UNIVERSITY OF GENOVA

PHD PROGRAM IN BIOENGINEERING AND ROBOTICS

Implicit communication to convey and perceive object properties for human robot interaction

by

Linda Lastrico

Thesis submitted for the degree of *Doctor of Philosophy* (35° cycle)

May 2023

Dr. Alessandro Carfí	Supervisor
Prof. Fulvio Mastrogiovanni	Supervisor
Dr. Francesco Rea	Supervisor
Dr. Alessandra Sciutti	Supervisor
Prof. Paolo Massobrio	Head of the PhD program

Thesis Jury:

Prof. Matej Hoffmann, *Czech Technical University, Prague*	External examiner
Prof. Lorenzo Jamone, *Queen Mary University of London*	External examiner
Prof. Maura Casadio, *University of Genova*	Internal examiner

Dibris

Department of Informatics, Bioengineering, Robotics and Systems Engineering

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Linda Lastrico
May 2023

Abstract

As human beings, having social connections plays a crucial role in our lives, and our interactions are based on communication. Much of the exchange of information occurs in a non-verbal way through what can be described as implicit communication. Indeed, we are exceptionally skilled at reading the unspoken from other people's actions: gaze direction, gestures rhythm, or body posture for instance reveal, even unintentionally, the intentions, mood, or urgency of others' actions. Also when manipulating objects, the way we do it can make explicit otherwise hidden characteristics, such as weight or fragility.

My Ph.D. research focuses on the implicit communication of object properties in the context of human-robot interaction. Collaborative robots can benefit significantly from exploiting implicit communication channels as humans do, allowing for more natural, spontaneous, and safe interactions. The main scientific question I aim to answer is whether it is possible to automatically detect features of the carried objects from human movements and, on the other hand, communicate the same information exploiting the robot embodiment. My approach exploits the kinematics modulations that naturally occur when transporting objects: it is precisely the way we interact with them that makes their characteristics understandable to an external observer, resolving at the same time all the possible misunderstandings due to shape, size, or occlusions.

To detect and study the kinematics in human actions, I resorted to and compared motion capture systems, inertial sensors, and cameras, exploiting machine learning algorithms to classify the observed motion. To reproduce communicative movements on the robot, I controlled the end-effector kinematics, by modulating its velocity to follow synthetic profiles, automatically generated after training on real human examples.

As regards the property to detect and express, I focused mainly on the carefulness associated with object manipulation, i.e., to assess and implicitly communicate if any caution is required to handle the carried item.

Findings prove that it is possible exploiting features such as the movement velocity, retrieved with various sensors, to classify online whether an action is careful or not. Exploiting a generative strategy to produce robot motions successfully delivers the intended carefulness and can be applied to different trajectories and robots, even those not humanoid. Moreover, the modulation in the robot's actions also induces a spontaneous motor adaptation in how participants perform their tasks, matching the robot's attitude. Such findings prove that information can be exchanged with robots through implicit cues embedded in the actions, opening a channel of communication that relies on a core human interaction ability.

Given the adaptability of the approach to different robots and the non-invasive sensing methods, an industrial field of application seems feasible, beyond the one of social robotics. The ability of the robot to automatically perceive and express the carefulness feature could improve the safety and efficiency of collaborative object manipulation tasks. Future developments of this work may include other object properties in the framework, such as the weight or the temperature, and could exploit additional cues to the kinematics, to expand the possible field of application.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

1.1 Motivation

Human-robot interaction (HRI) has become a rapidly growing field in recent years, as advancements in technology have led to the increasing presence of robots in various domains, including industry, healthcare, and education. However, the obstacles to integrating robots into our society are still widely present and difficult to overcome. Among the others, one issue is represented by the challenges of communicating smoothly and directly with an artificial agent.

Standard means of communication that are being implemented on robots use speech or explicit gestures, or others which can be mediated by additional devices such as tablets or augmented reality visors [23, 97, 133]. However, we communicate only partially through words and we tend to rely on a whole set of additional cues that convey information and help the coordination with the partner: gaze direction, walking pace, body posture, or gestures rhythms are just examples [141]. Resorting to this way of implicitly communicating can be intentional: in this case, it is commonly referred to as signaling [125]. It represents an intentional strategy that supports social interaction, especially when the context makes actions or intentions difficult to interpret. For example, when dancing, a slight tilt of the head or a weight shift can signal the next move.

Other adaptations occur naturally, even if we do not think about it and we are acting alone: for instance, we tend to direct our gaze towards the object before grasping it. Or we bend our knees and move slowly when carrying a particularly heavy package. These modulations are functional and optimize the completion of the action but, even if we do not explicitly want to communicate anything, an external observer can infer a discrete amount of information on our intentions, objectives, and internal state. This favors immediate coordination between the agents, that can efficiently collaborate to complete the task [83].

My Ph.D. thesis sits within this context. Knowing how human-human interaction works should guide us in designing artificial agents able to interact with us most spontaneously and naturally. Such an objective cannot overlook a whole sphere of information exchange that resides in implicit communication. Many works have investigated how to exploit robots to complete a task efficiently: a defined assembling task carried out by a robot is sometimes performed quicker, faster, and safer than a human will ever be able to do. However, if robots should truly coordinate with us and support our activities, especially in unconstrained environments, then the human aspect becomes central: coordination can emerge only from reciprocal understanding [149].

The focus of this work is centered in particular on common daily activities, where robots and humans can often be requested to collaborate, such as handling and transporting objects. In this case, non-verbal communication is associated with the properties of the object that is being moved: its weight, fragility, danger, or assembly state, for instance, may influence how it is carried around [22, 45, 181]. The application of this concept to robotics is twofold: both the perception of the objects' features and their expression to the human partner needs to be tackled. As regards understanding the characteristics of an unknown item, most of the approaches presented in Section 2.2 base the inference on its appearance: size, material, shape, and color can be informative of certain aspects, but they risk being misleading when deciding how to transport the object. The novelty of the proposed approach is inspired by human behavior and resides in using movement kinematics and modulation as a source of information on the item properties. Moreover, being detached from the exterior aspect allows generalization over novel and unknown objects. Inspiration from humans is also taken when it comes to generating expressive movements with the robots: communication is therefore seen as an additional layer that modulates the action without changing its goal [42, 82].

Using different methodologies and approaches, this thesis contributes to pursuing a final goal: to apply the implicit communication associated with carrying objects to interact with robots, drawing inspiration from the natural mechanisms that shape human behavior. This objective encompasses several challenges that can be summarized in three key points: i) study the human strategies to handle objects with variable features, focusing on the transport phase ii) find solutions to let the robot autonomously infer the properties of the objects when observing the action performed by the partner, ideally relying on its own sensors; iii) identify approaches to generate robots' actions which are expressive of specific object characteristics while being task oriented. This is also declined in studies investigating how the modulation of the robot motions and its embodiment impacts the interaction.

In the chapters that follow, I will go into detail about why each of these factors should be taken into account as well as the novel contributions this thesis aims to make in each case.

A recurrent case of use in this thesis is the manipulation and transport of items that required different levels of care to be moved and whose weight could be variable. Specifically, we constrained the concept of carefulness to that required to move a container full of liquid while avoiding spills, by assuming that appropriate orientation and grip force are applied and studying the modulations in the transport velocity. Although this narrows the field with respect to all the possible characteristics that a commonly used object can have, it still represents a good example for studying the effect of a specific property on how we handle the object. In fact, it is suitable to be used as a proof of concept to test the automatic detection of motion characteristics and at the same time its artificial expression through the robot's movements. In any case, the generalization potential of this approach will be proven by testing different sensors, robotic platforms, and use cases.

1.2 Research overview

Many works have studied the relevance of implicit communication as a part of the broader spectrum of non-verbal communication, and its potential applications in robotics, exploring the possibility for artificial agents to express or interpret non-verbal cues (for an extensive review, see Chapter 2).

This thesis aims to provide additional contributions to this field, particularly addressing the challenges of associating implicit communication to object transport and manipulation with robotic platforms. In this Section, I will provide a breakdown of my research path.

The first step of my work consisted in investigating human behavior in the context of object handling. Regularities in how the arm moves within specific tasks, e.g. writing, are well established [53, 87, 167, 126]. Several studies proved how informative our movements can be, to the point that the intention of drinking or pouring from a bottle can be anticipated by looking at a few frames of the action [26]. As regards the interaction with objects, kinematics modulations emerged for instance depending on the weight of the object lifted [16, 150]. My contribution to achieving a better understanding of how multiple features of the object influence the action started from the collection of a comprehensive dataset. The future application to HRI was planned from the beginning and, for this reason, the structure of the dataset was designed to be extremely controlled with respect to the objects employed (common plastic cups), while varying the direction and the considered combination of features (weight and carefulness required for the movement, demanded by the possible presence of liquid inside). Indeed, we privileged a controlled scenario to isolate the effect of the features of the object on the action, avoiding variability that objects with different shapes or uses could introduce. Moreover, the dataset involved a set of synchronized sensors, for later studying the typologies of information that could be extracted from each of them and their relevance in describing and understanding the motion. Inertial sensors, cameras, and motion capture system confirmed their usefulness in tracking the actions, with the latter being chosen as ground truth for the accuracy and detail. By analyzing the transport kinematics, we observed how the carefulness required, obtained by filling the cups with water, was responsible for most of the action modulation. Careful transport actions were characterized by longer duration, lower maximum velocities, and an asymmetric and prolonged deceleration phase.

From the dataset acquisition, we measured how object features impacted the motion. The next challenge was then to design and implement an architecture to automatically detect such properties relying on the kinematics of the action. We started by determining which features of the motion could be extracted depending on the sensor considered and we privileged motor descriptors validated in literature and easily interpretable also from the human eye. By exploiting the dataset acquired, several classifiers were tested offline to discriminate the weight of the carried object and the carefulness adopted by the handler, comparing the performances obtained with different sensors' training data. Our findings showed how the detection of the carefulness in the motion was reliable, even using a single noisy sensor such as the robot camera as a source of information. The weight perception was instead more challenging, and we ascribed it to the nature of the dataset where, as already mentioned, the presence of water required more adaptation in the actions than the weight difference. We confirmed that the velocity of the action is particularly informative on the attitude of the manipulation, which shall be interpreted not as a constant value, referred for instance to peak velocity, but as a time series with a specific dynamic and temporal evolution.

In light of these results, we decided to design a complete architecture for autonomously discriminating the carefulness online, relying exclusively on the robot camera, and testing it in different scenarios. Even if some limiting assumptions were made, such as the robot head being fixed, to the best of our knowledge this is the first attempt at detecting object properties through the robot sensors relying on the action kinematics. This represents a contribution towards robots being autonomously able to perceive and interpret implicit cues when interacting with a person. Without requiring cumbersome equipment to measure the action kinematics, our solution seems promising to be adapted to a more practical and unconstrained environment where humans and robots need to collaborate to handle objects, allowing for natural coordination.

In the first part of my Ph.D. I focused on the inference process that the robot could perform starting from human movement. However, I was well aware that communicating is a bi-directional process, which also applies when robots are involved. Therefore, I then addressed the dual problem of generating expressive robots' actions, always linked to object manipulation. Design and implementing robotic movements is a classic, well-known, and extremely broad question, that can be approached from many angles depending of course on the robotic platform, the context of use, the task to accomplish, and many other variables. Previous studies, when describing proficient human-robot collaboration, highlighted the importance of following bio-inspired rules in motion design such as minimizing the jerk. Regarding implicit communication, the relevance of generating legible robot movements turned for instance on appropriately deforming the trajectory to make the robot's intention and final goal clearer to an observer. All these insights are derived from deeply studying what characterizes the human motion pattern.

Inspired by human behavior, the aim of this work is to build up on existing control strategies to add an additional layer to enrich the robotic action with a human-like expressive attitude, without disrupting the designated tasks. According to my knowledge, this represents the first attempt at modulating how a robot transports an object to reveal its features implicitly. Based on our findings when studying careful transportation, our choice fell on controlling the velocity profile bell followed by the end-effector. However, we did not want to merely copy previously recorded human actions; instead, we wanted to learn what truly distinguishes a careful action from a not careful one. Hence, exploiting once more the dataset collected in the beginning, we resorted to generative models. After being trained on human examples, they can capture the distribution of the time series and then produce novel velocity profiles relevant to the desired attitude. We then tested this method by controlling with the synthetic data the end-effectors of different robots, even those not humanoid. Indeed, using the velocity profile to depict the motion makes our approach freely generalizable to different robotic platforms and trajectories, after appropriately rescaling them. After validating from a technical point of view that the robots' actions followed the planned profiles, it remained to validate whether indeed the modulation we implemented was expressive and communicative.

To understand how the generated actions were perceived, we resorted to two complementary strategies: conducting questionnaires and measuring how participants in our experiments performed their tasks. We conducted three studies, one involving iCub [109] and Baxter

[52], and the two others with a Kinova Gen3 manipulator. These latter were designed and conducted at the Institute for Systems and Robotics of Instituto Superior Técnico of Lisbon, where I spent a visiting research period. We found proof that the communicative intent was achieved and, even more interestingly, that a spontaneous motor contagion emerged: when the robot, either humanoid or not, displayed what we designed to be a careful action, also participants slowed down and diminished their maximum speed. In a human-to-robot handover task, a facilitation effect was measured when the robot adopted an expressive motion controller and adapted to the properties of the handled object.

To conclude, in this thesis I show how it is possible to open an additional channel of communication to interact with robots, exploiting implicit cues embedded in the movement. As the area of designated application, it was chosen the innovative one of perceiving and transmitting the properties of objects involved in transport actions. The leading idea and the whole framework of this work are inspired by human behavior and interaction abilities. This could be seen as a limitation for some specific application areas, where the speed of execution and strength of robots reach levels unimaginable for a human being. However, if we want to build robots able to share the social sphere with us and truly interact seamlessly, I believe there is no other possible way but to be inspired by what happens on the biological level.

1.3 Structure

The thesis is composed of 10 Chapters and is structured as follows:

This was **Chapter 1**, where I introduced the rationale and the overview of my Ph.D. work.

Chapter 2 provides a general background of the research, starting with an overview of what non-verbal communication means, with a special focus on implicit communication (see Section 2.1). Then, a survey of methods for understanding object properties is provided, together with a section on the generation of expressive actions with robots. Finally, in Section 2.4, I go through methods for quantitatively evaluating Human-Robot Interaction (HRI). In the various scenarios addressed, the aspect of an application in robotics is always kept in mind.

In **Chapter 3**, I present the organization and the content of a multimodal dataset acquired to study how humans deal with containers with different characteristics. In particular, I discuss the effect of properties such as the presence of a liquid inside the cup on the kinematics of human actions. This dataset will be extensively exploited in my research to study the perception and the generation of communicative movements associated with object transportation.

In **Chapter 4 and 5**, I address strategies to automatically detect features of the objects involved in the action by observing the human movements and their kinematics modulation. In Chapter 4, I assess different models and data sources in an offline study design. In Chapter 5 I present a computational architecture used to provide the robot iCub with the ability to autonomously detect online the carefulness in the observed gestures. By doing so, we replicate what happens in human-human interaction, and we show how the amount of information implicitly embedded in human movements can be exploited for practical tasks with a non-invasive approach.

In **Chapter 6 and 7** I address the generation of robots' actions which can implicitly convey the carefulness of the object transport, mimicking what happens in human transport motions. In Chapter 6, we present a generative architecture to produce synthetic velocity profiles after training on human examples. In Chapter 7, we expand the generative approach and show how to control different robots to follow the desired velocity profiles associated with careful and not careful object manipulation.

In **Chapter 8 and 9**, I study the effect and the impact of modulating robots' actions to convey implicitly the carefulness required by the object to be moved. In Chapter 8 I present an experimental work where the perception of the robots' movements and its effect on the interaction was studied through questionnaires and videos. Then, Chapter 9 describes two studies conducted during my period abroad at the Institute for System and Robotics in Lisbon. The approach of generating expressive movements is here applied to a robotic manipulator, investigating its effect on a dyadic interaction.

In conclusion, **Chapter 10** summarizes the findings and the limitations of the thesis and outlines potential future research areas based on the progress made in this work.

Chapter 2

Background

Given the multidisciplinary structure of this work, I have included references to the most relevant literature in each area of study. The background starts with an overview of non-verbal communication, distinguishing between explicit and implicit signals, reasoning on the biological basis of coordination, and presenting the available applications in human-robot interaction (Section 2.1). Then, the state of the art of the key topics addressed in the thesis will be presented (with reference to the specific application context), such as: object properties recognition (Section 2.2); robotic movements generation (Section 2.3) and strategies to assess the interaction with robotic agents (Section 2.4).

2.1 Non-verbal communication

In recent years, the use of robots in different fields, including industry, healthcare, and education, has increased. However, challenging barriers remain in place for integrating robots into our society. The difficulties in having direct and clear communication with an artificial agent stand out among the others. Standard communication mediums in robotics are based on speech [23], or explicit gestures [97, 95], while more recent approaches investigate new technologies such as augmented reality headsets [101, 133]. However, an often overlooked channel of communication is the non-verbal one. Indeed, as humans, we communicate only partially through words; we heavily rely on subtle cues to collect from our peers a spectrum of information, ranging from internal status to intentions or needs [141]. Broad psychology reviews [63, 102] define non-verbal communication as behavior that includes all communicative acts except speech, involving cues from many modalities such as face, voice, body, touch, and interpersonal space. A distinction needs to be made between explicit and implicit non-verbal communication. The first encloses a series of para-verbal signals that can integrate or substitute the spoken word; the latter, on the contrary, is made of all those modulations and cues that characterize our actions often against our conscious will, and that are present even without an observer. What they have in common is promoting a smooth, immediate, and effective interaction based on mutual understanding. In the following paragraphs, I will describe in greater detail these two aspects, focusing in particular on the knowledge regarding implicit communication, which is the main focus of this thesis.

2.1.1 Explicit communication and signaling

Explicit non-verbal communication refers to all those social signals that are *intentionally* performed to share information or make it more understandable. A classic example is a pointing gesture: it may substitute words or accompany a sentence, but it would never be performed by a person alone in a room: it is justified only by social interaction. Gestures either enrich or replace spoken communication, often serve as a shortcut, and even contribute to the comprehension of each other's interior state [97]. Para-verbal cues are used in conversations to indicate acceptance, comprehension, or continuous attention, providing silent feedback to the speaker as it happens in backchannelling, which effects have also been studied in robotics [18, 115]. Always in a conversation, when miscommunication happens, the speaker tends to raise the hands concurrently, and this can be exploited to automatically detect disfluencies in the communication [183]. These signals may be emphasized with a social purpose to make the action even more predictable and readable by a collaborator: this is referred to as the signaling theory [125]. It is the deliberate, yet often unconscious, change of one's own conduct to send information to another person, usually a co-actor. The signaler typically has the communicative aim of influencing the co-belief actor's (e.g., facilitating his comprehension of the signaler's goals) in combination with a pragmatic purpose (e.g., executing the joint task). Joint action optimization comes with a cost to pay in terms of action performance: for instance, exaggerating the trajectory to make it more predictable may require an additional effort. After all, cooperation is inherently part of our being, to the extent that we prefer to cooperate even when individual performance would be higher [35]. The effect of the intention to communicate has been proved, for instance, on grasping gestures, where the emerging kinematic pattern differed between individual and social conditions [143], or when playing a musical instrument, where modulations in the kinematics were used to simulate participants playing alone or in synchrony with others [107]. The signaling theory is also corroborated by the so-called *motionese* [19], the attitude that mothers show in modulating their infant-directed actions in ways that help infants process human action.

Implementing such mechanisms in robotics is a step toward a shared comprehension of a joint task with an artificial agent [41, 43, 106, 159].

2.1.2 Implicit communication

Implicit communication refers to all those modulations that occur in our gestures independently of our will or consciousness. Coordination emerges from the interaction precisely because of the implicit signals [83], giving rise to a series of mutual synchronization and anticipation phenomena that drastically reduce the need for complex verbal instructions and the resulting delays [15]. Before going into details and examples of how implicit cues, particularly kinematics ones, contribute to the flow of information, I will dedicate a paragraph to describe historic and well-established knowledge of the motion of the human arm. Indeed, the way we act grants reciprocal understanding: thanks to the product of motor resonance, observing an action triggers the same set of neurons, providing a common ground for understanding others [132].

The arm kinematics follow specific regularities when moving. The minimum-jerk model, presented by Flash and Hogan in 1985 [53], hypothesizes a relation between the trajectory and the velocity of the action: unconstrained point-to-point motions are approximately straight with bell-shaped tangential velocity profiles; curved motions show a reduction in the tangential velocity at points of high curvature instead. The optimization of the cost of the movement results in a trajectory with minimum jerk. The so-called Two-Third Power Law [87, 167] was firstly observed in drawing movements and states that the instantaneous tangential velocity of the hand is proportional to the third root of the radius of curvature of the trajectory: again, the result is that the velocity increases when the trajectory becomes straighter. These two findings correlate well [166] and are strictly linked to the kinematic configuration of the human arm. Human motor performance is also characterized by Fitts' law [51, 126], which relates the speed of a movement with its accuracy: when moving from target A to B, the time required depends on the distance between the points and their width. Intentional violations of this law have been used to implicitly transfer information to the partner in a joint task [163]. The understanding of action is greatly facilitated by the consistency with the constraints of these three laws (Fitts' law, minimum-jerk, and the two-thirds power law), and this also applies to robotics [68]; even passively following the trajectory of a manipulator requires less force exchange if the path respect the constraints of the Two Third power law, and the adaptation to other dynamics demands effort and training from the user [105]. Strategies to generate plausible robot actions will be addressed in detail in Section 2.3.

The human brain is remarkably effective at processing the biological motion it observes, and this ability starts at birth [153]. Even if only a point-light sequence is shown, as long as the motion respects some of the properties mentioned before of biological motion, the characteristics or purpose of the action can be interpreted [68]. It should be emphasized, however, that this ease is present only if one observes a movement, even if it was only a tenth of a second: from a static representation of point-lights, it is difficult to distinguish a meaningful representation [79]. As a result, there is a plethora of information encoded in the way individuals move that other humans automatically analyze during action observation or cooperation.

