly rearranged the positions of the objects on and around the table.

### 4.1.4. Data Processing

#### Data Synchronization

To ensure temporal synchronization of data from various sensor modalities used in our dataset collection, we developed a software tool called *CaptureComponent*. This software synchronously triggers recordings across different sensors deployed on multiple remote host computers. The component communicates

common signals via UDP packets to remote software components associated with each sensor. It manages recording-specific details such as filename, date, and duration. Each sensor is equipped with a dedicated remote software component that encapsulates its functionalities, enabling remote access and distributed processing regardless of the sensor interface's programming language or the operating system. The *CaptureComponent* also facilitates monitoring of sensor status and provides access to each remote component for status checks. Further synchronization is achieved by collecting and aligning timestamps from each sensor recording, ensuring precise coordination of data across all sensor modalities.

#### Master Motor Map (MMM) Framework

To facilitate accessibility of the collected data across different sensors for the research community, we build upon and extend our earlier work with the Master Motor Map (MMM) framework[9] (Mandery et al., 2016). This open-source framework offers a unified representation of human motions as well as their perception, visualization, reproduction and recognition. Further, it provides standardized data structures tailored for organizing and storing data within extensive motion databases. To align captured human motions with a reference kinematics and dynamics model of the human body, subject-specific parameters are employed to normalize and present the motions in a standardized format. This process ensures consistency and comparability across different datasets. For mapping these captured human motions onto a target robot embodiment, we minimize the squared distances between real markers attached to the subject's body at predefined anatomical landmarks and virtual markers on the MMM reference model, which also correspond to these anatomical landmarks. This alignment is achieved using non-linear optimization techniques, ensuring accurate and precise mapping of human movements to the virtual model. This approach is crucial for applications where human motion data needs to be transferred to robotic systems or simulations effectively.

In this study, we extend the mapping process by incorporating the subject's hand size, scaling the hand model within the MMM independently of the subject's height. The hand size is determined by measuring the distance from the subject's wrist to the tip of their middle finger. To further refine the accuracy of hand pose mapping, each hand marker's squared error is adjusted using a specific weighting parameter. This approach ensures that hand movements

[9] https://h2t.iar.kit.edu/752.php

are accurately represented relative to the MMM reference model. While we provide one predefined mapping of hand motions in this dataset, this mapping can be customized or replaced to suit specific application requirements. Additionally, the XML-based MMM data format and framework have been adapted to accommodate and manage data from all additional sensor modalities used in this dataset, such as IMUs, RGB, RGB-D, and data gloves. This adaptation utilizes an extendable, plugin-based sensor structure, enabling efficient storage, handling, and visualization of multi-modal data. The videos are stored in a suitable video container format (e. g., mp4) and are referenced within the MMM data format, ensuring comprehensive integration of visual data alongside other sensor modalities.

### Segmentation and Labeling

To facilitate comprehensive processing of recorded human movements, the MMM data format has been expanded to support both manual and automatic hierarchical segmentation. This segmentation operates on symbolic and subsymbolic levels, following the hierarchical segmentation approach presented in (Wächter et al., 2015). Additionally, the data can be annotated to provide detailed contextual information. Similar to the approach described in (Dreher et al., 2020), the motion recordings in this study are manually segmented and annotated based on the actions performed by each hand. This segmentation and annotation process enhances the dataset's utility by categorizing movements into meaningful segments, aiding in analysis and interpretation for various research applications.

Manipulation tasks typically involve multiple sequential actions such as approaching an object, lifting it, performing a manipulation, placing the object, and retracting the hand. Annotations in this dataset encompass various types of actions, including manipulation actions (e. g., scoop, wipe, peel), supporting actions (e. g., hold, move), and phases describing different stages of grasping (e. g., approach, lift, place, retreat). A segmentation example for scooping is illustrated in Figure 4.5. In this scenario, the manipulation task for the right hand is segmented into detailed actions such as approaching the ladle with the right hand, lifting the ladle, moving the hand with the ladle into the bowl, performing the scoop action, moving the hand with the ladle to the cup, pouring from the ladle into the cup, and so forth.

Each manipulation action, such as scoop, is further hierarchically segmented into fine-grained segments, which can be valuable when analyzing isolated partial

actions. Annotations also include relevant objects involved in the task. There are distinctions between a main object (the grasped object), a target object (with which the main object interacts), and a source object (if interaction occurs at the beginning of the manipulation task). For example, in the manipulation action "scooping from a bowl into a cup using a ladle," the bowl is categorized as the source object, the cup as the target object, and the ladle as the main object. These annotations provide a detailed context about the objects and actions involved in each recorded manipulation task.

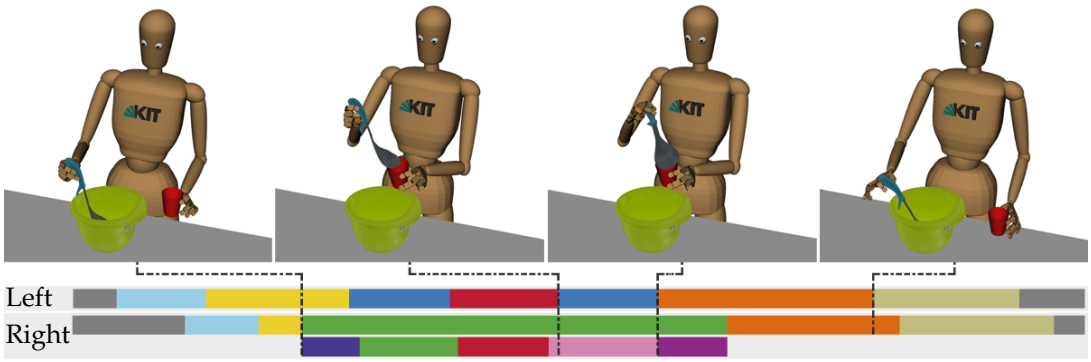



Figure 4.5.: Bimanual segmentation for scooping from a bowl to a cup. *Top:* Visualisation of the motion on the MMM reference model. *Bottom:* Segmentation tracks for both hands (Left, Right). Occuring actions are: *idle*, *approach*, *lift* , *hold* , *move*, *place*, *retreat*, *scoop* as well as additional actions *pre*, *pour*, *post* within the scooping task.
Source: Krebs and Meixner et al. (2021) © 2021 IEEE

## 4.2. Extension of the KIT Bimanual Actions Dataset

In the scope of this thesis, the previously existing *KIT Bimanual Actions Dataset* was extended. The *Bimacs* (Dreher et al., 2020) provides RGB-D data of 6 subjects performing 9 different actions. Each task was recorded 10 times using a PrimeSense Carmine 1.09 camera. The dataset additionally provides extracted 3D bounding boxes of objects and spatial relations of those. Since the original dataset barely includes any motion of the *symmetric* bimanual category, in the scope of this thesis it was extended with two further tasks as presented in Krebs and Leven et al. (2023). The extension consists of 120 recordings with a total duration of 1 hour and 2 minutes and is publicly available[10].

[10] https://bimanual-actions.humanoids.kit.edu/bimanual_categories

### 4.2.1. Sensor Setup

The data was recorded from a single-view using an RGB-D camera, positioned opposite the human to mimic a scenario where a robot observes a demonstration. The recordings were captured at 30 FPS with a resolution of 1920×1080 pixels for RGB and 640×576 pixels for depth data. The setup replicates the original sensor configuration described in (Dreher et al., 2020). However, instead of the PrimeSense used in the previous study, this research employs an Azure Kinect[11].

### 4.2.2. Actions and Objects

Two additional tasks in a household scenario were recorded: *set table* (see Figure 4.6) and *prepare dough* (see Figure 4.7). In the first task, subjects reach for various objects and place them on the opposite side of the table. Additionally, they must take a piece of bread from a bowl and place it on a plate. In the second task, subjects first empty a cup into a bowl and stir the contents with a whisk. Following this, they transfer the contents of the bowl onto the table and use a rolling pin to roll it out. In both tasks, the application of the *symmetric* bimanual category is promoted either by the presence of large objects, such as a bowl or plate or by the use of an object specifically designed for bimanual manipulation, such as a rolling bin. For those recordings six different objects are used: a mixing bowl, a cup, a plate, a rolling pin, a spoon and a whisk.

Figure 4.6.: Example frames of a *set table* recording. Based on Krebs and Leven et al. (2023).

[11] https://learn.microsoft.com/de-de/azure/kinect-dk/hardware-specification

Figure 4.7.: Example frames of a *prepare dough* recordings. Based on Krebs and Leven et al. (2023).

### 4.2.3. Recordings Procedure

As in the KIT Bimanual Actions Dataset, six subjects were recorded (3 male, 3 female; 5 right-handed, 1 left-handed) using an RGB-D camera. This study was approved by the ethics committee of the Karlsruhe Institute of Technology, Karlsruhe, Germany. The participants gave their written informed consent before the experiments that the data may be made publicly available for research purposes. In the collected video data faces are anonymized. The structure of the recordings was designed to match the data of the *Bimacs* dataset (Dreher et al., 2020). Two new household tasks were each recorded 10 times for each subject. Subjects were provided with a description of the overall goal to be achieved, but the precise execution was left to them. Initial object positions were varied within different recordings. The two new tasks are designed to include symmetric actions within a household context. Symmetric motions are expected for transferring big objects like a bowl or a plate and for using a rolling pin. In total, we collected 120 new recordings with a total duration of 60.2 minutes.

### 4.2.4. Data Processing

To detect objects within RGB video sequences, we use YOLOv7 (Wang et al., 2022), trained on a custom dataset specific to our objects of interest. The resulting 2D bounding boxes were utilized to extract relevant points from a corresponding point cloud generated from the depth images. These points were then filtered based on the color characteristics of the detected objects. Subsequently, a 3D bounding box, accurately reflecting the dimensions of the

detected object, was positioned at the centroid of the remaining filtered point cloud.

For hand tracking, we utilized the Azure Kinect Body Tracking SDK[12]. A 3D bounding box for each hand was constructed by determining the minimum and maximum coordinates of four key points: the wrist, the hand, the tip of the hand, and the thumb.

As a further step, we employed the methods outlined in (Kartmann et al., 2018) to extract 15 static and dynamic spatial relations between the identified 3D axis-aligned bounding boxes, as defined in (Ziaeetabar et al., 2018). These spatial relations include: *contact*, *above*, *below*, *left of*, *right of*, *in front of*, *behind*, *inside*, *surround*, *moving together*, *halting together*, *fixed moving together*, *getting close*, *moving apart*, and *stable*.

To establish a global coordinate system, ArUco markers (Garrido-Jurado et al., 2014) are used in conjunction with OpenCV (Bradski, 2000). The choice of ArUco markers is motivated by their high robustness to noise (Garrido-Jurado et al., 2014) and the comprehensive processing methods available in the OpenCV library. Furthermore, ArUco markers are cost-effective due to their ease of generation and printing. Two distinct markers are affixed to the front corners of the table: the left marker defines the origin of the global coordinate system, while the right marker serves as a backup in case the left marker is not detected. The configuration of the recording setup and the markers' coordinate system are illustrated in Figure 4.8.

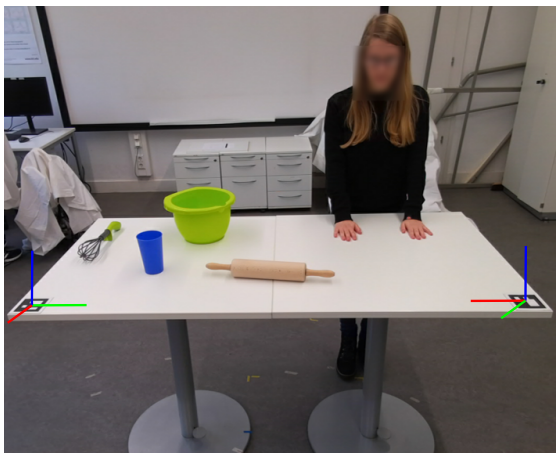


Figure 4.8.: Exemplary recording setup. The coordinate system defined by the ArUco markers are depicted.

[12] https://microsoft.github.io/Azure-Kinect-Body-Tracking

The labeling tool[13] used in (Dreher et al., 2020) was adapted to label the data with bimanual categories as defined in the taxonomy. The data from the two new tasks as well as the five kitchen tasks of the original KIT Bimanual Actions Dataset are labeled with bimanual categories.

## 4.3. Summary

In this chapter, we introduced the datasets collected for this thesis, which are intended to support both the research questions in this thesis and the broader scientific community. By making these datasets available, we aim to contribute to the advancement of research in bimanual manipulation. Therefore the KIT Bimanual Manipulation Dataset[14] as well as the Extension to the KIT Bimanual Manipulation Dataset[15] are publicly available. In the scope of this thesis, both datasets are used to develop methods for the recognition of the bimanual categories defined by the taxonomy which will be described in detail in Chapter 5.

**KIT Bimanual Manipulation Dataset** A new multi-modal dataset of bimanual manipulation actions has been created, featuring precise human whole-body motion data, comprehensive hand configurations, and the 6D pose and trajectories of all objects involved in the tasks. Data collection utilized five different sensor systems: a motion capture system, two data gloves, three RGB-D cameras, a head-mounted egocentric camera, and three inertial measurement units (IMUs). The dataset comprises 12 bimanual daily household activities performed by two healthy subjects, including a large number of intra-action variations and three repetitions of each action variation, totaling 588 recorded demonstrations. A total of 21 household items are used in various actions.

**Extension of the KIT Bimanual Actions Dataset** The pre-existing KIT Bimanual Actions Dataset (Dreher et al., 2020) is a well-established resource within the community, frequently utilized as a benchmark for action recognition (Morais et al., 2021; Xing and Burschka, 2022; Ziaeetabar et al., 2024). However, the original dataset rarely includes bimanual, symmetric actions where both hands grasp a common object. To address this critical gap, two additional tasks in

[13] https://git.h2t.iar.kit.edu/sw/bimanual-actions/action-labeller
[14] https://motion-database.humanoids.kit.edu/list/datasets/
[15] https://bimanual-actions.humanoids.kit.edu/bimanual_categories

a kitchen context, focusing on such symmetrical actions, were recorded. The new data maintain consistency with the original dataset by following the same recording structure – number of subjects and repetitions – and providing identical extracted features including 3D-bounding boxes of hands and objects and their spatial relationships.

# CHAPTER 5

# Recognition of Bimanual Categories in Human Demonstrations

To effectively utilize the bimanual categories defined in the Bimanual Manipulation Taxonomy, as introduced in Chapter 3, they must be reliably recognized in human demonstrations. Only then can these labels serve as the basis for an autonomously generated task representation derived from demonstration data.

This chapter presents both a rule-based and a learning-based approach for frame-wise classification of the bimanual categories defined by the taxonomy, with the latter employing Graph Neural Networks. A sliding window technique is employed to segment motion data into distinct bimanual categories based on the Bimanual Manipulation Taxonomy. Since we aim for such an online capable approach, we treat the categories *uncoord_bi* and *loosely* as a joint category which we refer to as *loosely*. This is due to the fact, that this distinction requires analyzing the role of both hands within a wider temporal context. The chapter begins by introducing the data used for evaluating both approaches, which consists of marker-based motion capture data and RGB-D recordings. Subsequently, both approaches are described in detail, followed by a comprehensive evaluation of their performance.

The rule-based approach for processing motion capture data was first introduced by Krebs and Asfour (2022), while the learning-based method was proposed in Krebs and Leven et al. (2023).

## 5.1. Evaluation Data

For the evaluation of the approaches, the data presented in Chapter 4 is used. This section provides a concise overview of the data used and outlines the additional preprocessing steps applied.

### 5.1.1. Motion Capture Data

For our analysis, we utilize a dataset comprising 120 recordings from the KIT Bimanual Manipulation Dataset (Krebs and Meixner et al., 2021), which is described in detail in Section 4.1. This dataset contains multimodal recordings of various bimanual household activities. Each motion sequence in the dataset starts and ends with both hands resting on the table, encapsulating a distinct bimanual manipulation task.

Our analysis is based on high-fidelity whole-body human and object motion data, captured using a marker-based Vicon motion capture system operating at 100 Hz. Additionally, hand motion data is recorded at 90 Hz using 18-DoF data gloves. This comprehensive dataset enables precise reconstruction of full-body movements, including detailed hand configurations and the 6D poses of all interacting objects.

Since both subjects in this dataset are right-handed, the data is mirrored along the sagittal plane to enhance dataset balance and mitigate potential bias. This results in a total number of 187 690 labeled frames. The distribution of bimanual categories in the dataset, which are used as ground truth data is depicted in Figure 5.1.

### 5.1.2. RGB-D Data

The second dataset comprises RGB-D data, which is more representative of the sensory input available to a robot due to its compact and mobile sensor configuration. For evaluation, we utilize the kitchen tasks of the *Bimacs* dataset (Dreher et al., 2020) and its extension presented in Section 4.2. The combined dataset consists of 420 recordings with a total duration of about 127 minutes which we refer to as the combined dataset in the remainder of this chapter. 53.6 % of the frames belong to the *Bimacs* dataset and 47.4 % are part of the extension. A single annotator manually labeled the data on a frame-wise basis using the categories of the Bimanual Manipulation Taxonomy.

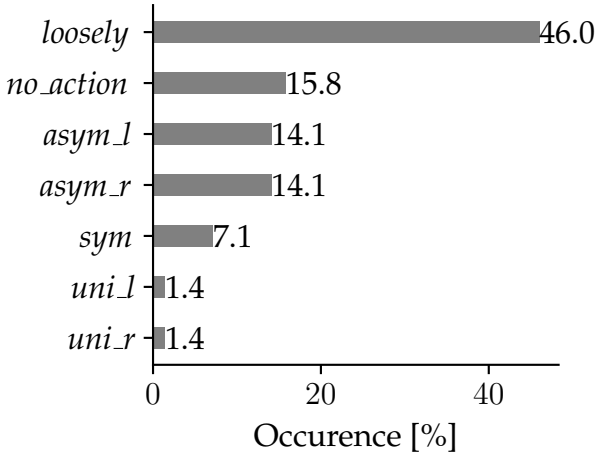



Figure 5.1.: Distribution of the categories in the ground truth motion capture data.

To mitigate the imbalance between right- and left-handed subjects, the complete dataset was mirrored, effectively doubling the data and ensuring a more balanced distribution. This augmented dataset consists of 457 092 frames. The distribution of bimanual categories in the dataset, which are used as ground truth data is depicted in Figure 5.2.

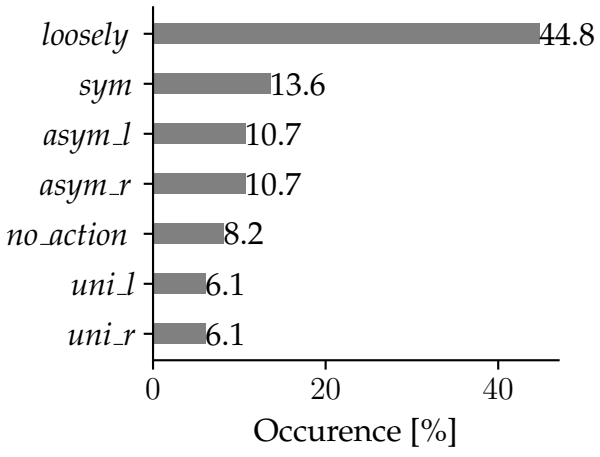



Figure 5.2.: Distribution of the categories in the ground truth RGB-D data.

## 5.2. Rule-based Approach

The objective of this study is to analyze the presence and manifestation of bimanual categories, as defined by our taxonomy, in human demonstrations of daily tasks. As an initial step, we employ a rule-based approach to verify whether these categories can be inherently distinguished in human behavior. A rule-based classification offers several advantages, including efficient data utilization and a reduced risk of overfitting. Additionally, it enhances interpretability by providing explicit decision criteria, facilitating introspection into the decision-making process, and improving the traceability of errors. The rule-based approach presented in this section was originally presented in Krebs and Asfour (2022).

### 5.2.1. Method

A contact graph is constructed based on the hands and objects present in the scene. For motion capture data, precise 3D models of the objects and hands are used, whereas in RGB-D data, these are approximated using axis-aligned bounding boxes. The extraction of contact relationships builds upon previous research on semantic relation extraction (Kartmann et al., 2018). The contact graph is generated frame-by-frame, capturing both contact relationships and their variations throughout the execution of a task. The classification of bimanual categories is derived from the topological structure of this graph.

Additionally, motion features of the hands and objects are of paramount importance. Each node of the graph is augmented with

- the global pose and velocity of the node,
- an object ID as unique identifier of the node, e. g., the object names *rolling_pin*, *sponge_small*, etc. and
- a group ID for *rightGroup*, *leftGoup*, *background* and *scene*.

The construction of the contact graph begins with the evaluation of contact relations in the scene, followed by the assignment of the objects to distinct groups. Objects that remain stationary and are not subject to manipulation, such as tables or walls, are manually assigned to the *background* group to prevent erroneous contact detections between hands and these objects. The *rightGroup* and *leftGroup* include the respective hand together with all objects in direct or indirect contact with it. The right hand and left hand are assigned to the

corresponding groups *rightGroup* and *leftGroup*, which are continuously updated based on changes of the topological structure of the contact graph. Objects in contact with elements of the *rightGroup* or *leftGroup* are appended to the corresponding group unless they are already assigned to another group. All remaining nodes are classified under the *scene* group. Node pose information is obtained by tracking the hands and objects, while node velocity is estimated through numerical differentiation.

While object information is utilized for constructing the contact graph, our approach remains object-agnostic, as the determination of bimanual categories relies solely on the topological structure of the graph. The overall approach is represented in Figure 5.3.

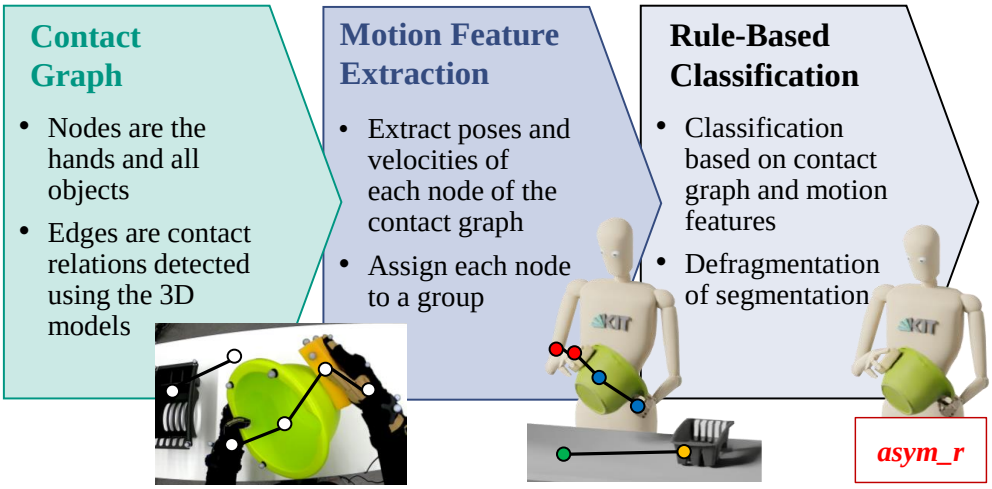



Figure 5.3.: Pipeline for extracting bimanual categories. The figure shows also an example of the contact graph with nodes associated to the *rightGroup* (red), *leftGroup* (blue), background (green) and scene (yellow) as well as the detected bimanual category as tightly coupled, asymmetric with a dominant right hand.
Source: Krebs and Asfour (2022) © 2022 IEEE

To determine the category of bimanual actions, we employ a rule-based classification approach, where decisions are made based on predefined and interpretable *if-else* rules given in Figure 5.4. Starting with the contact graph constructed for each frame of the motion, we employ sliding window approach with a window size of 10 frames for motion classification.

We start at the root node of the decision tree and determine whether there is contact between elements of the *rightGroup* and *leftGroup* in the contact graph,

indicating physical interaction between the hands. If no contact exists, we analyze the average vector norms of the hand velocities $\|\bar{v}_L\|$ and $\|\bar{v}_R\|$ to guide decisions at subsequent levels of the tree. A hand is deemed inactive if its velocity falls below a predefined threshold $v_{th}$ and it is not in contact with any object other than background objects. If only one hand is active, the motion segment is labeled as *uni_r* or *uni_l* depending on which hand is moving. If neither hand is active, the motion segment is classified as *no_action*, whereas if both hands are active, it is classified as *loosely*.