Body movement can help in prospective planning by revealing the intentions of others, and conveying information regarding why the action is being performed [141]. Kinematic features of reach-to-grasp actions, such as velocity, trajectory, and finger position, can vary based on the intended use of the grasped object, making it possible to distinguish between grasping to pour, to drink, or to place [26, 85]. Also the intention to initiate a handover can be predicted from implicit cues [155]. Social intentions, such as competition or cooperation, also impact the kinematics of the action [104], and information embedded in the action is so powerful that it can be used to distinguish even between different intentions (to teach or to perform a joint task) [107].

A classic example of an implicit exchange of information is gaze-mediated communication [24], since we cannot avoid looking, in an anticipated way, toward the area of interest. This can be used to simplify the interaction with robots with a reduced effort from the user side, to build mechanisms for shared attention [144], or to predict the intention of reaching for a particular object [94]. A recent work presents a model for a robot that can correctly understand human action from gaze signals, modify its gaze fixation based on human eye-gaze behavior, and convey nonverbal messages that match with the robot's own action intentions [127]. Findings also show that intention anticipation is often based on the combination of multi-modal cues: if the gaze is used at the onset of the action, then other motor cues such as hand pre-shape or arm trajectory contribute more to the inference [5].

Body movements also help to regulate the timing of the interaction, enabling two partners to get a common, coordinated rhythm [99] or revealing the position and the timing of the handover [156]. Moreover, the execution timing of a discrete movement is automatically influenced by observing the speed of others, even unrelated, as long as they respect biological motion laws [170]. Modulating the duration of a specific action with a deviation from its predicted duration can establish a channel of communication to pass otherwise inaccessible information to the partner [163];

Finally, automatic kinematics adaptations occur when transporting an object, reaching for something, or preparing to pass it and are useful for an observer to understand the properties of the item and consequently design an appropriate motion plan. Body motions can reveal hidden aspects of the item to be handled, such as the force required to raise it [135]. This can be explained with a straightforward example: imagine taking a carton of juice out of the refrigerator and discovering it is almost empty. This may lead to an awkward movement, as one's expectations of the container's weight are not met. The exterior of the carton does not provide information regarding its content, thus one may rely on prior experiences to infer the quantity of juice remaining. However, after observing someone else lifting it before pouring a glass, the amount of juice inside would become immediately clear.

The perception of an object's weight from the modifications that it causes to the handling has been extensively studied. When lifting an object, our velocity is linked to its mass, and if we observe someone else doing the same, we automatically use a similar weight-speed rule to assess how heavy the burden they are carrying [16, 134]. The weight and the size of unknown objects are therefore implicitly communicated to the observer by the action itself [22]. When inferring the weight of an object, observers seem to rely mostly on the duration of the lifting movement: the longest it takes, the heavier it is [64], and even children showed the ability to read the weight from motion alone [151]. When pretending to grasp an item, without actually touching it, the action still retains information about object weight, although to a lesser extent [6].

When an action occurs, it is also deeply influenced by the carefulness that the person decides to adopt. Many reasons may require to do so: from the context where the action takes place, to the properties of the object involved such as its fragility, precarious balance, or content about to spill (these implications will be further discussed in Chapter 8). Moving a cup very slowly while passing it to someone could convey that it has a high personal value: even the knowledge of the ownership status of objects influences the visuomotor process [32]. In any case, the effect on the movement is an overall slowdown with alterations in the velocity and acceleration profiles [45]. Given the difficulty in framing this characteristic, not many studies explicitly refer to it, although it is addressed, for example, in the handling of filled containers [111, 113, 123, 140, 178].

From this overview, we can understand how informative our actions are, especially from the kinematics point of view. The richness in the human motor repertoire lends itself to be exploited to understand more about the actions considered, by processing the embedded implicit cues.

2.2 Understanding object properties

In this Section, I will summarize possible strategies and techniques to identify object properties that can be applied in robotics. A distinction will be made between those that involve direct examination of the object, for instance visual, and those that rely instead on indirect cues for the inference process. Moreover, I will examine the sensors and models that enable feature classification.

When describing the interaction with objects, a key concept is the one of *affordances*. This term was first coined by Gibson [60]. Surveys from Jamone *et al.* [78] and, more recently, Hassanin *et al.* [66], also describe the application in robotics. Without going into details, the idea is that, when observing an object, the agent perceives the possibilities offered by it without the need for constructing a detailed model: what the agent can do, its motor capabilities, influence the perception, which happens in a self-centered and action-centered way through a sensorimotor learning process. Regarding object manipulation, object affordances were popularized in robotics and linked to *(i)* the action associated with the object, *(ii)* a physical property, or *(iii)* the type of behavior required to manipulate the object [78, 142]. The theory of affordances can be used as inspiration to guide the development of perceptual systems in robotics: the goal of perception is to allow the agent to act on the environment effectively and it is deeply influenced by the action itself. In this thesis, I explore the possibility of developing the robot perception regarding specific properties of the objects as mediated by their effect on human actions.

2.2.1 Inference methods

Object detection is a crucial issue in computer vision that has received a great amount of attention in recent decades (see [171, 180] for recent reviews). If a robot needs to interact with an object, it must first isolate it and locate it in space, then it needs to understand the object category and its affordances to perform adequate grasping and task-planning [30, 103]. All these steps demand solving many different problems, often requiring the availability of a labeled dataset, and the generalization over new and unseen items is not trivial [30, 84, 179].

Focusing on specific properties of an object to detect, a well-studied topic is the understanding of a container's capacity and filling, where visual or audio cues represent the main modality for the estimation. The recent CORSMAL challenge [173] aimed indeed at building a framework for robots to recognize physical objects' properties of previously unseen containers using multi-modal sensory data, to achieve smooth handover tasks from a human to a robot. Pang *et. al.* [123] proposed a framework to design secure human-to-robot handovers by estimating the physical features of novel cups and assessing hand position from images of a person handling the container, without using scanned 3D models or additional equipment such as motion capture markers. However, the cups' features are estimated from the appearance by predicting the content type (mass and density) and filling level for each frame and each view with a Convolutional Neural Network (CNN) classifier [111]. Other approaches to determine the filling level exploit audio modality [77] or combinations of audio and visual features [76]. Ishikawa *et al.* [77] rely on audio spectrogram data to classify the content type by using CNN, and the percentage of filling with a Long-Short Term Memory (LSTM) classifier; the estimation of container's capacity and shape is performed instead through RGB 3D point cloud. Iashin *et al.* [76] combined instead audio, RGB, infrared, and depth information and opted for the late fusion of the predictions from the different sensors, in a complex architecture including Recurrent-CNN (R-CNN), Gate Recurrent Units (GRU) and Random Forest classifiers. CNN and impact audio information have also been exploited to determine the surface material of an object, retrieved from human or robotic actuators hitting specific items [39]. Other works proposed to determine the container localization and empty mass estimation resorting to a CNN architecture trained on RGB-D data [9], or, by using a single RGB image and the 3D models derived from it, to estimate the volume, the content, and the prediction of the pouring action for different containers [113]. Finally, another approach for reactive handovers of unknown objects focuses not specifically on their properties, but on their shape and avoidance of collision with the human hand using point clouds obtained from RGB-D images [176]. Exploiting a 6-Degrees-of-Freedom (DOF) GraspNet, the system selects the best grasp and executes the handover [114].

The results obtained by the presented works are remarkable. However, they strongly rely on the availability of abundant and precise multi-modal data and often base the detection on the appearance of the object. This makes them prone to failure in case the container is opaque or translucent and can suffer from illumination variations and occlusions, caused either by the human or the robot. Such approaches can be adequate for certain contexts and tasks, but difficult to apply in other, more natural settings where some sensors may not be available or fail. Identifying general rules to act on a set of objects, independently from their use, shape, or material, helps to design the behavior of an artificial agent appropriately, and one possible strategy to do so is to observe and learn from human strategies.

Sanchez-Matilla *et al.* [140] introduced a benchmark where, using only two RGB cameras, a robotic manipulator can receive objects of unknown dimensions and handover location, assuming that the item is a symmetric cup initially on the table. The distance between the giver-receiver is modeled as Gaussian Mixture Model (GMM) trained on a dataset of human-to-human handover trajectories, and the end-effector is controlled to follow a human-like path. However, in this study, the inference of the cup mass or filling is out of scope, so they accept the risk that the controller can produce motions that spill the content. Another study based the handover framework on the analysis of human behavior, by tracking the skeleton and the object using the robot cameras alone to determine where, when, and how to grasp, assuming again that the object is initially placed on the table [28].

Reading human motion proved useful also to discriminate if a transported cup was full of liquid, therefore soliciting a careful motion, or empty. Two dynamical systems, modeled as GMM and trained on features such as the wrist velocity, describe careful and not-careful behavior; the wrist velocity during the manipulation is compared to the velocities of the models, for the classifier to provide a label: tracking the kinematics of the wrist, naturally modulated as a function of the cup content, allows for its discrimination [45]. A recent work analyzed full-body kinematics during common interaction with objects and related tasks, showing that the motion is highly dependent on the item's latent properties, which can be inferred from 3D skeleton sequences [181]. For instance, it was possible to classify through Convolutional-GRU fed with skeleton sequences, different levels of object weight, content amount, type of liquid, and even softness of a chair for seating actions.

The studies presented in this Section show different strategies for understanding object features, often designed to ensure safe and efficient human-to-robot handovers. Some rely on explicit information available, such as the appearance, or cues available during the manipulation, as the sound produced. Observing how humans deal with an object represents a valid shortcut: it dispenses with numerous sensors, a camera often being sufficient; it preserves the naturalness of interaction, and it ensures generality with respect to unknown objects, despite allowing to detect with less precision certain details.

2.3 Robotic movements generation

The problem of generating robotic human-like movements has been addressed from different perspectives. Some strategies resort to optimizing motion primitives such as the jerk [53, 65]; others make use of animation techniques or exploit learning from demonstration (see respectively [147] and [129] for a review). Two main approaches can be distinguished when generating robotic non-verbal behaviors: rule-based and data-driven. Rule methods require strategies to encode expert knowledge and perform well in specific contexts, but allow for limited adaptability [3]; directly learning from data allows for more variability but requires the acquisition of appropriate datasets [86].

In Section 2.1 we examined the importance of implicit cues in everyday communication, when time constraints often do not allow for explicit verbal negotiation or reduce its effectiveness. As pointed out by Pezzulo [125], an action may contain two levels of interpretation: a functional component that achieves the goal and a communicative one which, intended or not, expresses some information to the co-actor. Its interpretation leverages a shared ground that should be extended to artificial agents: introducing biologically realistic adjustments to movement kinematics in robots could make them more understandable and transparent, for humans to use their internal models to interpret the action and engage with more ease. In this sense, implicit communication is particularly relevant for robotics actions, because it allows adding an expressive layer without modifying the overall goal or requiring cumbersome training and adaptations from the human side.

Most of the approaches that will be presented build on the knowledge of how human-human interaction and implicit communication work, either prior or investigated for a specific task, to develop models and algorithms to produce expressive robotic motions.

The concept of legibility associated with robot actions was first introduced by Dragan *et al.* [43] and it is distinguished from predictability. A direct reaching motion is predictable, hence not communicative; a curved trajectory, instead, helps to identify the target of a reaching action in an anticipated way - it is *legible*. A legible motion designed to express the robot's intent grants more fluent collaborations than a motion planned to meet the collaborator's expectations [42]. There is a trade-off between the action efficiency and its communication bundle: the robot should be able to automatically define the range of the most comprehensible but still realistic paths, obtained through a functional gradient optimization in the space of trajectories. Additional formalizations inspired by this theory have been presented, describing how the communicative channel can be used on top of normal functional actions for robots to work more effectively with human partners [82]. Drawing upon human sensibility to physical movements and dynamic affordances, effectively-designed robot motions can communicate and engage, going beyond the robot's outward appearance or practical motion trajectories [71]. Venture and Kulić proposed a recent and complete survey on expressive robot motions summarizing how the embodiment can be used to generate expressive movements while achieving a task [162]; Lohan *et al.* [96] addressed how the adaptation of movement and behavior can favor communication in HRI. As pointed out by Sandini *et al.* [141], the challenge of generating plausible motions with a humanoid robot can be even greater: from one hand, their shape can solicit the same coordination mechanisms emerging in humans, but it also creates an expectation of competency that may be easily disrupted.

2.3.1 Joint actions

Considering joint actions, if the robotic agent contradicts human movement traits, such as by adopting incorrect speed, velocity profile, or timing, the anticipation from the human side will be compromised, resulting in a slow and inefficient interaction [34]. As a practical example, the domestic scenario of unloading dishes from a rack was studied to develop a computational coordination model, using an RGB-D camera to track the user's body joints, extracting features to describe the action and predicting the user's current state using K-Nearest Neighbor (KNN) algorithm. The framework was then implemented on a Kinova MICO robotic arm to perform the handover task, finding that the efficiency is improved if the awareness of the partner's task is incorporated into the artificial agent's planning [72]. The handover scenario has been extensively investigated also from the point of view of non-verbal communication. The initial part of the carrying phase indicates the intention to hand over, whereas the remaining movement sequence is used to position the object; this and many other associated findings are summarized in [156]. This work also highlights the steps to consider to design an effective robot-to-human handover. The robot should transport the object in a distinctive carrying posture while approaching the person, and look for signs of shared attention such as mutual gaze and availability. If the conditions are met, the robot should reach out with the object toward the torso of the person or, if the person is faster, the robot should act in response. External RGB-D cameras, robot's camera, and tactile force sensors were employed depending on the specific study. The manipulator reaching action could be controlled via joint velocities knowing the initial configuration of the arm and the target pose [110]. Rasch *et al.* [128] assumed that the robot should mimic humans as much as possible to maintain the feeling of safety and simplified the problem by focusing exclusively on the handover phase, considering the object properties as known. Joint angles movement primitives were elaborated with OpenPose and Convolutional Pose Machine from human-human handovers recorded through RGB-D cameras. Each primitive movement corresponded to a motion function that was re-mapped sometimes arbitrarily, after scaling, to the robot's joints, enhancing the perceived human-likeness and safety of a Kuka manipulator.

2.3.2 Manipulating objects

Coming to generating motions suitable for object manipulation, an interesting posture-based motion planning algorithm was recently proposed by Gulletta *et al.* [62] for the upper limb of anthropomorphic robots, optimizing a task-dependent hierarchy of spatial and postural constraints. During reach-to-grasp movements, human wrist velocity typically shows a single-peaked, bell-shaped profile, with a longer deceleration phase, especially for delicate and precise movements. By decomposing the interaction with the object in sub-phases (e.g. transport, approach, retreat, and more granular descriptions), the robot action can be appropriately planned, knowing the start position, velocity, and acceleration of each joint provided by the encoders and the pose and size of any object in the workspace provided by the vision system. The timing and duration of the action are instead optimized according to Fitts' law [51].

Timing is often overlooked with respect to trajectory planning; nevertheless, it represents an important constraint to respect to avoid the robot to adopt unrealistic velocities in its gestures. By keeping the path constant and arbitrarily modulating only the timing of the action (i.e. acting on the adopted velocity profile and eventual pauses), users interpreted the carrying of a cup as more or less confident, natural, animated, or perceived the object involved as heavier [182]. From spotlight observations, they built mathematical models that correlate with the perceptions of the users to express naturalness, confidence, or weight. The idea is that robots could optimize their timing, given a path, to purposefully convey a specific message. Additional studies correlated the expression of objects' weight with the action timing. In [150], it is shown that modulating the velocity profile of lifting actions, in particular associating an increase in weight with smaller vertical velocity, allows users to estimate the weight from humanoid action observation. Moreover, they exploit this implicit information to plan appropriate grasping forces.

2.3.3 Generative models for movements design

As previously discussed, the generation of movements can be seen as learning a set of fixed parameters, goal-oriented, together but independently from an additional layer of parameters that describe *how* the task should be accomplished. Motion Primitives are versatile units that can be combined to create complex behaviors, learned from given demonstrations and adapted to new instances of a task [145]. Generative approaches such as Task Conditioned Variational Autoencoder (TC-VAE) have been proposed in this sense to generate for instance pouring actions: if the pouring location must match the container and is a strong constraint for the action success, other features such as the duration of the action or the pouring angle can vary depending on the context and should be adjustable [122]. Generative algorithms, such as Generative Adversarial Networks (GANs) were first used to produce synthetic images or for data augmentation, however, they can also produce time-series [177, 48] and be applied to action generation to produce socially acceptable approach trajectories [175], continuous movements [119], or social interactions [21]. However, there are no previous mentions of generative models to create implicitly expressive robot movements related to object manipulation, which is one of the objectives of my thesis. The advantage of resorting to generative methods is that can produce synthetic non-stereotyped data, capturing the distribution of the training set and therefore accounting for the natural variance that affects human actions [61]. Variability plays a key role in a joint action indeed, to the point that it is a predictor of how fast the motor task will be learned, as long as the co-actor gestures preserve a certain degree of predictability [136].

2.4 Evaluating HRI

In the previous Sections, we examined methods for the perception of implicit cues, the generation of expressive actions, and their deployment in the context of robotics. Assessing the effect of these applications during the interaction with a robot is however not trivial. Indeed, if the employed algorithms and models can be extensively tested in terms of performance and success rate, the quality and efficacy of the interaction is a sum of multiple contributions difficult to measure quantitatively: fluency, acceptance, trust, legibility, resonance, safety, and much more. In the following paragraphs, I will present strategies to evaluate the outcomes of interacting with a robot, distinguishing between explicit and implicit measurements.

Probably the most widely adopted method to assess HRI is questionnaires. They represent a direct, explicit measure of users' perception yet, for this reason, they need to be attentively calibrated to avoid possible bias in the answers. The Godspeed questionnaire for instance, introduced by Bartneck *et al.* in 2008 [11], is a standardized measurement tool for HRI with scales to measure five key concepts in HRI: anthropomorphism, animacy, likeability, perceived intelligence and perceived safety. Another scale, the Trust-Perception Scale-HRI, was instead proposed to measure trust over time and across robotic domains [146]. Often, questionnaires are developed ad hoc to obtain an assessment of specific experimental conditions: to measure the attribution of emotional valence and trust towards movements of a non-humanoid robot [92], to study the effect on perceived predictability and trust of dominant and submissive movement strategies in a cooperation scenario [131], or yet to collect feedback on a handover task [176].

One weakness of questionnaires is that they only examine conscious judgments of robotic systems, making it difficult to quantify HRI thoroughly. For instance, they allow only for a posteriori evaluation, with the risk of missing important reactions and events that may occur during the interaction itself. For this reason, they are often paired with other objective metrics.

To evaluate the fluency of the interaction, experimenters used scales and general measures such as the idle time of the two agents, the robot functional delay, and the percentage of concurrent activity; these are detached from specific tasks and suitable for any shared-location collaborative activity [70]. In another study, they were interested in assessing if the robot mood displayed while executing a functional task could be recognized and produce a congruent mood state in the partner, through a contagion effect [174]; the affective state was expressed by arousal and valence values for both the agents - observed and self-reported - collected through questionnaires, while the task performance was measured as the percentage of correct imitations during each session. In an experiment where participants interacted with objects after seeing a humanoid robot do the same with expressive vitality forms, kinematics information such as the hand speed and the finger aperture were collected together with open comments about the experiment [160]. Kinematics metrics are indeed particularly insightful to assess the interaction and can prove useful to measure the engagement: body posture variation, head movements, synchronous events, imitation cues, and joint attention naturally arise in response to the robot's behaviors [8].

Physiological measures such as skin conductance, electromyogram, ocular or brain activity can be a precious source of information during the interaction but require of course a more structured and invasive framework. They have been proposed for instance, together with questionnaires, to estimate the level of ease of participants, declined as motion legibility, safety (level of stress or arousal), and physical comfort [36]. Functional Magnetic Resonance Imaging (FMRI) activation patterns represent an access point to the deep processes of the brain, allowing to compare at the most primitive level reactions to human and robotic behaviors. For instance, they have been used to verify whether observing vitality forms expressed by a humanoid robot solicited the same circuits of human-human interaction [38], or if listening to action and abstract verbs pronounced by a human or a robotic voice produced similar activation patterns [37].

Reasoning on the brain functioning associated with HRI, leads to a discussion of the concepts of motor resonance and interference. Motor resonance is the automatic activation of the observer's motor control system during action perception [132]. When observing an action, the two motor brains "sound" together because they share a similar motor knowledge, which is a fundamental component of human social behavior [148]. In connection to this, observing another individual performing an action facilitates the execution of a similar action, whereas interference in completing a task may occur when observing an incongruent gesture, for instance in the form of delay or diminished performance. Awareness of these phenomena has often been leveraged to have implicit measures of spontaneous adaptations, since it allows investigating the unconscious human responses to robotic agents.

Such mechanisms are particularly suitable for assessing the plausibility of generated robot movements. Chaminade *et al.* [27] reported an increase in the variance of action execution, a measure of motor interference, when participants observed a humanoid agent performing incongruent actions only when it followed a realistic model of human kinematics. Similar findings emerged when observing pick-and-place actions: participants' motion velocity was influenced by both the human and the robot partner, as long as the artificial agent respected the biological laws of motion. Such motor contagion might modulate the spontaneity and the pleasantness of the interaction, whatever the nature of the communication partner [17]. Reaction time is also a revealing measure of action facilitation. To enhance the importance of adopting robots' biologically plausible motion, a comparison between human-human and human-robot handovers was carried out, where the robot adopted either a conventional trapezoidal joint velocity profile or a minimum-jerk optimization at the end-effector [75]; the reaction time with the human-inspired controller was significantly shorter and also the reported sense of safety was higher in this case. Interestingly, the human-human interaction was optimized in the post-handover phase, not in the reaction time.