In the event of contact between the hand groups, we compare the average distance $\|\overline{x_R - x_L}\|$ between the hands during motion execution against the distance at the beginning of the respective window $\|x_{R,0} - x_{L,0}\|$ (right sub-tree in Figure 5.4). If the difference is below a given threshold $x_{sym,th}$ , the motion segment is classified as *sym*. Additionally, we validate contact persistence even with minimal model inflation, a necessary requirement for motions classified as *sym*.

If the motion does not fit the *sym* category, we compare the velocities of the hands. Following the concept of Guiard (1987), the hand with the higher mean velocity is deemed the dominant hand, resulting in classification as either *asym_r* or *asym_l* at the leaf node of the tree.

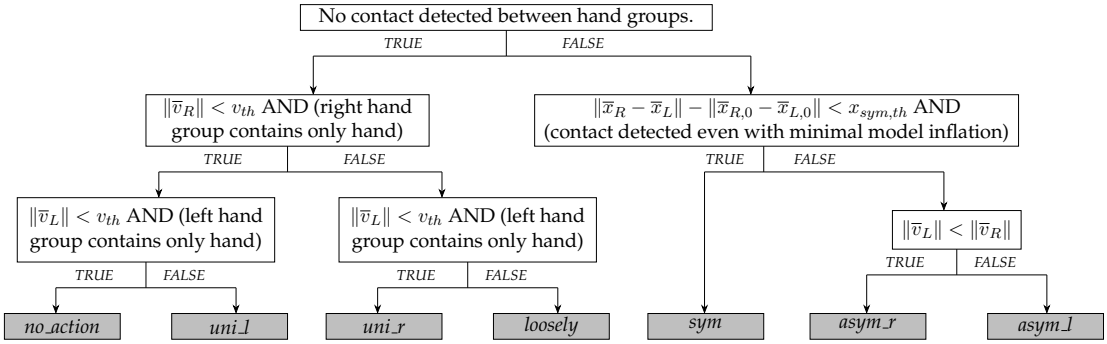



Figure 5.4.: Decision tree for the rule-based classification. Abbreviations are used as defined in Table 3.1.
Source: Krebs and Asfour (2022) © 2022 IEEE

To apply these procedures to longer action sequences, we utilize a sliding window approach, where each window independently yields a category as a classification result. Adjacent windows assigned to the same category are merged into a single segment. A subsequent defragmentation step eliminates any resulting small segments. In essence, the approach not only classifies bimanual actions into predefined categories but also segments the demonstration accordingly.

The threshold parameters used for classification are listed in the Appendix (see Table A1).

## 5.2.2. Evaluation

We evaluate the proposed approach using the data described in Section 5.1. First, we analyze motion capture data from the KIT Bimanual Manipulation dataset, which provides high-precision measurements of human and object movements. Next, we apply the rule-based approach to features extracted from RGB-D data. While RGB-D data is inherently less precise, it offers a more practical and realistic representation for robotic applications.

### Motion Capture Data

The manually segmented reference data was compared frame-by-frame to the automatically generated labels. Figure 5.5 presents the confusion matrix for all motions.

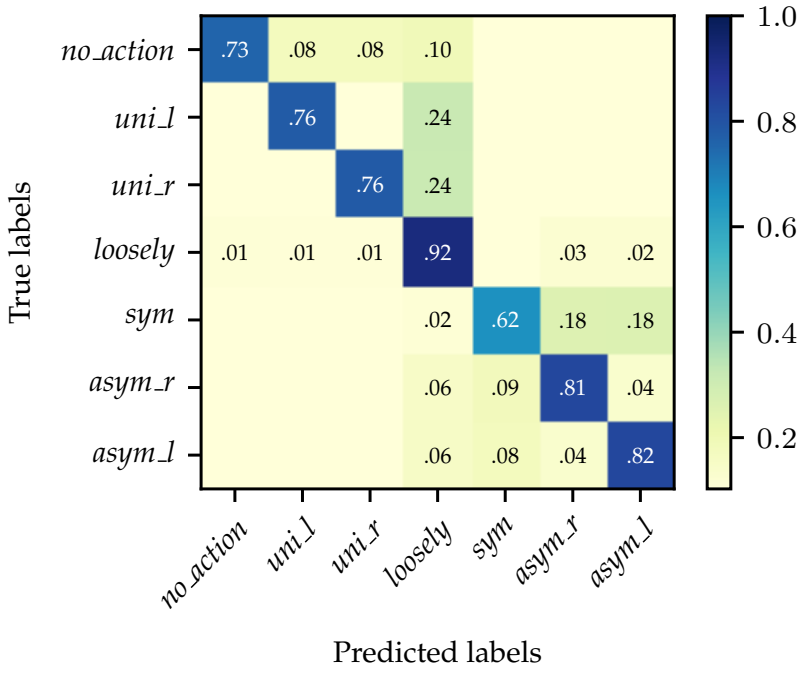



Figure 5.5.: Normalized confusion matrix using the rule-based approach applied on motion capture data.

Confusion frequently arises also within *tightly-coupled* actions. One contributing factor is that the conditions required for rule-based classification are not always strictly satisfied. For example, the condition of constant offset between the hands during rolling is occasionally violated due to subtle changes in hand poses during execution. Furthermore, it is important to acknowledge that, although manually labeled data is used as ground truth, it is still subject to human errors and imprecisions. In some cases, the key points generated by the automatic segmentation may even be more accurate than the manually labeled ones. For example, at the end of the *wiping* action, a segment is often labeled as *asym_l* because the wiping action on the plate has stopped, and the plate is being moved away in a transitional motion to place it down while still in contact with the sponge. Furthermore, the onset of hand motions might be labeled too early or too late, leading to confusion between unimanual and *loosely* frames. Table 5.1 details precision, recall and the $F_1$-score for the different categories.

Table 5.1.: Metrics of the rule-based approach applied on motion capture data.

| Category | Precision | Recall | $F_1$-score |
|---|---|---|---|
| *no_action* | 0.92 | 0.73 | 0.82 |
| *uni_l* | 0.53 | 0.76 | 0.62 |
| *uni_r* | 0.53 | 0.76 | 0.62 |
| *loosely* | 0.93 | 0.92 | 0.92 |
| *sym* | 0.79 | 0.62 | 0.70 |
| *asym_r* | 0.72 | 0.81 | 0.76 |
| *asym_l* | 0.72 | 0.82 | 0.76 |
| Micro avg. | 0.83 | 0.83 | 0.83 |
| Macro avg. | 0.73 | 0.77 | 0.74 |
| Weigh. avg. | 0.83 | 0.83 | 0.83 |

As also evident from the confusion matrix, the recall values for the categories *no_action*, *uni_l*, and *uni_r* are relatively similar. However, their precision differs considerably, with *no_action* exhibiting a substantially higher precision. This discrepancy is likely due to misclassifications occurring during transitions between *no_action* and *uni_l*/*uni_r*, as well as between these categories and *loosely*. Notably, this effect is more pronounced for the unimanual categories due to their considerably lower frequency in the dataset (see Figure 5.1).

### RGB-D Data

In the case of RBG-D data, we leverage features that are extracted from the raw RBG-D data including bounding boxes of hands and objects, along with their spatial relationships. However, compared to the application of the approach on motion capture data, minor adaptations are required. Specifically, the consideration of orientation is omitted, as axis-aligned bounding boxes are utilized. Furthermore, heuristics were implemented to handle objects not detected in intermediate frames and objects only detected in few isolated frames. Threshold parameters are adapted to better suit the less precise data. Further, in the RGB-D dataset spatial relations instead of only contacts are available. In order to extract the contact relations from the spatial relations we consider the attributes *contact*, *fixed moving together*, *halting together*, *moving together*, *surround* and *inside* as indicators for contact. As for the GNN-based approach the original and mirrored data is considered. Figure 5.6 shows the confusion matrix for the rule-based classification based on the combined dataset including both the original *Bimacs* dataset and the extension and Table 5.2 the corresponding metrics.

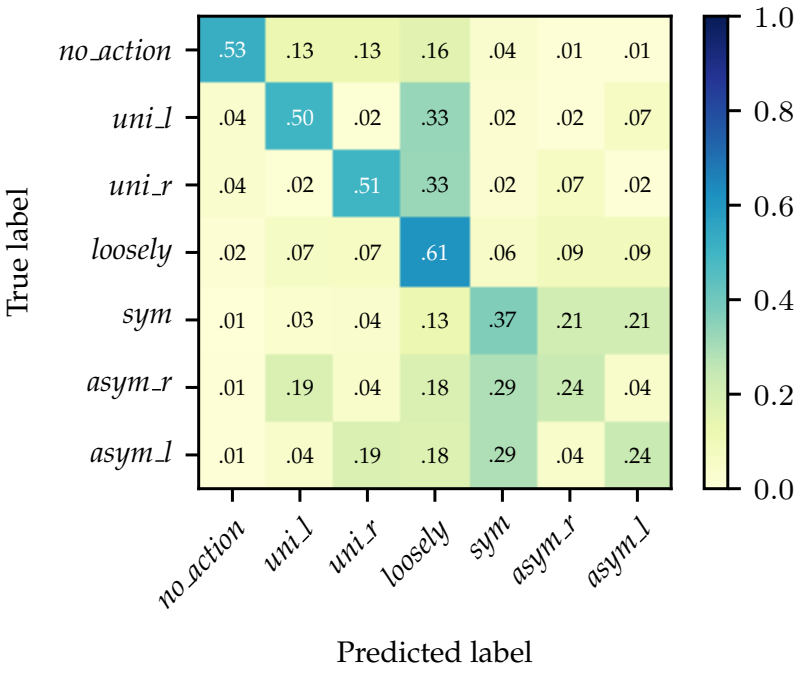



Figure 5.6.: Normalized confusion matrix using the rule-based approach applied on RGB-D data.

The performance is significantly worse for RGB-D data which can be observed e. g., based on the weighted $F_1$-score of 0.49 compared to a value of 0.83 for the motion capture data. This is due to the reduced quality of RGB-D data compared to accurate motion capture data used in Section 5.2.2. Notably, predictions are particularly imprecise for the tightly coupled categories. This indicates that for most cases, the contact-based differentiation still works reasonably well, but the motion-based differentiation within the tightly coupled categories is not working anymore. This is due to the fact that a rule-based approach with fixed thresholds performs worse for the information extracted from the noisy RGB-D data. In addition, not considering orientations due to using axis-aligned bounding boxes might increase the effect. The high number of frames that are wrongly classified as *loosely* supports this hypothesis. There are some categories whose true labels belong to the tightly coupled categories indicating that a misclassification occurred due to failure in contact detection. However, there are even more frames with true labels that are either *unimanual* or *no_action*. This is likely to be either due to missing objects or imprecise motion data.

Table 5.2.: Metrics of the rule-based approach applied on RGB-D data.

| Category | Precision | Recall | $F_1$-score |
|---|---|---|---|
| *no_action* | 0.73 | 0.53 | 0.61 |
| *uni_l* | 0.30 | 0.50 | 0.38 |
| *uni_r* | 0.30 | 0.51 | 0.38 |
| *loosely* | 0.71 | 0.61 | 0.66 |
| *sym* | 0.35 | 0.37 | 0.36 |
| *asym_r* | 0.24 | 0.24 | 0.24 |
| *asym_l* | 0.24 | 0.24 | 0.24 |
| Micro avg. | 0.48 | 0.48 | 0.48 |
| Macro avg. | 0.41 | 0.43 | 0.41 |
| Weigh. avg. | 0.52 | 0.48 | 0.49 |

## 5.3. Learning-Based Approach

While the rule-based approach demonstrated reasonable performance for motion capture data, its effectiveness for RGB-D data proved insufficient. To address this limitation, we explore alternative methodologies commonly employed in action recognition. As discussed in Section 2.2.2, deep neural networks are widely utilized for this purpose. In this section, we introduce the neural network

architecture used in our study, along with its training process and evaluation for both motion capture and RGB-D data

### 5.3.1. Method

The selected network architecture must satisfy several requirements. On the one hand, the network must be capable of dealing with variable input sizes due to the variable structure of the considered scene resulting from different numbers of objects and spatial relations. On the other hand, the network should be able to process graph-based representations similar to the scene graphs used in our previous rule-based approach. GNNs are predestined for such tasks as they fulfill both requirements. Thus, GNNs are selected for the recognition of bimanual manipulation categories.

As described in (Dreher et al., 2020) and originally defined in (Battaglia et al., 2018), we define a graph $G$ as a 3-tuple $G = (u, V, E)$, with $u$ being the global attribute of the graph, $V$ the set of nodes in the graph and $E$ the set of edges. The set of nodes $V$ consists of the node attributes $v_a \in V$ and the edges $E$ of 3-tuples $e = (e_a, s, r) \in E$. Within the edges, $e_a$ represents the edges attributes and $s$ and $r$ are the sender and receiver nodes in $V$. In our case, the input graph is constructed based on the extracted features in each frame, where nodes are the object/hand instances and edges encode the spatial relations between them. Furthermore, the scene graph of the current and the last nine frames are concatenated by temporal edges connecting the node of a specific object instance between consecutive frames. The global attribute $u$ is not used in the input graph but encodes the determined category as one-hot-encoding in the output graph.

Given the similarity of the problem in Dreher et al. (2020), which involves essenetially the same input graph, a similar network structure is used. The model consists of two independent graph network blocks for the encoder and decoder, respectively, and one full graph network block for the core. The architecture follows an *encode-process-decode* configuration. In all blocks Multilayer Perceptrons (MLPs) were used as update functions with the sum function used as an aggregation function. A hyperparameter search was performed to determine the number of layers and neurons per layer for all MLPs, as well as the batch size, learning rate, history size and processing steps of the core.

### 5.3.2. Evaluation

We evaluate the proposed approach using the datasets described in Section 5.1. The results for both motion capture and RGB-D data are presented and compared to those obtained with the rule-based approach. Additionally, we analyze and compare the resulting motion segmentations.

#### Motion Capture Data

To employ a GNN for frame-wise classification of marker-based motion capture data, the same features used for classifying RGB-D recordings are extracted from the recordings analyzed in the rule-based approach, including the mirrored data. These features encompass bounding boxes for hands and objects, derived from object poses, as well as the computation of their spatial relationships. Testing the performance of GNNs on this type of data remains a preliminary investigation, as the dataset size is relatively small for a learning-based approach. A leave-one-subject-out cross-validation scheme is employed for evaluation. In this setup, the model is trained using the data from one subject, with every sixth recording reserved for validation, and subsequently evaluated on the data from the other subject. To achieve a more balanced distribution in the training dataset (see Figure 5.1), only every third frame labeled as *loosely* is considered.

We train GNNs with two different architectures: i) the same hyperparameters that have been determined for the RGB-D data as given in Table 5.5 and ii) a model with reduced size. The reduced model differs to the large model in regard of having only one layer, 64 neurons and 5 processing steps.

The resulting $F_1$-scores are presented in Table 5.3, demonstrating that both architectures achieve superior performance compared to the rule-based classification. Notably, they perform very similar but the smaller model exhibits a slight performance advantage over the larger one. The confusion matrix for the smaller model is shown in Figure 5.7 and the larger one in the Appendix (see Figure A1). When comparing Figure 5.7 with the confusion matrix of the rule-based approach for the same dataset (Figure 5.5), it becomes evident that the rule-based method frequently misclassifies the symmetric category as asymmetric, whereas this issue does not occur with the GNN-based approach. However, both models exhibit problems in correctly classifying unimanual samples, which are often misidentified as *loosely*.

Table 5.3.: Comparison of the rule-based and GNN-based classification of motion capture data. The Micro $F_1$-score corresponds to the accuracy.

| | $F_1$-score | | |
|---|---|---|---|
| | Micro | Macro | Weighted |
| Rule-based | 0.83 | 0.71 | 0.83 |
| GNN (large model) | 0.90 | 0.83 | 0.90 |
| GNN (small model) | 0.92 | 0.85 | 0.92 |

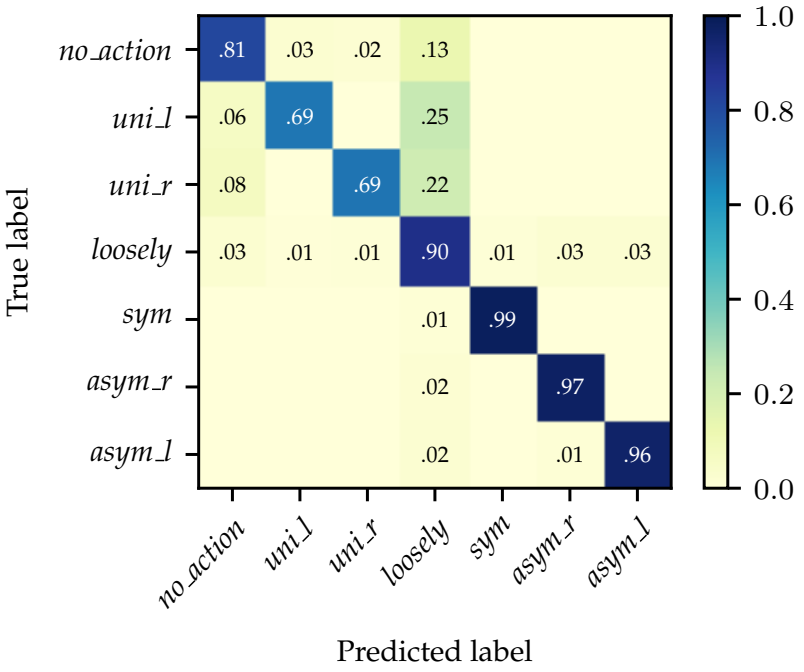



Figure 5.7.: Normalized confusion matrix using the GNN-based approach (small model) applied on motion capture data.

Table 5.4 shows the overall high performance of the GNN-based approach applied on motion capture data with a weighted average $F_1$-score of 0.92.

Table 5.4.: Metrics of the GNN-based approach applied on motion capture data.

| Category | Precision | Recall | $F_1$-score |
|---|---|---|---|
| *no_action* | 0.81 | 0.81 | 0.82 |
| *uni_l* | 0.63 | 0.69 | 0.66 |
| *uni_r* | 0.64 | 0.69 | 0.67 |
| *loosely* | 0.94 | 0.90 | 0.92 |
| *sym* | 0.98 | 0.99 | 0.98 |
| *asym_r* | 0.91 | 0.97 | 0.94 |
| *asym_l* | 0.91 | 0.96 | 0.94 |
| Micro avg. | 0.92 | 0.92 | 0.92 |
| Macro avg. | 0.83 | 0.86 | 0.85 |
| Weigh. avg. | 0.92 | 0.92 | 0.92 |

### RGB-D Data

As the evaluation of the rule-based approach on RGB-D data yielded suboptimal results, with a weighted average $F_1$-score of only 0.49, learning-based approaches offer significant potential, particularly for RGB-D data. Such methods may be better suited for handling the inherent noise and imprecision of this data. For training and evaluation, the RGB-D dataset described in Section 5.1.2 is used. To account for the overrepresentation of *loosely* coupled bimanual actions (see Figure 5.2) with over 40 % of all frames for the training of the GNN, two of three frames labeled as *loosely* are skipped. To identify the optimal parameters for the GNN network structure, a systematic evaluation was conducted based on the data of a single subject, beginning with the initial parameter settings reported in Dreher et al. (2020). As shown in Table 5.5, the highlighted row corresponds to the parameters yielding one of the highest $F_1$ scores and was selected as the final configuration. Although configurations with marginally higher $F_1$ scores were observed—achieved through either larger history sizes or an increased number of processing steps—these were excluded due to the substantially longer training times required.

The dataset was split into a training set and a testing set. Testing sets contain all recordings from one subject (one subject of each dataset including its mirrored motions), while training sets contain all remaining recordings. Additionally, before training, one out of the ten repetitions for each task in the training set was put aside as a validation set. For the quantitative evaluation of the classifier, a leave-one-subject-out cross-validation was performed to obtain six folds of training and testing sets.

Table 5.5.: Parameter evaluation of the MLPs was conducted on a single subject. The highlighted row indicates our final parameter selection. Source: Krebs and Leven et al. (2023) © 2023 IEEE

| Layers | Neurons | Batch size | Learning rate | History size | Process steps | Macro $F_1$-score |
|---|---|---|---|---|---|---|
| 2 | 256 | 256 | 0.001 | 10 | 10 | 0.7086 |
| **1** | 256 | 256 | 0.001 | 10 | 10 | 0.6870 |
| **3** | 256 | 256 | 0.001 | 10 | 10 | 0.6996 |
| 2 | **128** | 256 | 0.001 | 10 | 10 | 0.6892 |
| 2 | **512** | 256 | 0.001 | 10 | 10 | 0.7045 |
| 2 | 256 | **32** | 0.001 | 10 | 10 | 0.6874 |
| 2 | 256 | **128** | 0.001 | 10 | 10 | 0.6851 |
| 2 | 256 | **512** | 0.001 | 10 | 10 | 0.6980 |
| 2 | 256 | 256 | **0.01** | 10 | 10 | 0.5029 |
| 2 | 256 | 256 | **0.0001** | 10 | 10 | 0.6734 |
| 2 | 256 | 256 | 0.001 | **1** | 10 | 0.6463 |
| 2 | 256 | 256 | 0.001 | **5** | 10 | 0.6886 |
| 2 | 256 | 256 | 0.001 | **20** | 10 | 0.7027 |
| 2 | 256 | 256 | 0.001 | 10 | **5** | 0.6882 |
| 2 | 256 | 256 | 0.001 | 10 | **20** | 0.7101 |

A combined evaluation of the six test sets results in a weighted average $F_1$ score of 0.72. The confusion matrix is shown in Figure 5.8. The overfitting of the *loosely* category is visible and is caused by the fact that the *loosely* category is the category with the highest occurrence in the training data. However, we hypothesize that this distribution is legitimate for the training data, as both datasets suggest that the category *loosely* is indeed more prevalent in natural movements than other categories.

Table 5.6 shows the macro metrics obtained in the evaluation for the different categories. It can be observed that particularly the tightly-coupled categories (*asym_r*, *asym_l*, *sym*) are detected best. The precision and $F_1$ score are also high for *loosely*. However, the recall is significantly lower, since as shown in the confusion matrix in Figure 5.8 many unimanual motions are falsely detected as *loosely*.

In the described scenario, the node IDs of the input graphs are consistently associated with a specific hand or object across the entire dataset. While the GNN does not explicitly encode the semantic properties of an object (e. g., recognizing that a rolling pin is used for rolling), it does capture the correlation between the object's use and its associated category. For instance, the GNN frequently recognizes the symmetric category when a rolling pin is involved in rolling. During inference,

the GNN can leverage the object ID to aid in classification, effectively utilizing object-specific information as part of its knowledge.

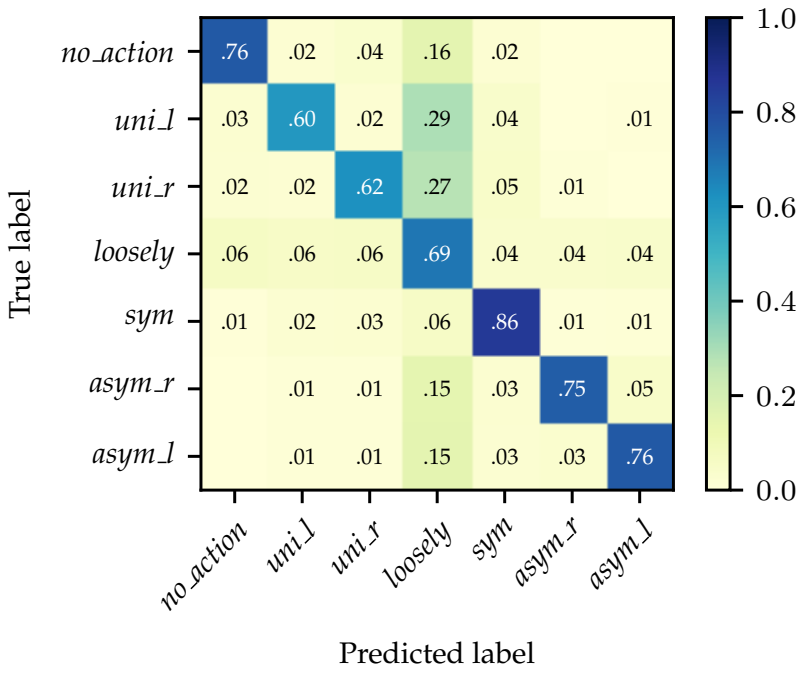



Figure 5.8.: Normalized confusion matrix using GNN-based approach with object knowledge applied on RGB-D data.