A sign of having generated a successful interaction with a robot is its similarity with a human-human exchange. This also means that the same level of information should flow and be available to the two agents. For instance, provided a well-thought robot's action, humans can use the estimated weight of the object, retrieved from observing the robot moving, for planning their own lifting action appropriately [150].

Part I

Recognizing object properties from human manipulation

Chapter 3

On the effects of object characteristics on human manipulations: a multimodal dataset

The following is an elaboration and integration of an article submitted to:
International Journal on Robotics Research

Designing social robots requires considering all the possible forms of communication that humans exploit to exchange information, which add to the verbal one. Section 2.1 extensively remarked how an implicit communication method is central to human social abilities. Aiming at a more seamless and natural interaction with artificial agents, an effort is needed to integrate non-verbal cues associated with daily activities such as object handling. For these reasons, one of the first objectives of my thesis was to investigate the adaptation of human movements to the properties of manipulated objects. Gaining a better understanding of how the kinematics of the gestures is modulated as a function of the object characteristics is crucial to developing algorithms to interpret our actions. The core idea is that a robot, simply observing how the human acts, can obtain essential information on the object involved in the scene, that may be otherwise inaccessible or misleading due to obstructions, shape, and appearance.

The first step towards this goal is to study how humans handle objects with different properties. Previous works investigated the effect of weight on human actions, showing that the weight of an object directly relates to the speed adopted to lift and transport it, therefore influencing the kinematics of the movement [16]. In particular, the duration of the early part of the lift movement contributes most to the observer's judgment of the weight of the lifted object [64]. Recently, there has been a growth of research interest in solutions to estimate the physical characteristics of objects when humans manipulate them. The main insight is that a correct estimate of object properties would allow a robot to interact with a human more appropriately, especially when physical interactions are involved, e.g., an handover. Researchers explored using deep neural networks to estimate container capacity, dimension and weight while observing human manipulations with RGB-D [9, 168] or even simple RGB cameras [123]. Since most of these studies rely on data-driven approaches, datasets must ground the learning process. Moreover, a comprehensive dataset is needed to detect object properties based on how humans manipulate them. The literature presents different examples of datasets for the study of object manipulations [73], most of the time focusing on the interaction between the humans and the objects in specific applications, e.g., activities of daily living [74], kitchen-related actions [158, 154, 118], or handovers [25]. A recently published dataset, the CORSMAL Container Manipulation [173], collects actions such as pouring and handover initiations with containers with various shapes, materials, and content recorded with RGB-D cameras and microphones.

The motivation to acquire a novel dataset was brought by the interest in investigating the human behavior associated with the transport of objects of different weights and which required, or not, carefulness in being handled due to their content. Such property, the carefulness, is particularly intriguing because, as we will see, it is difficult to define unambiguously or measure and, nevertheless, it strongly affects the way we move. Then, it is challenging for a robot to detect it automatically, yet quite crucial to be able to collaborate in an unconstrained environment. Specifically, in the unfolding of this thesis we will address in detail the concept of carefulness associated with the transport of water-filled containers. Differently from the datasets already available, the one proposed narrows the focus on human pick-and-place actions of transparent containers, with or without water filling and balanced weights, including synchronized complementary sensors and the robot point of observation on the scene.

Research questions and objectives

For the reasons mentioned above, before designing any intelligent system to detect object properties from human manipulation, I decided to acquire an organized multimodal dataset *(i)* to investigate how humans actions are affected by the carried items, *(ii)* to assess how eventual kinematics modulation can be captured by different sensors and, finally, *(iii)* to start making assumptions on how the findings can be applied to improve human-robot interaction. The proposed dataset involves multiple subjects performing pick-and-place tasks and can also be suitable for tackling more general learning problems in the hand-object interaction domain with applications in the robotic field. The data has been collected thanks to a multimodal set-up composed of multiple cameras, observing the actions from different perspectives, a motion capture system, and a wrist-worn inertial measurement unit. All the objects manipulated in the experiments are identical in shape, size, and appearance but differ in weight and carefulness required for their handling. I chose to favor an organized structure with movements in multiple directions over a large variability of objects to keep differences in weight and carefulness manageable.

In this Chapter, I introduce such novel dataset involving pick-and-place actions performed with a glass of different weights and water fillings, recorded with a multisensory setup, i.e., motion capture system (MoCap), cameras, and wrist-worn inertial measurement units (IMUs). Section 3.1.1 describes the study design, while the data acquisition process with the details on the used sensors can be found in Section 3.1.2. In Section 3.1.3 we present a description of the dataset and its organization. Finally, in Section 3.1.4 we assess the quality of the acquired data. To conclude, Section 3.2 opens to possible scenarios where the acquired data can be exploited.

3.1 Methods

3.1.1 Study Design

As already anticipated, the concept of carefulness associated with motions has many facets, and objects may demand careful manipulation for different reasons. A glass full of water requires carefulness to avoid spilling, while a ceramic vase requires carefulness to avoid breaking it. Likewise, the different reasons inducing carefulness also influence its physical manifestation. The careful manipulation of a glass of water would manifest in slow motions with constant orientation and appropriate gripping force. Instead, carefully manipulating a ceramic vase would maximize the distance to nearby objects to avoid collisions. To summarize, some properties of the object can be affected and irreparably changed by the manipulation: an agent displaying careful behavior should adopt from time to time the most suitable pose, speed, trajectory, and strength so as not to alter them.

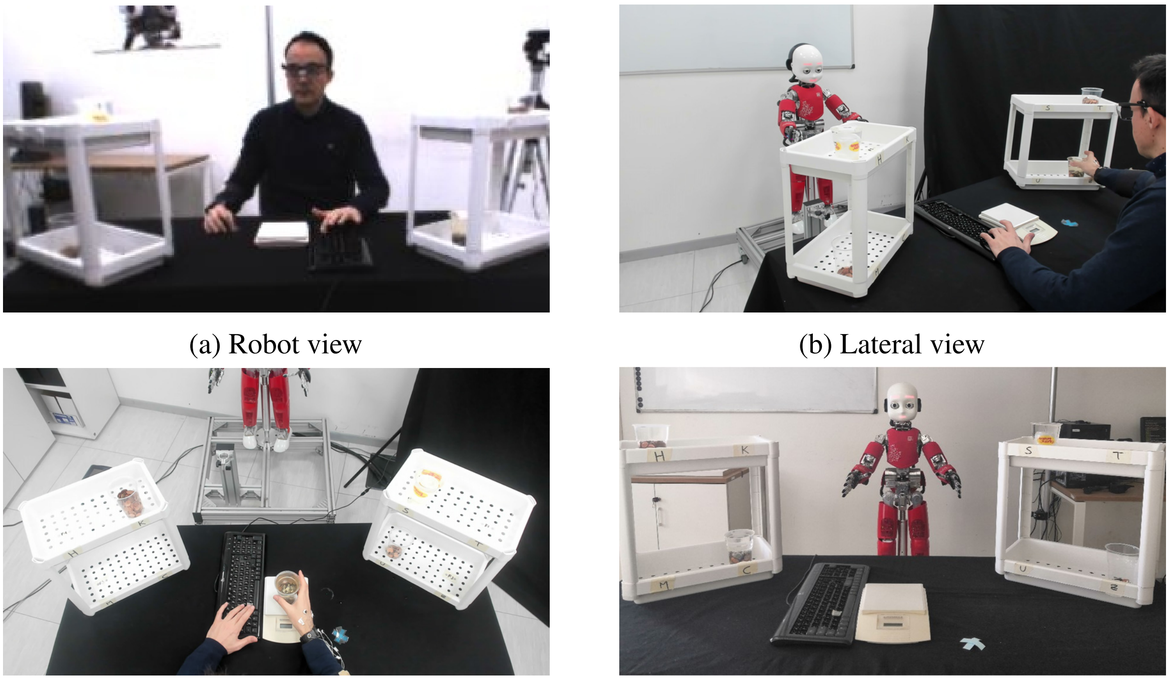
Given the limitations of previous research in the field, we decided to narrow the study limiting the notion of carefulness to the one induced by the need to move a container filled with a liquid while avoiding spills, hence by adopting appropriate pose and transport velocity. In particular, the human actions recorded in the dataset consist of reaching, transportation, and departing movements involving four possible objects. In order to allow for simple reproduction of the experiments, we chose plastic glasses, which are easy to manipulate and of everyday use. The glasses, identical in shape and material, differed in their contents. In order to induce careful behavior in the recorded actions, we filled two of the glasses with water to the brim so that they required a high level of carefulness. Two different weights are considered: light (W1: 167 grams) and heavy (W2: 667 grams). Such values were determined by the fact that we wanted the light and heavy objects to be consistently different (500 grams) while balancing the presence or the absence of water in containers with the same volume. The desired weights were obtained by adding screws and coins inside the glasses, until reaching W1 or W2, balancing the presence of water. In this way, we defined four classes of actions, depending on the properties of the manipulated object, namely light and not careful (W1-NC), light and careful (W1-C), heavy and not careful (W2-NC), heavy and careful (W2-C). See Table 3.1 for a summary of the object properties.

Table 3.1 Glasses characteristics and corresponding abbreviations

Abbreviation	**Weight (gr)**	**Carefulness level**
W1-NC	167	low – no water
W2-NC	667	low – no water
W1-C	167	high – full of water
W2-C	667	high – full of water

The sequence of performed actions is the same for every participant. It is designed to alternate the manipulation of the four categories of objects together with the direction of the movements.

At the beginning of the experiment, the volunteer seats at a table, with their hands resting on it. On the table, covered with a black cloth, there is one shelf at each end, a scale right in front of the subject, and a keyboard on the left side, see Figure 3.1 and Figure 3.2. Each shelf has four possible positions, denoted by a letter marked on the frontal edge of the shelf, where the glasses could be positioned, i.e., two on the bottom level and two on the upper one, see Figure 3.1d. The shelves measure 36 × 23 cm, the top level is 36 cm above the bottom one and there is a border of 6 cm, delimiting each shelf and constituting an obstacle. The two positions on each shelf level are indicatively 18 cm apart. A blue cross on the table marks the resting position that the participants' dominant hand should reach after each movement, see Figure 3.1a. The distance between the starting position and the shelves is indicatively 50 cm. The humanoid robot iCub [109] is placed in front of the table and passively records the scene with its left camera. The robot is not introduced to the participants, and it shows no



(a) Robot view

(b) Lateral view

(c) Back view

(d) Frontal view

Figure 3.1 The three points of view of the experiment: the resting position as seen by the iCub robot (3.1a), a transportation movement towards the right shelf (3.1b), and the positioning of a glass on the scale (3.1c). In (3.1d) the labels which identify the 8 positions on the shelves.

interactive behavior. As in the CORSMAL dataset [173], we are interested in acquiring a realistic view of the operating area from the robot perspective for future applications.

Participants did not receive any specific instruction on how to perform the pick-and-place actions with the different objects. They were asked to follow the indications provided from time to time by a synthetic voice from two audio speakers, which specified from which position to take the cup and, later, where to put it down. The underlying assumption is that motions performed this way are natural and automatically modulated according to the glass filling. Given the passive role of the robot, which can be treated as an element of the laboratory, we believe that participants perform the actions – autonomously – as if they were on their own. The additional modulation of the actions with communicative intent, predicted by the signaling theory in a collaborative scenario [125], should therefore not be solicited; thus, the pick-and-place is influenced only by the cups' properties. The volunteers interacted with the items with their right hand and received instructions on the next movement to perform by pressing a key on the keyboard with their left hand. Hence, they perform a series

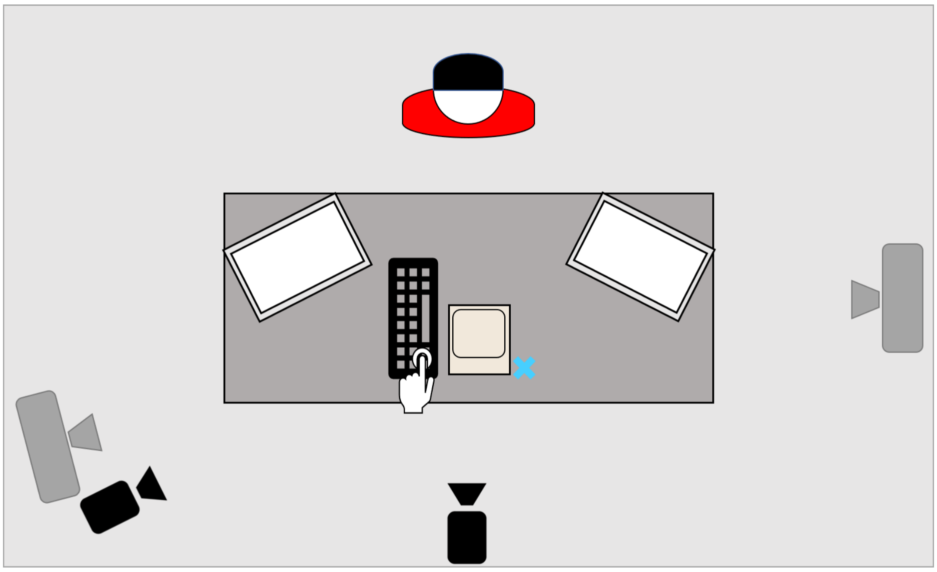


Figure 3.2 Top view of the experimental setup. In dark grey the cameras from the Optotrak motion capture system used to detect the active markers, while in black the two high resolution cameras.

of reaching, transportation, and departing movements of the four glasses. The experiment is set up as summarized in Figure 3.3:

1. The experiment starts with the volunteer in the resting position and with the four objects distributed on the shelves.

2. When a key of the keyboard is pressed, a synthetic voice indicates the position on the shelf of the object to be transported. The position is referred to using the corresponding letter.

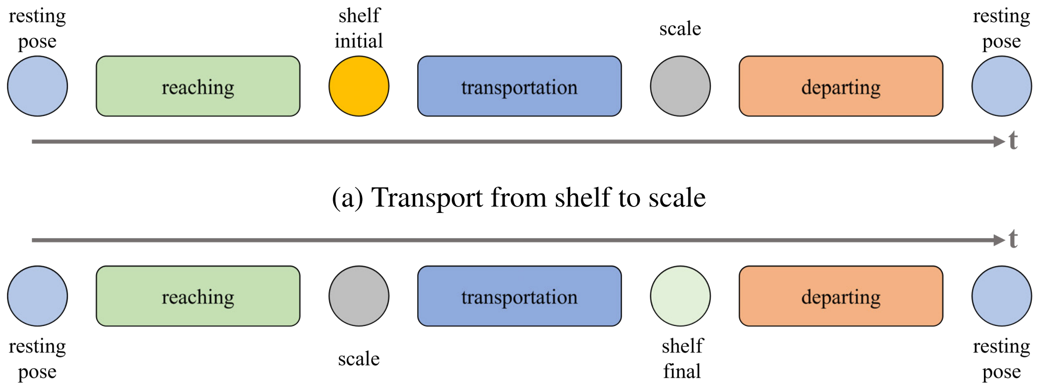
3. The volunteer reaches for the glass and grasps it in the specified location (reaching phase), as in Figure 3.1b.

4. In the transportation phase, the volunteer moves the glass from the shelf (shelf initial position) to the scale.

5. The volunteer releases the glass and returns with the dominant hand to the resting position.

6. The volunteer presses the key a second time, and the synthetic voice suggests where the glass should be placed on the other shelf. The shelf spot chosen this time is vacant.

7. The volunteer reaches for the scale and takes the glass, see Figure 3.1c.



(a) Transport from shelf to scale

(b) Transport from scale to shelf

Figure 3.3 The structure of each trial falls within two possibilities, whereby the main action is the glass transportation: in (3.3a) are shown the steps for taking a glass from the shelves and placing it on the scale, while in (3.3b) those for putting back the glass from the scale to the shelf.

8. The volunteer moves the glass from the scale to its final location on the shelf, performing a transportation action.

9. The volunteer places the glass down and returns to the resting position.

The order in which the volunteers performed the experiment is detailed in Table 3.2. The first sixteen trials were used as a practice before the main experiment started. In the main experiment, each volunteer performed 64 reaching movements, 64 transportation movements, and 64 departing movements to the resting position. After the 16th and the 48th trial, the objects' position is changed by an experimenter to maintain the properties of the manipulated objects and the initial position of the glasses equally balanced.

Table 3.2 Movements sequence performed by each participant. The first column refers to the ID used to identify each trial. The second and third columns specify the characteristics of the object handled in the trial, which could be light (*W1*) or heavy (*W2*) and require carefulness (*C*) or not (*NC*). The last two columns refer to object positions: they indicate whether the glass to grab was placed on the shelf or on the scale, and the shelf position where the glass needed to be taken from or put down.

	Trial ID	**Weight**	**Care**	**Object Initial Position**	**Shelf Position**
Practice	1	W1	NC	Shelf	Z
	2	W1	NC	Scale	K
	3	W2	NC	Shelf	H
	4	W2	NC	Scale	U
	5	W1	C	Shelf	S
	6	W1	C	Scale	M
	7	W2	C	Shelf	C
	8	W2	C	Scale	T
	9	W1	NC	Shelf	K
	10	W1	NC	Scale	Z
	11	W2	NC	Shelf	U
	12	W2	NC	Scale	H
	13	W1	C	Shelf	M
	14	W1	C	Scale	S
	15	W2	C	Shelf	T
	16	W2	C	Scale	C
1st Session	17, 33	W1	NC	Shelf	H
	18, 34	W1	NC	Scale	U
	19, 35	W2	NC	Shelf	Z
	20, 36	W2	NC	Scale	K
	21, 37	W1	C	Shelf	M
	22, 38	W1	C	Scale	S
	23, 39	W2	C	Shelf	T
	24, 40	W2	C	Scale	C
	25, 41	W1	NC	Shelf	U
	26, 42	W1	NC	Scale	H

	27, 43	W2	NC	Shelf	K
	28, 44	W2	NC	Scale	Z
	29, 45	W1	C	Shelf	S
	30, 46	W1	C	Scale	M
	31, 47	W2	C	Shelf	C
	32, 48	W2	C	Scale	T
2nd Session	49, 65	W1	NC	Shelf	S
	50, 66	W1	NC	Scale	M
	51, 67	W2	NC	Shelf	C
	52, 68	W2	NC	Scale	T
	53, 69	W1	C	Shelf	U
	54, 70	W1	C	Scale	H
	55, 71	W2	C	Shelf	K
	56, 72	W2	C	Scale	Z
	57, 73	W1	NC	Shelf	M
	58, 74	W1	NC	Scale	S
	59, 75	W2	NC	Shelf	T
	60, 76	W2	NC	Scale	C
	61, 77	W1	C	Shelf	H
	62, 78	W1	C	Scale	U
	63, 79	W2	C	Shelf	Z
	64, 80	W2	C	Scale	K

3.1.2 Data Acquisition

The data collection process involved 15 healthy right-handed participants (8 males, 7 females, age: 28.6 ± 3.9). The participants are part of our research organizations, but none of them is directly involved in this research. Liguria Regional Ethical Committee approved the research protocol for this study (protocol 396REG2016 of July 25th, 2019), and all participants provided written informed consent to publish the collected data. The framework adopted during the experiment was designed to ensure a high degree of automation in the acquisition phase. Most sensors were directly interfaced with the YARP middleware [108], allowing timestamp synchronization, whereas the wrist-worn IMU was using a Robot Operating System (ROS)-YARP interface.

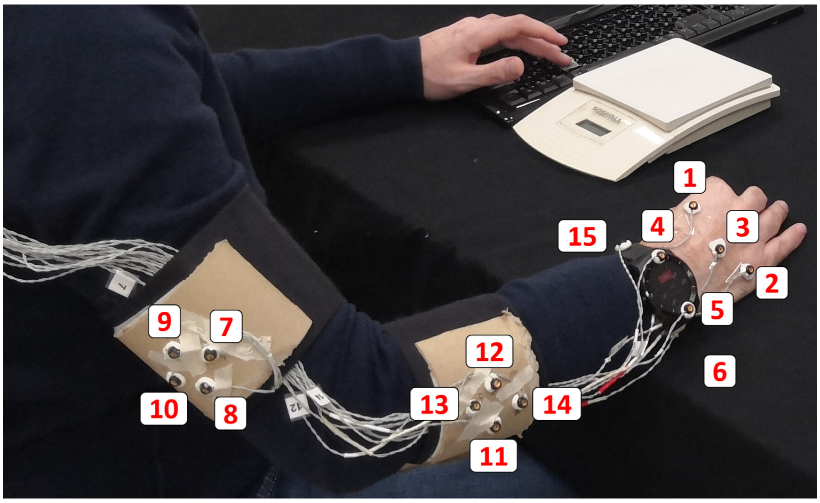


Figure 3.4 Marker positions on the participant's right arm and hand. The smartwatch equipped with inertial sensors is visible on the wrist.

The data have been segmented to separate each action presented in Figure 3.3. The MoCap and the IMU segmentation have been performed automatically, exploiting the participants' pressures on the key. Instead, the camera images were organized offline into separate folders according to the saved timestamps. For each participant, we saved a log file containing information on the experiment. These log files contain the YARP timestamp of each instruction communicated by the synthetic voice and the time for each key's pressures.

Motion Capture System As a motion capture system (MoCap), we used the Optotrak Certus®, NDI, with active infrared markers. In total, we recorded the signal from 15 markers. As pictured in Figure 3.4, the five markers on the hand were placed, respectively, on the metacarpophalangeal joints of the index and the little finger, on the diaphysis of the third metacarpal, and on the smartwatch in correspondence to the radial and ulnar styloid. Additionally, two markers were positioned on the watch strap, one per side, to better characterize wrist movements. Even though the main focus of the recording was to acquire hand and wrist motions, we decided to position a few markers on the participants' arm and forearm. We used two rigid cardboards where we put four markers each. See Figure 3.4 for reference. The frequency of the acquisition is 100 Hz. For every frame, the three-dimensional coordinates of every marker (in millimeters) were saved into the file associated with the trial. Moreover, the timestamp at the beginning and the end of each trial was saved. This was used to retrieve the timestamp corresponding to each frame by applying linear interpolation.

Inertial Sensors On the right wrist of the volunteers, we mounted an LG G Watch R smartwatch equipped with a 6-axis IMU. The sampling rate was 71 Hz. As for the MoCap data, a separate file was created for each trial, whenever the key on the keyboard was pressed at the end of the departing movement. The file was saved in `json` format containing the ROS and YARP timestamps at each sample, the internal Android timestamp, and the three components of the linear acceleration in m/s^2 and those of the angular velocity in rad/s. The data were published by the Android app on ROS for the smartwatch acquisition and then saved. Through the ROS-YARP interface, the corresponding YARP timestamp was sent to ROS at each key's pressure and written on the `json` file.

Cameras Two cameras with a resolution of 1920×1080 pixels were positioned in the room where the experiments took place. The former was placed at the back of the participant's chair and recorded the scene from an overhead viewpoint, remaining elevated by 130 *cm* with respect to the table. The latter was on the left side of the participants, with an oblique point of view, 65 *cm* higher than the table with a distance from the hand starting position of *circa* 140 *cm*. The reader is referred to the scheme in Figure 3.2 for reference. The frame rate was set to 30 Hz. The last sensor used in the experiment was iCub's left camera. The robot was located opposite the table, in front of the volunteer, with a complete perspective of the table and the shelves. The camera's resolution was 320×240 pixels, and the frame rate was 22 Hz. As previously mentioned, the images acquired with the cameras were not automatically segmented during the acquisition. Segmentation occurred afterward using the YARP timestamps, which were saved for each image from each camera in a log file. Indeed, each saved camera frame was associated with a YARP timestamp, making it possible to relate the acquired images with the events triggered by the key's pressure.

3.1.3 Data Records

The described dataset has been made available on Kaggle[1], while software utilities can be found on GitHub[2]. While we used MATLAB as a reference software to create the utility functions, we chose to make sensory data available in a non-proprietary format (`log`, `csv`, `json`, or `jpg`, depending on the sensor). The provided utilities allow to simply import and visualize the data of the desired sensors. A comprehensive table with a summary of performed movements, with the details about their direction and the properties of the object involved is

[1] https://tinyurl.com/5jzzf9p3

[2] https://github.com/lindalastrico/objectsManipulationDataset

available in the utilities, also with a machine-readable structure. Table 3.2 summarizes the characteristics of each trial. The same sequence was performed by each participant.

In the repository, the data are organized into separate folders on a sensor basis. Inertial and MoCap recordings can be found in folders `data/inertial` and `data/mocap`, respectively. While cameras recording are divided in three folders according to the following naming:

- `data/cam_0`: low resolution frontal images from the iCub left camera,
- `data/cam_1`: high resolution lateral camera,
- `data/cam_2`: high resolution images from behind.

The main folder also includes one folder `data/log`, which contains the experiments log files. They report the YARP timestamp corresponding to the key's pressures at the end of each trial, together with the information relative to the transport movement: the instruction given by the synthetic voice is reported as "sending speech: *Prendi/Metti* in *{Position letter}*". The verbs "Prendi" and "Metti" mean, respectively, "Take from" and "Put in" in Italian, and were used to tell participants where to take the glass from or where to put it back on the shelf, using the letter corresponding to the position, as in Figure 3.1d.

Each one of the described folders is organized into sub-folders named from `P001` to `P015`, containing the files associated with each participant. In the case of MoCap and inertial sensors, the files are sequentially named from trials 1 to 80.

Instead, the camera folders are divided, for each subject, into 80 subfolders (one for each trial), e.g., the folder `data/cam_1/P001/P001_Trial_001` contains the images from the lateral high resolution camera for the first trial of the first subject. The same sub-folders also contain a *data.txt* file reporting the correspondence between the YARP timestamps and image names for the specific trial.

The MoCap raw data are organized in `csv` files, with 62 columns and as many rows as the samples in the trial. The first column contains a progressive ID, and the second one the YARP timestamp for each sample. As previously mentioned, this timestamp was computed after the acquisition by linearly interpolating the timestamps indicating the start and the end of the trial. The remaining columns represent triplets of three-dimensional trajectories (x, y, z coordinates) followed by an additional column for each marker containing 0, if the marker was visible in that particular frame, or 1 if not. The order in which the markers appear in the files follows the numbering introduced in Figure 3.4, therefore the first triplet being the coordinates of the marker on the metacarpophalangeal joint of the index, the second triplet the marker on the joint of the little finger, and so on.

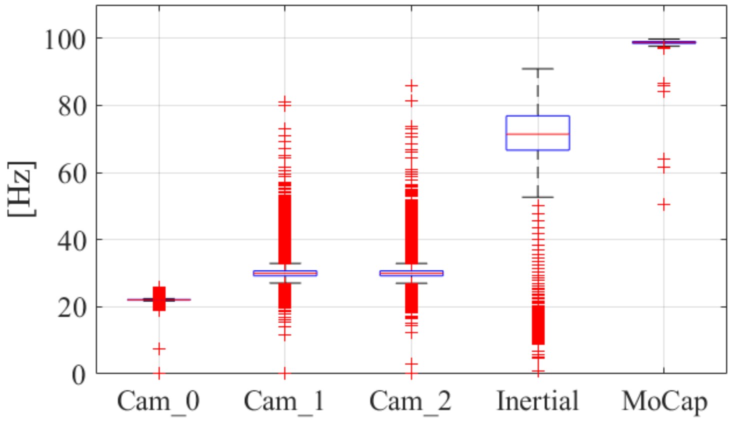


Figure 3.5 Box plots of each sensor recording frequency. The red lines represent the medians, the blue rectangles limit the 25th and 75th percentiles.

Regarding the inertial sensor raw files, they are saved in `json` format. For each sample, the available information is the ROS timestamp, the YARP timestamp, the Android one, and the three components of linear acceleration and angular velocity.

3.1.4 Data Assessment

A first evaluation of the quality of the provided data is related to their frequency. As already mentioned, the acquisition frequency depends on the specific sensor, and it is 100 Hz for the MoCap, 71 Hz for the inertial sensor, 30 Hz for the cameras, and 22 Hz for the robot camera. Figure 3.5 represents the frequencies for the sensors considering the whole experiment for all the participants. Even though some outliers are present, a possible solution could be downsampling the data to the lowest frame rate (22 Hz), allowing for automatically cleaning them and keeping a frequency still well above the range of human motion, which is generally below 10 Hz [81] and lies in the interval $[0.3, 4.5]$ Hz for hand motion [172]. In any case, in the dataset available online, the original sample rates are preserved.

An in-depth kinematics analysis can be carried out to study how the properties of the different objects affect the movement of the arm and condition how the transport action is completed. Our dataset allows not only to calculate different kinematic parameters but also to compare the information which can be retrieved from the different sensors. Proof of the richness of representations allowed by our dataset is shown in Figure 3.6. As an exemplification, we chose a W1-NC trial and a W2-C one, from the scale toward the upper shelf on the participant's left. The same pick-and-place action is represented by the 3D hand trajectory from the MoCap system, by the norm of the linear acceleration computed from

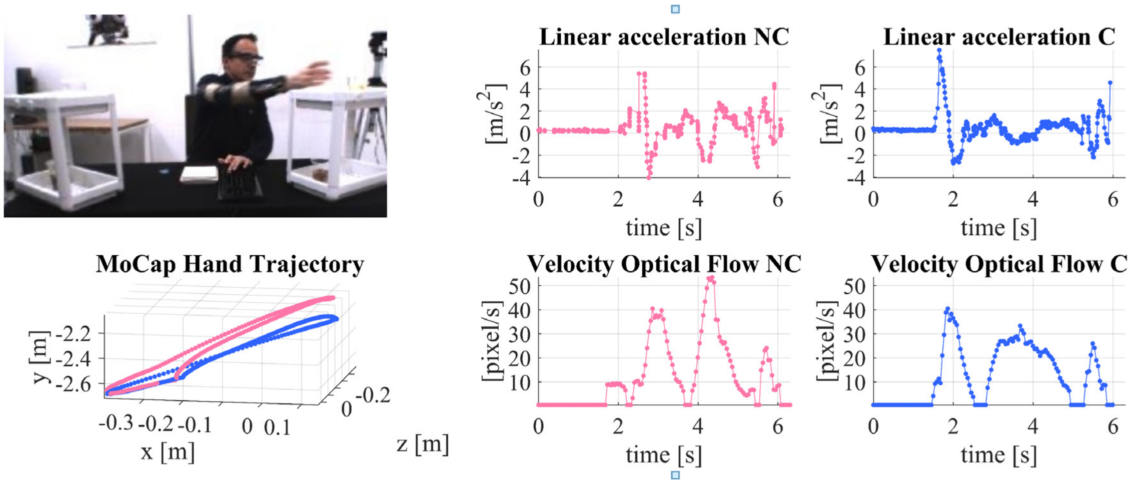
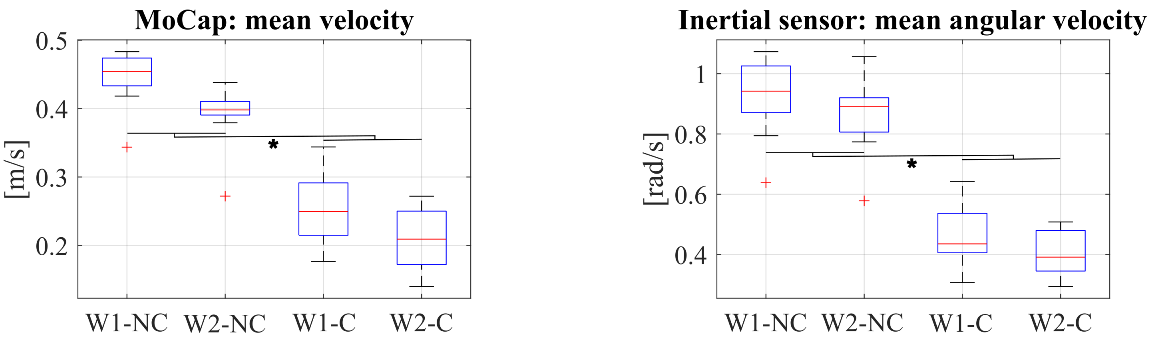


Figure 3.6 Synchronized data representations from MoCap, IMU sensor, and optical flow extracted from robot camera for two sample trials. The hand trajectories recorded with an infrared marker, the norm of the linear acceleration from the inertial sensor on the wrist, and the apparent velocity extracted from the optical flow, are shown together with the robot perspective on the scene.

the triaxial components of the IMU on the wrist, and by the velocity of the region in motion extracted by applying an optical flow algorithm to the scene recorded with the robot camera (see Chapter 4 and Section 4.1.1 for more details on Optical Flow computation).

Figure 3.7 compares instead the kinematics parameters extracted from the synchronized MoCap and inertial sensors during the transportation of the four glasses. In detail, in Figure 3.7a are shown the median, the 25th and the 75th percentiles of the hand velocity calculated deriving the 3D position of one of the markers placed on the volunteer's hand (markers 1 to 4, see Figure 3.4 for reference). According to the Kruskal-Wallis test for non-normal distributions, we found a significant difference in the velocity adopted for transporting the glasses between those filled with water (W1-C and W2-C) and those empty, with simply a weight difference (W1-NC and W2-NC). Even though a trend is visible, no significant difference was found concerning weight. The same results emerge in Figure 3.7b as well, where the wrist's mean angular velocities are recorded with the inertial sensor in the smartwatch. Again, a significant difference in the magnitude of the angular velocity appears between the careful and not careful transport motions. These findings can be qualitatively appreciated by the representation in Figure 3.8, with the hand velocities acquired through the Mocap system, as described before. To create such a global representation, the mean of the hand velocities of all the trials, separated into the four classes of motion, was computed for every time instant, together with its standard deviation. The reaching phase, represented by



(a) Mean velocity of the participants' hand while manipulating the four different glasses, acquired through active infrared markers.

(b) Mean angular velocity during the glasses transportation as recorded by the smartwatch on the participants right wrist.

Figure 3.7 Kinematics parameters during the transport movements for the four different classes, retrieved for motion capture data (3.7a) and inertial sensor data (3.7b). The red lines represent the medians, the blue rectangles limit the 25th and 75th percentiles, and $*$ indicates a significant difference according to the Kruskal-Wallis test.

the first peak of velocity, is uniform for all the glasses, either full, empty, heavy, or light; also the direction of the reaching action, which was performed towards the scale on the table in front of the participant or towards the different positions on the shelves, does not influence the hand velocity according to its standard deviation. The difference in how the glasses were manipulated is confirmed, as in Figure 3.7, between the glass with and without water, so for that movement feature that we interpret as carefulness. It was already clear (see Figure 3.7 that the speed of the movement was reduced in presence of water and also, to a smaller extent, by the weight. However, in Figure 3.8, it can be noted how not only the peak of velocity is reduced when transporting the W1-C and W2-C glasses, but also the shape of the velocity profile is modulated. Indeed, where for glasses without water the velocity is bell-shaped, the careful transport movements present an anticipated peak with a prolonged deceleration phase, especially for the full, heavy container. Participants tended to be especially cautious in the final phase of the movement, gently leaning the glass so as not to spill the content. In the initial phase of the movement, the derivative of the velocity is by far steeper for the empty glasses, indicating higher acceleration in moving the object only slightly modulated by the weight.

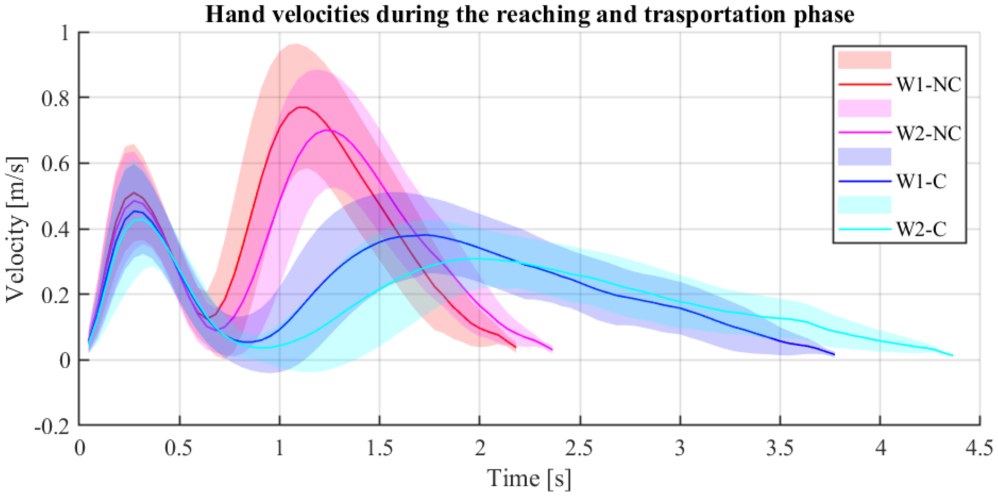


Figure 3.8 Hand mean velocities and standard deviation, in transparency, associated with the reaching (first peak) and transportation movements (second peak) of the four different glasses.

3.2 Applications

The presented multimodal dataset contains human object pick-and-place tasks under different experimental conditions. In particular, the dataset describes the effect of object weight and associated carefulness on human motions during a pick-and-place. The dataset is collected in a controlled environment, where the variability in the direction of the actions (height and regions of grasping and delivering) and the synchronization between multiple sensors is privileged. These qualities make the dataset useful for a series of possible applications. The data are acquired through multiple cameras with different perspectives, IMU, and MoCap sensors, thus providing the possibility of integrating and comparing the diverse sensing modalities. Each sensor presents some specific advantages that can be exploited depending on the context of application: the IMU is not affected by occlusions, whereas the motion capture represents the ground truth in the study of human motion, and the different resolutions offered by the webcams and the robot camera can represent a useful baseline for computer vision applications. For example, having data from multiple cameras and a motion capture system, can be used to study view-invariant action recognition or to compare marker-based and marker-less methods for motion analysis. It can also be exploited to build classification models to distinguish the weight and the associated care in the handling, namely carefulness, of an object or to discriminate between different actions (reaching for, grasping, transporting, dropping) and appropriate labeling is provided in this sense. The features of the manipulated items, for example, can be automatically detected by training models with the available data acquired during the transportation movements performed by humans, either relying on robot cameras, motion capture system, or inertial sensors. The human motions during the pick and place task can also be used as a reference for robot motion generation with the appropriate attitude, granting consistency with the context and the properties of the object and exploring the potentiality of robotic implicit communication. Furthermore, given the variety in the directions and elevations of the gestures, the dataset is suitable for creating predictive models of human behavior; for instance, to anticipate the direction of hand movements allowing the robot to react quickly and efficiently to human actions. Finally, the dataset could be used to evaluate humans' capability to anticipate action or understand the final goal or direction of an action when presented with a video picturing a portion of a motion.

Chapter 4

Observation of human movements to estimate objects properties

Elaboration and integration of two articles
Submitted to: International Journal on Robotics Research
Published in: International Workshop on Human Friendly Robotics (HFR 2020) [89], Best Paper Nominee

Given how much humans rely on implicit communication to interact with each other, in HRI a consistent effort is directed toward developing robots' ability to understand implicit signals and subtle cues that naturally characterize our movements. This becomes of critical importance for robots to support human partners in daily activities, especially in unconstrained environments as in manufacturing, helping human operators to lift loads, or assisting the elderly. One of the most common tasks we accomplish every day is interacting with objects. Humans can easily manipulate items they have never used before: at first, by inferring their properties such as the weight, the stiffness, and the dimensions, also from the observation of others manipulating them; at a later time, using tactile and force feedback to improve the estimation. If we consider a scenario where the robot acts as a partner, it acquires great importance to endow it with the ability to correctly estimate the characteristics of the handled objects; consequently, the robot can plan a safe and efficient motion action. Replicating this behavior in robotic systems is challenging, and a complete framework would require integrating vision, force, and tactile sensors, allowing the robot to inspect the object independently. Preliminary results have been achieved in estimating objects physical properties, relying on inference-based vision strategies [46, 140]. We propose an approach to automatically detect object features inspired by what happens in human-human interaction. Indeed, when we do not have direct access to the item, the most reliable source of information on its characteristics is watching another person lifting and transporting it. Not only that, observing how another person handles the object allows overcoming all those misconceptions related to the exterior appearance of the object, such as shape or size. In this chapter, I present how a robot can automatically assess an object's features just by seeing it transported by a human partner. Inferring those properties from the human kinematics during the manipulation of the objects, rather than from their appearance, grants the ability to generalize over previously unseen items. Of course, the robot could refine its assumption in a second moment by directly touching the item. Still, this approach mediated by the partner's actions is safe, immediate, and inspired by human behavior. In particular, I focus on the features of human motor actions that communicate insights on the *weight* of the object and the *carefulness* required in its manipulation. From the dataset presented in Chapter 3, I explore how the different sensors can be used to detect the mentioned properties offline, extracting information on the human kinematics and exploiting machine learning algorithms to perform the classification. This represents the first step towards automatically detecting the degree of care required in object handling and the discrimination of the item weight on a robotic platform, during a dyadic interaction. Indeed, multiple steps need to be accomplished to have the robot discriminate the carefulness or the weight, possibly simply by using its camera to observe the human, granting the most natural interaction. In the first place, it requires to be assessed which kinematic features are obtainable from the different sensors, how they can be extracted, and their reliability and quality with respect to the task. Then, different machine learning methods will be tested, and the classification outcomes will be discussed. According to the findings, the online implementation of the framework will be presented in the next Chapter.

Research questions and objectives

Suppose to transport a glass full to the brim with water: safely manage it without spilling a drop requires accurate planning in terms of grasping strategy, orientation and velocity adopted. If we want a robot to perform the same action, the first step would be to give the robot the capability of recognizing the intrinsic difficulty of the task; if we consider a hand-over task between a human and a robot, the latter should be aware that it is about to receive an object that requires a certain degree of carefulness in the handling. Moreover, assessing the object's weight would allow an efficient lift. These features could be estimated from the human motion and ideally should be available before the end of the observed action, to trace the human abilities and to allow the robot to prepare for the possible handover.

Differently from the weight, the concept of carefulness is not trivial. Previous studies have dealt with delicate objects but focused more on robotic manipulation: the difficulty in the addressed tasks was given by the item's stiffness or deformability; tactile sensors were used for estimating the necessary force to apply a proper grasp [157, 139]. Our study considers the carefulness necessary to move an item from a broader perspective. Indeed, not only the object's stiffness but also its fragility, the content about to be spilled, or its sentimental value may lead a person to perform a particularly careful manipulation. In those real-life cases, we would like the robot to successfully estimate the carefulness required just by observing the human kinematics. Using the dataset presented in Chapter 3, which consists of transportation movements involving glasses that vary in weight and level of care, we developed kinematic features representative of human motion. These features were used to train classifier algorithms. We obtained multiple features from a motion capture system, a robot camera, and an inertial sensor worn on the wrist. The goal is to determine which sensing approach is best suited for understanding the properties of objects from human movement and whether fusing information from different sensing modalities can help. To do this, we conducted an offline comparison of different classification approaches and sensor performances, with the ultimate goal of implementing these techniques online on the robot. The questions we aim to answer are: ***(H1)*** can we detect the presence or absence of care associated with the object transport, and ***(H2)*** can we detect the two levels of weight that the glasses can have?

4.1 Methods

These studies aim at evaluating and comparing the performances of different sensors and classifiers to infer the weight and the carefulness required during the transport motions of four different glasses. We focus on transport motions because of the direct influence of the objects' physical characteristics, given by the contact presence, on how the action is performed. We begin by evaluating how the same classifier model performs in discriminating the carefulness with various sources of data used for the training, comparing motion capture system, inertial sensor, robot camera, and their combination. Then, we compare the performances of different classifiers trained on the same subset of features extracted from specific sensors.