Table 5.6.: Metrics of the GNN-based approach with object knowledge applied on RGB-D data.

| Category | Precision | Recall | $F_1$-score |
|---|---|---|---|
| *no_action* | 0.65 | 0.76 | 0.70 |
| *uni_l* | 0.52 | 0.60 | 0.56 |
| *uni_r* | 0.50 | 0.62 | 0.55 |
| *loosely* | 0.78 | 0.69 | 0.73 |
| *sym* | 0.79 | 0.86 | 0.82 |
| *asym_r* | 0.76 | 0.75 | 0.76 |
| *asym_l* | 0.75 | 0.76 | 0.75 |
| Micro avg. | 0.72 | 0.72 | 0.72 |
| Macro avg. | 0.68 | 0.72 | 0.70 |
| Weigh. avg. | 0.73 | 0.72 | 0.72 |

To also investigate the case of unknown objects, we keep the object IDs the same only within one recording. This corresponds to the case where the object is unknown but can be tracked through a demonstration. For different recordings,

the IDs are assigned differently, so that it is not possible for the model to learn a relation as described above. This results in a lower weighted average $F_1$ score of 0.62, the confusion matrix depicted in Figure 5.9 and detailed metrics can be found in the Appendix (Table A6). As can be seen from the confusion matrix, the results are worse across all categories showing the relevance and advantage of using object knowledge.

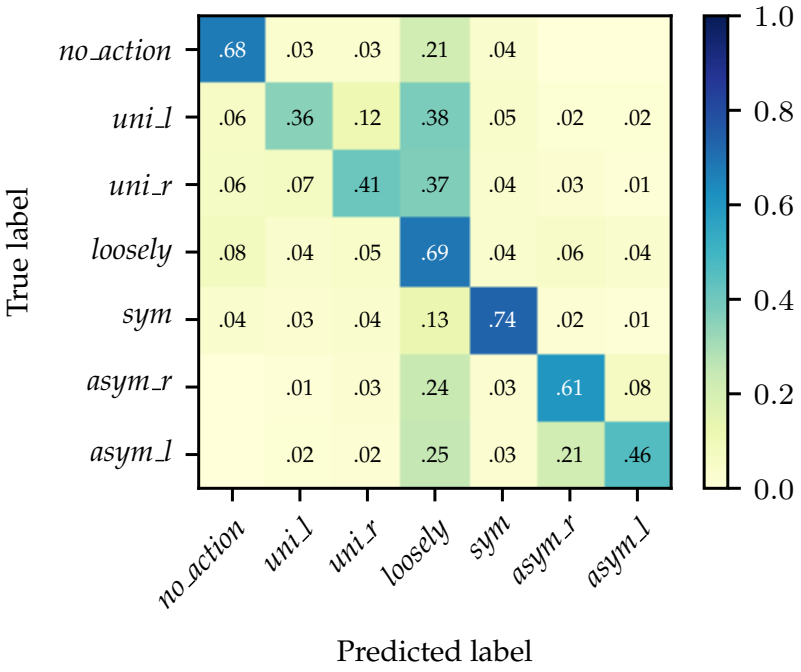



Figure 5.9.: Normalized confusion matrix using GNN-based approach without object knowledge applied on RGB-D data.

Compared to the results of the rule-based approach applied to RGB-D data (see Table 5.2 and Figure 5.6), the GNN-based method demonstrates a substantial improvement in performance. The weighted average $F_1$ score increased from 0.49 for the rule-based approach to 0.72 for the model incorporating object knowledge. This indicates that while category recognition is feasible using RGB-D data, a more sophisticated model than a rule-based approach is required to achieve higher accuracy.

**Ablation Study** The presented results demonstrate that GNNs significantly outperform the rule-based approach when applied to imperfect data and features extracted from RGB-D. In this section, we analyze the specific characteristics of the GNN and their influence on the network's performance. The GNN uses spatial relations instead of only contact relations as done in the rule-based

approach in (Krebs and Asfour, 2022). For comparison, we also train a GNN by only considering contact relations. Furthermore, we consider a version where the input graph contains information from one frame only and there are no temporal edges connecting object instances in the scene graphs over multiple frames. This is evaluated both with spatial relations and only contact relations. We also add the version without object knowledge for comparison. The resulting macro scores are shown in Table 5.7 and the weighted scores can be found in the Appendix (see Table A7).

As expected, the suggested approach yields the best performance. However, interestingly, omitting the temporal edges decreases the performance less than considering only contact relations. This could be due to certain temporal information being encoded within dynamic spatial relations e. g., *halting together*, *moving apart*. As expected, the lowest scores are obtained for the model that only considers contacts without any temporal edges. However, even this variant outperforms the rule-based approach (see Table 5.2).

Table 5.7.: Ablation study comparing the macro metrics.

| Training data | | | Results | | |
|---|---|---|---|---|---|
| Spatial Relations* | Temporal Edges | Object Knowledge | Precision | Recall | $F_1$-score |
| ✓ | ✓ | ✓ | **0.68** | **0.72** | **0.70** |
| | ✓ | ✓ | 0.54 | 0.58 | 0.55 |
| ✓ | | ✓ | 0.63 | 0.68 | 0.65 |
| | | ✓ | 0.48 | 0.55 | 0.50 |
| ✓ | ✓ | | 0.56 | 0.56 | 0.56 |

*In case of no spatial relations only contact relations are considered.

**Comparison to State-Of-the-Art Methods** As discussed in Section 2.2.2, several studies have explored activity recognition using RGB-D data. To facilitate a comparative analysis, we retrain selected state-of-the-art approaches for the specific task of bimanual category recognition and evaluate their performance[1].

For our learning-based approach, we employ GNNs, as they are inherently well-suited for modeling structural dependencies in data that can be represented as graphs, particularly when both temporal and spatial constraints are present, as is in our case. Our GNN operates on predefined features that can be extracted from

[1]This analysis was performed in the bachelor's thesis of Christian Diehm.

multiple sensor modalities, specifically motion capture and RGB-D data. We hypothesize that this prior feature selection enhances generalization by reducing the risk of overfitting to irrelevant factors such as background information. Consequently, we opted against approaches that directly process raw visual data, such as STIGPN (Wang et al., 2021) and ASSIGN (Morais et al., 2021) (see Table 2.2).

In Table 5.8, we compare our approach against the methods discussed Section 2.2.2. PGCN (Xing and Burschka, 2022) is excluded from the evaluation, as its implementation is not publicly available, preventing a rigorous and reproducible comparison. The evaluation follows the same methodology as for the GNN, utilizing a leave-one-subject-out cross-validation approach.

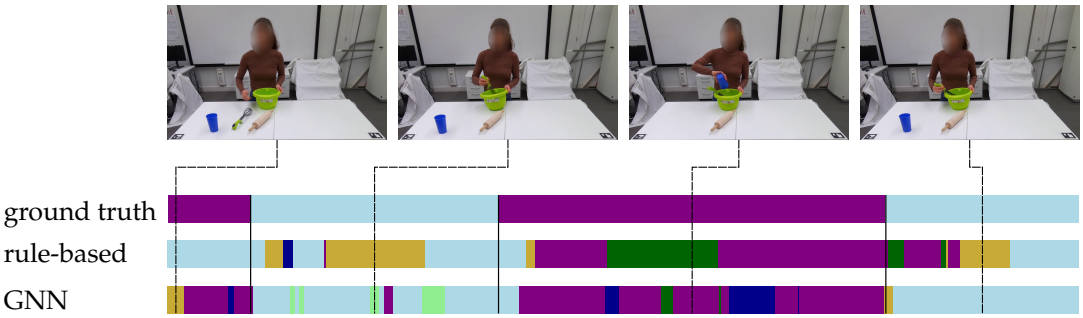
| Approach | Macro $F_1$ | Micro $F_1$ |
|---|---|---|
| **Ours** | **0.70** | **0.72** |
| STIGPN (Wang et al., 2021) | 0.65 | 0.71 |
| ASSIGN (Morais et al., 2021) | 0.78 | 0.82 |
| ISTA-Net (Wen et al., 2023) | 0.70 | 0.75 |

Table 5.8.: Comparison of our GNN-based approach applied on RGB-D data with other state-of-the-art methods for activity recognition.

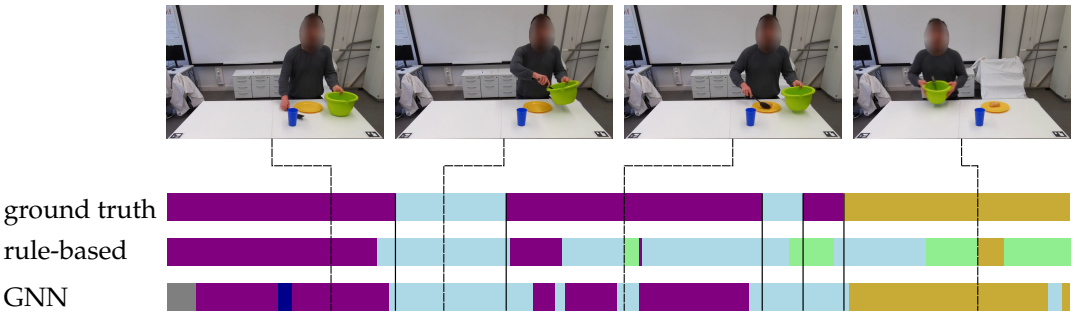
The performance of our model is comparable to existing approaches in action recognition. It slightly outperforms STIGPN (Wang et al., 2021) and achieves results similar to ISTA-Net (Wen et al., 2023). Only ASSIGN (Morais et al., 2021) demonstrates a notably superior performance. However, as discussed in Section 2.2.2, ASSIGN is unsuitable for online applications, as it requires the entire motion sequence to be recorded before classification and segmentation can be performed. This limitation is particularly relevant for potential deployment on a humanoid robot, where near real-time processing is essential.

### Segmentation Results

While the previous section mainly considered the classification, in this section we focus on segmentation. Considering the segmentation of RGB-D data, an exemplary segmentation for an extract of the newly recorded task *prepare dough* is shown in Figure 5.10a and for the task *set table* in Figure 5.10b.




(a) Example segmentation of the *prepare dough* task.



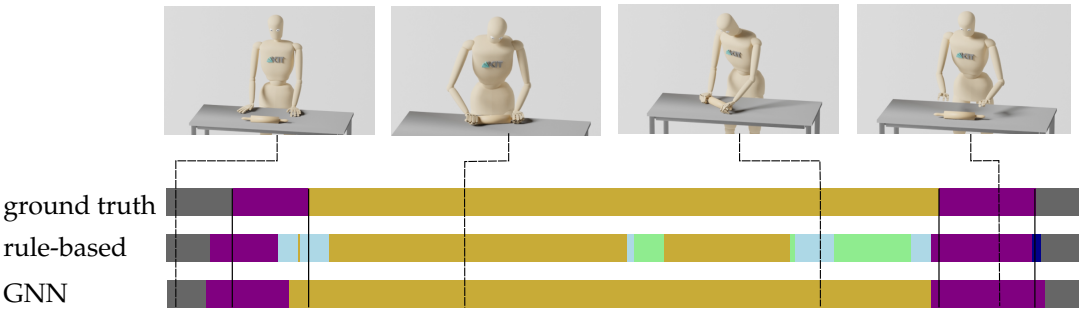

(b) Example segmentation of the *set table* task.

Figure 5.10.: Example segmentation of the *prepare dough* and *Set table* tasks which are part of the RGB-D dataset. The top bar visualizes the ground truth, the middle bar is the segmentation of the rule-based approach, and the bottom bar is the segmentation of the GNN-based approach. Categories: *no_action*, *uni_r*, *uni_l*, *loosely*, *sym*, *asym_r*, *asym_l*
Source: Krebs and Leven et al. (2023) © 2023 IEEE

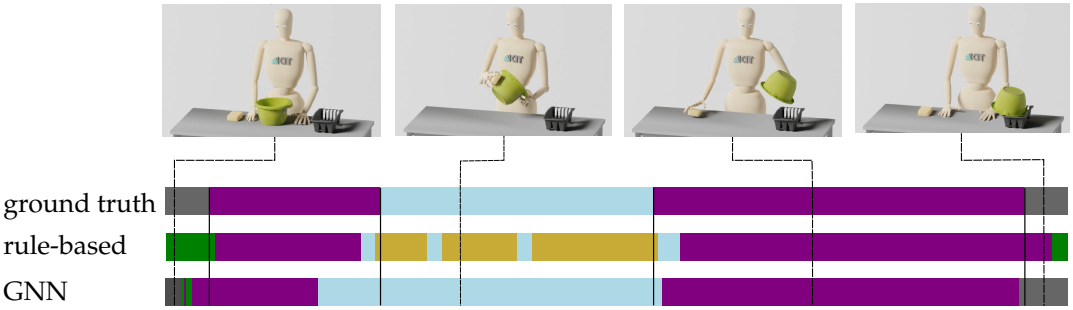
The manually annotated ground truth segmentation is compared against the rule-based and GNN-based approach. Compared to the rule-based approach the segmentation points of the GNN are quite close to the ground truth data. During the *loosely* actions in Figure 5.10a there are some segments of unimanual actions in both approaches which means, that the activity of one hand was not properly detected. For the rule-based approach, the *asym_r* actions also have a high misclassification rate because the threshold for symmetric motions leads to the wrong label *sym*, and the failure to recognize contact relations results in the wrong label *loosely*. This is also evident in Figure 5.10b where particularly the rule-based approach is erroneous within the *loosely* segment and hardly detects the *sym* category at all.

Figure 5.11 shows example segmentations for results for the motion capture data. For both examples, it is clearly evident that the GNN outperforms the rule-based method. In these examples particularly the problems of the rule-based

approach to differentiate between tightly-coupled actions (here *asym_r* and *sym*) is evident.




(a) Example segmentation of the *roll* task.




(b) Example segmentation of the *wipe* task.

Figure 5.11.: Example segmentation of the *roll* and *wipe* tasks which are part of the motion capture dataset and represented using the MMM model. The top bar visualizes the ground truth, the middle bar the segmentation of the rule-based approach, the bottom bar the segmentation of the GNN-based approach. Categories: *no_action*, *uni_r*, *uni_l*, *loosely*, *sym*, *asym_r*, *asym_l*

## 5.4. Summary

In this chapter, we demonstrate the recognition of bimanual categories, as defined by the Bimanual Manipulation Taxonomy in the context of human bimanual manipulation.

We first show that a rule-based approach achieves good results on high-quality marker-based motion capture data, with a weighted average $F_1$ score of 0.83. However, this modality is impractical for humanoid robots. Therefore, we also consider RGB-D data, which can be recorded using onboard cameras. In this case, rule-based methods perform notably worse, with a weighted average $F_1$ score of 0.49. To overcome this limitation, we develop methods that are tailored to work with lower-quality features extracted from RGB-D data.

Building on existing research in human-object interaction recognition and human activity recognition (HAR), we introduce learning-based methods leveraging GNNs to tackle this challenge. These methods substantially improve classification performance for RGB-D data, achieving a weighted average $F_1$ score of 0.72. This demonstrates that bimanual categories can be effectively recognized in human demonstrations.

In this study, axis-aligned bounding boxes were used as features for RGB-D data. We hypothesize that incorporating more precise motion tracking techniques and 6D pose estimation (e. g., Labbé et al. (2022)) could further enhance classification performance.

# CHAPTER 6

# Taxonomy-driven Execution of Bimanual Tasks

This chapter introduces the formulation of a bimanual task model for humanoid robots based on bimanual categories (Section 6.1), as defined in the Bimanual Manipulation Taxonomy described in Chapter 3. The chapter is structured to systematically address the different levels of the task model. Section 6.2 provides a detailed explanation of the low-level task impedance controller. Subsequently, Section 6.3 explores motion generation based on bimanual categories, supported by experimental validation on a humanoid robot. Finally, Section 6.4 presents a conceptual discussion on the sequencing of Bimanual Action Categories.

## 6.1. Taxonomy-Driven Task Model

A task model for bimanual manipulation must effectively represent actions and their associated constraints to ensure smooth and coordinated execution. We assume that the sequence of actions for both hands is given, either learned from human demonstrations or generated by a symbolic planner. The task model should not only capture the sequence of bimanual actions but also incorporate the constraints inherent to each category, such as relative hand positioning, force interactions, and temporal dependencies. By integrating these constraints, the model ensures that execution remains consistent with the constraints defined by the bimanual categories.

Within the **taxonomy-driven task model**, we define a task $T$ as a sequence of $N$ bimanual action categories $BAC_n$, with $n \in \{1, \ldots, N\}$ and can be described as

$$T = \{BAC_1, BAC_2, \ldots, BAC_N\}. \tag{6.1}$$

A bimanual action category $BAC_n$ is given by the actions performed by both hands as well as the constraints specified by each bimanual action category. Thus, we can describe a $BAC_n$ as a triple

$$BAC_n = (l_n, a_L, a_R), \tag{6.2}$$

where $l_n$ is the bimanual category label, corresponding to a specific bimanual category as defined in the Bimanual Manipulation Taxonomy and $a_L/a_R$ are the actions to be executed by the left and right hand.

For successful execution, actions must be mapped to controllers with specific parameters that define their key characteristics, including start and end configuration, duration and other relevant motion attributes. In this work, we employ via-point movement primitives (Zhou et al., 2019) for action execution. This approach allows for smooth and flexible motion generation by defining intermediate waypoints that guide the trajectory while maintaining task-specific constraints. Formally, an action $a$ is represented by a triple

$$a = (mp, c, \varphi), \tag{6.3}$$

where $mp$ represents the motion primitive, $c$ is the parametrization of the motion primitive (including start, end, and duration), and $\varphi$ is the current temporal state of the motion primitive.

The resulting structure of the taxonomy-driven task model and hierarchical controller for execution is shown in Figure 6.1.

1. **Task Level**: This level is responsible for high-level planning, i. e., the generation of plans as sequences of actions to achieve the goal of a task such as *setup a table for two persons*. This high-level symbolic planning is not addressed in the thesis as this goes beyond the scope of the work.
2. **Bimanual Action Category Level**: This level represents specific bimanual action categories, which describe actions performed by both hands and their corresponding constraints such as *place a bowl on the table*. It does not only receive this information from the task model but also keeps its internal $S_{BAC}$.

3. **Category-Based Motion Generation Level**: This level translates bimanual action categories into task-space targets based on the system's current state and information from the higher level.
4. **Control Level**: This level implements the low-level robot controller to map task-space targets to robot control commands.

The Task Level contains the task model $T$ including the sequence of bimanual action categories. During task execution, one bimanual action category $BAC_n$ is processed at a time. Each $BAC_n$ is sent to the Bimanual Action Category Level, which, in turn, sends the following information

$$g = (l, \xi_{L,d}, \xi_{R,d}, \xi_{rel}),$$

to the Category-Based Motion Generation Level, where

- $l$ is the current bimanual category and
- $\xi_{L,d}, \xi_{R,d}, \xi_{rel}$ are task-space goals for both arms and the relative pose of the hands as computed by the current actions.

The Category-Based Motion Generation Level outputs task-space goals for both arms. These goals are then converted into joint torque setpoints at the Control Level and sent to the robot.

From a bottom-up perspective, the robot continuously transmits its current state $r$ to the Control Level, which subsequently provides the current poses of the robot's end-effectors $\xi_{R,c}, \xi_{L,c}$ to the Category-Based Motion Generator Level. This evaluates whether the executed motions are being tracked with sufficient accuracy. If deviations exceed given thresholds, stopping signal $\gamma_R, \gamma_L$ are issued. Based on this evaluation and its current internal state $S_{BAC}$, the Bimanual Action Category Level updates its internal state and determines whether a transition to the next bimanual action category is possible. If so, the decision is communicated to the Task Level.

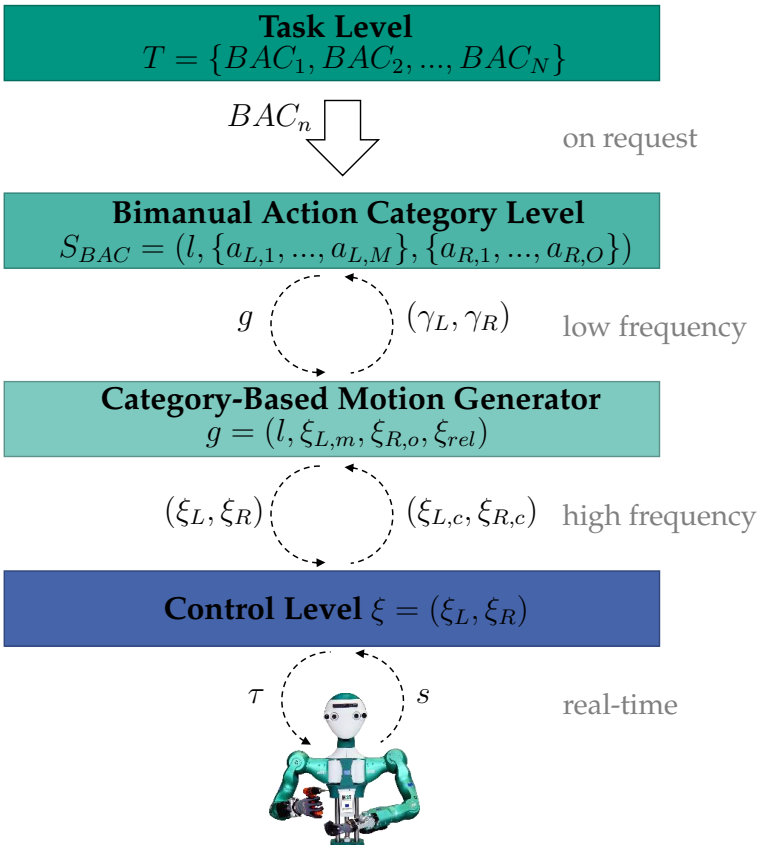



Figure 6.1.: Hierachical task model based on categories of the bimanual manipulation taxonomy.
$T$: task, $BAC_n$: bimanual action category, $S_{BAC}$: internal state of Bimanual Action Category Level, $g$: tuple including bimanual category label and desired end-effector poses, $l$: bimanual category label, $a$: action, $\gamma$: indication of phase stopping, $\xi$: robot end-effector poses, $\tau$: torque commands, $s$: current robot state, indices $L/R$ denote the left/right hand respectively.

The thesis focuses on the Bimanual Action Category Level and the Category-Based Motion Generation Level, which serve as intermediary layers for translating symbolic bimanual categories into effective robot control strategies. In the following, we start by describing the low-level robot controller (Section 6.2), which serves as the basis for the realization of the category-based controller (Section 6.3). Finally, we describe how transitions between bimanual action categories are implemented (Section 6.4).

## 6.2. Low-Level Controller

To ensure safe human-robot interaction, we implement an impedance-control-based scheme that specifically enforces two key safety constraints: self-collision avoidance and joint limit protection[1]. These constraints are particularly critical in bimanual scenarios, where the risk of arm collisions and the reduction of available nullspace due to closed-kinematic chain configurations must be addressed. The control approach must be reactive and capable of operating in real-time. To achieve a cycle time of 1 kHz, we focus on methods that utilize artificial potential fields (APFs).

In this section, we provide an overview of the key components of the controller, focusing on self-collision avoidance as presented in (Dietrich et al., 2012b) (Section 6.2.1) and joint limit avoidance based on (Eckhoff et al., 2023) (Section 6.2.2). Additionally, we explain how these components are integrated with task space impedance and joint space control to form a hierarchical control structure (Section 6.2.3). A more comprehensive discussion of related work on these topics can be found in appendix B.

### 6.2.1. Self-Collision Avoidance

The implementation of self-collision avoidance is based on prior work by (Dietrich et al., 2012b). The authors show that this approach can effectively avoid self-collisions for a dual-arm robot in real-time. In this section, we recall the main mathematical concepts of their work, details can be found in their original paper.

The core concept can be summarized as the generation of repulsive forces at contact points on robot links that are in close proximity, based on minimum distance criteria. These forces are then mapped to joint torques to prevent collisions.