4.1.1 Data pre-processing

The data presented in Chapter 3 needed some pre-processing before being used as training data for the object properties classifier. The following paragraphs will present the specifics of each sensor data and the pre-processing techniques which involve segmentation, feature extraction, data filtering, resampling, data normalization, and padding.

Motion capture system data The data acquired by the motion capture system consisted of the tridimensional coordinates of each marker with respect to the coordinate reference frame of the Optotrak. Occlusions limited the MoCap visibility for specific parts of the human movement. In our experiment, the main source of occlusion was given by the presence of the shelves, in particular for the lower right positions. To partially overcome this problem, after a preliminary analysis, we chose to consider for each trial the most visible marker among the subset of four placed on the hand of the participants as representative of the movement. Two different cubic interpolations, inpaintn [56] and interp1 of MATLAB ver. R2019b, have been used to reconstruct the data missing because of the occlusions, respectively for the initial part of the trials and the central one. The data was filtered with a second-order low-pass Butterworth filter with a cutoff frequency of 10 Hz. Some trials have been excluded from the data set because of inconsistencies in the segmentation among the acquired sensors or because of errors of the subjects in pressing the key at the right moment, i.e. when their right hand was lying on the table in the resting position. Overall only 1.25% of the total acquired trials have been removed. Since our hypothesis is that it is possible to distinguish the features of the object being transported, it was necessary to isolate the transportation movement in every trial. To do so, we took advantage of the experiment design. Indeed, as shown in

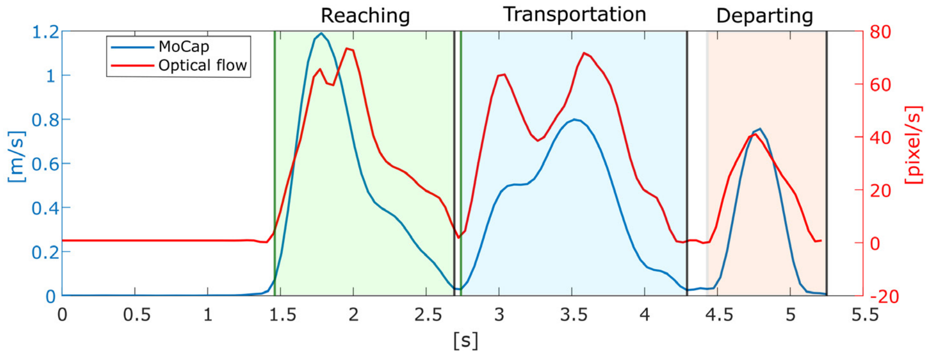


Figure 4.1 Example of the velocity patterns from motion capture (in blue) and optical flow data (in red). The peaks characterizing the three phases of the trial (reaching, transportation and departing) are visible.

Figure 4.1, each trial contains a pick-and-place action divided into three identifiable phases: a reaching action, from the resting pose to the position occupied by the glass (either on the shelf or on the scale), a transportation movement and finally the departing. At the beginning and the end of the transportation, the subject hand stops to grasp and release the object. Our segmentation assumed that the start and end of the transportation phase are associated with a lower norm velocity of the hand, with a peak in between. Therefore, the segmentation was performed by placing a threshold of 5% on the peak of the norm of the velocity extracted from the MoCap data, after filtering it with a fourth-order filter with a cutoff frequency of 5 Hz. The timestamps corresponding to the onset and the offset of the transport actions were saved to later segment the inertial sensors and camera data. After isolating the transport motion of the glass in every trial, we chose to compute the first and second derivative of the marker tri-dimensional trajectory, obtaining the hand triaxial linear velocity and acceleration.

Camera data and optical flow extraction As motion descriptor, from the saved raw images of the robot camera (see Fig. 4.2 for an example) we chose to compute the Optical Flow (OF), following an approach already tested [165]. In this method, the optical flow is computed for every time instant using a dense approach [49], which estimates the apparent motion vector for each pixel of the image. The magnitude of the optical flow is thresholded to consider only those parts of the image where the change is significant. A temporal description of the motion happening in the derived region of interest is then computed by averaging the optical flow components. On the velocity extracted, a second-order low-pass Butterworth filter with a cutoff frequency of 4 Hz was applied to remove the noise (see Fig. 4.1). The data described by the optical flow was segmented using the timestamps related to the start

(a) View from the iCub perspective (b) OF moving towards the right of the image (c) OF moving towards the left of the image

Figure 4.2 Example of iCub view of the scene and the extracted OF. The colors codify for the direction of the movement: red is for motion towards the right part of the image (4.2b), blue for motion towards the left (4.2c).

Table 4.1 Motion features computed from motion capture (u, v, z components of the velocity) and optical flow data (u, v components of the velocity)

Motion feature	Analytical expression
Tangential velocity	$\mathbf{V}_i(t) = (u_i(t), v_i(t), z_i(t), \Delta_t)$
Tangential velocity magnitude	$V_i(t) = \sqrt{u_i(t)^2 + v_i(t)^2 + z_i(t)^2 + \Delta_t^2}$
Acceleration	$\mathbf{A}_i(t) = (u_i(t) - u_i(t-1), v_i(t) - v_i(t-1), z_i(t) - z_i(t-1), 0)$
Curvature	$C_i(t) = \frac{\lVert \mathbf{V}_i(t) \times \mathbf{A}_i(t) \rVert}{\lVert \mathbf{V}_i(t)) \rVert^3}$
Radius of curvature	$R_i(t) = \frac{1}{C_i(t)}$
Angular velocity	$A_i(t) = \frac{V_i(t)}{R_i(t)}$

and end of the transportation, obtained from the motion capture system. Indeed, even if an independent segmentation of the movement from the optical flow was achievable, we preferred to use the ground-truth data to avoid introducing noise and variability for the sake of comparison.

From the optical flow, it was then necessary to further pre-process the data to obtain a motor descriptor of the action. As set of features representing the region in movement, we chose: the velocity $V_i(t)$, the curvature $C_i(t)$, the radius of curvature $R_i(t)$ and the angular velocity $A_i(t)$. Their analytical expression is stated in Table 4.1 and two components of the velocity on the screen (horizontal u and vertical v) were computed. As shown in [165], these data representations have been successfully used to discriminate online between biological and non-biological motion and to facilitate coordination in human-robot interaction [130]. In addition, kinematics properties, such as velocity, have been shown to be relevant in human perception of object weight [16]. The proposed descriptors can be updated at every time instant, and allows to progressively gather an increasing amount of information about the observed movement. This would then grant the robot the ability to discriminate online the characteristics of the object handled by the human partner.

Inertial sensor A smartwatch equipped with a 6-axis IMU provided inertial data at a sampling rate of 71 Hz. As for the images from the robot's camera, the transport action was isolated by exploiting the YARP timestamps retrieved from the threshold segmentation applied to the motion capture data. As features for later training the classifier, we used the raw data, i.e., linear acceleration and angular velocity. The resulting sequences are then scaled using min-max normalization to decrease the difference in scale between the features. For the IMU, we selected as maximum and minimum values the full-scale range of the sensors, i.e., ± 2g for accelerations and $\pm 8.73\,\mathrm{rad/s}$ (± 500 deg/s) for the angular velocities.

We applied to every temporal sequence from the 3 considered sensors a first-order Butterworth filter with a threshold frequency equal to the original sampling rate of the sensor (i.e., 71 Hz for the IMU, 100 Hz for the MoCap, and 22 Hz for the camera). To simplify the comparison and easily combine information from the different sources, we resampled the IMU and MoCap data to match the camera sampling frequency. This choice also finds support in previous research suggesting that 20 Hz is an ideal sampling frequency for the perception of human daily activities [7]. To perform the resampling, we interpolated the data and used the camera timestamp to extrapolate the new data.

Due to the characteristics of the glasses and the intrinsic variability of the task, the duration of the transport movement varied consistently among the trials (i.e. the duration of the movement is consistently longer when the moved glass is full of water, belonging to the high carefulness class). For this reason, to exploit the ability of certain models to handle temporal sequences of different lengths and use batch training, we adopted common zero-padding and masking techniques. Therefore, the shorter temporal sequences were completed with zero values, which were then ignored during the training, while the length of the longest transport movements was preserved. In particular, we used pre-padding since it is considered more robust to the noise introduced by the zeros [47]. Using a padding technique allows also for a smoother transition to an online application, where the beginning and the end of the transportation phase are not available a priori, making the re-sampling of the time-series difficult to handle.

The actions of the dataset presented in Chapter 3 involved four identical plastic transparent cups which differed in weight and attention required to be moved. Two glasses were full of water to the brim, requiring carefulness, and weighted either 167 *gr* (W1-C) or 667 *gr* (W2-C). As we have detailed before some sequences had to be removed for inconsistencies in the segmentation. This led to a slightly unbalanced data set, containing more examples for specific classes. Indeed, class W1-NC had 235 sequences, class W2-NC 239, class W1-C 238, and class W2-C had 236. Although cardinally the difference is minimum, to preserve the balance of the dataset we decided to fix the maximum number of sequences for each class to 235 and we have randomly selected the sequences for W2-NC, W1-C, and W2-C. Notice that the four classes were characterized only by the weight and the carefulness level.

Table 4.2 Summary of the implemented classifiers. For each model, are indicated the sensors used as data source for the training set and the corresponding time-series considered as features.

	Model	**Classification**	**Data source**	**Features (time-series)**	**#**
Classifier 1	LSTM	Carefulness	MoCap	Triaxial velocity and triaxial acceleration	6
			IMU	Triaxial angular velocity and triaxial linear acceleration	6
			OF	Curvature, Radius of Curvature, Velocity, Angular Velocity	4
			IMU + OF	Combination	10
Classifier 2	CNN-LSTM-DNN	Carefulness or Weight	MoCap OF	Curvature, Radius of Curvature, Velocity, Angular Velocity	4
Classifier 3	LSTM-DNN	Carefulness or Weight	MoCap OF	Curvature, Radius of Curvature, Velocity, Angular Velocity	4

Therefore other variables, such as the initial and final position of the glass and the direction of the movement, are not considered in the classification.

4.1.2 Classifiers

The first attempt to tackle the detection of objects' features from implicit cues focused on carefulness. It involved a simple model, *Classifier 1*, which learned from the three available sensors or their combination. After preliminary findings, we investigated two possible architectures in more detail, trying to discriminate not only the carefulness but also the weight of the glasses. *Classifier 2* relied on re-sampled features, while *Classifier 3* used the original data with variable lengths and padding techniques. As for the data sources, we compared the motion capture system as ground truth and the robot camera, which would be the preferred choice in an online interaction, being available on the robot and not invasive at all. Moreover, to better compare the performance achieved by *Classifiers 2 and 3* and have more precise control over the information used during the learning process, we decided also to uniform the features extracted from the two sensor typologies and used in the training phase. From the robot camera images, we used the four features extracted from the optical flow on the image plane: the norm velocity, the angular velocity, the curvature, and the radius of curvature. The same subset was also computed from the hand tri-dimensional trajectory recorded with the moCap system, where the third component z was considered (see Table 4.1 for reference). Refer to Table 4.2 for a comprehensive overview of the implemented classifiers, the data used to train them, and the selected features.

Classifier 1 **– Long-Short-Term-Memory model** We trained the same binary classifier with multiple training set: one for each sensor (i.e., camera, IMU and MoCap) and the fourth by using both IMU and camera data. The idea of combining IMU and camera data arose to determine if an autonomous robot could leverage its standard perception system, the camera, with a wearable sensor, unaffected by occlusions, for a more reliable perception. Previous research has suggested that Long-Short Term Memory (LSTM) networks are effective models for learning long-term temporal dependencies and dynamics; this is because they are a type of Recurrent Neural Network (RNN) that can handle a full sequence of data and are well-suited for classifying and processing time-series data [40, 69, 117].

We designed a simple model implemented in Python using layers provided by Keras[1] libraries. The LSTM layer has 64 hidden units and an input shape equal to $[sequence\_length \times n\_features]$, where the sequence length is fixed to 132 samples (maximum sequence length after padding the data). The number of features varies according to the data source: 6 for MoCap (triaxial linear acceleration and velocity) and IMU (triaxial linear acceleration and angular velocity), 4 for the camera (norm of the velocity, angular velocity, curvature and radius of curvature, see Table 4.1) and 10 for IMU plus camera. The following layer is fully connected, with 32 neurons, and it is preceded and followed by dropout layers with a value of 0.5. The output layer is another fully connected one with two output neurons corresponding to the two classes: careful and not careful. Given the double output, we chose a softmax function for the activation and the categorical cross-entropy to evaluate the loss. An L1-L2 kernel regularization was added to the last layer to prevent the model from overfitting with $L1 = 0.001$ and $L2 = 0.001$ as parameter values. The chosen optimization algorithm was AdamOptimizer, with a learning rate of 0.0002 and batch size of 16.

For each of the four models, we carried out the training and testing phases by adopting the Cross-Validation with a Leave-One-Out approach to test the ability of the model to generalize over different participants. Therefore, to split the 1200 sequences ($15\ participants \times 80\ sequences$) into training, validation, and test sets, we adopted the following procedure. One at a time, the data of each participant were used as a test set, and the remaining 14 were further divided, 80% for the training and 20% for the validation. The validation set has been picked randomly from the 14 volunteers. The models have been trained for 100 epochs using an early stopping on the validation loss with patience set to 5 epochs to avoid overfitting.

[1] https://keras.io/

Classifier 2 – Convolutional, Long-Short-Term-Memory and Deep Neural Network

The combined use of Convolutional Neural Network (CNN), Long-Short Term Memory (LSTM) and Deep Neural Networks (DNN) is a good solution for classifying time-dependent data, such as speech or motion kinematics [137, 117]. This model was inspired by [137] and consisted of two time distributed 1-D convolutional layers (that took as input 4 subsequences of 8 frames each), a max pooling and flatten layers, a 100 neurons LSTM, a 100 neurons Dense layer, and a 2 neurons output layer with a sigmoidal activation function. A Leave-One-Out approach was adopted, to test the ability of the model to generalize over different participants. Thus, for each one of the 15 folds, the data from 14 subjects were used as training set and the data of the fifteenth participant as test set. The 20% of the data for each training set was kept for validation, and early stopping was implemented according to the validation loss function (with a patience parameter set to 5 epochs): this allowed to obtain good accuracy without incurring in overfitting. The batch size was fixed to 16. The model was fit with ADAM optimization algorithm and categorical cross-entropy as loss function. With respect to the model described in [137] some regularizers were added to avoid overfitting and make the network less sensitive to specific neurons weights. A L1-L2 kernel regularization was added to the two 1D convolutional layers $(l1 = 0.001, l2 = 0.002)$ and a L2 kernel regularizer $(l2 = 0.001)$ was added to the fully connected DNN layer; moreover, 0.5 dropouts were introduced. Since the required masking layer was not supported by the Keras implementation of the CNN layer, we resorted to re-sample to obtain sequences with the same number of samples for each trial. We used the interp1 function of MATLAB and the number of samples was selected considering the class associated with the shorter duration of the transport phase, W1-C, and computing the median value among all its trials. The resulting value was 32. Therefore, the dataset to train this model was composed of two data structures: one derived from the MoCap data and the other one from the OF. Both structures had dimensions $940 trials \times 32 frames \times 4 features$

Classifier 3 – Long-Short-Term-Memory and Deep Neural Network

This model strictly resembles *Classifier 1*, with a 64 neurons LSTM, followed by a 32 neurons dense layer and a 2 neurons output layer with a sigmoidal activation function; also in this case L1-L2 regularization and 0.5% dropout were used to avoid overfitting. The optimization algorithm, the loss function and the validation approach with early stopping were the same as before. We used zero-padded data in a structure: $940 trials \times 132 frames \times 4 features$. This model grants the possibility of learning independently from the length of the temporal sequence and that is why it is suitable to be implemented on the robot; indeed, no previous knowledge of the duration of the movement would be required to perform the classification, since the model is trained on variable temporal sequences.

4.2 Results

The results obtained for the carefulness and the weight detection are presented separately, commenting on the performances obtained by the tree classifiers.

4.2.1 Carefulness level

We trained the model presented as *Classifier 1* separately on kinematic features extracted from the motion capture system, optical flow, inertial sensor, and a combination of the last two. The results for these four classifiers are shown in Figure 4.3 using box plots that represent the accuracy in classification and show the median, average, and distribution of values for each data source. The overall performance of the models is comparable, with 91.3% for the IMU, 91.4% for the camera, 91.6% for the MoCap, and 91.1% for IMU plus camera. This suggests that the four data sources are nearly equivalent for classification purposes. However, if we exclude outliers, the minimum values shown in the chart seem to indicate some differences between the models. The lowest accuracy achieved using camera and MoCap data is around 84%, while the lowest accuracy achieved using IMU data is approximately 82%. The combination of IMU and camera data had the lowest accuracy at 75%. This suggests that combining IMU and camera data may not be particularly effective for estimating motion carefulness, as it increases the variability of the classification. However, all the single classifiers achieved high performances, and it may be possible to use an autonomous system that combines the results of different classifiers with a voting system, to obtain more stable and reliable inferences. The accuracy obtained with the optical flow was particularly promising, as it would allow for property detection during one-to-one interaction without requiring the human to wear any sensors, enabling natural and spontaneous collaboration.

The results obtained with *Classifier 1* confirmed already the possibility of discriminating ***(H1)*** the carefulness level associated with the object offline; we then assessed two additional classifier models to understand how the same set of interpretable kinematics features, extracted from different sensors, would impact on their performance. Indeed, for comparison purposes, all of the following results are based on the same set of kinematic features (see

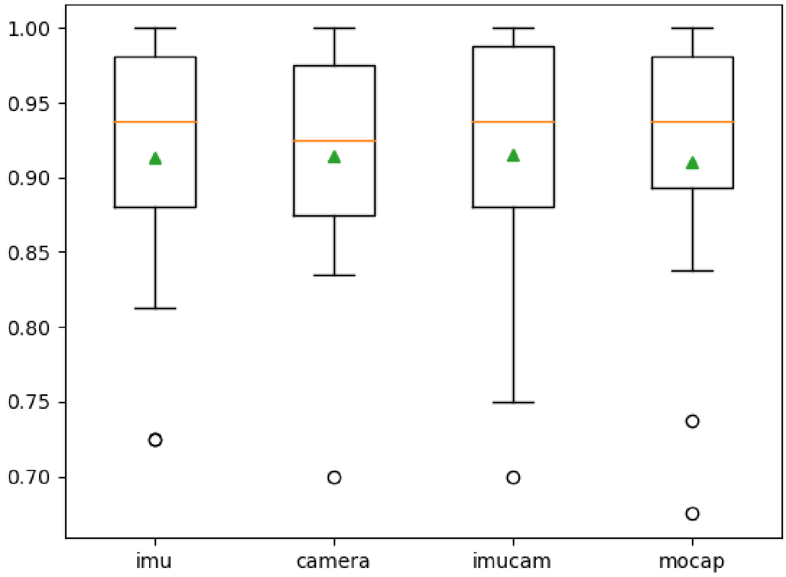


Figure 4.3 Boxplots expressing the results of *Classifier 1* trained with data coming from different sensors and combinations in terms of accuracy. The yellow line represents the median, while the green arrow indicates the average.

Table 4.1) extracted from the optical flow or the motion capture system, which we considered as ground truth.

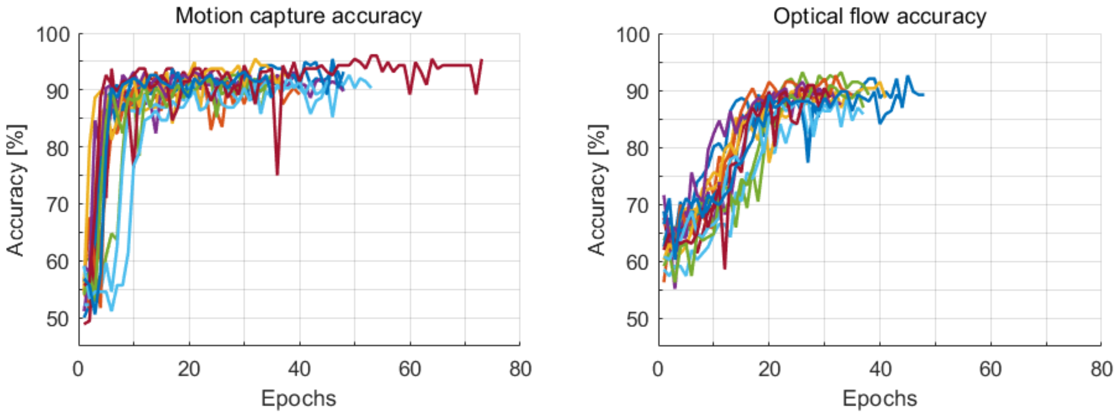
Table 4.3 shows the performance of *Classifier 2*, the Convolutional LSTM model, in detecting the level of care when trained on resampled time-series data. When performing the Leave-One-Out cross validation we noticed that the classification accuracy on the test set associated with volunteer 8 was significantly lower than the average accuracy obtained in the remaining folds with other participants' data as test set (**MoCap test** *without vol8:* 91.68 ± 5.00, *vol8:* 51.62; **OF test** *without vol8:* 90.54 ± 6.46, *vol8:* 77.42). Examining the experiment videos we noticed that volunteer 8 was particularly careful even when handling the glasses not containing water. Our impression was confirmed after computing the median duration of the not careful movements among the subjects. The duration for volunteer 8 (2.04 ± 0.18 seconds, median and median absolute deviation) differed significantly from the ones of the other participants, as for the rest of the group the median duration resulted in 1.47 ± 0.15 seconds (Kruskal-Wallis test: $\chi^2(14, N = 480) = 136.8$, $p < .01$). In Table 4.3 we have reported in brackets the results when including this subject in the training set and using its data as test set. As can be observed, when this outlier was included, the classification

Table 4.3 Model accuracy (%, mean and standard deviation) on carefulness level classification with *Classifier 2*, the CNN-LSTM-DNN model. In brackets are the results when volunteer 8 was included in the data set

	Motion capture	**Optical flow**
Training	$92.15(92.00) \pm 2.14(3.42)$	$94.03(92.18) \pm 1.05(1.00)$
Test	$91.68(90.97) \pm 5.00(11.12)$	$90.54(89.43) \pm 6.56(7.59)$

accuracy was lower, and the variance on the test increased significantly for each of the sensing modalities.

Figure 4.4 shows the trend of the accuracy over the epochs for the validation set of each one of the folds. Comparing the graphs for the two sources of data ((a) motion capture, (b) optical flow) it can be noticed how the first one reaches an accuracy above the 80% in less than 10 epochs, while, using the features from the optical flow, more training is necessary to reach the same level of accuracy (over 20 epochs). Furthermore, the accuracy trend of the motion capture features is more stable.



(a) MoCap accuracy

(b) OF accuracy

Figure 4.4 Carefulness classification accuracy over epochs on the validation set for each one of the 15 folds, obtained with *Classifier 2* based on a CNN-LSTM-DNN model. Accuracy from motion capture (4.4a) and from optical flow features (4.4b).

Similarly, the performance in detecting the carefulness of *Classifier 3*, a simpler LSTM model fed with the original temporal sequences of variable lengths, is shown in Table 4.4. As before, the variability in the test accuracy was reduced when volunteer 8 was excluded from the dataset, and the overall accuracy improved for both the sensing modalities. With this model, compared to the values in Table 4.3, the accuracy achieved with the MoCap data is higher, while the one of the OF is slightly reduced.

Table 4.4 Model accuracy (%, mean and standard deviation) on carefulness level classification for *Classifier 3*, based on a simpler LSTM-DNN model. In brackets the results considering volunteer 8

	Motion capture	**Optical flow**
Training	96.57(94.32) ± 1.19(1.77)	92.10(90.39) ± 4.58(2.56)
Test	95.17(92.66) ± 5.46(8.49)	88.38(86.50) ± 8.68(10.75)

4.2.2 Weight

We used the same models to investigate the ability to discriminate between the weight level, again training them with resampled data (for *Classifier 2*) or padded data (for *Classifier 3*) from the robot camera or hand tracking markers. In Table 4.5 are shown the results for the classification of the weight achieved with re-sampled data on *Classifier 2*. In this case, volunteer 8 did not present any peculiarity and was therefore included in the dataset. As we can observe in Table 4.5, the accuracy with the motion capture data is above 60% and is higher than the one obtained from the optical flow.

We have noticed that, despite adopting the same approach, the accuracy of the weight classification is not as satisfying as the one achieved for carefulness. A possible explanation of these results could be related to the different effects of weight on different transport movements. Possibly the weight influence varies if the transportation is from top to bottom or vice-versa. Furthermore, the presence of water in some of the glasses may have led the subjects to focus mainly on the carefulness feature, unconsciously overlooking the weight difference. Therefore, we add two specifications of the second hypothesis ***(H2)*** about the feasibility of detecting the two levels of weight: ***(H2.1)*** the influence of the weight during the transportation is dependent on the trajectory of the motion; ***(H2.2)*** when an object is associated with a high level of carefulness, the weight has a limited influence on the transportation movement. Both hypotheses were tested with *Classifier 2*, which gave better results for the weight classification. Concerning the first hypothesis, we reduced the variability in the movements and tried to discriminate the weight in the subset of transport movements from the scale towards the shelves (**MoCap:** *Tr*: 68.90 ± 2.68 *Test*: 63.42 ± 8.96; **OF:** *Tr*: 59.10 ± 4.27 *Test*: 55.17 ± 6.24); there is a slight improvement for both the data

Table 4.5 Models accuracy (%, mean and standard deviation) on weight classification for the CNN-LSTM-DNN model (*Classifier 2*), fed with re-sampled data, and for the LSTM-DNN model (*Classifier 3*), trained on zero-padded data

Model		**Motion Capture**	**Optical Flow**
Classifier 2	*Training*	64.10 ± 2.34	55.24 ± 2.37
	Test	61.83 ± 7.16	54.47 ± 4.29
Classifier 3	*Training*	54.95 ± 2.66	55.30 ± 1.95
	Test	54.75 ± 5.27	53.29 ± 3.59

sources compared to the values in Table 4.5. Notice that the trajectories still have a discrete amount of variability since the position to reach on the shelf could be left or right, high or low. The second hypothesis was investigated by testing the weight discrimination within the subset of objects which required the same carefulness level: low (**MoCap:** *Tr*: 64.49 ± 5.24 *Test*: 61.93 ± 6.86; **OF:** *Tr*: 62.52 ± 3.53 *Test*: 56.84 ± 6.77) or high (**MoCap:** *Tr*: 62.72 ± 3.65 *Test*: 59.03 ± 8.73; **OF:** *Tr*: 57.92 ± 1.31 *Test*: 53.48 ± 7.63). The results for both tests are inconclusive since the classification accuracies have not changed much with respect to the ones reported in Table 4.5. It should be noted though that the dimension of the dataset used to validate hypotheses ***(H2.1)*** and ***(H2.2)*** halved, which has an impact on the statistical relevance of the results.

4.3 Discussion

In this chapter, we introduced various methods for detecting the properties of transported objects. While the sources of information and the specific classifiers may vary, the underlying concept behind our approach is the ability to infer characteristics of the objects indirectly, by observing the spontaneous changes and modulation in hand kinematics during the action.

The first attempt to classify the carefulness in the observed actions was successful for all the tested sensors as sources of information, with scores above 90% as reported in Figure 4.3. Differently from what was hypothesized, the combination of features from IMU and robot's camera did not improve the classification accuracy of *Classifier 1*, leading instead to an increase in its variance. However, when using one sensor at a time, the performance was totally comparable.

Two classifiers, one based on a CNN-LSTM model trained with resampled features (*Classifier 2*) and one on a LSTM trained with padded sequences (*Classifier 3*), were then tested to separately detect the care and the weight. In this case, we used the same time-series (velocity norm, angular velocity, curvature, and radius of curvature) extracted from the optical flow and the moCap for a better comparison of the models. Regarding the carefulness feature, as reported in Table 4.3, *Classifier 2* can correctly discriminate if the transportation of the object requires carefulness or not, independently from the sensing modality used. Considering the performance of the data coming from the two sources, no significant difference is detected between them. Therefore, not only by using an accurate system, such as motion capture, that integrates sensory inputs from different locations to estimate the position in the space of the target but also using the camera of the robot (single point of view), it is possible to extract features to discriminate between careful and not careful motions. Figure 4.4 shows insight on how the learning process advanced for the two data sources. Even though the final performances are comparable, it can be appreciated how the model trained with the features from the motion capture converges quicker to an accuracy value above 80%.

The approach adopted with *Classifier 3* is more general, in the sense that data are not re-sampled to have the same dimension. Still, the variability in their duration is taken into account. Even though this model is simpler, with just one LSTM and one dense layer, the performance on the carefulness classification considering the MoCap data increased (see Table 4.4 for reference). Although the accuracy using the optical flow is slightly lower, we consider this a promising step toward implementing the same algorithm on the robot.

The accuracy achieved for both sensing modalities and both models in discriminating between weights was lower than the accuracy achieved for detecting the level of care (see Table 4.5 for reference). In Section 4.2.2, we formulated two additional hypotheses to explain this outcome. ***(H2.1)*** was inspired by [150], where it was proposed that the vertical component of velocity during the manipulation of an object is perceived by humans as informative about its weight. Since the trials in our dataset explored a variety of directions and elevations, this introduced a great deal of variability in the vertical component of velocity. ***(H2.2)***, on the other hand, was based on the idea that the greatest challenge for the volunteers during the experiment was to safely handle glasses full of water, and that the difference in weight between the objects was not as noticeable as the stark contrast between objects with and without water. As mentioned in Section 4.2.2, *Classifier 2* was tested against these hypotheses, but no significant improvements in accuracy were achieved. Based on the results of our experiment, we cannot validate hypothesis ***(H2)***. Still, since we have only explored a subset of possible kinematic features, we cannot argue against it either. Objects can have multiple concomitant features, which do not always have the same effect on kinematics. It may be the case that the interaction between two or more of these features leads to the attenuation of their effect as compared to when considered individually. The proposed approach supports research in human-robot interaction, particularly in the context of addressing complex situations in realistic settings (e.g., industrial environments, construction sites, home care assistance, etc.). In these situations, the robot can autonomously leverage insights inferred from implicit signals, such as the level of care required to move an object, to facilitate cooperation with the human partner. Moreover, in a collaborative setting, human actions can become even more expressive of the object properties to favor mutual understanding, thanks to the phenomenon of signaling [41, 125], and this may improve the classification performance.

Given the promising results obtained offline in detecting the level of care in actions, in the next Chapter we will present an online application of the same approach, using the robot camera and a LSTM model trained with padded sequences. We will not address the problem of weight classification, as the available dataset is mostly representative of care in motion.

Chapter 5

Online recognition of carefulness from human kinematics

Elaboration of an article published in:
International Conference of Social Robotics (ICSR 2021) [88]

When manipulating objects, humans fine-tune their motions to the characteristics of what they are handling. This allows an attentive observer to infer hidden properties of the manipulated object, such as its weight, temperature, and even whether it requires special care in manipulation. This chapter presents a step towards endowing a humanoid robot with this capability. Specifically, we study how a robot can infer online, from vision alone, whether or not the human partner is being careful when moving an object. We define carefulness as the caution and attention that humans exercise when handling an object. This qualitative property is influenced both by the object's physical characteristics (e.g., fragility) and by other factors such as emotional attachment or economic value. For example, consider a robot that is asked to receive a glass of water from a human. It should recognize the human's carefulness in manipulating the glass without spilling water. The ability to promptly recognize the level of care in a human's movement by observing their actions can allow robots to adapt their actions accordingly, making them safer and, in the future, display the same level of care as their human partners. Carefulness has been explored in studies of human-human handovers to teach robots how to correctly transfer objects (e.g., [140, 45]), using motion capture sensors to monitor human movements. In Chapter 4, we demonstrated that it is possible to train a classifier to distinguish between *careful (C)* and *non-careful (NC)* human motions using data from a low-resolution camera like the one on a robot. However, our carefulness recognition method was tested offline on precisely segmented data in a single experimental scenario. In this chapter, we will present: (i) an online implementation of our method for carefulness recognition, (ii) a study to demonstrate its online performance, and (iii) a study to evaluate the generalization of the method in new scenarios. Although we are aware that carefulness only partially accounts for all possible properties of an object, this work is an important step towards a global approach for robots to interpret human movements using vision alone

5.1 Methods

The objective of this paper is to prove that a robot, in particular the humanoid iCub, can distinguish online and in different scenarios whether a human is performing a careful transport motion or a not careful one, relying on the kinematics modulation in the human gesture perceived through its camera.

5.1.1 Software Architecture

Aiming at the presented goal, we developed, using the YARP middleware [108], the software architecture shown in Figure 5.1.

As first step, the robot camera captures images from the scene with a resolution of 320×240 pixels and a 22 Hz frame rate. Then, the following module computes the optical flow (OF) using a dense approach [49], and applies a threshold on the OF magnitude to consider only the parts of the image where the change is significant. This choice introduces the strong assumption that, in the robot's field of view, relevant motions are the ones that generate the largest OF. However, choosing the OF to characterize human motion, grants the system robustness to small changes in the point of view. The OF is a suitable tool for human motion description, for common daily activities such as cooking [58, 130], but also for understanding the meaning of hand gestures [100, 29].

The components of the motion velocity (horizontal u and vertical v) are extracted from the OF, as described by Vignolo *et al.* [165], and used to compute the norm of the tangential velocity, as in Eq. 5.1. The architecture extracts this feature with a frequency of 15 Hz.

$$V(t) = \sqrt{u(t)^2 + v(t)^2 + \Delta_t^2} \tag{5.1}$$

The segmentation module implements a heuristic threshold mechanism to consider only significant data: it detects the start of a motion when the velocity $V(t)$ overcomes a threshold τ and the end when the velocity becomes lower than τ. Once the end of the movement is detected, the segmentation module has two alternatives. If the temporal length is below 1

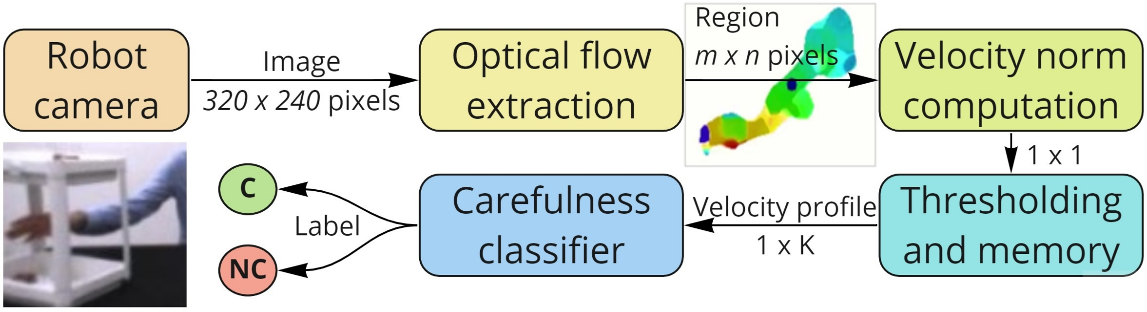


Figure 5.1 The system's architecture structure gathers images from the robot camera and extracts features from the computed optical flow to discriminate between careful (C) and not careful (NC) motions.

second, the motion is discarded. Otherwise, the temporal sequence of size $1 \times K$ is fed to the classifier. The minimum duration was set to 1 second since in the training set NC movements, which were the shortest, had a median duration of 1.2 seconds and the minimum duration was 1.1 seconds.

5.1.2 Model training and dataset description

The classifier model is inspired by the work presented in the previous Chapter 4 where a Long-Short Term Memory (LSTM) neural network showed promising results for the classification of temporal sequences of tangential velocity between careful and not careful motions. In this study, we adopted a neural network with one hidden layer followed by an output layer. The hidden layer is a 32-neuron bidirectional LSTM, while the output layer has two neurons and a sigmoidal activation function. The training has been performed using the ADAM optimization algorithm, binary cross-entropy loss function, exponential decay of the learning rate, and a batch size of 30. An early stopping condition on the validation loss, i.e., patience set to 5, has been introduced to prevent over-fitting. A zero-padding and masking technique has been adopted for the training to handle sequences with different temporal lengths.

The dataset, used to train and preliminarily test the model, had been collected asking 14 volunteers to displace four glasses in front of iCub. It is the same dataset presented in Chapter 3 and exploited for the training of the models in Chapter 4, where the outlier data from one participant was discarded. The glasses differed in weight, light (167 gr) or heavy (667 gr), and content, since two of them were filled with water till the brim, to induce careful motions. Even though we consider the carefulness in the gesture as the feature to be detected, the 500 grams weight difference was kept to increase the dataset variance. The dataset contains 878 segmented sequences, 438 for each class (C and NC). Preserving the class balance, we used 72% of the data for the training, 8% for the validation, and 20% for the test. The trained model got an accuracy of 95.14% on the test set, in line with the results of our previous work (90.5%). Furthermore, following a statistical analysis of the available data, we determined the threshold value τ for the segmentation module as $5.25\,pixels/s$.

5.1.3 System evaluation

Given the system presented in Section 5.1.1 for the discrimination of careful and not careful motions, we performed new experiments to test its performance. In particular, the objectives to assess are:

O1 The possibility for the system to work online, providing the C/NC label when a human completes a transportation motion.

O2 The ability of the system to generalize over unknown human subjects.

O3 The possibility for the system to generalize over new kinds of transportation motions.

Eleven healthy subjects, members of our organizations, voluntarily agreed to participate in the data collection (7 females, 4 males, age: 28.0 ± 2.4); none of them is author of this research. All participants used their dominant hand in the experiment and only one was left-handed. We divided the volunteers into two groups $G1$ (4 females, 1 male, age 27.8 ± 3.6, one left-handed) and $G2$ (3 females, 3 males, age 28.2 ± 1.3). We purposely chose different participants from those included in our training set to grant a wider variability in the new data collection and assess O2.

The experiment consists of a series of structured transportation movements of four glasses performed by the participants while sitting at a table. The robot iCub is positioned on the opposite side of the table and perceives the scene through its camera; replicating the same setup presented in Chapter 3, iCub is passive, does not show any interactive behavior and is not introduced to the participants. We use four glasses identical to those in the training set, representative of two classes, namely C and NC, according to the presence or absence of water inside. Throughout the experiment, a synthetic voice from two nearby external speakers instructs the participant on which object to grasp and where to place it. Placing positions in the scenario are identified with letters (see Figure 5.2). To receive instruction on the next transportation, the participant presses a key on a keyboard with their non-dominant hand. In between each transport motion, the volunteer rests their hands on the table. To investigate the system's ability to generalize over new transportation trajectories (O3), we have designed three experimental scenarios, namely: *Shelves*, *Simple Table* and *Advanced Table*.

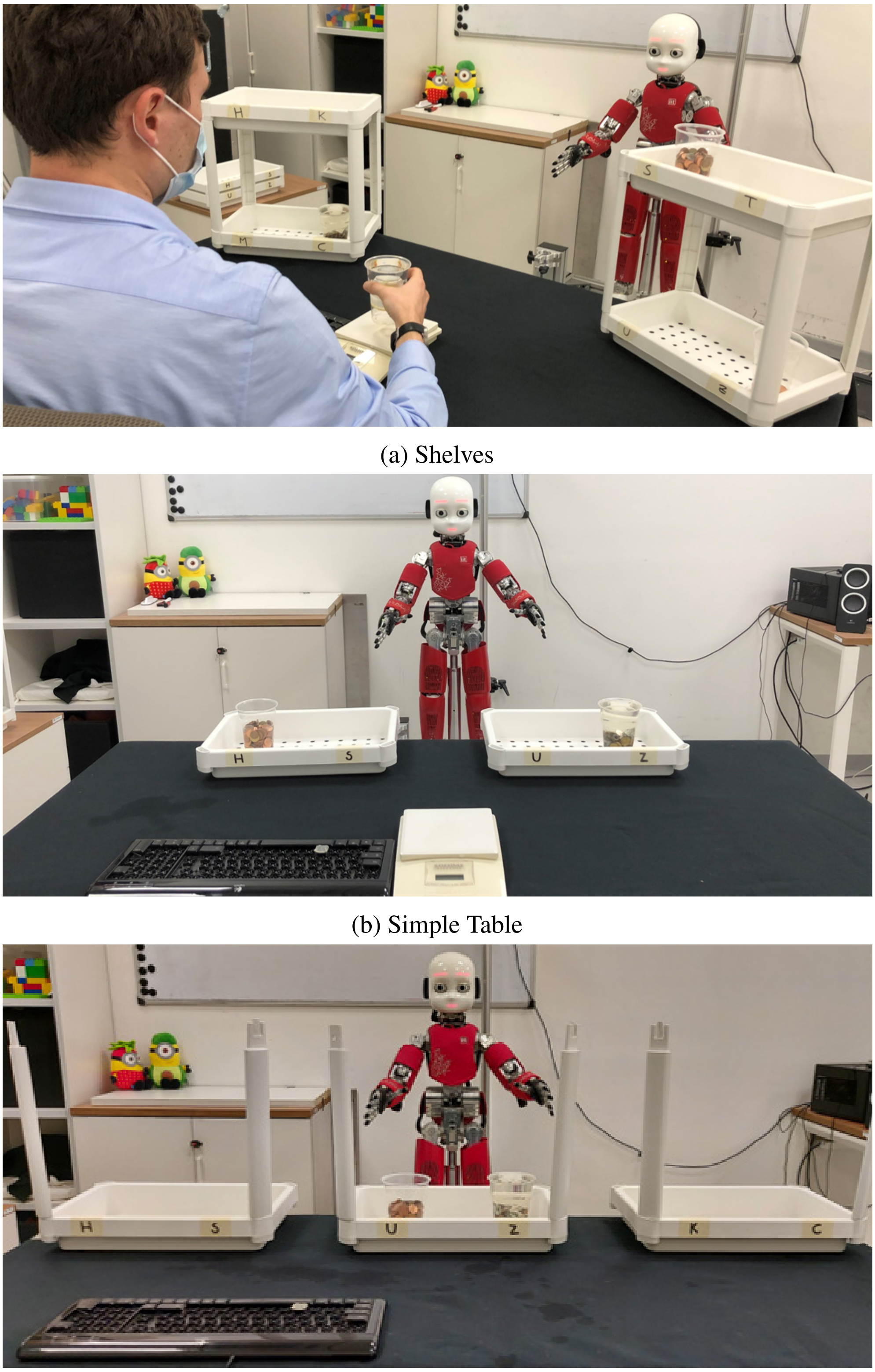
Shelves. The first scenario replicates the one used to collect the training set. This scenario allows for testing the online performance of the classifier (O1) and the system's generalization over new subjects (O2). The objects are transported back and forth from a fixed position on the table, delimited by a scale, to two shelves located on the right and left side of the table (see Figure 5.2a). Eight positions where the objects can be grasped or placed are defined on the two shelves. Both $G1$ and $G2$ completed the experiment in this scenario, and each participant performed 32 transport movements (16 careful and 16 not careful).

Simple Table. This scenario is aimed at assessing the system's capability to generalize on a new set of movements (O3) and has been performed only by the 5 volunteers in $G1$ group. The glasses are moved from the scale in front of the participant to four positions on the table, delimited by a container, or vice-versa (seen Figure 5.2b). Each volunteer performed 32 transport movements (16 for each class).