Based on the maximum repulsive force $F_{max}$, a specified distance $d_0$ and the distance $d_{i,j}$ between the points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, the repulsive force $F_{rep}$ is defined as

$$F_{rep}(d_{i,j}) = -\frac{\partial V_{rep,i,j}}{\partial d_{i,j}} = \begin{cases} \frac{F_{max}}{d_0^2}(d_{i,j} - d_0)^2 & \forall d_{i,j} \leq d_0 \\ 0 & \forall d_{i,j} > d_0. \end{cases} \tag{6.4}$$

[1]This control scheme was implemented as part of the bachelor's thesis of Jan Fenker, which was co-supervised by Jianfeng Gao.

In order to map the forces at specific contact points into joint space, Jacobians are projected into the direction of collision. The Jacobian $\boldsymbol{J}_{x,i}$ at the point $\boldsymbol{x}_i$ is defined as

$$\boldsymbol{J}_{x,i}(\boldsymbol{q}) = \frac{\partial \boldsymbol{x}_i(\boldsymbol{q}_i)}{\partial \boldsymbol{q}_i}. \tag{6.5}$$

Based on this, the projected Jacobian is computed as

$$\boldsymbol{J}_i(\boldsymbol{q}) = \boldsymbol{e}_i^T \boldsymbol{J}_{x,i}(\boldsymbol{q}), \tag{6.6}$$

where $\boldsymbol{e}_i$ is a unit vector pointing from $\boldsymbol{x}_i$ to $\boldsymbol{x}_j$. If both contact points are on the same arm, the repulsive force is only applied to the point that is further away from the robot's root. A key consideration in dissipating kinetic energy during motion is the introduction of damping. In the context of self-collision avoidance, this implies that limbs should be maintained outside critical zones, while avoiding excessive displacement beyond what is necessary for safety. As for the repulsive force, damping is applied for all considered contact pairs.

The configuration dependent damping $\boldsymbol{D_{d,i,j}}$ applied on point $\boldsymbol{x}_i$ is computed based on the *Double Diagonalization* method (Albu-Schäffer et al., 2004).

Overall, as discussed in (Dietrich et al., 2012b), the repulsive force and the damping force of $n_p$ contact points can be combined and transformed into torque commands

$$\boldsymbol{\tau}_{coll} = \sum_{i=1}^{n_p} \begin{pmatrix} \boldsymbol{J}_i^T(\boldsymbol{q}) \\ \boldsymbol{J}_j^T(\boldsymbol{q}) \end{pmatrix} \left( \begin{pmatrix} -F_{rep}(d_{i,j}) \\ F_{rep}(d_{i,j}) \end{pmatrix} - \boldsymbol{D}_{d,i,j} \begin{pmatrix} \dot{d}_i \\ \dot{d}_j \end{pmatrix} \right). \tag{6.7}$$

### 6.2.2. Joint-Limit Avoidance

The objective is to incorporate joint-limit avoidance into the same computational framework as self-collision avoidance, utilizing a methodology akin to (Dietrich et al., 2012b). Building on the approach of (Eckhoff et al., 2023), we outline a method for avoiding joint limits in revolute joints. While this approach leverages similar concepts as those used in self-collision avoidance (see Section 6.2.1), the relevant quantities (joint angles) already reside in joint space, thereby obviating the need for a mapping between task space and joint space.

We denote the position of the joint as $q_k$, where $k$ represents a specific joint of the robot. For most revolute joints, mechanical constraints define both a lower and upper limit, such that $q_{lim,low} \leq q_k \leq q_{lim,high}$. The distance of a joint position from its respective limits plays a role analogous to the distance between contact points in the framework used for self-collision avoidance.

Since these limits can be different for each joint $k$, it is not beneficial to set a fixed distance to define the zone in which joint limit avoidance is applied. Instead, the borders of the buffer zone are defined based on a parameter $\eta$, which defines a proportion of the total area $q_{k,range}$.

$$q_{k,range} = q_{lim,max} - q_{lim,min} \tag{6.8}$$

$$q_{k,0} = \eta q_{k,range} \tag{6.9}$$

Based on $q_{k,0}$ the upper limit of the lower buffer zone $q_{0,min}$ and the lower limit of the upper buffer zone $q_{0,max}$ can be computed. The following equations concern a single joint, however the subscript $k$ is omitted.

$$q_{0,low} = q_{lim,min} + q_0 \tag{6.10}$$

$$q_{0,high} = q_{lim,max} - q_0 \tag{6.11}$$

Similar to Equation (6.4), but directly in joint space, the repulsive torque can be computed as

$$\tau_{rep,k}(q) = \begin{cases} \frac{\tau_{max}}{q_0^2}(q_{0,low} - q)^2 & \forall q \leq q_{0,low} \\ 0 & \forall q_{0,low} \leq q \leq q_{0,high} \\ \frac{\tau_{max}}{q_0^2}(q - q_{0,high})^2 & \forall q_{0,high} \leq q. \end{cases} \tag{6.12}$$

The local stiffness $K_{d,k} = \frac{\partial \tau_{rep,k}(q)}{\partial q}$ can be computed based on the derivative of the repulsive torque.

Using local stiffness the joint space damping is computed based on *Double Diagonalization* (Albu-Schäffer et al., 2004). To obtain the joints space inertia matrix the *Composite-Rigid-Body Algorithm* described in (Featherstone, 2020) is used. The diagonal inertia element is computed with respect to the entire subtree of the kinematic chain. As a result, it completely captures the inertia associated with any movement around the corresponding joint axis.

For each joint this results in

$$\tau_{jl,k} = \begin{cases} \tau_{rep,k} - D_{d,k}\dot{q} & \text{if } q \leq q_{0,low} \\ 0 & \text{if } q_{0,low} < q < q_{0,high} \\ -(\tau_{rep,k} + D_{d,k}\dot{q}) & \text{if } q_{0,high} \leq q. \end{cases} \tag{6.13}$$

The overall torque command $\boldsymbol{\tau}$ is obtained by assembling the individual elements $\tau_{jl,k}$ into a vector:

$$\boldsymbol{\tau}_{jl} = \left(\tau_{jl,1}\ \tau_{jl,2}\ \cdots\ \tau_{jl,k}\right)^T, \tag{6.14}$$

where $\tau_{jl,k}$ represents the torque contribution from the $k$-th joint.

### 6.2.3. Hierarchy

To achieve a hierarchical controller, differently prioritizing goals, we employ the continuous nullspace projection method proposed in (Dietrich et al., 2012a), which is capable of dealing with dynamic hierarchies and unilateral constraints by designing a suitable redundancy resolution method for the system. This allows the activation of certain constraints, thereby locking some DoFs, only when they become relevant, such as near joint limits or during potential self-collision. In these cases, only the behavior in the critical direction is altered. Further, this approach leads to a continuous transition to ensure the stability of the system. The approach to join limit avoidance based on (Eckhoff et al., 2023) described in the previous section is integrated into the hierarchical framework by (Dietrich et al., 2012a).

Based on (Dietrich et al., 2012a) the nullspace is defined as

$$\boldsymbol{N} = \boldsymbol{I} - \boldsymbol{J}^T\boldsymbol{J}^{\dagger T}, \tag{6.15}$$

with the generalized inverse

$$\boldsymbol{J}^{\dagger} = \boldsymbol{W}^{-1}\boldsymbol{J}^T(\boldsymbol{J}\boldsymbol{W}^{-1}\boldsymbol{J}^T)^{-1}, \tag{6.16}$$

with a weighting matrix $\boldsymbol{W}$. The commonly used Moore-Penrose pseudoinverse corresponds to the case where the weighting matrix is the identity matrix.

This can be used to compute the control torques $\boldsymbol{\tau}_{cmd}$ for including the torques of a secondary task goal $\boldsymbol{\tau}_{sec}$ in the nullspace of a primary task goal $\boldsymbol{\tau}_{prim}$:

$$\boldsymbol{\tau}_{cmd} = \boldsymbol{\tau}_{prim} + \boldsymbol{N}\boldsymbol{\tau}_{sec}. \tag{6.17}$$

By actively shaping the nullspace $\boldsymbol{N}$, the task-relevant directions in which the primary control input, $\boldsymbol{\tau}_{prim}$, operates can be constrained, preventing the secondary control input, $\boldsymbol{\tau}_{sec}$, from affecting the execution of the primary task. This ensures that the secondary task does not interfere with the higher-priority task, establishing a strict task hierarchy. This approach is known as statically consistent nullspace projection (Dietrich et al., 2012a).

As demonstrated in (Dietrich et al., 2012b), applying singular value decomposition (SVD) for Equation (6.15) results in

$$\boldsymbol{N} = \boldsymbol{I} - \boldsymbol{V}\boldsymbol{S}^T\boldsymbol{U}^T(\boldsymbol{V}\boldsymbol{S}^+\boldsymbol{U}^T)^T \tag{6.18}$$

$$= \boldsymbol{I} - \boldsymbol{V}\boldsymbol{S}^T\boldsymbol{U}^T\boldsymbol{U}\boldsymbol{S}^{+T}\boldsymbol{V}^T \tag{6.19}$$

$$= \boldsymbol{I} - \boldsymbol{V}\underbrace{\boldsymbol{S}^T\boldsymbol{S}^{+T}}_{\boldsymbol{A}}\boldsymbol{V}^T, \tag{6.20}$$

with the unitary matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ and the diagonal matrix $\boldsymbol{S}$ containing all singular values $\sigma_i$. The activation matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ can be described as

$$\boldsymbol{A} = diag(a_1, a_2, ..a_m, \mathbf{0}_{1\times(n-m)}), \tag{6.21}$$

with

$$a_i = \begin{cases} 0 & \text{if } \sigma_i < \epsilon \\ 1 & \text{otherwise.} \end{cases} \tag{6.22}$$

When unilateral constraints are considered, the nullspace projector $\boldsymbol{N}$ can be dynamically adapted based on the constraint. For instance, in the case of self-collision avoidance, the secondary task proceeds uninterrupted as long as no collision risk is present. However, upon detecting a potential collision, the nullspace projector is engaged to constrain the DoFs in the direction of the collision. The transition between the inactive and active states of the projector must be smooth to ensure system stability during the shift. To enable a smooth transition we use the methods proposed in Dietrich et al. (2012a).

They define the activator for a diagonal element in $\boldsymbol{A}$ as

$$a_{des} = 1 - N_{des}(z), \tag{6.23}$$

parameterized by variable z which can, for example, indicate the distance between contact pairs.

For the continuous transition Dietrich et al. (2012a) propose a third-order polynomial

$$g(z) = c_1 z^3 + c_2 z^2 + c_3 z + c_4. \tag{6.24}$$

For the case where for $z < z_1$ (e. g., close to collision), the secondary torque should be completely disabled. To ensure a smooth transition between $z_1$ and $z_2$ the desired nullspace projector is defined as follows:

$$N_{des}(z) = \begin{cases} 0 & \text{if } z < z_1 \\ g(z) & \text{if } z_1 \leq z \leq z_2 \\ 1, & \text{otherwise.} \end{cases} \tag{6.25}$$

The parameters $c_1$ - $c_4$ can be computed based on the boundary constraints for a smooth transition

$$g(z_1) = 0,\ g(z_2) = 1,\ \dot{g}(z_1) = 0,\ \dot{g}(z_2) = 0. \tag{6.26}$$

### Nullspace Self-Collision Avoidance

As shown in (Dietrich et al., 2012a) this method can be used to integrate self-collision avoidance into a hierarchical framework. In this case, the parameter $z$ corresponds to the distance between contact pairs. $z_1$ and $z_2$ define the transition zone between completely locking the respective direction and unconstraint movements in regard to the lower priority tasks. However, while the unilateral constrained is defined in task space, the corresponding nullspace projection matrix for each collision pair can be computed as

$$\boldsymbol{N}_{coll,i} = \boldsymbol{I} - \hat{\boldsymbol{J}}_i^T a_{des} \hat{\boldsymbol{J}}_i, \tag{6.27}$$

where $\hat{\boldsymbol{J}}_i = \frac{\boldsymbol{J}_i}{||\boldsymbol{J}_i||}$ is the normalized projected Jacobian in collision direction in the contact point $i$ and $a_{des}$ is defined as described in Equations 6.23 - 6.25.

The overall nullspace $N_{coll}$ projection matrix is computed by multiplying the projection matrices resulting from all considered contact points:

$$\boldsymbol{N}_{coll} = \prod_{i=1}^{n_p} \boldsymbol{N}_{coll,i}. \tag{6.28}$$

### Nullspace Joint Limit Avoidance

In contrast to self-collision avoidance, where constraints are typically formulated in task space, joint limit avoidance is inherently defined in joint space for revolute joints, as considered in this work. Further, adaptions of Equation (6.25) are required to account for both the lower and upper joint limits. Therefore, two transition zones parameterized by $q_{1,\text{low}}, q_{2,\text{low}}, q_{1,\text{high}}, q_{2,\text{high}}$ are introduced.

$$N_{des,jl,k}(q) = \begin{cases} 0 & \text{if } q < q_{1,low} \text{ or } q > q_{1,high} \\ g_{low}(q) & \text{if } q_{1,low} \leq q \leq q_{2,low} \\ g_{high}(q) & \text{if } q_{2,high} \leq q \leq q_{1,high} \\ 1, & \text{otherwise} \end{cases} \tag{6.29}$$

As observable in Figure 6.2, $q_{1,\text{low}}, q_{2,\text{low}}$ denote transition area for the lower bound and $q_{1,\text{high}}, q_{2,\text{high}}$ the transition area of the higher bound.

$q_{lim,min}$ and $q_{lim,max}$ denote the physical limits of the respective joint and $q_{0,low}$ and $q_{0,high}$ (see Section 6.2.2) define where the repulsive potential field is active. $q_{1,\text{low}}, q_{2,\text{low}}$, $q_{1,\text{high}}$ and $q_{2,\text{high}}$ are computed similar to $q_{0,low}$ and $q_{0,high}$ (see Sec-

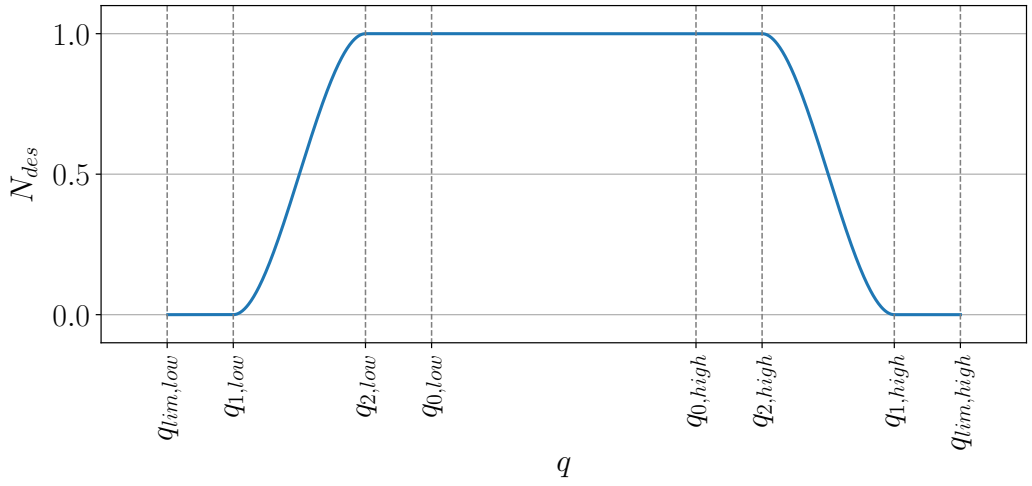



Figure 6.2.: Nullspace projector $N_{des}$ for joint limit avoidance.

tion 6.2.2 - 6.11). Based on the ratios $\eta_{z1}$ and $\eta_{z2}$ with $\eta_{z1} < \eta_{z2}$ they are computed as

$$q_{1,low} = q_{lim,low} + \eta_{z1} q_{k,range} \qquad q_{1,high} = q_{lim,high} - \eta_{z1} q_{k,range} \tag{6.30}$$

$$q_{2,low} = q_{lim,low} + \eta_{z2} q_{k,range} \qquad q_{2,high} = q_{lim,high} - \eta_{z2} q_{k,range}. \tag{6.31}$$

Based on the computed values of $N_{des,jl,k}(q)$ for each joint $k$ the matrix $\boldsymbol{N}_{jl}$ can be constructed as a diagonal matrix

$$\boldsymbol{N}_{jl} = \begin{pmatrix} N_{des,jl,1} & 0 & 0 & \cdots & 0 \\ 0 & N_{des,jl,2} & 0 & \cdots & 0 \\ 0 & 0 & N_{des,jl,k} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & N_{des,jl,K} \end{pmatrix}, \tag{6.32}$$

where $K$ corresponds to the number of actuated joints.

### Overall Hierarchy

Overall the hierarchical control framework can be described by

$$\boldsymbol{\tau}_{cmd} = \underbrace{\boldsymbol{\tau}_g + \boldsymbol{\tau}_{coll} + \boldsymbol{\tau}_{jl}}_{\text{Priority 1}} + \underbrace{\left(\prod_{i=1}^{n_p} \boldsymbol{N}_{coll,i}\right) \boldsymbol{N}_{jl}(\boldsymbol{\tau}_{imp}}_{\text{Priority 2}} + \underbrace{\boldsymbol{N}_{imp}\boldsymbol{\tau}_{js})}_{\text{Priority 3}}. \tag{6.33}$$

Self-collision avoidance and joint-limit avoidance are assigned the highest priority. In the nullspace of priority 1, task space impedance targets are followed as priority 2. Finally, in the nullspace of priority 1 and 2, joint space targets are considered. The meaning of all variables is described in Table 6.1.

To clarify the working mechanisms of this structure we consider two exemplary cases.

- In case 1 the current robot state is far enough from joint limits and self-collision, so that none of the respective potential fields is active. For this case $\boldsymbol{\tau}_{coll} = \boldsymbol{\tau}_{jl} = \mathbf{0}$ and $\boldsymbol{N}_{coll,i} = \boldsymbol{N}_{jl} = 1$, resulting in the overall equation

  $$\boldsymbol{\tau}_{cmd} = \boldsymbol{\tau}_g + \boldsymbol{\tau}_{imp} + \boldsymbol{N}_{imp}\boldsymbol{\tau}_{js}, \tag{6.34}$$

  which corresponds to a standard task space impedance control law.

- To also consider the other extreme, in case 2 the robot is not in a state where it is not close to joint limits but very close to a self-collision. For this case $\boldsymbol{\tau}_{jl} = \mathbf{0}$ and $\boldsymbol{N}_{jl} = \mathbf{1}$ resulting in

$$\boldsymbol{\tau}_{cmd} = \boldsymbol{\tau}_g + \boldsymbol{\tau}_{coll} + \left( \prod_{i=1}^{n_p} \boldsymbol{N}_{coll,i} \right) (\boldsymbol{\tau}_{imp} + \boldsymbol{N}_{imp} \boldsymbol{\tau}_{js}). \tag{6.35}$$

  For the directions of the collision pairs $n_p$ only the gravity compensation torque $\boldsymbol{\tau}_g$ and the self-collision avoidance torque $\boldsymbol{\tau}_{coll}$ remain.

| Symbol | Explanation |
|---|---|
| $\boldsymbol{\tau}_g$ | Gravity compensation torque |
| $\boldsymbol{\tau}_{coll}$ | Self-collision avoidance torque |
| $\boldsymbol{\tau}_{jl}$ | Joint limit avoidance torque |
| $\boldsymbol{\tau}_{imp}$ | Torque of impedance control |
| $\boldsymbol{\tau}_{js}$ | Torque of joint space control |
| $\boldsymbol{N}_{coll}$ | Self-collision nullspace projector |
| $\boldsymbol{N}_{js}$ | Joint limit nullspace projector |
| $\boldsymbol{N}_{imp}$ | Impedance control nullspace projector |

Table 6.1.: Explanation of symbols used in the control law equations

## 6.3. Category-Based Controller

The bimanual categories defined by the bimanual manipulation taxonomy serve as a basis for selecting control strategies that enforce category-specific constraints. Our goal is to develop an approach suitable for robots operating in close proximity to humans, such as assistive robots in household environments. This necessitates a compliant and reactive control framework capable of adapting to external disturbances while maintaining a robust task execution. Addressing this challenge requires balancing precise execution with responsiveness to unexpected interactions. To achieve this, we introduce a category-based controller that leverages the constraints associated with a category. The concepts discussed in this section were previously presented in Krebs and Asfour (2024).

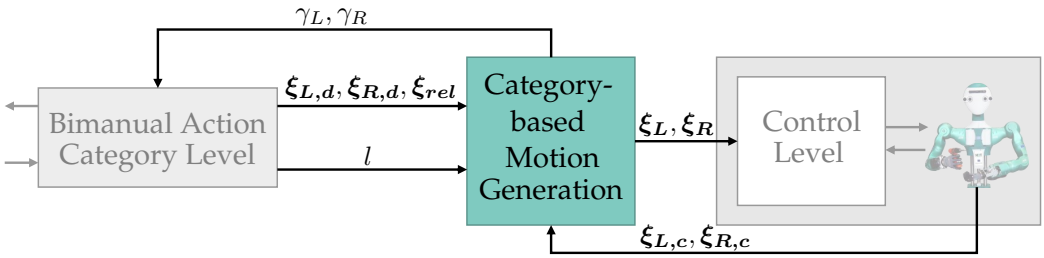



Figure 6.3.: Framework for category-based robot control. The meaning of variables is given in Table 6.2.

Table 6.2.: Global variables for the descriptions of task space goals.

| Variable | Meaning |
|---|---|
| $\xi_L, \xi_R$ | computed target hand pose |
| $\xi_{L,d}, \xi_{R,d}$ | desired hand pose for the current state |
| $\xi_{rel}$ | desired relative pose (right hand in frame of left) |
| $\xi_{L,c}, \xi_{R,c}$ | actual measured hand pose for the current state |

### 6.3.1. Methods

We introduce a framework that ensures category-specific spatial and temporal constraints are met while being reactive to external perturbations. This section focuses on single-category bimanual segments, excluding transitions between bimanual categories. The framework operates with minimal environmental knowledge and does not rely on extensive force-torque sensing.

This functionality corresponds to the **Category-Based Motion Generation** level in the overall task model (Figure 6.1). As illustrated in Figure 6.3 (with variables explained in Table 6.2), the module receives desired relative and absolute hand poses along with the category label from higher-level components. Using this input and the robot's current end-effector poses, it computes task space targets, which are then converted into control commands via an impedance controller to ensure precise execution.

The category-specific behavior is designed based on the constraints imposed on each hand, as detailed in Section 3.3.1. This approach considers the categories outlined in Table 3.2 and parameterizes their trajectories while accounting for the most relevant constraints.

**Spatial constraints:** The task model provides the currently desired poses of both hands as input. Using this information, reactive target poses for each

hand are formulated based on the specified category. Poses are represented as homogeneous 4×4 matrices, $\boldsymbol{\xi}$, defined relative to the global coordinate system. Each hand's target pose can be defined either directly using its global target pose

$$\boldsymbol{\xi}_L = \boldsymbol{\xi}_{L,d} \qquad \boldsymbol{\xi}_R = \boldsymbol{\xi}_{R,d} \tag{6.36}$$

or indirectly based on the current pose of the other hand and the specified relative trajectory. The relative trajectory can be derived from the initial, unaltered trajectories. Relative poses, denoted as $\boldsymbol{\xi}_{rel}$, represent the pose of the right hand relative to the left hand. Consequently, global target poses based on the relative pose are computed as:

$$\boldsymbol{\xi}_R = \boldsymbol{\xi}_{L,c} \cdot \boldsymbol{\xi}_{rel} \qquad \boldsymbol{\xi}_L = \boldsymbol{\xi}_{R,c} \cdot \boldsymbol{\xi}_{rel}^{-1}. \tag{6.37}$$

For different categories, target poses are combined in various ways, as outlined in Table 3.2.

- In the *asym_r* case, the left hand acts as the *leader*, with its pose defined as $\boldsymbol{\xi}_L = \boldsymbol{\xi}_{L,d}$, while the right hand takes on the role of the *follower*, defined by $\boldsymbol{\xi}_R = \boldsymbol{\xi}_{L,c} \cdot \boldsymbol{\xi}_{\text{rel}}$. Conversely, for the *uncoord_bi* case, both hands are described globally as $\boldsymbol{\xi}_L = \boldsymbol{\xi}_{L,d}$ and $\boldsymbol{\xi}_R = \boldsymbol{\xi}_{R,d}$.
- The most complex case is *sym*, where relative poses are prioritized, but global poses are followed when possible. To achieve this, the framework alternates between the two asymmetric categories, selecting the perturbed hand as the leader (i. e., the non-dominant hand). Switching the roles of leader and follower is triggered only when the error calculated for the follower exceeds the leader's error by a predefined value.

**Temporal constraints** are described through the Petri net templates shown in Figures 3.4–3.7. This implementation enables uncoordinated motions to advance independently while ensuring synchronization when needed. In the event of a perturbation, the corresponding trajectory progression is paused. This behavior aligns with the concept of *phase stopping* of movement primitives.

- For tightly coupled categories, start synchronization is implemented, ensuring that a segment begins only when both hands are ready, i. e., when the preceding segment is completed for both hands.

- In symmetric cases, hand motions are described within a shared temporal system, meaning that a perturbation to one hand triggers *phase stopping* for the other hand, thereby achieving implicit goal synchronization.
- In asymmetric cases, phase stopping is applied to both hands only if the follower experiences a perturbation.

Threshold values must be adjusted based on the task-space controller and its tracking accuracy.

## 6.3.2. Evaluation

The objective of this section is to validate that the designed control modes satisfy the constraints associated with each category. To this end, we conduct experiments in both simulation and real-robot settings.

### Simulation Experiments

To quantitatively assess the effectiveness of the proposed method in satisfying spatial constraints, we conduct experiments using the humanoid robot ARMAR-6 (Asfour et al., 2019) within a MuJoCo simulation environment (Todorov et al., 2012). Simulations provide a controlled setting for applying perturbations in a reproducible manner, thereby facilitating the generation of comparable results.

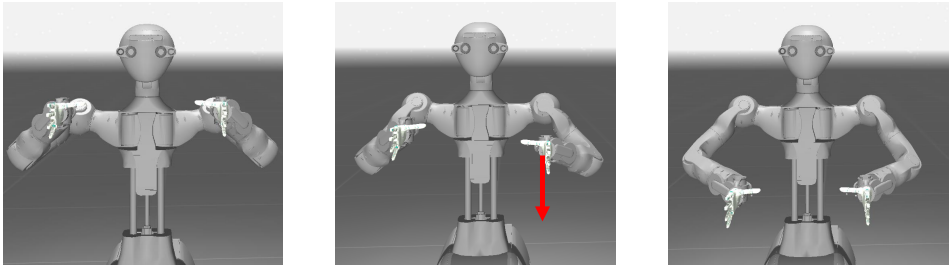


Figure 6.4.: Robot experiments in a MuJoCo simulation. The hands move downwards with a constant offset. For $t \in [3, 7]$ seconds, a constant force of 100 N in $z$-direction is applied on the left wrist.
Sources: Krebs and Asfour (2024) © 2024 IEEE

As illustrated in Figure 6.4, the robot executes a simple downward motion with a constant offset between its hands. The motion spans a total duration of 10 seconds. During the interval $t \in [3, 7]$ seconds, a perturbation force of 100 N

is applied to the left arm in the negative z-direction, as indicated by the red arrow.

We configure the controller based on the defined bimanual control categories. Figure 6.5 presents the position data for three distinct scenarios: *uncoord_bi*, *asym_l* and *sym*. In the *uncoord_bi* case, the motion of the right hand remains unaffected by the perturbation applied to the left hand.

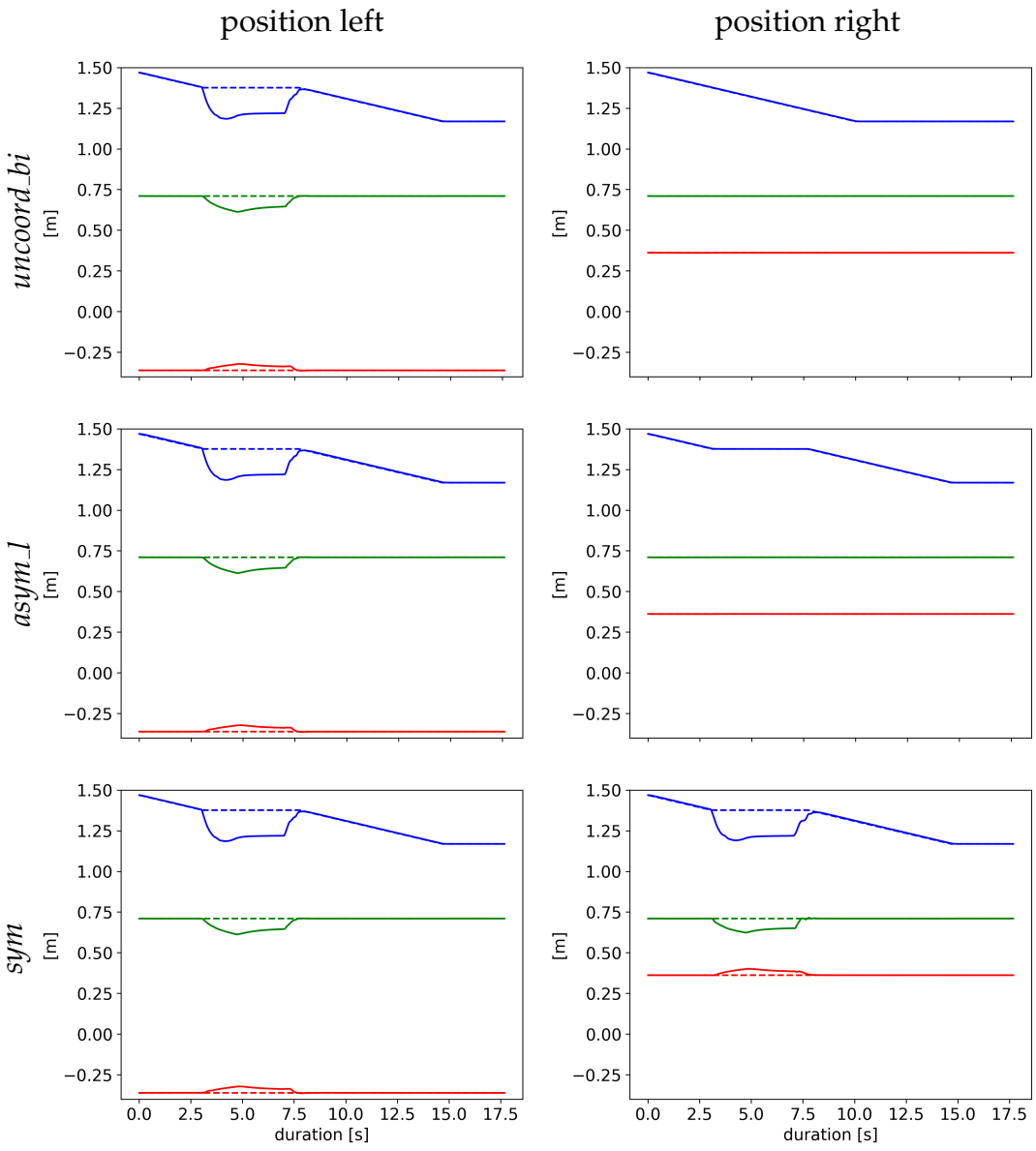



Figure 6.5.: Positions for both hands in the case of a perturbation of the left hand for $t \in [3, 7]$ s. Color legend: actual x (—), actual y (—), actual z (—), computed x (- -), computed y (- -), computed z (- -).
Sources: Krebs and Asfour (2024) © 2024 IEEE

In the *asym_l* scenario, the right hand does not adapt its spatial trajectory but halts its movement. Lastly, in the *sym* case, the right hand not only spatially

adjusts but also pauses its trajectory execution while the left hand is subjected to perturbation. Figure 6.5 additionally illustrates perturbations in both the $x$- and $y$-directions, as the force is applied at the wrist while the position is measured at the tool center point (TCP), located near the center of the robot's palm.

Each experiment is conducted five times using different simulation seed values. We measure both the absolute and relative positional errors of the robot's hands under unperturbed and perturbed conditions. The errors are calculated as the translational discrepancies between two frames. Specifically, the absolute error is calculated as the Euclidean difference between $\boldsymbol{\xi_{R,d}}$ and $\boldsymbol{\xi_{R,c}}$, while the relative error is determined as the Euclidean difference between $\boldsymbol{\xi_{L,c}^{-1}\xi_{R,c}}$ and $\boldsymbol{\xi_{rel}}$.

The average errors are presented in Table 6.3, with the most relevant errors (according to Table 3.2) for each category highlighted in gray. For the unperturbed case, the errors are in general rather small.

Table 6.3.: Average errors in mm of 5 repetitions performed in MuJoCo simulation.

| | unperturbed | | | perturbed | | |
|---|---|---|---|---|---|---|
| | $e_{a,L}$ | $e_{a,R}$ | $e_{rel}$ | $e_{a,L}$ | $e_{a,R}$ | $e_{rel}$ |
| *uncoord_bim* | 1.33 | 1.33 | 0.63 | 43.62 | 0.99 | 68.49 |
| *asym_l* | 2.49 | 1.33 | 1.33 | 44.32 | 1.00 | 43.54 |
| *asym_r* | 1.33 | 2.50 | 1.25 | 43.38 | 63.98 | 3.02 |
| *sym* | 1.33 | 2.50 | 1.25 | 43.38 | 43.87 | 3.02 |

The overall highest errors of larger than 2 mm only occur for the less relevant errors. In the perturbed case across all categories, the absolute error of the left hand is notably high due to the direct perturbation, which cannot be controlled by our approach. However, examining the highlighted absolute errors for the right hand and the relative errors reveals that they remain relatively small, indicating that the relevant constraints are successfully enforced. The only exception is in the *asym_l* case, where the relative error is high. This is expected because the conflicting nature of $e_{a,R}$ and $e_{rel}$ arises when the left hand is perturbed, leading to this behavior.

### Real Robot Experiments

We designed four representative household tasks to demonstrate the feasibility of the proposed approach in real-world scenarios, using the ARMAR-6 robot (Asfour et al., 2019). These scenarios, depicted in Figures 6.6-6.8, represent different control categories. Videos of the robot experiments were published with the original publication[2] (Krebs and Asfour, 2024).

**Unimanual** In this scenario, the robot operates in a kitchen, holding a box of tea bags in its right hand while its left hand obstructs a drawer. A human can move the robot's left hand aside to access the drawer, demonstrating how the exploitation of undefined aspects of the task formulation can be leveraged for secondary goals.

**Uncoordinated** The robot holds a box of tea in each hand and places them on a table. As shown in Figure 6.6, the left hand continues its placing motion even if the right hand is stopped. The deviation between the actual and desired hand positions during perturbation matches the predefined threshold for detecting external disturbances. In the uncoordinated case, the motion of the left hand is not unnecessarily prolonged.

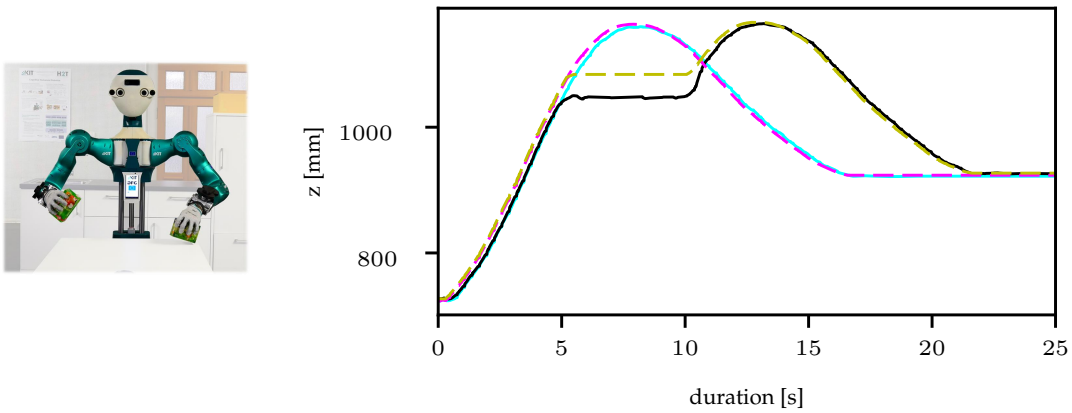



Figure 6.6.: Perturbation of the right hand at $t = 5\,\mathrm{s}$ for an *uncoord_bi* reaching motion. Color legend: actual right (—), desired right (- -), actual left (—), desired left (- -). Based on Krebs and Asfour (2024).

[2] https://www.youtube.com/watch?v=MQjPv_ELvtE

**Asymmetric** In this scenario, the robot holds a bowl with its left hand while stirring with a ladle in its right hand. The stirring action continues relative to the pot even if the bowl-holding hand is disturbed (see Figure 6.7), illustrating the robot's capacity to maintain the most relevant task constraints despite external perturbation.

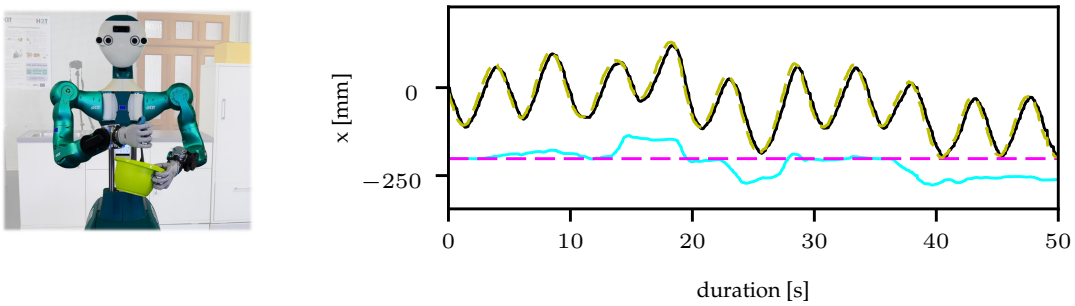



Figure 6.7.: Perturbation of the left hand during an *asym_r*stirring motion. Color legend: actual right (—), desired right (- -), actual left (—), desired left (- -). Based on Krebs and Asfour (2024).

**Symmetric** The robot grasps a long tube and moves it upwards. During the motion, the left hand is perturbed first, followed by the right hand (see Figure 6.8. Since the controller is configured with the right hand following the left, the right hand immediately adjusts when the left is perturbed. If the difference between the desired and actual poses exceeds a set threshold, the motion is paused until the perturbation is resolved. When the right hand is subsequently disturbed, the system switches to having the left hand follow the right, again pausing the motion if a critical threshold is reached.

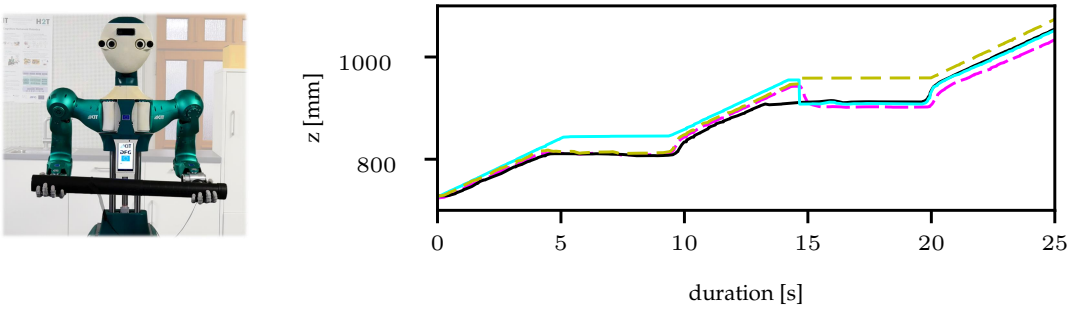



Figure 6.8.: Perturbation of the left hand at $t = 4\,\mathrm{s}$ and of the right at $t = 13\,\mathrm{s}$ for *sym*. Color legend: actual right (—), desired right (- -), actual left (—), desired left (- -). Based on Krebs and Asfour (2024).

Overall, the robot exhibited expected behaviors for each task: Relaxed coordination constraints allowed the robot to incorporate additional considerations,

such as moving out of the way in the *unimanual* scenario and minimizing delays in the *uncoordinated* task. In the *asymmetric* case, the task could proceed despite perturbations, while in the *symmetric* case, the robot successfully maintained a secure grip on the tube.

## 6.4. Transitions Between Bimanual Action Categories

This section explores how transitions between bimanual action categories can be realized by emphasizing the conditions governing these transitions while adhering to spatial and temporal constraints. To illustrate the practical implementation of the transition behavior, we present exemplary cases demonstrating how different bimanual categories interact within a sequence.

The transition behavior is governed by the temporal constraints outlined in Section 3.3.2 and illustrated in Figures 3.4-3.7. When the *completed* place of the Petri net templates is reached for a bimanual action category $BAC_n$, a transition occurs to the *ready* place of the subsequent bimanual action category $BAC_{n+1}$. The process then either continues immediately or, in the case of *asym_r*, *asym_l*, and *sym*, may wait for the other hand.

The representation of bimanual categories as a sequence of Petri nets does not explicitly account for the spatial behavior in cases where one hand remains in the previous bimanual action category while the other has already transitioned to the next. This results in a limited number of possible transition combinations. Notably, this situation does not arise when transitioning into *asym_r* or *asym_l*, as both hands must synchronize upon entering these segments, as illustrated in Figure 3.6. Likewise, due to the inherent input and output synchronization, this issue does not occur for *sym*.

Consequently, only specific transitions allow for overlapping transitions, where one hand shifts to the next category while the other remains temporarily in the previous one. The eligible transitions for such behavior are:

- *uni_l* →*uni_r* (or *uni_r* →*uni_l*)
- *uni_r*/*uni_l* →*uncoord_bi*
- *uncoord_bi* →*uni_r*/*uni_l*
- *asym_r* →*uni_r* (or *asym_l* →*uni_l*)

- *asym_r* →*uncoord_bi* (or *asym_l* →*uncoord_bi*)

In transition periods where one hand remains in the previous category segment while the other enters the next, the overall category is *uncoord_bi* for all those cases. Another case occurs when, due to these transitions, *uncoord_bi* motions begin earlier for one hand, causing it to finish earlier as well. Once only one hand remains active, the category switches to *uni_r* or *uni_l*, depending on the active hand. These temporal category assignments align with the spatial constraints defined in Table 3.3 and are discussed based on illustrative examples in the next section.

**Example Cases** We investigate the role of temporal constraints and the associated transition behavior between two bimanual action categories. Therefore, we take a closer look at some of the possible combinations listed in the previous section and point out their potential effects on the task execution.

At its core, the task begins with the initial bimanual action category $BAC_1$. After defining the respective hand goals, the system assesses its readiness for transitioning to the next bimanual action category. A transition becomes possible when at least one hand has already *completed* its current action. In a *unimanual* scenario, for example, the *inactive* hand is immediately available for a subsequent action. Once the system is deemed ready, the next bimanual action category is requested from the Task Level. Upon reaching the appropriate transition conditions, the system updates its internal state according to $BAC_2$.

**Transitions without external perturbations:** We first conduct an exemplary analysis of the transition from *uni_r* to *uni_l*. In this case, we assume that *uni_l* initiates only after a predefined temporal delay following *uni_r* (e.g., Action 2 begins 2 s *after* Action 1 started). This demonstrates the model's ability to incorporate such constraints seamlessly. The resulting spatial constraints, as defined in Figure 3.2, are illustrated in Figure 6.9.

The sequence starts with *uni_r*, during which a *global* trajectory for the right hand is initialized, while the left hand remains in an *inactive* state. Once the predefined temporal offset elapses, *uni_l* begins. At this point, the system verifies that the left hand is ready for execution and subsequently initiates its *global* trajectory. Since the right hand completes its action earlier than the left hand, it transitions to an *inactive* state towards the end of the sequence.

Based on this representation of spatial constraints, the effective bimanual categories at different points in time can be inferred, as depicted in Figure 6.10,

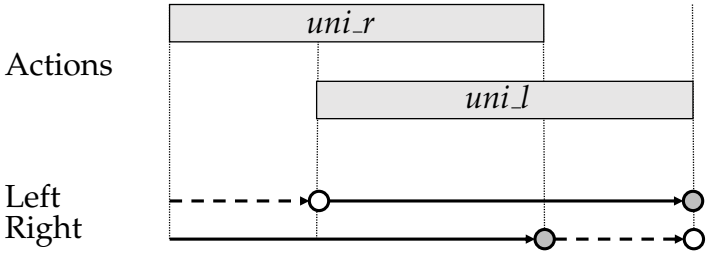



Figure 6.9.: Transition behavior from *uni_r* to *uni_l*. The top section depicts the executed actions, the bottom section represents the spatial task representation.

following the principles outlined in Section 3.3.1 (see in particular Table 3.3). This results in *uni_r* at the beginning (when only the right hand is active), *uncoord_bi* in the middle (when both actions overlap), and *uni_l* towards the end (when only the left hand is active).

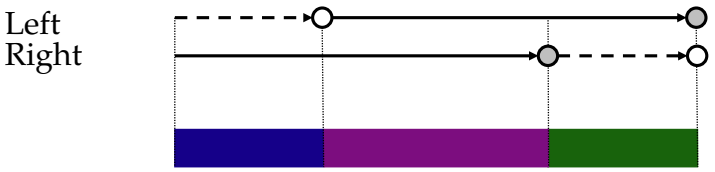



Figure 6.10.: Transition behavior from *uni_r* to *uni_l*. The top section depicts the spatial task representation, the bottom section represents the emerging bimanual categories.
Categories: *uni_r*, *uni_l*, *uncoord_bi*.

An overall representation of this example is provided in Figure 6.11. The right side of Figure 6.11 illustrates an additional scenario without perturbations, specifically the transition from *uni_r* to *uncoord_bi*. In this instance, the left hand executes the motion defined by the *uncoord_bi* category while the right hand still follows the trajectory of *uni_r*, transitioning to *uncoord_bi* only after completing its initial motion.

Figure 6.11 clearly demonstrates that, in both cases, the task execution time is significantly reduced compared to a strictly sequential approach, where the second action starts only after the first has ended. This acceleration is achieved while maintaining the constraints imposed by the respective bimanual categories.

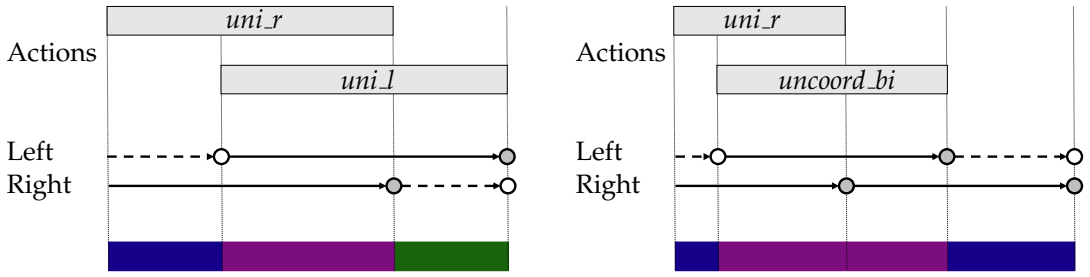



Figure 6.11.: Illustrative examples of transition behavior between two bimanual categories in the absence of perturbations. The top section depicts the executed actions, the middle section represents the spatial task representation, and the bottom section shows the resulting bimanual categories. Categories: *uni_r*, *uni_l*, *uncoord_bi*.

**Transitions influenced by external perturbations:** In a second set of example cases, we examine transitions influenced by external perturbations, specifically from *uncoord_bi* to *uni_r*, and from *asym_r* to *uni_r*. These cases are represented in Figure 6.12 in the same manner as the previous examples, with an additional red section indicating perturbations affecting the left hand and preventing its motion.

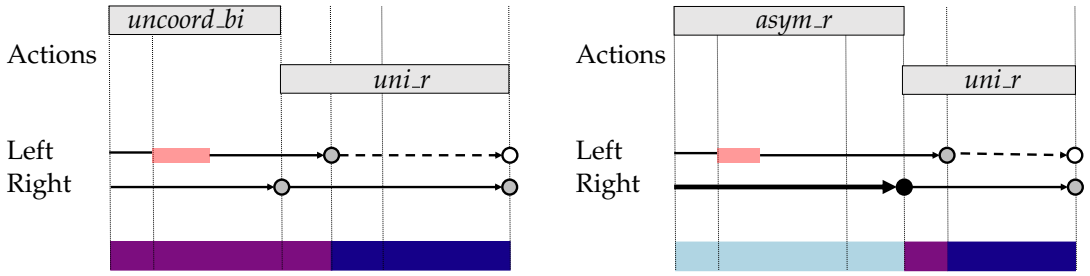



Figure 6.12.: Illustrative examples of transition behavior between two bimanual categories with perturbations (indicated in red). The top section depicts the executed actions, the middle section represents the spatial task representation, and the bottom section shows the resulting bimanual categories. Categories: *uni_r*, *uni_l*, *uncoord_bi*, *asym_r*.