Advanced Table. This setup tests the system's capability to generalize over more ample and complex transport movements (O3). In this scenario, the glasses are moved between positions defined on the table, i.e., the scale is removed. In this way, the transportation motion is no more towards and away from the volunteer. Three containers are placed on the table, with two possible positions each, and columns are mounted on their frontal corners (see Figure 5.2c). The columns obstacle the transportation, making the experiment more challenging. This more complex setup was designed after a preliminary analysis of the classification results with the Simple Table task, therefore only volunteers from $G2$ experimented with this scenario, and each of them performed 16 transport movements (8 for each class).

5.2 Results

Throughout all the experiments described, the recognition architecture described in Section 5.1.1 was running, recognizing careful and not careful motions. We analyze these results for each scenario, focusing on the system accuracy and the recognition time (i.e., the time between the motion end and the system recognition). Furthermore, we performed a statistical analysis of the velocities extracted from the OF to highlight possible differences between the three scenarios.



(a) Shelves

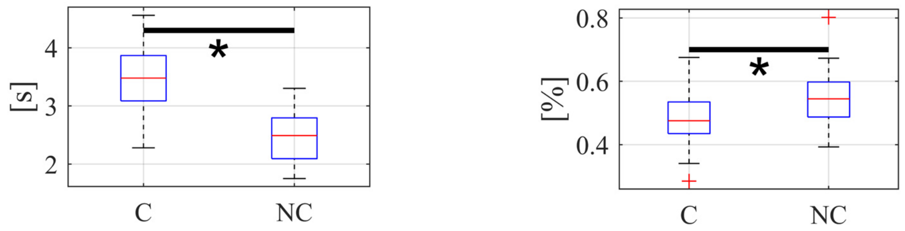
(b) Simple Table

(c) Advanced Table

Figure 5.2 Setups of the different scenarios explored for the system evaluation. The Shelves scenario replicates the training condition (5.2a). Simple Table (5.2b) and Advanced Table (5.2c) scenarios are introduced to evaluate the generalization performance.

Table 5.1 *Shelves*. Confusion matrix for the transportation movements performed by the 11 volunteers. The dark grey cell shows the overall accuracy

	Target class		
Output class	NC	C	Precision
NC	**163** - 46.3%	**109** - 31.0%	59.9% - 40.1%
C	**13** - 3.7%	**67** - 19.0%	83.8% - 16.2%
Recall	92.6% - 7.4%	38.1% - 61.9%	65.3% - 34.7%



(a) MD

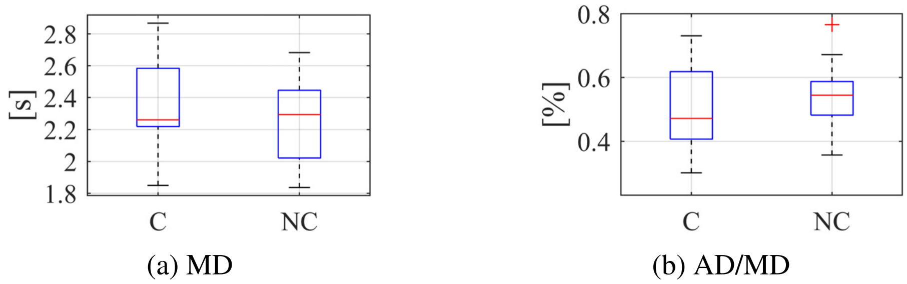
(b) AD / MD

Figure 5.3 *Shelves*. Box plots of the Movement Duration (5.3a) and the Acceleration Duration over the Movement Duration of the velocity profiles (5.3b) for careful (C) and not careful (NC) transport motions. The red lines represent the medians, the blue rectangles limit the 25th and 75th percentiles, and $*$ indicates a significant difference according to the Wilcoxon test.

5.2.1 Shelves Task

We report in Table 5.1 the confusion matrix related exclusively to the glasses transportation movements performed by the 11 participants, with a F1-Score of 72.9%. In this scenario, the classifier has been invoked correctly for all the 352 transport movements (32 movements of 11 volunteers) with a median recognition time below 150 *ms* (136.6 ± 18.8 *ms* - median and median absolute deviation). However, because of the system design, the classifier was called not only when a transport movement happened, but every time a velocity above threshold persisted at least for more than one second. Indeed, 300 more movements were detected and classified as NC 89.3% of the times. These movements are those that the volunteer performs to reach the glass and go back to the resting position. Since these movements are not transportations, it is reasonable that the majority of them are classified as NC; however, they were not included in the confusion matrix results.

Finally, we characterized the velocity profiles using two metrics, i.e., the transport movement duration (MD, proposed as significant to investigate the carefulness by [45]), and the asymmetry of the velocity peak (AD/MD, see Eq. 5.2). This last metric is expressed as



(a) MD

(b) AD/MD

Figure 5.4 *Simple Table*. Box plots of the Movement Duration (5.4a) and the Acceleration Duration over the Movement Duration of the velocity profiles (5.4b) for careful (C) and not careful (NC) transport motions. The graphical conventions are the same as in Figure 5.3.

the acceleration duration (AD) over the movement duration (MD), and it is widely used to characterize arm movements [116, 59].

$$AD/MD = \frac{index_{Vmax}}{MD} \tag{5.2}$$

Since the populations were not normally distributed, in order to test if these two metrics showed any significant differences between C and NC motions, we used a Wilcoxon Signed Rank test. Considering all the 11 participants who performed the Shelves Task, we report for the MD a $p-value$: $< .01$, while for the AD/MD a $p-values$: $< .05$. In Figure 5.3 are shown the corresponding ranges of movement duration and velocity asymmetry.

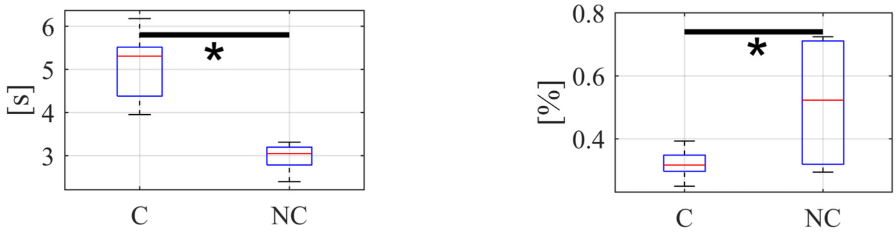
5.2.2 Simple Table Task

This scenario entailed movements that differed from those included in the training set, and only $G1$ experienced it. The online classifier did not achieve a good performance. We report an F1-Score of 66.09% with 96.25% recall and 50.33% precision values. The system tended to classify as not careful most movements, correctly identifying only 2.5% of the careful trials. However, the classifier was rightfully called at the end of every one of the 160 transport movements, with a median recognition time of $137.8 \pm 21.4\,ms$. Regarding the motions detected beyond the transport ones, the classifier was called 77 times, giving an NC label in 96.1% of the cases.

Interestingly, analyzing the MD and AD/MD metrics (see Figure 5.4), which we use as distance measures between the careful and not careful movements, the Wilcoxon Rank Signed test reported *p-values* $> .2$ for both. Thus, according to the chosen metrics, no significant difference in the velocity profiles was detected between the C and NC groups in this scenario. These results suggest that for short transportations (about $40\,cm$) with no

Table 5.2 *Advanced Table*. Confusion matrix for classifying transport movements performed by *G*2 in the generalization task. The dark grey cell shows the overall accuracy.

		Target class		
		NC	C	Precision
Output class	NC	**42** - 43.8%	**12** - 12.5%	77.8% - 22.2%
	C	**6** - 6.3%	**36** - 37.5%	85.7% - 14.3%
	Recall	87.5% - 12.5%	75.0% - 25.0%	81.3% - 18.7%



(a) MD

(b) AD/MD

Figure 5.5 *Advanced Table*. Box plots of the movement duration (5.5a) and asymmetry of the velocity profiles (5.5b) for careful (C) and not careful (NC) transport motions. The graphical conventions are the same as in Figure 5.3.

obstacles, the kinematics properties do not change significantly between careful and not careful motions.

5.2.3 Advanced Table Task

This scenario was designed to test the generalization capability of the model further. Glasses handling has been made more difficult by introducing obstacles and forcing longer paths between the grasping and release positions. In Table 5.2 is shown the confusion matrix for the transportation movements in this scenario, where our system reaches an F1-Score of 82.4%. The classifier output was available for every one of the 96 glass manipulations with a recognition time of $145.3 \pm 16.3\,ms$ (median and median absolute deviation). Regarding the 143 other movements that the classifier evaluated, the given label was NC for 97.9% of them. Finally, concerning the parametric measures (shown in Figure 5.5), both differences between C and NC were statistically significant (MD: $p < .01$, AD/MD: $p < .05$).

5.3 Discussion

The proposed approach allows a robot to identify, online, whether an object is being handled with care simply by observing the human movements. Optical flow is an effective motion descriptor for this purpose because it provides a comprehensive evaluation of the entire movement and is resistant to small, quick occlusions caused by obstacles such as shelves (see Figure 5.1). However, when the movements are slow, such as when handling glasses full of water, image obstructions may be prolonged and have a greater impact. The proposed architecture generated a classifier output for every glass transportation, ensuring that no transport movements went undetected. The model output was available at the end of the transportation, with a median recognition time of $135.9 \pm 17.9\,ms$ for all tasks. The system also detected movements beyond the transport ones, such as reaching and departing actions to grasp the glass or return to the resting position. In these instances, since no object was being carried, it is reasonable that the classifier returned a "not careful" label in 92.7% of the occurrences. This result implies that when the system returns the "careful" label, it has high confidence. In the Shelves scenario, which replicates the training conditions, the performance of the overall online classifier is lower than that obtained with offline testing (which gave 90.5%, see Chapter 4). However, considering the novel testing conditions (e.g., different lighting and perspective) and the fact that motion velocities were segmented online, these results suggest that our system can work online (O1) while generalizing over new subjects (O2).

In contrast, our architecture did not perform well in the Simple Table scenario. We attribute this to the setup design, which required shorter movements without any obstacles compared to the Shelves and Advanced Table scenarios (see Figure 5.2 for reference). This result leads us to speculate that the boundary conditions of the external environment can emphasize the impact of carefulness. Therefore, it may be easier to detect the presence of carefulness in more complex scenarios. This hypothesis is supported by analyzing the distance metrics of the velocity profiles shown in Figure 5.2.2. In the Simple Table scenario, there was no significant difference in movement duration (MD) or the asymmetry of the velocity peaks (AD/MD). These results leave us with two potential explanations: (i) volunteers did not act with particular care when transporting the glasses full of water in the Simple Table scenario, or (ii) tangential velocity alone is not sufficient to distinguish between careful and non-careful motions in this case, and additional data such as the actor's gaze pattern may be needed.

Finally, we obtained the best results when monitoring a completely new scenario (see Table 5.2). As we hypothesized, this result is linked to the extra care the volunteer needs to take in transporting the glass of water in a more complex setup. This is supported by the significant difference in the MD and AD/MD metrics (see Figure 5.5) between the two classes. These results demonstrate the ability of our system to work online (O1) and to generalize to new subjects (O2). We also showed that the system could generalize to new scenarios if transportation carefulness is evident (O3). It's worth noting that we tested our system with non-interactive actions (i.e., the participants perform the task alone, with the robot acting as an observer). An interactive context might facilitate carefulness recognition, as participants would more explicitly convey this information, similar to how humans signal to each other during tasks [43, 41, 125].

To determine the fragility of objects, we used the information naturally embedded in human kinematics during manipulation, extracted using vision alone. This approach is intended for use when the robot collaborates with a human partner, such as during handover tasks, where the human's movements can be observed. During the interaction, the robot can detect the care taken by the human partner and adjust its own manipulations accordingly without needing prior knowledge of the object or visual detection of its physical properties. Since the robot's ability to detect carefulness is independent of the external appearance of the objects, it can be generalized to previously unseen objects. In order to improve the interaction, it would also be useful for the robot to adapt online and modulate its movements to be consistent with the properties of the object, mimicking natural human behavior and conveying the same information about the object's features. This would greatly facilitate natural, implicit communication between humans and robots and address the dual problem of generating communicative robot actions in the following chapters

Part II

Design of Communicative Robot Movements

Chapter 6

A generative approach for property-aware robot object manipulation

Elaboration of an article published in:
International Conference on Development and Learning (ICDL 2021) [57]
Best Paper Nominee

The previous chapters presented possible approaches to automatically detect object properties by relying on the natural kinematics modulations occurring during their transport motion. The second part of my thesis will focus on the dual problem of making robots' actions more legible, appropriate, and communicative when coming to object manipulation. Indeed, when transporting an object, we unconsciously adapt our movement to its properties, for instance by slowing down when the item is fragile. The most relevant features of an object are immediately revealed to a human observer by the way the handling occurs, without any need for verbal description. In a social context, such kind of natural, implicit information can be purposely emphasized to make our actions or future intentions more readable by the partner [125]. It would greatly facilitate collaboration to enable humanoid robots to perform movements that convey similar intuitive cues to the observers. Indeed, in human-robot collaboration settings, it is relevant that also robots master this form of communication, modulating their goal-oriented actions to make them more intuitively legible to the human partner [43]. In particular, if we consider the scenario where humans and robots interact with the same set of objects, a robot should be able to both infer object properties from the observation of how the partner manipulates them and to select and perform the suitable manipulation action.

We focused then on how to generate robot motion adapted to the hidden properties of the manipulated objects, such as their weight and fragility. The core idea behind the implementation is to achieve a modulation of their gesture similar to the one that occurs in humans without necessarily copying and scaling the action to adapt to the robot kinematic chain. This would help to achieve a two-fold outcome: *(i)* each object would be handled accordingly to its features, therefore mimicking natural human behavior, and *(ii)* the gesture performed by the robot would convey the same information *about* the object features, being, therefore, more transparent and readable for the human partner. To allow robots to communicate with their actions, we decided to modulate the strategy to transport objects with the end-effector, specifically shaping its velocity to replicate natural human behavior. Indeed, from our previous studies on how humans move objects, it was the velocity in the action which was mostly influenced by their properties: not only in the absolute magnitude but also in its shape, depending on the acceleration and deceleration phases, to the point that this feature can be used to automatically detect the carefulness in human actions (see Chapter 4 and 5). In the studies presented in the previous Chapters, we evaluated if the complete velocity profile of the action is communicative of the carefulness; this is in line with other works considering the same kinematic feature in the context of object manipulations [45, 46]. Actually, it is not completely clear what the communicative part of the velocity profile is: it may be sufficient to modulate on the robot only the average speed of the motion, or the duration of the action. However, having in mind as our ultimate goal to create movements communicative of the carefulness with different robotic platforms, and aware of the importance of biological plausibility [17, 27, 75, 148], we chose to follow an approach as human-inspired as possible. For this reason, we decided to have the robots' end-effector move following human-like trajectories, taking care to fully reproduce human-inspired time-series. This choice should also ensure the final achievement of our first objective. Indeed, if humans successfully handle an object by adopting a certain velocity profile, the same mechanism should be mimicked to allow a robot to do the same. The first problem to address was how to create velocity profiles to later use to control the robot end-effector, which resembled the human distribution while being novel and not stereotyped. We decided to leverage a generative approach to produce new and consistent movement patterns with a distribution of velocity profiles comparable to the human one, applicable over different trajectories while preserving the informative content of the actions. In particular, we exploit Generative Adversarial Networks (GANs), which can synthesize new actions coherent with the object's properties after being trained on human examples. Human motion kinematics are an example of a time-series, from which it is possible to extract descriptive features like velocity, acceleration profile, or the curvature radius associated with the trajectory. We posit that generative models can be used to "learn" such time-dependent patterns, where the model is not only tasked with capturing the distributions of features within each time point, but it should also capture the potentially complex dynamics of those variables across time [177]. In this Chapter, we present the training of GANs to produce velocity profiles associated with object manipulation and their assessment through qualitative and quantitative metrics.

6.1 Methods

As training data for the generative networks, we exploited once again the dataset described in Chapter 3. The goal of this work is to design a method to automatically generate synthetic velocity profile norms consistent with the ones generated by a person manipulating an object with specific properties. The experimental setup intrinsically allows for great variability in the performed gestures: the trajectories could be directed towards the left or the right side, towards lower or higher shelves and the direction of the motion could be either abductive or adductive. This We considered 1001 total object transportation movements (W1-NC: 248, W2-NC: 251, W1-C: 254, W2-C: 248 trials), where the transportation phase was isolated in each trial by placing a threshold equal to the 5% of the corresponding velocity peak, as already described in the Methods of Chapter 4. The sampling rate of the acquisition was 22 Hz.

In order to obtain a first corroboration for our hypotheses, for each transport movement, we considered the norm of the velocity, computed as in (6.1), from the three velocity components $V_x(t)$, $V_y(t)$, $V_z(t)$ at each time instant.

$$V(t) = \sqrt{V_x(t)^2 + V_y(t)^2 + V_z(t)^2}. \tag{6.1}$$

Considering the norm as a feature, we trained a TimeGAN[177] on all the transportation actions in the dataset. The choice of using the norm of the velocity grants the detachment from any particular trajectory, therefore the generality of the further generated movements; moreover, such feature had already proved to be insightful when detecting the characteristics of the manipulated object, being representative of the action kinematics. We trained four different models, one for each object class, generating as many velocity profiles as the ones fed to each GAN. Since the duration of the movements was characterized by limited variability, we decided to pad the time series in each class with zeros to the length of the longest sequence. Although more sophisticated approaches are possible, this simple solution seems adequate in our case. This resulted in velocity profiles of length 55 for (W1-NC), 50 for (W2-NC), 127 for (W1-C) and 131 for (W2-C). All the TimeGAN models were implemented with 3-layer Gate Recurrent Units (GRU), each one with 28 neurons (an extensive explanation of the architecture is provided by [177]). The training phase lasted 2000 epochs, and was performed with batches of size 15.

6.1.1 Metrics for generated data assessment

To assess the quality of the generated data, we computed different metrics, inspired by those proposed by [177, 48]:

1. *Data distribution:* the synthetic data population should be distributed to match real data samples.
2. *Data discriminating power:* a classifier trained to discriminate among the different object typologies on synthetic data should perform equally well when tested on real data and vice versa (i.e., train-on-synthetic, test-on-real approach).
3. *Features preservation:* from synthetic data, it should be possible to extract kinematics features similar to the ones originating from real data.

Data Distribution

In order to understand how the time series of the different classes could be globally visualized, we performed a manifold analysis to reduce their dimensions. This approach shows the inherent structure of the data by representing them in a two-dimensional space, allowing us to qualitatively evaluate whether the generated data are consistent with the original examples. More specifically, flattening the temporal dimension, we computed Principal Component Analysis (PCA), which is a linear projection of the data along the directions of their maximum variance, and t-distributed Stochastic Neighbor Embedding (t-SNE), which highlights non-linear local structures.

Data Discriminating Power

To quantitatively measure the similarity between original and generated data for each category, we decided to assess the performance of a classifier in discriminating the velocity profiles associated with the transport movements of *careful/not careful* and *light/heavy* glasses. To this aim, and referring to [48], we introduce two evaluation methods. The first metric is called *Train on Real, Test on Synthetic* (TRTS), meaning that the model is trained and validated on real data and then tested on artificial data to discriminate either the carefulness or the weight of the transported glasses. Mirror-wise, the second evaluation metric consists in training and validating the model on the synthetic dataset, then testing it on the original dataset, and it is referred to as *Train on Synthetic, Test on Real* (TSTR). For the sake of comparison, all the classification tasks were addressed by using the same Long Short-Term Memory (LSTM) model with an architecture having an input layer, a bidirectional LSTM with 64 hidden units and 2 fully connected layers, followed by an output layer with softmax as activation function. All the sequences were restored to their original length, by removing the values under a velocity threshold of 0.005 m/s. For each training session, we split the training data into Training Set *plus* Validation Set by using a 5-folds cross-validation. In Table 6.1 we report in parentheses the performance of the Validation Set. The Test Set was instead considered as *all real data* (TSTR) or *all synthetic data* (TRTS).

Features Preservation

Another interesting aspect that must be considered when generating new velocity examples is feature preservation. Even though deep learning is widely applied for its ability to extract features in an unsupervised way, our goal is to generate data that preserve specific characteristics (features) of the movements. Indeed implicit information about the object we are manipulating is naturally embedded in the kinematics of our movements. The final goal of generating synthetic velocity profiles inspired by humans is to endow robots with the same communicative ability while performing their movements. Therefore, it seems paramount that synthetic data preserve those features associated with specific object properties.

Similarly to [45], we want to generate data preserving such characteristics as *Movement Duration* (MD), *Peak Amplitude* (PA), and *asymmetry* in the velocity profiles. This last parameter, which we will refer to as AD/MD, represents the Acceleration Duration (AD) in relation to the full movement duration and is commonly used to describe arm movement kinematics [116, 59]. It gives a percentage value for how far the velocity peak is ahead of or behind in relation to the MD, therefore indicating a longer or shorter deceleration period.