In a strictly sequential execution, the second action can only commence once the first action has been fully completed, resulting in an overall task delay due to the induced perturbation. In contrast, the examples demonstrate that, in certain cases, more flexible sequencing approaches enable task ex-

ecution without prolongation, even in the presence of delaying perturbations.

Since the relative timing and therefore relative poses of the arms might change compared to the demonstration, the consideration of self-collision avoidance is essential. This is currently ensured by the previously introduced controller (see Section 6.2).

## 6.5. Summary

This chapter presents an approach for incorporating bimanual categories from the proposed taxonomy into a comprehensive task model suitable for deployment on humanoid robots.

We begin by detailing the representation of the task model as well as the hierarchical control architecture with its layers and their interactions. Subsequently, we adopt a bottom-up perspective to examine key aspects of these layers. First, we describe the underlying task-space impedance controller, which hierarchically accounts for self-collision avoidance and joint limit constraints. Next, we demonstrate how spatial and temporal constraints can be integrated at the level of individual categories to enable reactive behavior that preserves these constraints even in the presence of external perturbations—an essential feature for the successful execution of bimanual tasks. Finally, we investigate how individual bimanual action categories can be combined by enabling transitions that adhere to the previously defined temporal constraints. Through illustrative examples, we highlight the benefits of this approach, particularly its positive impact on the overall execution time of a task.

# CHAPTER 7

# Conclusion

This thesis contributes to the advancement of bimanual robotic manipulation by introducing a novel taxonomy that categorizes bimanual actions based on coordination constraints. The taxonomy serves as a foundation for comprehensive task representations and facilitates the selection of appropriate control strategies. The work demonstrates how the taxonomy's categories can be derived from motion data and automatically segmented using human motion recordings. Furthermore, the thesis illustrates how representing tasks as a sequence of bimanual categories enables the enforcement of suitable coordination constraints during task execution. To achieve this, a reactive control system is implemented, enabling adaptive and compliant execution of bimanual tasks. These contributions provide a foundation for improving both learning bimanual manipulation tasks from human demonstrations and their execution.

## 7.1. Scientific Contributions

This section summarizes the four main chapters of this thesis and their respective scientific contributions.

**Bimanual Manipulation Taxonomy:** Chapter 3 introduced a novel taxonomy for bimanual manipulation in robotics, grounded in the principles of coordination, interaction, and hand roles. Inspired by research in robotics, neuroscience,

and rehabilitation, this taxonomy addresses the complexity of bimanual manipulation and informs the development of effective control strategies for humanoid robots. The taxonomy and its categories have been systematically described, and the associated temporal and spatial constraints have been formalized. The taxonomy was first published in Krebs and Asfour (2022), with the constraint formalization presented in Krebs and Asfour (2024).

**Datasets of Human Bimanual Manipulation:** Programming by Demonstration (PbD) for bimanual robotic manipulation relies on human manipulation datasets. This thesis contributes to this field by providing two comprehensive datasets, detailed in Chapter 4. The *KIT Bimanual Manipulation Dataset* offers a multimodal collection of bimanual manipulation actions in kitchen and household settings, including high-accuracy motion capture data. It captures variations in object locations, object types, and hand coordination strategies. Additionally, this thesis extended the existing *KIT Bimanual Actions Dataset*, which provides RGB-D data from a third-person perspective. Two new tasks were recorded to include actions where both hands manipulate a common object, increasing the diversity of bimanual coordination strategies.

Both datasets are publicly available, including raw data, extracted features, and segmentation labels. The KIT Bimanual Manipulation Dataset was published in Krebs and Meixner et al. (2021), and the extension of the KIT Bimanual Actions Dataset in Krebs and Leven et al. (2023).

**Recognition of Bimanual Categories in Human Demonstrations:** Chapter 5 presented methods for the simultaneous classification and segmentation of bimanual categories in human motion data, using motion capture data from the *KIT Bimanual Manipulation Dataset* and single-view RGB-D data from the extended *KIT Bimanual Actions Dataset*. A rule-based approach proved effective for precise motion capture data but performed poorly on features extracted from RGB-D data. To address this, we proposed a learning-based approach using Graph Neural Networks (GNNs), which improved recognition performance across both datasets. Using the proposed methods, we successfully recognized bimanual categories in human demonstrations.

The rule-based approach for motion capture data was published in Krebs and Asfour (2022), while the RGB-D recognition using rule-based and learning-based methods was presented in Krebs and Leven et al. (2023).

**Leveraging Bimanual Categories for Robot Control** In Chapter 6, the application of previously introduced constraints for bimanual manipulation, derived from bimanual categories, to humanoid robotics and their integration into a comprehensive task model was demonstrated. To achieve this, the underlying impedance controller was first described, incorporating joint limit avoidance and self-collision avoidance. Subsequently, the integration of category-specific constraints into an action associated with a particular category was illustrated, and the resulting behavior was validated through real-robot experiments. Finally, a conceptual discussion on the application of these constraints to a sequence of bimanual manipulation actions was provided.

The results presented in Section 6.3, including the description of the desired robot behavior within specific categories and the corresponding experimental validation, have been published in Krebs and Asfour (2024).

## 7.2. Discussion and Future Work

This thesis has introduced the *Bimanual Manipulation Taxonomy* and demonstrated how the resulting categories can be recognized from human demonstrations and reproduced on humanoid robots. As a key step in this direction, it establishes a foundation for further developments. This section outlines potential research directions and possible extensions to enhance and refine the proposed approach.

**Correlations Between Grasping and Bimanual Manipulation** In an exploratory study presented in Haas and Endrikat et al. (2024), we investigated correlations between grasp types and bimanual coordination patterns in a small video dataset of realistic housekeeping activities. The findings indicate, among other insights, a correlation between *tightly coupled symmetric* bimanual coordination and *power grasps*. Future research could extend this analysis by utilizing a larger, higher-quality dataset to further examine the intricate relationship between unimanual grasp types and bimanual coordination. Additionally, identifying such correlations could enhance the recognition of grasps and manipulation strategies, supporting activity recognition and informing the design of robotic systems optimized for specific manipulation tasks.

**Consideration of Force Constraints** While this thesis focuses on the formulation of temporal and spatial constraints between the hands, force constraints also play a crucial role, particularly in contact-rich interactions. These constraints are relevant for all bimanual actions classified as *tightly coupled*. Their importance spans from the integration of existing approaches for *symmetrical* tasks, such as non-prehensile object transport (Gao et al., 2018b), to more task-specific operations like food scooping (Pais Ureche and Billard, 2018) or manual assembly tasks, as seen in watchmaking (Yao et al., 2021). To ensure accurate reproduction on humanoid robots, methods for extracting these constraints, ideally through visual perception, along with suitable control strategies for their execution, should be developed.

**Integration of Bimanual Categories in Learning Frameworks** The approach presented in this thesis explicitly incorporates constraints between the hands based on the extracted bimanual action categories. While this provides a data-efficient framework, it cannot fully match the performance of task-specific methods. To overcome this limitation, future work could explore integrating the proposed approach with reinforcement learning. Specifically, the extracted categories could serve as an initial task hypothesis or be incorporated into the reward function. This integration would enable task-specific fine-tuning while simultaneously addressing additional task constraints beyond bimanual coordination.

**Symbolic Planning Using Bimanual Categories** While this thesis does not explicitly address symbolic planning, the findings demonstrate that integrating symbolic categories for bimanual constraints into the control strategy enhances both robustness and execution efficiency. This suggests that incorporating constraint defined by the bimanual categories into higher-level symbolic planning could further improve task planning for bimanual tasks. In this context, such coordination categories could play a crucial role in task allocation and role assignment, enabling more efficient and adaptable collaboration between multiple robot arms. Future work should explore how the defined bimanual categories can be taken into account in planning domain description.

# Appendix

## A. Additional Evaluations on the Recognition of Bimanual Categories

This section contains some additional information on Chapter 5. In Table A1 the parameters are specified that are used in the rule-based approach.

Table A1.: Parameters used for the rule-based approach. The exact utilization of the parameters can be examined through the implementation[1].

| Parameter | Motion Capture Data | RGB-D Data |
|---|---|---|
| velocity threshold | 300 mm/s | 300 mm/s |
| offset threshold | 1 mm | 13 mm |
| defragmentation threshold | 0.1 s | 0.33 s |
| defragmentation threshold min | 0.02 s | 0.17 s |
| model padding | 8 mm | - |
| model padding min | 3 mm | - |

Table A2 to Table A5 show the classification metrics for the individual subject. The tables show that there is no major difference between subjects.

[1] https://git.h2t.iar.kit.edu/sw/bimanual-category-classification

Table A2.: Classification results using the rule-based approach on motion capture data for individual subjects.

| | $F_1$-score | | |
|---|---|---|---|
| | Micro | Macro | Weighted |
| Subject 1 | 0.81 | 0.71 | 0.80 |
| Subject 2 | 0.84 | 0.77 | 0.84 |
| Total | 0.83 | 0.74 | 0.83 |

Table A3.: Classification results using the GNN-based approach on motion capture data for individual subjects.

| | $F_1$-score | | |
|---|---|---|---|
| | Micro | Macro | Weighted |
| Subject 1 | 0.91 | 0.84 | 0.91 |
| Subject 2 | 0.92 | 0.85 | 0.92 |
| Total | 0.92 | 0.85 | 0.92 |

Table A4.: GNN-based classification results for RGB-D data for individual subjects. Each subjects results correspond to the combination of one subject in the orginal *Bimacs* dataset and one subject of its extension.

| | $F_1$-score | | |
|---|---|---|---|
| | Micro | Macro | Weighted |
| Subject 1 | 0.76 | 0.69 | 0.76 |
| Subject 2 | 0.73 | 0.71 | 0.73 |
| Subject 3 | 0.71 | 0.68 | 0.72 |
| Subject 4 | 0.74 | 0.73 | 0.73 |
| Subject 5 | 0.70 | 0.70 | 0.71 |
| Subject 6 | 0.70 | 0.69 | 0.71 |
| Total | 0.72 | 0.70 | 0.72 |

Further, we provide some additional data for the evaluation of the GNN-based approach. Figure A1 shows the confusion matrix of the larger GNN for motion capture data which was not depicted in Chapter 5 due to its similarity with the smaller model (see Figure 5.7).

Table A5.: Rule-based classification results for RGB-D data for individual subjects. Each subjects results correspond to the combination of one subject in the original *Bimacs* dataset and one subject of its extension.

| | $F_1$-score | | |
|---|---|---|---|
| | Micro | Macro | Weighted |
| Subject 1 | 0.45 | 0.36 | 0.47 |
| Subject 2 | 0.46 | 0.39 | 0.47 |
| Subject 3 | 0.52 | 0.42 | 0.54 |
| Subject 4 | 0.47 | 0.40 | 0.47 |
| Subject 5 | 0.45 | 0.39 | 0.46 |
| Subject 6 | 0.54 | 0.45 | 0.55 |
| Total | 0.48 | 0.41 | 0.49 |

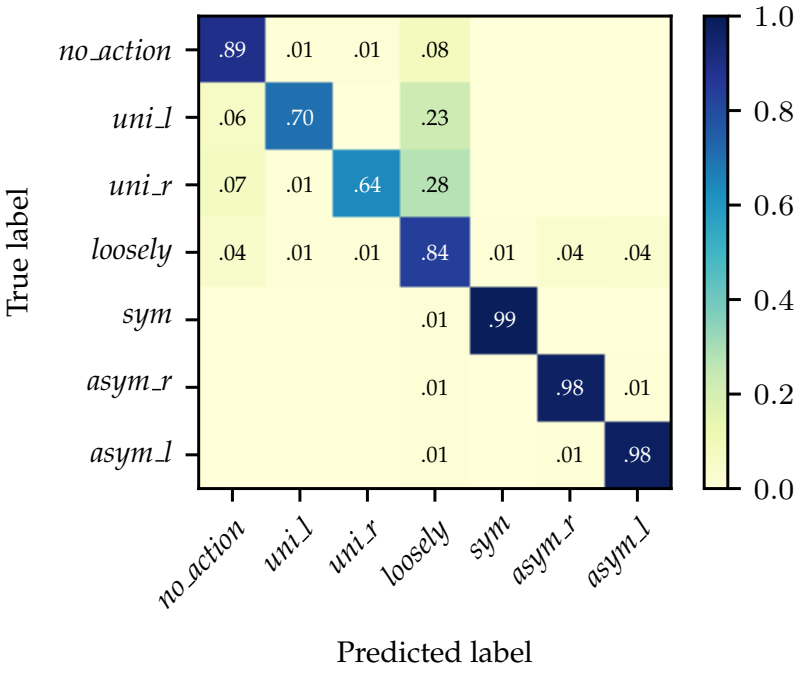



Figure A1.: Normalized confusion matrix using the GNN-based approach with the larger model for motion capture data.

Lastly, we provide additional data for the GNN-based approach on motion capture data. Table A6 shows the classification metrics for the GNNs without object knowledge and the ablation study results as weighted average metrics in Table A7.

Table A6.: Metrics of the GNN-based evaluated on RGB-D data without object knowledge.

| Category | Precision | Recall | $F_1$-score |
|---|---|---|---|
| *no_action* | 0.53 | 0.68 | 0.60 |
| *uni_l* | 0.39 | 0.36 | 0.37 |
| *uni_r* | 0.36 | 0.41 | 0.38 |
| *loosely* | 0.70 | 0.69 | 0.69 |
| *sym* | 0.76 | 0.74 | 0.75 |
| *asym_r* | 0.54 | 0.61 | 0.58 |
| *asym_l* | 0.63 | 0.46 | 0.53 |
| Micro avg. | 0.62 | 0.62 | 0.62 |
| Macro avg. | 0.56 | 0.56 | 0.56 |
| Weigh. avg. | 0.63 | 0.62 | 0.62 |

Table A7.: Ablation study comparing the weighted average metrics

| Training data | | | Results | | |
|---|---|---|---|---|---|
| Spatial Relations* | Temporal Edges | Object Knowledge | Precision | Recall | $F_1$-score |
| ✓ | ✓ | ✓ | 0.73 | 0.72 | 0.72 |
| | ✓ | ✓ | 0.62 | 0.60 | 0.60 |
| ✓ | | ✓ | 0.70 | 0.68 | 0.68 |
| | | ✓ | 0.58 | 0.53 | 0.53 |
| ✓ | ✓ | | 0.63 | 0.62 | 0.62 |

*In case of no spatial relations only contact relations are considered.

# B. Related Work Control Framework

In the following section, we introduce general work for incorporating self-collision avoidance and joint-limit avoidance into the control framework. Rather than focusing solely on scenarios where both robotic arms are utilized, we emphasize methods based on torque control.

## B.1. Self-Collision Avoidance

We review key approaches to self-collision avoidance, including optimization-based methods - often combined with neural networks - and artificial potential fields.

### Optimization- and Learning-based Approaches

Quiroz-Omaña and Adorno (2019) propose a self-collision avoidance method using Vector Field Inequalities (VFI) for both velocity and torque control. The robot is modeled using a combination of spheres and cylinders, with a line-static-line constraint based on Plücker lines to prevent collisions between the torso and the arm. This approach integrates self-collision avoidance, joint limit avoidance, and static environment collision avoidance, utilizing optimization techniques such as Quadratic Programming (QP) and Constrained Quadratic Programming (CQP). However, a key limitation is the lack of a hierarchical structure, as all tasks are treated as constraints. In Ren et al. (2024), optimization is combined with a neural network to learn the self-collision boundaries, utilizing Hierarchical Quadratic Programming (H-QP) for multi-objective optimization. Tasks are incorporated as constraints within the optimization problem, akin to the approach in Quiroz-Omaña and Adorno (2019). Both methods achieve real-time calculation times around 6 ms, which fall short of the requirements for a 1 kHz control cycle.

Koptev et al. (2021) model the self-collision boundaries for the humanoid robot iCub as smooth functions in joint space using a neural network and a Support Vector Machine (SVM). These learned boundary functions are integrated into the controller as constraints within a QP-based Inverse Kinematics solver, enabling real-time generation of collision-free motions. However, neural networks require suitable training data and can be challenging to interpret, which poses drawbacks for safety-critical applications such as human-

robot interaction. Furthermore, they are typically customized for a specific robot.

One other method for collision avoidance is the usage of Control Barrier Functions (CBFs) (Ames et al., 2019). This allows to mathematically proof the absence of collisions. However, since they rely on optimization, they can conflict with hard real-time constraints. Khazoom et al. (2022) propose a whole-body controller based on CBFs combined with the QP framework. Despite using an iterative approach, the solve time of the quadratic programs is very fast compared to other optimization-based approaches.

#### Artificial Potential Fields

Artificial Potential Fields (APFs) (Khatib, 1986) are widely employed for collision avoidance. Dietrich et al. (2012b) employ APFs for self-collision avoidance by integrating them into a task hierarchy with null space projections. This method benefits from the efficiency of lightweight mathematical functions and a geometric primitives-based collision model. Eckhoff et al. (2023) further extend this framework to address collisions with the environment by incorporating repulsive forces. APFs offer a low computation time which allows the integration into the real-time control loop. A potential disadvantage of APFs, as pointed out in Khazoom et al. (2022), is that it remains active near obstacles (below the threshold distance) even if moving out of the constraint (e. g., away from the joint limit), which can lead to undesirable behavior.

## B.2. Joint-Limit Avoidance

A broad body of research investigates joint limit avoidance through diverse methodolgies, including Weighted Least Norm (WLN) methods (Chan and Dubey, 1995; Huang et al., 2014) and the use of reference planes to determine feasible arm ranges, keeping joints within their limits (Oh et al., 2017). However, these approaches are predominantly designed for velocity-based controllers. In Charbonneau et al. (2016) and Pasandi and Pucci (2023), joint limits are managed by parameterizing the feasible joint positions and velocities using exogenous states which are additional variables that transform the control problem from a constrained to an unconstrained one. However, this method is only applicable in combination with a desired trajectory and not within zero torque mode.

In Xu and Sun (2018), a neural network is integrated into the torque control design, with a cost function based on joint positions that encourages the manipulator to stay near the midpoint of its limits. Meanwhile, Papageorgiou et al. (2016) employ a passivity-based torque signal to ensure joint limit avoidance in a redundant manipulator, using a performance function to assess how close the joint is to its bounds. In Han and Park (2013), an activation function is introduced that applies a torque input as a joint approaches its limit. To ensure smooth and continuous transitions, a buffer zone is incorporated when the constraint is activated, using a piecewise-defined transition function based on the joint limits to generate the repulsive torque. Similarly, Eckhoff et al. (2023) also produce a torque to avoid joint limits. They adapt the framework from Dietrich et al. (2012b) to generate repelling torques instead of forces, converting the mathematical equations from Cartesian space to joint space.

## B.3. Contraint Hierarchy

The most basic approach is to execute tasks at the same priority level. For example, in Eckhoff et al. (2023), the torques from impedance control and self-collision avoidance are combined additively. Similarly, in Quiroz-Omaña and Adorno (2019), all sub-tasks are incorporated as independent constraints within an optimization problem. However, conflicting objectives can cause undesired robot behavior. To prevent this, many approaches use a task hierarchy, prioritizing tasks like collision avoidance over trajectory accuracy.

The H-QP-based optimization method described in Ren et al. (2024) integrates the framework developed in Kanoun et al. (2011) to enforce a strict priority hierarchy, with self-collision avoidance taking precedence over other sub-tasks. The solver begins by addressing the constraints at the highest priority level and then iteratively optimizes according to the lower-level constraints within the solution set defined by the higher-priority levels. This ensures that the constraints of higher-priority tasks are fully satisfied before any consideration of lower-priority tasks. The work by Xu and Sun (2018) prioritizes trajectory tracking over joint limit avoidance, employing null space projections to ensure that joint limit avoidance does not interfere with motion in the task space. In contrast, Dietrich et al. (2012b) present a clear task hierarchy structure where lower-priority tasks are projected into the null space of higher-priority tasks. They utilize a successive projection method, which ensures smooth transitions when unilateral constraints, such as self-collision avoidance, are activated, thereby preserving a continuous control law.

# List of Figures

# List of Tables

# Acronyms

***Bimacs*** *KIT Bimanual Actions Dataset*
***KIT BMD*** *KIT Bimanual Manipulation Dataset*

**ACT** Action Chunking with Transformers
**ASSIGN** Asynchronous-Sparse Interaction Graph Networks

**BiRNN** Bidirectional Recurrent Neural Network

**CNN** Convolutional Neural Network

**DMP** Dynamical Movement Primitive
**DoF** Degree of Freedom

**GN** Graph Networks
**GNN** Graph Neural Network

**HAPPO** Heterogenous-Agent Proximal Policy Optimization
**HAR** Human Activity and Recognition

**ISTA-Net** Interactive Spatiotemporal Token Attention Networks

**MAPPO** Multi-Agent Proximal Policy Optimization
**MARL** Multi-Agent Reinforcement Learning
**MAT** Multi-Agent Transformer
**MLP** Multilayer Perceptron
**MMM** Master Motor Map
**MP** Movement Primitive

**PGCN** Pyramid Graph Convolutional Network
**PPO** Proximal Policy Optimization
**ProMP** Probabilistic Movement Primitive

**RL** Reinforcement Learning

**SAC** Soft-Actor Critic
**STIGPN** Spatio-Temporal Interaction Graph Parsing Network

**VMP** Via-Point Movement Primitive

# Bibliography

Aggarwal, J. K. and Ryoo, M. S. (2011). Human activity analysis: A review. *ACM Comput. Surv.*, 43(3):1–43. Cited on page 15.

Aksoy, E. E., Abramov, A., Dörr, J., Ning, K., Dellen, B., and Wörgötter, F. (2011). Learning the semantics of object–action relations by observation. *Int. J. of Robotics Research*, 30(10):1229–1249. Cited on page 25.

Albu-Schäffer, A., Ott, C., and Hirzinger, G. (2004). A passivity based cartesian impedance controller for flexible joint robots-part ii: Full state feedback, impedance design and experiments. In *IEEE Int. Conf. on Robotics and Automation*, volume 3, pages 2666–2672. Cited on pages 102 and 103.

Allen, J. F. (1983). Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843. Cited on page 24.

Amadio, F., Colomé, A., and Torras, C. (2019). Exploiting symmetries in reinforcement learning of bimanual robotic tasks. *IEEE Robotics and Automation Letters*, 4(2):1838–1845. Cited on page 29.

Ames, A. D., Coogan, S., Egerstedt, M., Notomista, G., Sreenath, K., and Tabuada, P. (2019). Control barrier functions: Theory and applications. In *European Control Conference (ECC)*, pages 3420–3431. IEEE. Cited on page 132.

Andrychowicz, M., Wolski, F., Ray, A., Schneider, J., Fong, R., Welinder, P., McGrew, B., Tobin, J., Abbeel, P., and Zaremba, W. (2017). Hindsight experience replay. *Advances in Neural Information Processing Systems*, 30. Cited on page 33.

Arapi, V., Della Santina, C., Averta, G., Bicchi, A., and Bianchi, M. (2021). Understanding human manipulation with the environment: A novel taxonomy for video labelling. *IEEE Robotics and Automation Letters*, 6(4):6537–6544. Cited on page 12.

Asfour, T., Wächter, M., Kaul, L., Rader, S., Weiner, P., Ottenhaus, S., Grimm, R., Zhou, Y., Grotz, M., and Paus, F. (2019). ARMAR-6: A high-performance humanoid for human-robot collaboration in real world scenarios. *IEEE Robotics & Automation Magazine*, 26(4):108–121. Cited on pages 112 and 115.

Avigal, Y., Berscheid, L., Asfour, T., Kröger, T., and Goldberg, K. (2022). Speedfolding: Learning efficient bimanual folding of garments. In *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, pages 1–8. Cited on pages 29 and 32.

Barral, J., Debû, B., and Rival, C. (2006). Developmental changes in unimanual and bimanual aiming movements. *Developmental Neuropsychology*, 29(3):415–429. Cited on page 8.

Batinica, A., Nemec, B., Ude, A., Raković, M., and Gams, A. (2017). Compliant movement primitives in a bimanual setting. In *IEEE-RAS Int. Conf. on Humanoid Robots*, pages 365–371. Cited on page 28.

Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R., et al. (2018). Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*. Cited on pages 18, 19, 85, and 135.

Billard, A. and Kragić, D. (2019). Trends and challenges in robot manipulation. *Science*, 364(6446). Cited on pages 1 and 7.

Billard, A. G., Calinon, S., and Dillmann, R. (2016). Learning from humans. *Springer Handbook of Robotics*, pages 1995–2014. Cited on page 33.

Boehm, J. R., Fey, N. P., and Fey, A. M. (2021). Online recognition of bimanual coordination provides important context for movement data in bimanual teleoperated robots. In *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, pages 6248–6255. Cited on pages 14 and 16.

Bombile Bosongo, M. and Billard, A. (2022). Dual-arm control for coordinated fast grabbing and tossing of an object. *IEEE Robotics & Automation Magazine*. Cited on page 26.

Borràs, J. and Asfour, T. (2015). A whole-body pose taxonomy for loco-manipulation tasks. In *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, pages 1578–1585. Cited on pages 2, 12, and 46.

Bradski, G. (2000). The OpenCV Library. *Dr. Dobb's Journal of Software Tools*. Cited on page 72.

Bullock, I., Feix, T., and Dollar, A. (2014). The Yale human grasping dataset: Grasp, object, and task data in household and machine shop environments. *Int. J. of Robotics Research*, 34:251–255. Cited on page 44.

Bullock, I., Ma, R., and Dollar, A. (2012). A hand-centric classification of human and robot dexterous manipulation. *IEEE Trans. on Haptics*, 6(2):129–144. Cited on pages 2 and 12.

Cai, M., Kitani, K. M., and Sato, Y. (2017). An ego-vision system for hand grasp analysis. *IEEE Trans. on Human-Machine Systems*, 47(4):524–535. Cited on page 38.

Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S., and Sheikh, Y. A. (2019). OpenPose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Trans. on Pattern Analysis and Machine Intelligence*. Cited on page 38.

Carmona, D. and Yu, H. (2024). Bicap: A novel bi-modal dataset of daily living dual-arm manipulation actions. *Int. J. of Robotics Research*. Cited on pages 42 and 44.

Cashmore, L., Uomini, N., and Chapelain, A. (2008). The evolution of handedness in humans and great apes: a review and current issues. *Journal of Anthropological Sciences*, 86:7–35. Cited on page 10.

Chan, T. F. and Dubey, R. V. (1995). A weighted least-norm solution based scheme for avoiding joint limits for redundant joint manipulators. *IEEE Trans. on Robotics and Automation*, 11(2):286–292. Cited on page 132.

Charbonneau, M., Nori, F., and Pucci, D. (2016). On-line joint limit avoidance for torque controlled robots by joint space parametrization. In *IEEE-RAS Int. Conf. on Humanoid Robots*, pages 899–904. Cited on page 132.

Chen, Y., Wu, T., Wang, S., Feng, X., Jiang, J., Lu, Z., McAleer, S., Dong, H., Zhu, S.-C., and Yang, Y. (2022). Towards human-level bimanual dexterous manipulation with reinforcement learning. *Advances in Neural Information Processing Systems*, 35:5150–5163. Cited on pages 34, 35, and 36.

Chi, C., Xu, Z., Feng, S., Cousineau, E., Du, Y., Burchfiel, B., Tedrake, R., and Song, S. (2023). Diffusion policy: Visuomotor policy learning via action diffusion. *Int. J. of Robotics Research*, page 02783649241273668. Cited on page 31.

Chiacchio, P., Chiaverini, S., and Siciliano, B. (1996). Direct and inverse kinematics for coordinated motion tasks of a two-manipulator system. *Journal of Dynamic Systems, Measurement, and Control*. Cited on page 26.

Chitnis, R., Tulsiani, S., Gupta, S., and Gupta, A. (2020). Efficient bimanual manipulation using learned task schemas. In *IEEE Int. Conf. on Robotics and Automation*, pages 1149–1155. Cited on pages 33, 34, and 36.

Chu, K., Zhao, X., Weber, C., Li, M., Lu, W., and Wermter, S. (2024). Large language models for orchestrating bimanual robots. In *IEEE-RAS Int. Conf. on Humanoid Robots*, pages 328–334. Cited on pages 30 and 37.

Cutkosky, M. R. (1989). On grasp choice, grasp models, and the design of hands for manufacturing tasks. *IEEE Trans. on Robotics and Automation*, 5(3):269–279. Cited on pages 2, 12, and 45.

Damen, D., Doughty, H., Farinella, G. M., Furnari, A., Kazakos, E., Ma, J., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al. (2022). Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *IEEE Int. Conf. on Computer Vision*, pages 1–23. Cited on pages 38 and 44.

Das, P., Xu, C., Doell, R. F., and Corso, J. J. (2013). A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 2634–2641. Cited on page 38.

De la Torre, F., Hodgins, J., Bargteil, A., Martin, X., Macey, J., Collado, A., and Beltran, P. (2009). Guide to the Carnegie Mellon University Multimodal Activity (CMU-MMAC) Database. Cited on pages 40 and 44.

Dehio, N., Smith, J., Wigand, D. L., Xin, G., Lin, H.-C., Steil, J. J., and Mistry, M. (2018). Modeling and control of multi-arm and multi-leg robots: Compensating for object dynamics during grasping. In *IEEE Int. Conf. on Robotics and Automation*, pages 294–301. Cited on page 27.

Deniša, M., Gams, A., Ude, A., and Petrič, T. (2015). Learning compliant movement primitives through demonstration and statistical generalization. *IEEE/SMC Trans. on Mechatronics*, 21(5):2581–2594. Cited on page 29.

Dietrich, A., Albu-Schäffer, A., and Hirzinger, G. (2012a). On continuous null space projections for torque-based, hierarchical, multi-objective manipulation. In *IEEE Int. Conf. on Robotics and Automation*, pages 2978–2985. Cited on pages 104, 105, and 106.

Dietrich, A., Wimbock, T., Albu-Schäffer, A., and Hirzinger, G. (2012b). Integration of reactive, torque-based self-collision avoidance into a task hierarchy. *IEEE Trans. on Robotics*, 28(6):1278–1293. Cited on pages 101, 102, 105, 132, and 133.

Dometios, A. C., Zhou, Y., Papageorgiou, X. S., Tzafestas, C. S., and Asfour, T. (2018). Vision-based online adaptation of motion primitives to dynamic surfaces: application to an interactive robotic wiping task. *IEEE Robotics and Automation Letters*, 3(3):1410–1417. Cited on page 40.

Dreher, C. R. G. and Asfour, T. (2022). Learning temporal task models from human bimanual demonstrations. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. Cited on page 24.

Dreher, C. R. G. and Asfour, T. (2024). Learning symbolic and subsymbolic temporal task constraints from bimanual human demonstrations. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5160–5167. Cited on page 24.

Dreher, C. R. G., Wächter, M., and Asfour, T. (2020). Learning object-action relations from bimanual human demonstration using graph networks. *IEEE Robotics and Automation Letters (RA-L)*, 5(1):187–194. Cited on pages 17, 18, 19, 20, 22, 23, 39, 40, 44, 59, 68, 69, 70, 71, 73, 76, 85, and 88.

Drolet, M., Stepputtis, S., Kailas, S., Jain, A., Peters, J., Schaal, S., and Amor, H. B. (2024). A comparison of imitation learning algorithms for bimanual manipulation. *IEEE Robotics and Automation Letters*, 9(10):8579–8586. Cited on pages 31 and 32.

Eckhoff, M., Knobbe, D., Zwirnmann, H., Swikir, A., and Haddadin, S. (2023). Towards connecting control to perception: High-performance whole-body collision avoidance using control-compatible obstacles. In *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, pages 2354–2361. Cited on pages 101, 102, 104, 132, and 133.

Fan, Z., Taheri, O., Tzionas, D., Kocabas, M., Kaufmann, M., Black, M. J., and Hilliges, O. (2023). ARCTIC: A dataset for dexterous bimanual hand-object

manipulation. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 12943–12954. Cited on pages 41 and 44.

Fathi, A., Ren, X., and Rehg, J. M. (2011). Learning to recognize objects in egocentric activities. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 3281–3288. Cited on page 38.

Featherstone, R. (2020). The Composite-Rigid-Body Algorithm. In Ang, M. H., Khatib, O., and Siciliano, B., editors, *Encyclopedia of Robotics*, pages 1–4. Springer, Berlin, Heidelberg. Cited on page 103.

Feix, T., Romero, J., Schmiedmayer, H.-B., Dollar, A. M., and Kragic, D. (2015). The grasp taxonomy of human grasp types. *IEEE Trans. on Human-Machine Systems*, 46(1):66–77. Cited on pages 2, 12, and 45.

Florence, P., Lynch, C., Zeng, A., Ramirez, O. A., Wahid, A., Downs, L., Wong, A., Lee, J., Mordatch, I., and Tompson, J. (2022). Implicit behavioral cloning. In *Conference on Robot Learning*, pages 158–168. PMLR. Cited on page 31.

Franzese, G., de Souza Rosa, L., Verburg, T., Peternel, L., and Kober, J. (2023). Interactive imitation learning of bimanual movement primitives. *IEEE/SMC Trans. on Mechatronics*. Cited on pages 29 and 32.

Fu, R., Zhang, D., Jiang, A., Fu, W., Funk, A., Ritchie, D., and Sridhar, S. (2024a). Gigahands: A massive annotated dataset of bimanual hand activities. *arXiv preprint arXiv:2412.04244*. Cited on pages 39 and 44.

Fu, Z., Zhao, T. Z., and Finn, C. (2024b). Mobile ALOHA: Learning bimanual mobile manipulation using low-cost whole-body teleoperation. In *Conference on Robot Learning*. Cited on pages 30, 32, and 42.

Gams, A., Nemec, B., Ijspeert, A. J., and Ude, A. (2014). Coupling movement primitives: Interaction with the environment and bimanual tasks. *IEEE Trans. on Robotics*, 30(4):816–830. Cited on page 28.

Gao, J., Jin, X., Krebs, F., Jaquier, N., and Asfour, T. (2024). Bi-KVIL: Keypoints-based visual imitation learning of bimanual manipulation tasks. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 16850–16857. Cited on page 27.

Gao, J., Zhou, Y., and Asfour, T. (2018a). Projected force-admittance control for compliant bimanual tasks. In *IEEE/RAS International Conference on Humanoid Robots (Humanoids)*, pages 607–613. Cited on page 27.

Gao, J., Zhou, Y., and Asfour, T. (2018b). Projected force-admittance control for compliant bimanual tasks. In *IEEE-RAS Int. Conf. on Humanoid Robots*, pages 1–9. Cited on page 126.

Garrido-Jurado, S., Muñoz-Salinas, R., Madrid-Cuevas, F. J., and Marín-Jiménez, M. J. (2014). Automatic generation and detection of highly reliable fiducial markers under occlusion. *Pattern Recognition*, 47(6):2280–2292. Cited on page 72.

Goyal, R., Ebrahimi Kahou, S., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., et al. (2017). The "something something" video database for learning and evaluating visual common sense. In *IEEE Int. Conf. on Computer Vision*, pages 5842–5850. Cited on pages 38 and 44.

Grannen, J., Wu, Y., Belkhale, S., and Sadigh, D. (2022). Learning bimanual scooping policies for food acquisition. In *Conference on Robot Learning*. PMLR. Cited on pages 32 and 34.

Grannen, J., Wu, Y., Vu, B., and Sadigh, D. (2023). Stabilize to act: Learning to coordinate for bimanual manipulation. In *Conference on Robot Learning*, pages 563–576. PMLR. Cited on pages 31, 32, and 37.

Grotz, M., Shridhar, M., Chao, Y.-W., Asfour, T., and Fox, D. (2024). Peract2: Benchmarking and learning for robotic bimanual manipulation tasks. In *CoRL 2024 Workshop on Whole-body Control and Bimanual Manipulation: Applications in Humanoids and Beyond*. Cited on pages 31 and 32.

Gu, Y. and Demiris, Y. (2024). Learning bimanual manipulation policies for bathing bed-bound people. In *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, pages 8936–8943. Cited on pages 35 and 36.

Guiard, Y. (1987). Asymmetric Division of Labor in Human Skilled Bimanual Action. *Journal of Motor Behavior*, 19(4):486–517. Cited on pages 9, 10, 11, 47, 48, 63, and 80.

Haarnoja, T., Zhou, A., Hartikainen, K., Tucker, G., Ha, S., Tan, J., Kumar, V., Zhu, H., Gupta, A., Abbeel, P., et al. (2018). Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*. Cited on page 34.

Haas, J., Endrikat, M., Krebs, F., and Asfour, T. (2024). An exploratory study on the relation between grasp types and bimanual categories in manipulation ac-

tivities. In *IEEE/RAS International Conference on Humanoid Robots (Humanoids)*, pages 388–395. Cited on page 125.

Han, H. and Park, J. (2013). Robot control near singularity and joint limit using a continuous task transition algorithm. *Int. J. of Advanced Robotic Systems*, 10(10):346. Cited on page 133.

Huang, S., Peng, Y., Wei, W., and Xiang, J. (2014). Clamping weighted least-norm method for the manipulator kinematic control: Avoiding joint limits. In *Proceedings of the 33rd Chinese Control Conference*, pages 8309–8314. IEEE. Cited on page 132.

Huang, Y. and Sun, Y. (2019). A dataset of daily interactive manipulation. *Int. J. of Robotics Research*, 38(8):879–886. Cited on pages 40 and 44.

Hung, Y.-C. and Zeng, W. (2020). Accuracy constraints improve symmetric bimanual coordination for children with and without unilateral cerebral palsy. *Developmental Neurorehabilitation*, 23(3):176–184. Cited on page 8.

Ijspeert, A., Nakanishi, J., and Schaal, S. (2002a). Learning attractor landscapes for learning motor primitives. *Advances in Neural Information Processing Systems*, 15. Cited on pages 28 and 57.

Ijspeert, A. J., Nakanishi, J., and Schaal, S. (2002b). Movement imitation with nonlinear dynamical systems in humanoid robots. In *IEEE Int. Conf. on Robotics and Automation*, volume 2, pages 1398–1403. Cited on page 53.

Ionescu, C., Papava, D., Olaru, V., and Sminchisescu, C. (2014). Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339. Cited on page 40.

Ivry, R., Diedrichsen, J., Spencer, R., Hazeltine, E., and Semjen, A. (2004). A cognitive neuroscience perspective on bimanual coordination and interference. In *Neuro-Behavioral Determinants of Interlimb Coordination*, pages 259–295. Springer. Cited on pages 9 and 10.

James, S., Ma, Z., Arrojo, D. R., and Davison, A. J. (2020). Rlbench: The robot learning benchmark & learning environment. *IEEE Robotics and Automation Letters*, 5(2):3019–3026. Cited on page 31.

Jang, J., Kim, D., Park, C., Jang, M., Lee, J., and Kim, J. (2020). Etri-activity3d: A large-scale RGB-D dataset for robots to recognize daily activities of the elderly.

In *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, pages 10990–10997. Cited on pages 39 and 44.

Jia, B., Chen, Y., Huang, S., Zhu, Y., and Zhu, S.-C. (2020). LEMMA: A Multi-view Dataset for LEarning Multi-agent Multi-task Activities. In *European Conference on Computer Vision*, pages 767–786. Cited on pages 39 and 44.

Kamakura, N., Matsuo, M., Ishii, H., Mitsuboshi, F., and Miura, Y. (1980). Patterns of static prehension in normal hands. *American Journal of Occupational Therapy*, 34(7):437–445. Cited on pages 2, 12, and 45.

Kanoun, O., Lamiraux, F., and Wieber, P.-B. (2011). Kinematic control of redundant manipulators: Generalizing the task-priority framework to inequality task. *IEEE Trans. on Robotics*, 27(4):785–792. Cited on page 133.

Kantak, S., Jax, S., and Wittenberg, G. (2017). Bimanual coordination: A missing piece of arm rehabilitation after stroke. *Restorative Neurology and Neuroscience*, 35(4):347–364. Cited on pages 11, 12, 14, 15, 45, 49, and 135.

Kartmann, R. and Asfour, T. (2023). Interactive and incremental learning of spatial object relations from human demonstrations. *Frontiers in Robotics and AI*, 10:1–14. Cited on page 26.

Kartmann, R., Liu, D., and Asfour, T. (2021). Semantic scene manipulation based on 3d spatial object relations and language instructions. In *IEEE/RAS International Conference on Humanoid Robots (Humanoids)*, pages 306–313. Cited on page 26.

Kartmann, R., Paus, F., Grotz, M., and Asfour, T. (2018). Extraction of physically plausible support relations to predict and validate manipulation action effects. *IEEE Robotics and Automation Letters (RA-L)*, 3(4):3991–3998. Cited on pages 72 and 78.

Kartmann, R., Zhou, Y., Liu, D., Paus, F., and Asfour, T. (2020). Representing spatial object relations as parametric polar distribution for scene manipulation based on verbal commands. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8373–8380. Cited on pages 25 and 26.

Kataoka, S., Ghasemipour, S. K. S., Freeman, D., and Mordatch, I. (2022). Bimanual manipulation and attachment via sim-to-real reinforcement learning. *arXiv preprint arXiv:2203.08277*. Cited on pages 34 and 36.

Kelso, J. (1984). Phase transitions and critical behavior in human bimanual coordination. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, 246(6):R1000–R1004. Cited on page 9.

Kelso, J., Southard, D. L., and Goodman, D. (1979). On the coordination of two-handed movements. *Journal of Experimental Psychology: Human Perception and Performance*, 5(2):229. Cited on page 9.

Khaire, P. and Kumar, P. (2022). Deep learning and RGB-D based human action, human–human and human–object interaction recognition: A survey. *Journal of Visual Communication and Image Representation*, 86:103531. Cited on page 17.

Khatib, O. (1986). Real-time obstacle avoidance for manipulators and mobile robots. *Int. J. of Robotics Research*, 5(1):90–98. Cited on page 132.

Khazatsky, A., Pertsch, K., Nair, S., Balakrishna, A., Dasari, S., Karamcheti, S., Nasiriany, S., Srirama, M. K., Chen, L. Y., Ellis, K., et al. (2024). Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*. Cited on page 42.

Khazoom, C., Gonzalez-Diaz, D., Ding, Y., and Kim, S. (2022). Humanoid self-collision avoidance using whole-body control with control barrier functions. In *IEEE-RAS Int. Conf. on Humanoid Robots*, pages 558–565. Cited on page 132.

Kim, H., Ohmura, Y., and Kuniyoshi, Y. (2021). Transformer-based deep imitation learning for dual-arm robot manipulation. In *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, pages 8965–8972. IEEE. Cited on pages 30 and 32.

Kim, H., Ohmura, Y., and Kuniyoshi, Y. (2024). Goal-conditioned dual-action imitation learning for dexterous dual-arm robot manipulation. *IEEE Trans. on Robotics*, pages 2287–2305. Cited on pages 27 and 32.

Kimmerle, M., Ferre, C. L., Kotwica, K. A., and Michel, G. F. (2010). Development of role-differentiated bimanual manipulation during the infant's first year. *Developmental Psychobiology: The Journal of the International Society for Developmental Psychobiology*, 52(2):168–180. Cited on pages 1, 8, and 11.

Koppula, H. S., Gupta, R., and Saxena, A. (2013). Learning human activities and object affordances from RGB-D videos. *Int. J. of Robotics Research*, 32(8):951–970. Cited on pages 22, 39, 40, and 44.

Koptev, M., Figueroa, N., and Billard, A. (2021). Real-time self-collision avoidance in joint space for humanoid robots. *IEEE Robotics and Automation Letters*, pages 1240–1247. Cited on page 131.

Kratzer, P., Bihlmaier, S., Midlagajni, N. B., Prakash, R., Toussaint, M., and Mainprice, J. (2020). Mogaze: A dataset of full-body motions that includes workspace geometry and eye-gaze. *IEEE Robotics and Automation Letters*, 6(2):367–373. Cited on page 41.

Krebs, F. and Asfour, T. (2022). A bimanual manipulation taxonomy. *IEEE Robotics and Automation Letters (RA-L)*, 7(4):11031–11038. Cited on pages 45, 48, 75, 78, 79, 80, 92, and 124.

Krebs, F. and Asfour, T. (2024). Formalization of temporal and spatial constraints of bimanual manipulation categories. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1302–1309. Cited on pages 50, 51, 52, 54, 55, 56, 57, 109, 112, 113, 115, 116, 124, and 125.

Krebs, F., Leven, L., and Asfour, T. (2023). Recognition of bimanual manipulation categories in RGB-D human demonstration. In *IEEE/RAS International Conference on Humanoid Robots (Humanoids)*, pages 1–8. Cited on pages 63, 69, 70, 71, 75, 89, 94, and 124.

Krebs, F., Meixner, A., Patzer, I., and Asfour, T. (2021). The KIT bimanual manipulation dataset. In *IEEE/RAS International Conference on Humanoid Robots (Humanoids)*, pages 499–506. Cited on pages 42, 44, 60, 63, 64, 69, 76, and 124.

Kuba, J. G., Chen, R., Wen, M., Wen, Y., Sun, F., Wang, J., and Yang, Y. (2021). Trust region policy optimisation in multi-agent reinforcement learning. *arXiv preprint arXiv:2109.11251*. Cited on page 34.

Kwon, T., Tekin, B., Stühmer, J., Bogo, F., and Pollefeys, M. (2021). H2O: Two hands manipulating objects for first person interaction recognition. In *IEEE Int. Conf. on Computer Vision*, pages 10138–10148. Cited on pages 22, 39, and 44.

Labbé, Y., Manuelli, L., Mousavian, A., Tyree, S., Birchfield, S., Tremblay, J., Carpentier, J., Aubry, M., Fox, D., and Sivic, J. (2022). Megapose: 6D pose estimation of novel objects via render & compare. *arXiv preprint arXiv:2212.06870*. Cited on page 96.

Laghi, M., Maimeri, M., Marchand, M., Leparoux, C., Catalano, M., Ajoudani, A., and Bicchi, A. (2018). Shared-autonomy control for intuitive bimanual tele-manipulation. In *IEEE-RAS Int. Conf. on Humanoid Robots*, pages 1–9. Cited on page 12.

Lee, A., Chuang, I., Chen, L.-Y., and Soltani, I. (2024). Interact: Inter-dependency aware action chunking with hierarchical attention transformers for bimanual manipulation. *arXiv preprint arXiv:2409.07914*. Cited on pages 30 and 32.

Li, Y., Pan, C., Xu, H., Wang, X., and Wu, Y. (2023). Efficient bimanual handover and rearrangement via symmetry-aware actor-critic learning. In *IEEE Int. Conf. on Robotics and Automation*, pages 3867–3874. Cited on pages 33 and 36.

Lin, Y., Church, A., Yang, M., Li, H., Lloyd, J., Zhang, D., and Lepora, N. F. (2023). Bi-touch: Bimanual tactile manipulation with sim-to-real deep reinforcement learning. *IEEE Robotics and Automation Letters*, pages 5472–5479. Cited on pages 34 and 36.

Liu, I., Arthur, C., He, S., Seita, D., and Sukhatme, G. (2024a). Voxact-b: Voxel-based acting and stabilizing policy for bimanual manipulation. *arXiv preprint arXiv:2407.04152*. Cited on pages 31, 32, and 37.

Liu, J., Chen, Y., Dong, Z., Wang, S., Calinon, S., Li, M., and Chen, F. (2022). Robot cooking with stir-fry: Bimanual non-prehensile manipulation of semi-fluid objects. *IEEE Robotics and Automation Letters*, 7(2):5159–5166. Cited on page 32.

Liu, J., Sim, H., Li, C., Tan, K. C., and Chen, F. (2023). BiRP: Learning robot generalized bimanual coordination using relative parameterization method on human demonstration. In *IEEE Conf. on Decision and Control (CDC)*, pages 8300–8305. Cited on pages 27 and 32.

Liu, J., Wang, G., Hu, P., Duan, L.-Y., and Kot, A. C. (2017). Global context-aware attention LSTM networks for 3D action recognition. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1647–1656. Cited on page 17.

Liu, Y., Yang, H., Si, X., Liu, L., Li, Z., Zhang, Y., Liu, Y., and Yi, L. (2024b). TACO: Benchmarking generalizable bimanual tool-action-object understanding. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 21740–21751. Cited on pages 34, 41, and 44.

Luh, J. and Zheng, Y. (1987). Constrained relations between two coordinated industrial robots for motion control. *Int. J. of Robotics Research*, 6(3):60–70. Cited on page 25.

Mahmood, N., Ghorbani, N., Troje, N. F., Pons-Moll, G., and Black, M. J. (2019). AMASS: Archive of Motion Capture as Surface Shapes. In *IEEE Int. Conf. on Computer Vision*, pages 5442–5451. Cited on page 40.

Mandery, C., Terlemez, O., Do, M., Vahrenkamp, N., and Asfour, T. (2015). The KIT whole-body human motion database. In *International Conference on Advanced Robotics (ICAR)*, pages 329–336. Cited on pages 40, 43, and 44.

Mandery, C., Terlemez, O., Do, M., Vahrenkamp, N., and Asfour, T. (2016). Unifying representations and large-scale whole-body motion databases for studying human motion. *IEEE Transactions on Robotics*, 32(4):796–809. Cited on page 67.

Maurice, P., Malaisé, A., Amiot, C., Paris, N., Richard, G.-J., Rochel, O., and Ivaldi, S. (2019). Human movement and ergonomics: An industry-oriented dataset for collaborative robotics. *Int. J. of Robotics Research*, 38(14):1529–1537. Cited on pages 40 and 44.

Mehta, D., Sridhar, S., Sotnychenko, O., Rhodin, H., Shafiei, M., Seidel, H.-P., Xu, W., Casas, D., and Theobalt, C. (2017). Vnect: Real-time 3D human pose estimation with a single RGB camera. *ACM Transactions on Graphics (TOG)*, 36(4):1–14. Cited on page 38.

Meixner, A., Krebs, F., Jaquier, N., and Asfour, T. (2023). An evaluation of action segmentation algorithms on bimanual manipulation datasets. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4912–4919. Cited on page 60.

Miller, A. and Wade, E. (2021). Classifying unimanual and bimanual upper extremity tasks in individuals post-stroke. In *Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 6301–6305. Cited on page 16.

Mirrazavi Salehian, S. S., Figueroa, N., and Billard, A. (2018). A unified framework for coordinated multi-arm motion planning. *Int. J. of Robotics Research*, 37(10):1205–1232. Cited on page 25.

Morais, R., Le, V., Venkatesh, S., and Tran, T. (2021). Learning asynchronous and sparse human-object interaction in videos. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 16041–16050. Cited on pages 21, 22, 23, 73, 93, and 135.

Nakamura, Y. C., O'Sullivan, C. A., and Pollard, N. S. (2019). Effect of object and task properties on bimanual transport. *Journal of Motor Behavior*, 51(3):245–258. Cited on page 10.

Nicora, E., Goyal, G., Noceti, N., Vignolo, A., Sciutti, A., and Odone, F. (2020). The MoCA dataset, kinematic and multi-view visual streams of fine-grained cooking actions. *Scientific Data*, 7(1):432. Cited on page 44.

Obhi, S. S. (2004). Bimanual coordination: An unbalanced field of research. *Motor Control*, 8(2):111–120. Cited on page 9.

Oh, J., Bae, H., and Oh, J.-H. (2017). Analytic inverse kinematics considering the joint constraints and self-collision for redundant 7dof manipulator. In *IEEE International Conference on Robotic Computing (IRC)*, pages 123–128. Cited on page 132.

Oh, J.-H., Espinoza, I., Jung, D., and Kim, T.-S. (2024). Bimanual long-horizon manipulation via temporal-context transformer RL. *IEEE Robotics and Automation Letters*, pages 10898–10905. Cited on pages 33, 35, and 36.

Osa, T., Pajarinen, J., Neumann, G., Bagnell, J. A., Abbeel, P., Peters, J., et al. (2018). An algorithmic perspective on imitation learning. *Foundations and Trends® in Robotics*, 7(1-2):1–179. Cited on page 30.

O'Neill, A., Rehman, A., Maddukuri, A., Gupta, A., Padalkar, A., Lee, A., Pooley, A., Gupta, A., Mandlekar, A., Jain, A., et al. (2024). Open X-Embodiment: Robotic learning datasets and RT-X models. In *IEEE Int. Conf. on Robotics and Automation*, pages 6892–6903. Cited on page 42.

Pairet, È., Ardón, P., Mistry, M., and Petillot, Y. (2019). Learning and composing primitive skills for dual-arm manipulation. In *Towards Autonomous Robotic Systems (TAROS)*, pages 65–77. Springer. Cited on page 27.

Pais, A. L. and Billard, A. (2014). Encoding bi-manual coordination patterns from human demonstrations. In *ACM/IEEE Int. Conf. on Human-Robot Interaction (HRI)*, pages 264–265. Cited on page 12.

Pais Ureche, L. and Billard, A. (2018). Constraints extraction from asymmetrical bimanual tasks and their use in coordinated behavior. *Robotics and Autonomus Systems*, 103:222–235. Cited on pages 12, 15, 27, 41, and 126.

Papageorgiou, D., Atawnih, A., and Doulgeri, Z. (2016). A passivity based control signal guaranteeing joint limit avoidance in redundant robots. In *Mediterranean Conference on Control and Automation (MED)*, pages 569–574. Cited on page 133.

Paraschos, A., Daniel, C., Peters, J., Neumann, G., et al. (2013). Probabilistic movement primitives. *Advances in Neural Information Processing Systems*. Cited on pages 28 and 57.

Pardowitz, M., Knoop, S., Dillmann, R., and Zöllner, R. D. (2007). Incremental learning of tasks from user demonstrations, past experiences, and vocal comments. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 37(2):322–332. Cited on page 25.

Park, H. A. and Lee, C. G. (2015). Extended cooperative task space for manipulation tasks of humanoid robots. In *IEEE Int. Conf. on Robotics and Automation*, pages 6088–6093. Cited on page 26.

Park, H. A. and Lee, C. G. (2016). Dual-arm coordinated-motion task specification and performance evaluation. In *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, pages 929–936. Cited on pages 14 and 26.

Pasandi, V. and Pucci, D. (2023). Torque control with joints position and velocity limits avoidance. In *IEEE Int. Conf. on Robotics and Automation*, pages 5310–5316. Cited on page 132.

Paulius, D., Huang, Y., Meloncon, J., and Sun, Y. (2019). Manipulation motion taxonomy and coding for robots. In *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, pages 5596–5601. Cited on page 12.

Pauwels, K. and Kragic, D. (2015). Simtrack: A simulation-based framework for scalable real-time object pose detection and tracking. In *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, pages 1300–1307. Cited on page 38.

Pek, C., Schuppe, G. F., Esposito, F., Tumova, J., and Kragic, D. (2023). Spatial: monitoring and planning of robotic tasks using spatio-temporal logic specifications. *Autonomous Robots*, 47(8):1439–1462. Cited on page 24.

Pereira, D., De Pra, Y., Tiberi, E., Monaco, V., Dario, P., and Ciuti, G. (2022). Flipping food during grilling tasks, a dataset of utensils kinematics and dynamics, food pose and subject gaze. *Scientific Data*, 9(1):5. Cited on page 41.

Perrig, S., Kazennikov, O., and Wiesendanger, M. (1999). Time structure of a goal-directed bimanual skill and its dependence on task constraints. *Behavioural Brain Research*, 103(1):95–104. Cited on page 9.

Pham, C. and Olivier, P. (2009). Slice&dice: Recognizing food preparation activities using embedded accelerometers. In *European Conference on Ambient Intelligence*, pages 34–43. Cited on page 44.

Pieropan, A., Salvi, G., Pauwels, K., and Kjellström, H. (2014). Audio-visual classification and detection of human manipulation actions. In *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, pages 3045–3052. Cited on pages 39 and 44.

Pnueli, A. (1977). The temporal logic of programs. In *Annual Symposium on Foundations of Computer Science*, pages 46–57. IEEE. Cited on page 25.

Pomerleau, D. A. (1988). Alvinn: An autonomous land vehicle in a neural network. *Advances in Neural Information Processing Systems*, 1. Cited on page 31.

Puranic, A., Deshmukh, J., and Nikolaidis, S. (2021). Learning from demonstrations using signal temporal logic. In *Conference on Robot Learning*, pages 2228–2242. PMLR. Cited on page 25.

Quiroz-Omaña, J. and Adorno, B. (2019). Whole-body control with (self) collision avoidance using vector field inequalities. *IEEE Robotics and Automation Letters*, 4(4):4048–4053. Cited on pages 131 and 133.

Rakita, D., Mutlu, B., Gleicher, M., and Hiatt, L. M. (2019). Shared control–based bimanual robot manipulation. *Science Robotics*, 4:eaaw0955. Cited on pages 14, 16, and 17.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 779–788. Cited on page 38.

Ren, Y. et al. (2024). Enabling versatility and dexterity of the dual-arm manipulators: A general framework toward universal cooperative manipulation. *IEEE Trans. on Robotics*, 40:2024–2045. Cited on pages 131 and 133.

Roebuck-Spencer, T. M., Mattson, S. N., Marion, S. D., Brown, W. S., and Riley, E. P. (2004). Bimanual coordination in alcohol-exposed children: Role of the corpus callosum. *Journal of the International Neuropsychological Society*, 10(4):536. Cited on page 8.

Rogez, G., Supancic, J. S., and Ramanan, D. (2015). Understanding Everyday Hands in Action from RGB-D Images. In *IEEE Int. Conf. on Computer Vision*, pages 3889–3897. Cited on page 39.

Roggen, D., Calatroni, A., Rossi, M., Holleczek, T., Förster, K., Tröster, G., Lukowicz, P., Bannach, D., Pirkl, G., Ferscha, A., et al. (2010). Collecting complex activity datasets in highly rich networked sensor environments. In *International Conference on Networked Sensing Systems (INSS)*, pages 233–240. IEEE. Cited on pages 40 and 44.

Rohrbach, M., Rohrbach, A., Regneri, M., Amin, S., Andriluka, M., Pinkal, M., and Schiele, B. (2016). Recognizing fine-grained and composite activities using hand-centric features and script data. *International Journal of Computer Vision*, 119(3):346–373. Cited on page 38.

Ross, S., Gordon, G., and Bagnell, D. (2011). A reduction of imitation learning and structured prediction to no-regret online learning. In *International Conference on Artificial Intelligence and Statistics*, pages 627–635. JMLR Workshop and Conference Proceedings. Cited on page 31.

Roşu, G. and Bensalem, S. (2006). Allen linear (interval) temporal logic–translation to ltl and monitor synthesis. In *International Conference on Computer Aided Verification*, pages 263–277. Springer. Cited on page 25.

Sainburg, R. L. (2002). Evidence for a dynamic-dominance hypothesis of handedness. *Experimental Brain Research*, 142(2):241–258. Cited on page 11.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*. Cited on page 34.

Schwarke, C., Klemm, V., van der Boon, M., Bjelonic, M., and Hutter, M. (2023). Curiosity-driven learning of joint locomotion and manipulation tasks. In *Conference on Robot Learning*, volume 229, pages 2594–2610. Cited on pages 33, 34, and 36.

Shahbazi, M., Lee, J., Caldwell, D., and Tsagarakis, N. (2017). Inverse dynamics control of bimanual object manipulation using orthogonal decomposition: An analytic approach. In *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, pages 4791–4796. Cited on page 27.

Shao, L., Migimatsu, T., Zhang, Q., Yang, K., and Bohg, J. (2020). Concept2robot: Learning manipulation concepts from instructions and human demonstrations. In *Robotics: Science and Systems (RSS)*, volume 40, pages 1419–1434. Cited on page 38.

Shi, L., Zhang, Y., Cheng, J., and Lu, H. (2019). Skeleton-based action recognition with directed graph neural networks. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 7912–7921. Cited on page 17.

Shirota, C., Jansa, J., Diaz, J., Balasubramanian, S., Mazzoleni, S., Borghese, N. A., and Melendez-Calderon, A. (2016). On the assessment of coordination between upper extremities: Towards a common language between rehabilitation engineers, clinicians and neuroscientists. *Journal of Neuroengineering and Rehabilitation*, 13(1):1–14. Cited on page 11.

Shridhar, M., Manuelli, L., and Fox, D. (2023). Perceiver-actor: A multi-task transformer for robotic manipulation. In *Conference on Robot Learning*, pages 785–799. PMLR. Cited on page 31.

Smith, C., Karayiannidis, Y., Nalpantidis, L., Gratal, X., Qi, P., Dimarogonas, D. V., and Kragic, D. (2012). Dual arm manipulation - a survey. *Robotics and Autonomus Systems*, 60(10):1340–1353. Cited on pages 7, 26, and 45.

Song, S., Lan, C., Xing, J., Zeng, W., and Liu, J. (2017). An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In *AAAI Conference on Artificial Intelligence*, volume 31. Cited on page 17.

Starke, J., Eichmann, C., Ottenhaus, S., and Asfour, T. (2018). Synergy-based, data-driven generation of object-specific grasps for anthropomorphic hands. In *IEEE/RAS International Conference on Humanoid Robots (Humanoids)*, pages 327–333. Cited on page 61.

Stavridis, S., Falco, P., and Doulgeri, Z. (2021). Pick-and-place in dynamic environments with a mobile dual-arm robot equipped with distributed distance sensors. In *IEEE-RAS Int. Conf. on Humanoid Robots*, pages 76–82. Cited on pages 27 and 50.

Stein, S. and McKenna, S. J. (2013). Combining embedded accelerometers with computer vision for recognizing food preparation activities. In *ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 729–738. Cited on pages 39 and 44.

Stepputtis, S., Bandari, M., Schaal, S., and Amor, H. B. (2022). A system for imitation learning of contact-rich bimanual manipulation policies. In *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, pages 11810–11817. Cited on pages 29 and 32.

Sun, Z., Ke, Q., Rahmani, H., Bennamoun, M., Wang, G., and Liu, J. (2022). Human action recognition from various data modalities: A review. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 45(3):3200–3225. Cited on pages 17 and 18.

Surdilovic, D., Yakut, Y., Nguyen, T. M., Pham, X. B., Vick, A., and Martin-Martin, R. (2010). Compliance control with dual-arm humanoid robots: Design, planning and programming. In *IEEE-RAS Int. Conf. on Humanoid Robots*, pages 275–281. Cited on pages 13, 15, and 135.

Swinnen, S. P. (2002). Intermanual coordination: From behavioural principles to neural-network interactions. *Nature Reviews Neuroscience*, 3(5):348–359. Cited on pages 1, 9, and 10.

Swinnen, S. P. and Wenderoth, N. (2004). Two hands, one brain: Cognitive neuroscience of bimanual skill. *Trends in Cognitive Sciences*, 8(1):18–25. Cited on page 8.

Taheri, O., Ghorbani, N., Black, M. J., and Tzionas, D. (2020). GRAB: A dataset of whole-body human grasping of objects. In *European Conference on Computer Vision*, pages 581–600. Cited on pages 41 and 44.

Tang, Y., Liu, X., Yu, X., Zhang, D., Lu, J., and Zhou, J. (2022). Learning from temporal spatial cubism for cross-dataset skeleton-based action recognition. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18(2):1–24. Cited on page 17.

Tarbouriech, S., Navarro, B., Fraisse, P., Crosnier, A., Cherubini, A., and Sallé, D. (2022). An admittance based hierarchical control framework for dual-arm cobots. *Mechatronics*, 86:102814. Cited on page 27.

Tenorth, M., Bandouch, J., and Beetz, M. (2009). The TUM kitchen data set of everyday manipulation activities for motion tracking and action recognition. In *IEEE Int. Conf. on Computer Vision*, pages 1089–1096. Cited on pages 39 and 44.

Todorov, E., Erez, T., and Tassa, Y. (2012). Mujoco: A physics engine for model-based control. In *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, pages 5026–5033. Cited on page 112.

Toyokura, M., Muro, I., Komiya, T., and Obara, M. (1999). Relation of bimanual coordination to activation in the sensorimotor cortex and supplementary motor area: analysis using functional magnetic resonance imaging. *Brain Research Bulletin*, 48(2):211–217. Cited on page 10.

Toyokura, M., Muro, I., Komiya, T., and Obara, M. (2002). Activation of pre–supplementary motor area (sma) and sma proper during unimanual and bimanual complex sequences: an analysis using functional magnetic resonance imaging. *Journal of Neuroimaging*, 12(2):172–178. Cited on page 10.

Tracy, J., Faro, S., Mohammed, F., Pinus, A., Madi, S., and Laskas, J. (2001). Cerebellar mediation of the complexity of bimanual compared to unimanual movements. *Neurology*, 57(10):1862–1869. Cited on page 10.

Uchiyama, M. and Dauchez, P. (1992). Symmetric kinematic formulation and non-master/slave coordinated control of two-arm robots. *Advanced Robotics*, 7(4):361–383. Cited on pages 25 and 26.

Ude, A., Gams, A., Asfour, T., and Morimoto, J. (2010). Task-specific generalization of discrete and periodic dynamic movement primitives. *IEEE Trans. on Robotics*, 26(5):800–815. Cited on page 28.

Umlauft, J., Sieber, D., and Hirche, S. (2014). Dynamic movement primitives for cooperative manipulation and synchronized motions. In *IEEE Int. Conf. on Robotics and Automation*, pages 766–771. Cited on page 28.

Volkmar, R., Dosen, S., Gonzalez-Vargas, J., Baum, M., and Markovic, M. (2019). Improving bimanual interaction with a prosthesis using semi-autonomous control. *Journal of Neuroengineering and Rehabilitation*, 16:140. Cited on pages 14, 15, 16, and 58.

Wächter, M., , and Asfour, T. (2015). Hierarchical Segmentation of Manipulation Actions based on Object Relations and Motion Characteristics. In *International Conference on Advanced Robotics (ICAR)*, pages 549–556. Cited on page 68.

Walke, H. R., Black, K., Zhao, T. Z., Vuong, Q., Zheng, C., Hansen-Estruch, P., He, A. W., Myers, V., Kim, M. J., Du, M., et al. (2023). Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning*, pages 1723–1736. PMLR. Cited on page 42.

Walker, I. D., Freeman, R. A., and Marcus, S. I. (1991). Analysis of motion and internal loading of objects grasped by multiple cooperating manipulators. *Int. J. of Robotics Research*, 10(4):396–409. Cited on page 27.

Wang, C.-Y., Bochkovskiy, A., and Liao, H.-Y. M. (2022). YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv preprint arXiv:2207.02696*. Cited on page 71.

Wang, N., Zhu, G., Zhang, L., Shen, P., Li, H., and Hua, C. (2021). Spatio-temporal interaction graph parsing networks for human-object interaction recognition. In *ACM International Conference on Multimedia*, pages 4985–4993. Cited on pages 22, 23, and 93.

Wen, M., Kuba, J., Lin, R., Zhang, W., Wen, Y., Wang, J., and Yang, Y. (2022). Multi-agent reinforcement learning is a sequence modeling problem. *Advances in Neural Information Processing Systems*, 35:16509–16521. Cited on pages 35 and 36.

Wen, Y., Tang, Z., Pang, Y., Ding, B., and Liu, M. (2023). Interactive Spatiotemporal Token Attention Network for Skeleton-Based General Interactive Action Recognition. In *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, pages 7886–7892. Cited on pages 20, 22, 23, and 93.

Wiesendanger, M. and Serrien, D. J. (2001). Toward a physiological understanding of human dexterity. *Physiology*, 16(5):228–233. Cited on page 10.

Wimböck, T., Ott, C., Albu-Schäffer, A., and Hirzinger, G. (2012). Comparison of object-level grasp controllers for dynamic dexterous manipulation. *Int. J. of Robotics Research*, 31(1):3–23. Cited on page 27.

Xie, F., Chowdhury, A., De Paolis Kaluza, M., Zhao, L., Wong, L., and Yu, R. (2020). Deep imitation learning for bimanual robotic manipulation. *Advances in Neural Information Processing Systems*, 33:2327–2337. Cited on page 29.

Xing, H. and Burschka, D. (2022). Understanding spatio-temporal relations in human-object interaction using pyramid graph convolutional network. In *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, pages 5195–5201. Cited on pages 20, 22, 23, 73, and 93.

Xing, H. and Burschka, D. (2024). Understanding human activity with uncertainty measure for novelty in graph convolutional networks. *Int. J. of Robotics Research*, page 02783649241287800. Cited on page 20.

Xu, Q. and Sun, X. (2018). Adaptive operation-space control of redundant manipulators with joint limits avoidance. In *Int. Conf. on Advanced Computational Intelligence (ICACI)*, pages 358–363. IEEE. Cited on page 133.

Yang, F., Wu, Y., Sakti, S., and Nakamura, S. (2019). Make skeleton-based action recognition model smaller, faster and better. In *ACM Multimedia Asia*, pages 1–6. Cited on page 17.

Yang, Z., Han, Y., and Ravichandar, H. (2024). Asymdex: Leveraging asymmetry and relative motion in learning bimanual dexterity. In *CoRL Workshop on Whole-body Control and Bimanual Manipulation: Applications in Humanoids and Beyond*. Cited on pages 34, 36, and 37.

Yao, K., Sternad, D., and Billard, A. (2021). Hand pose selection in a bimanual fine-manipulation task. *Journal of Neurophysiology*, 126(1):195–212. Cited on pages 13 and 126.

Yu, C., Velu, A., Vinitsky, E., Gao, J., Wang, Y., Bayen, A., and Wu, Y. (2022). The surprising effectiveness of ppo in cooperative multi-agent games. *Advances in Neural Information Processing Systems*, 35:24611–24624. Cited on page 35.

Yu, D., Xu, H., Chen, Y., Ren, Y., and Pan, J. (2024). BiKC: Keypose-conditioned consistency policy for bimanual robotic manipulation. *arXiv preprint arXiv:2406.10093*. Cited on pages 31 and 32.

Zakour, M., Nath, P. P., Lohmer, L., Gökçe, E. F., Piccolrovazzi, M., Patsch, C., Wu, Y., Chaudhari, R., and Steinbach, E. (2024). Adl4d: Towards a contextually rich dataset for 4d activities of daily living. *arXiv preprint arXiv:2402.17758*. Cited on pages 41 and 44.

Zhan, W. and Chin, P. (2024). Safe multi-agent reinforcement learning for bimanual dexterous manipulation. In *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, pages 12420–12427. Cited on pages 35 and 36.

Zhan, X., Yang, L., Zhao, Y., Mao, K., Xu, H., Lin, Z., Li, K., and Lu, C. (2024). OAKINK2: A dataset of bimanual hands-object manipulation in complex task completion. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 445–456. Cited on pages 41 and 44.

Zhang, M., Jian, P., Wu, Y., Xu, H., and Wang, X. (2021). Dair: Disentangled attention intrinsic regularization for safe and efficient bimanual manipulation. *arXiv preprint arXiv:2106.05907*. Cited on pages 33, 34, and 36.

Zhang, T., Li, D., Li, Y., Zeng, Z., Zhao, L., Sun, L., Chen, Y., Wei, X., Zhan, Y., Li, L., et al. (2024). Empowering embodied manipulation: A bimanual-mobile robot manipulation dataset for household tasks. *arXiv preprint arXiv:2405.18860*. Cited on page 42.

Zhao, T. Z., Kumar, V., Levine, S., and Finn, C. (2023). Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*. Cited on pages 30, 31, 32, and 42.

Zhou, B., Yuan, H., Fu, Y., and Lu, Z. (2024). Learning diverse bimanual dexterous manipulation skills from human demonstrations. *arXiv preprint arXiv:2410.02477*. Cited on pages 34 and 36.

Zhou, Y., Do, M., and Asfour, T. (2016). Coordinate change dynamic movement primitives - a leader-follower approach. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, page 5481–5488. Cited on pages 25 and 28.

Zhou, Y., Gao, J., and Asfour, T. (2019). Learning via-point movement primitives with inter- and extrapolation capabilities. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4301–4308. Cited on pages 28, 57, and 98.

Ziaeetabar, F., Kulvicius, T., Tamosiunaite, M., and Wörgötter, F. (2018). Recognition and prediction of manipulation actions using enriched semantic event chains. *Robotics and Autonomus Systems*, 110:173–188. Cited on pages 25, 26, and 72.

Ziaeetabar, F., Tamosiunaite, M., and Wörgötter, F. (2024). A hierarchical graph-based approach for recognition and description generation of bimanual actions in videos. *IEEE Access*. Cited on pages 15, 20, and 73.

Zöllner, R., Asfour, T., and Dillmann, R. (2004). Programming by demonstration: Dual-arm manipulation tasks for humanoid robots. In *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, pages 479–484. Cited on pages 13, 14, 16, 25, and 53.