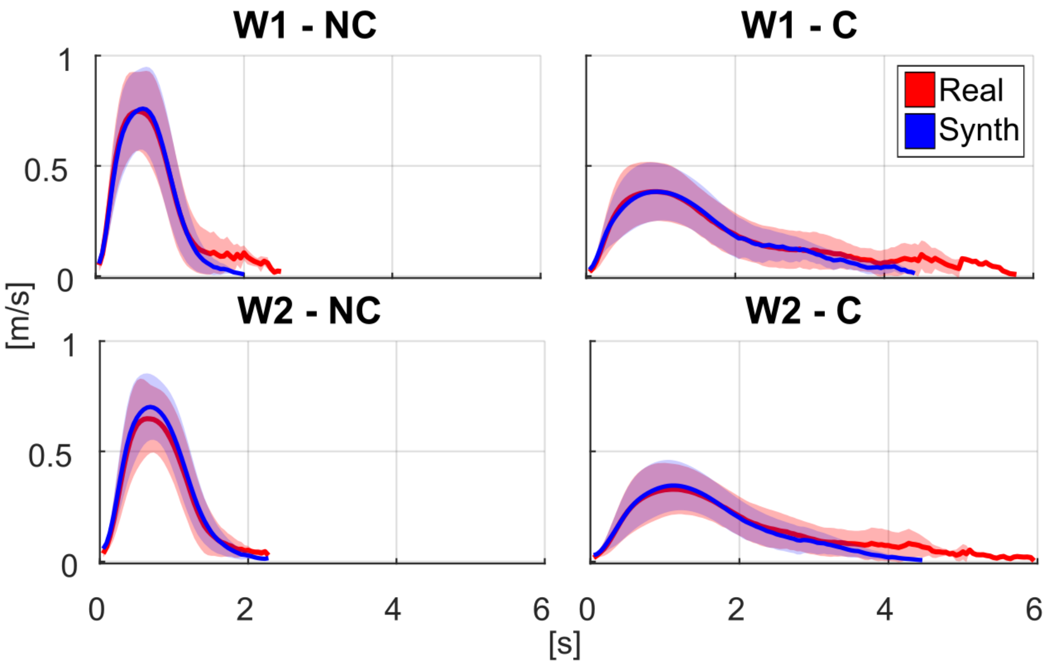


Figure 6.1 Velocity profiles of real (red) and synthetic (blue) data for each object class. The thick line represents the mean value, while the coloured area the standard deviation.

6.2 Results

In Figure 6.1 are shown global velocity profiles, obtained by calculating the mean and the standard deviation over the samples of every trial, for each of the four object classes. We have based all the subsequent analyses on these data. The tails visible on the original velocity profiles are due to an overall number of 21 trials (equal to 2.10% of the entire dataset) which differed in the movement duration from the mean of each class (W1-NC: 1.44 ± 0.17, W2-NC: 1.61 ± 0.19, W1-C: 2.62 ± 0.63, W2-C: 3.04 ± 0.69 [s]) for more than three times the respective standard deviation. Therefore, such movements can be considered outliers, and it is fair to assume that the GANs did not capture them, which is the reason why the generated velocity profiles (in blue) are more consistent with each other.

Data Distribution

In Figure 6.2 are shown Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE) projections for real and synthetic data, divided according to the weight and the carefulness associated with the transported glasses. The synthetic data for each class are those generated by the corresponding GAN. This representation allows appreciating how similar and overlapping the original and the synthetic distribution are.

Table 6.1 Accuracy values training a classifier on real data and then testing it on synthetic data (TRTS) and vice versa (TSTR). TRTS and TSTR analysis has been carried out also on subsets of the original dataset: considering separately only the heavy and the light objects or the C and NC movements.

Dataset	**Detection**	**Classifier**	**Validation (%)**	**Test (%)**
All	Weight	TSTR	97.40 ± 1.08	62.76 ± 1.01
		TRTS	70.53 ± 3.04	75.18 ± 6.17
	Carefulness	TSTR	98.10 ± 1.64	92.97 ± 1.36
		TRTS	96.40 ± 1.92	92.53 ± 1.57
Heavy	Carefulness	TSTR	99.00 ± 0.71	87.29 ± 4.42
		TRTS	96.99 ± 2.36	94.15 ± 3.17
Light		TSTR	98.21 ± 0.84	94.13 ± 2.31
		TRTS	96.81 ± 1.31	93.47 ± 1.07
Careful	Weight	TSTR	87.65 ± 3.91	61.39 ± 5.29
		TRTS	71.51 ± 2.60	74.74 ± 5.41
Not Careful		TSTR	95.99 ± 2.36	73.03 ± 3.05
		TRTS	78.35 ± 3.89	79.60 ± 6.77

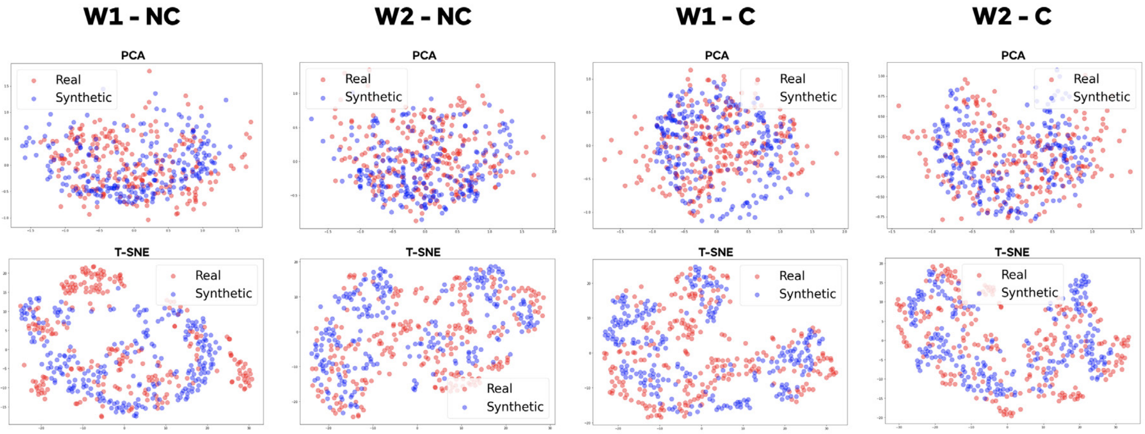


Figure 6.2 **Data distribution:** PCA and t-SNE projections for each movement category. The real and the synthetic data are superimposed.

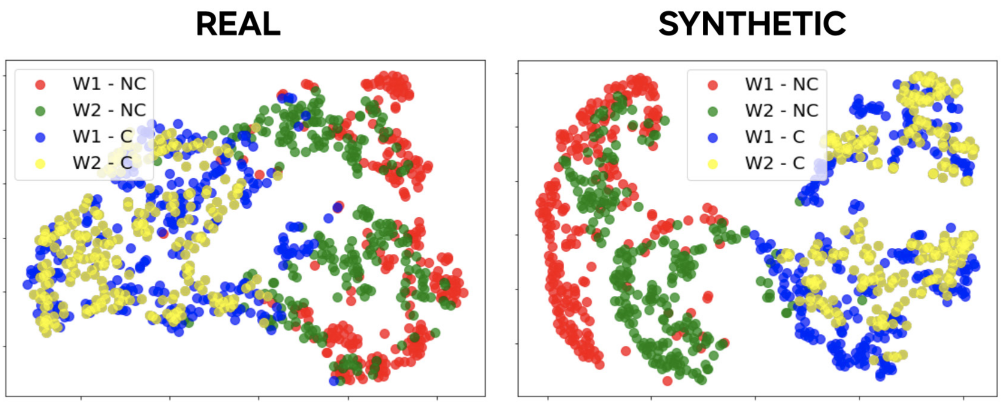
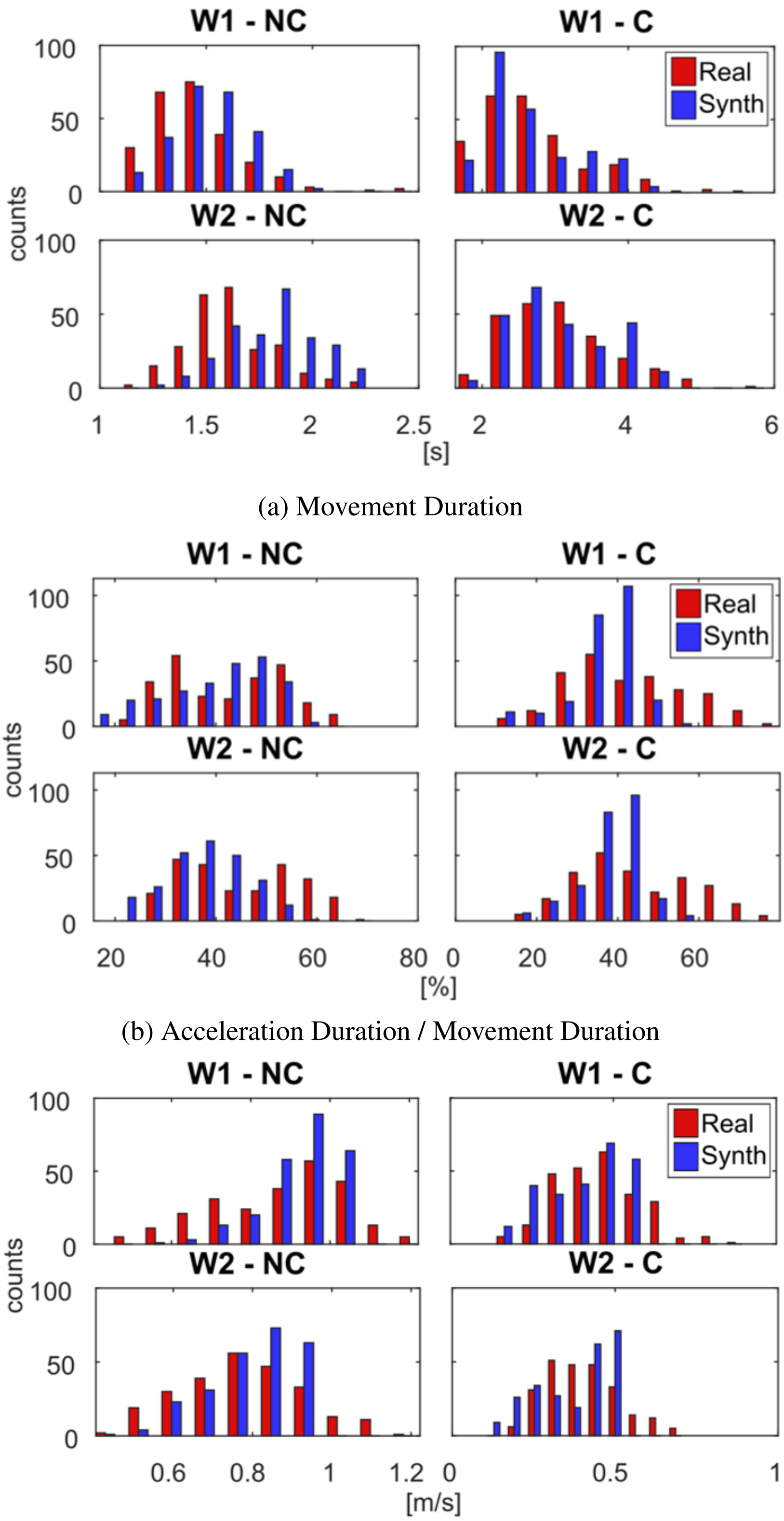


Figure 6.3 **t-SNE** representation for the two data sources. The four colors depict the glasses' properties: red and green for the objects which did not require care (red - light, green - heavy), blue and yellow for those to be moved carefully (blue - light; yellow - heavy).

To assess the separability among the different movements, we again used the t-SNE projection; the whole dataset (synthetic and real) is represented in Figure 6.3, where the different colors highlight the four distinct classes of movements. This allows evaluating which classes are more similar to each other according to their velocity profiles and whether the generated data reflect the same tendency. It can be noticed how the clusters linked to the care in the gestures are generally clearly separated. At the same time, those which differ only for the weight are largely overlapped, especially considering the portion of careful motions. Synthetic data appear even more strongly separable according to carefulness, and considering only the NC movements, the light (in red) and heavy (in green) objects seem to be distinguishable.



(a) Movement Duration

(b) Acceleration Duration / Movement Duration

(c) Velocity Peak Amplitude

Figure 6.4 **Kinematics features distribution** for each object category for the Real (in red) and Synthetic (in blue) data. The y-axis represents the number of trials with a given MD (6.4a), AD/MD (6.4b) or Velocity Peak (6.4c) value.

Data discriminating power

The results reported in Table 6.1 refer to multiple classification tasks performed using either the real data as training set (TRTS) or the synthetic data (TSTR) for the same purpose. As explained before, two main features simultaneously characterized the handled glasses: the weight and the carefulness required by the presence of water. Even though the care in the manipulation seemed the prevalent feature, we decided to test the possibility of discriminating the weight or the care in the gestures separately.

We trained the same binary classifier model using the whole dataset (real or synthetic) to discriminate the carefulness or the weight, then testing it on the dual source of data. The corresponding accuracies are reported in Table 6.1. The performance is better when discriminating the carefulness (above 92%), with comparable results on both the test sets. The results on the weight classification are worse, however satisfying when training on real and testing on synthetic data, with an overall accuracy above 70%. The TSTR approach on the weight discrimination reveals a remarkable disparity between the accuracy obtained on the synthetic validation set ($97.40 \pm 1.08\%$) and the one on the real data used as test set ($62.76 \pm 1.01\%$).

We also assessed the separability using as training set a halved dataset, to reduce a variability factor: the ability to discriminate the weight was evaluated using separately only the careful or the not careful velocity profiles; symmetrically, the possibility of classifying the carefulness was verified exploiting only the heavy or the light glasses. The second half of Table 6.1 shows the matching results. In particular, concerning the carefulness classification, the mean accuracy slightly improved for every combination, except for the TSTR when using only the simulated manipulations of heavy glasses. When discriminating the weight relying on the subset of not careful movements, the performances improved in particular for the TRTS, reaching a $79.60 \pm 6.77\%$ accuracy. The most critical result, even though above chance level ($61.39 \pm 5.29\%$), is the one related to the weight discrimination on the careful subset when training on synthetic data.

Features preservation

The last analysis we performed is meant to evaluate the persistence of meaningful properties. We assessed how close the real and synthetic distributions were according to some parameter characteristics of the kinematics of the movement: the MD, the AD/MD (i.e., the asymmetry in the velocity profile), and the Peak Amplitude. We report in Figure 6.4 the histograms corresponding to each one of the parameters. In detail, Figure 6.4a shows the duration of the transport movements. This feature cannot be generally considered as significant to detect the carefulness in a gesture or the weight of the item, since it is strongly dependent on the length of the trajectories: a glass full of water can be moved on a very short path, therefore the duration of that careful movement can be shorter than the one corresponding to the transportation of an empty glass for a more extended trajectory. However, in our dataset the paths covered with the four glasses were altogether the same, therefore we can consider the MD as characteristic of the object type. For the same distance covered, the durations for careful movements are markedly longer.

Concerning the AD/MD parameter reported in Figure 6.4b, the difference among the four classes is less striking; it can be noticed that for the careful gestures, the distribution of the synthetic data is less wide than the one of the original dataset, presenting a marked peak slightly before the 50% of the profile. This means that the generated data, in this case, tend to show a more symmetrical acceleration and deceleration phase. Finally, observing the histograms for the Velocity Peak Amplitude (Figure 6.4c), it can be noticed how such a parameter shows lower values in the careful manipulations and that the synthetic data can capture such a tendency.

6.3 Discussion

From a qualitative point of view, it is possible to notice that the synthetic velocity profiles are generated consistently with the real data distribution. This is supported by the mean velocity profiles in Figure 6.1, by the linear and non-linear manifold analysis in Figure 6.2, and by the kinematic parameters distributions of Figure 6.4. However, it must be stressed that the purpose of the work is not aiming at a perfect overlapping, which would mean a trivial copy of the data. Still, we are interested in a coherent representation of the distribution of the original data. The scope is to allow in the future the autonomous generation of appropriately communicative manipulation movements by robots, such as the humanoid iCub. When all the classes are visualized using a t-SNE representation, in Figure 6.3, it is possible to notice that both real and synthetic data share the same structure: careful actions (**C**) and not careful actions (**NC**) appear to be separable with overlapping in their sub-classes representing the carrying of light (**W1**) and heavy (**W2**) objects. It is worth considering that each class was generated independently from the others, training the GAN model using only the corresponding subset of original trials, without any knowledge of the velocity profiles of the other classes. The overlapping patterns emerging from the manifold analysis of the independently generated data strictly resemble the ones of the original dataset, meaning that the information was embedded in the data themselves. The proposed GANs were able to capture the characteristics of the original dataset, without any need to force the learning in a particular direction.

The difference between **C/NC** classes in synthetic data may be due to the fact that our model learned the most salient characteristics of these actions, overlooking the variability typical of the real data. A clearer separability seems to be present, according to the t-SNE representation of Figure 6.3, even between the generated movements performed with heavy (green) or light (red) objects in the NC class. It is reasonable to assume that the original transportation movements were largely influenced by the presence or absence of water inside the glasses; when this "disturbing" factor was not present, that is when the gestures were not careful, it was the weight that played a role in influencing the movement kinematics. The GANs have captured this occurring, and the corresponding W1-NC and W2-NC velocity profiles are less overlapped in the synthetic manifold representation. The easier discrimination of weight when the gestures are not careful is also supported by the classifier results presented in Table 6.1. As suggested by the t-SNE representation, the easier separability between light and heavy objects in the synthetic dataset results in a validation accuracy of $95.99 \pm 2.36\%$, which drops to $73.03 \pm 3.05\%$ when testing the model on the NC original data, which are more overlapped. After these considerations, we hypothesize that the real actions embed a degree of variability that is not descriptive of the class, but which may be ascribed to the natural variations in how humans perform a repeated gesture. Indeed, Table 6.1 shows, in general, how models trained on real samples are robust when tested on synthetic actions. When the classifier is instead trained on the synthetic data, there is a drop in the test accuracy over the original data, which are naturally noisier. This claim is also supported by the distributions of the kinematics features in Figure 6.4, where it can be noticed how the real data present a broader base with less pronounced peaks.

The proposed strategy aims to enrich the robots' movements with a communicative intention. Indeed, even a child-shaped robot like iCub would be able to move objects without difficulties in the range of weights we considered. However, our purpose is to deliberately design legible movements to improve human-robot interaction. Other strategies could be investigated and may prove effective in expressing carefulness in robotic actions, such as identifying and strictly reproducing only the key communicative factors in the kinematics. However, GANs can fully capture and reflect the global distribution of human kinematics during different object manipulations and seem appropriate to reproduce it in an original way. GANs represent a human-inspired approach that grants biological plausibility, and they allow preserving in the robot moves the implicit information on the properties of the handled object.

Thus, by making the robot movements not only effective but also readable and informative for the human partner. The following Chapters will present the application of the generative approach on different robots, exploring the implementation of the movements and the effect of such modulation of the movements during the interaction.

Chapter 7

Synthesis and Execution of Communicative Robotic Movements with Generative Adversarial Networks

Manuscript under preparation

Multiple communication channels can be used to share information between robots and humans: synthetic speech, light-based, digital display, mixed or augmented reality [54, 101, 133]. However, we believe that whenever it is possible, depending on the context and the task to accomplish, the most spontaneous form of implicit communication should be pursued, which is the one mediated by movement. In the context of human-robot collaboration in object manipulation tasks, it would improve safety and efficiency if the robot could adjust its motor actions based on the item's properties. Humans naturally modulate their movements to reveal information about the context or characteristics of the objects they interact with. Even though planning movements which are goal-oriented and optimized for efficiency, they still convey additional information to observers.

In the previous Chapter, we focused on how to create synthetic velocity profiles, suitable for different object characteristics, using Generative Adversarial Networks (GANs): for each object feature considered, namely the weight and the carefulness required, we trained a separate model, resulting in four separate generative networks able to produce velocities associated with the transport of *light/heavy* × *careful/not* objects. The core idea was that, after adequate training with movements acquired during the manipulation of objects by humans, GANs could generalize and autonomously generate sufficiently communicative movements, belonging to the desired class, without having to copy human ones exactly.

GANs have been commonly used in several research domains, especially in the computer vision one [169], since they can create novel realistic data after being trained on a set of real samples. However, their application in the generation of multivariate time series, as human motion kinematics in our case, has been explored only to a limited extent [20, 177, 48, 112]. In particular, there are a few examples of GANs being used to generate motion in the context of human interaction [119, 21, 175].

In this Chapter, (i) we introduce a novel *conditional* generative architecture for synthesizing velocity profiles associated with the manipulation of objects with different properties; in particular, the transportation of glasses which required careful (C) or not (not careful, NC) handling, given the presence or absence of water inside them. We focused on the carefulness feature to investigate the potential of having a single network producing, if appropriately conditioned, two distinct classes of data. The same network can be exploited to generate intermediate and completely new data with respect to the two classes used to train the network. The communicative effectiveness at the time of movement execution will be evaluated in the following Chapters, where we will assess if the reproduced movements are communicative enough of the carefulness property, if the reactions elicited depend on the modulation of the velocity profile, and finally, how the different embodiment of the robots influence the interaction. Moreover, (ii), we focus here on the control aspects, transferring the planned movements on two different robots: iCub [109] and Baxter [52]. These two platforms show remarkable differences in their appearance and dimension, in the distribution of the degrees of freedom (kinematic chain), in typologies of actuators and control strategies; however, they are both provided with two arms and offer the possibility of transporting objects with their end-effectors. For these reasons, they are particularly suitable for testing the replicability of our approach, comprising both motor action generation and motor control and its generalization potential. We evaluate the performances of the two robots in following the desired trajectories while modulating the velocity of their end-effector according to the selected velocity profile. In this thesis, we do not explore the control of the secondary joints of the kinematic chain, but focus exclusively on the trajectories of the end-effector. This simplification, while it may have an impact on communication, allows us to quickly extend our approach to different robotic platforms and replicate what happens in human motion.

Why a generative approach

A natural interaction with robots requires them to have a rich motor repertoire. This is because stereotyped and repetitive movements can make the interaction less natural and fluid. If we take for example any action performed with an arm, it can be executed with many trajectories and different velocity profiles. Although in the literature there are already works that have attempted to generate movements with different trajectories [119, 175, 21], to the best of our knowledge, there are no studies on producing movements purposely generating the velocity profiles, especially if the velocity modulation is aimed to implicitly describe an emotional state or the properties of the manipulated object. One possible approach to generate communicative movements would be to perform pre-recorded human movements on the robots. However, this solution would not make the robots truly "autonomous". For this reason, we tried to fill this gap by developing a generative neural network capable of synthesizing new velocity profiles that are descriptive of the properties of an object manipulated by the robot. The first goal of our generative model is to obtain synthetic data which preserves the temporal dynamics of the original time-series. To produce velocity profiles that reflect the two original classes, C and NC, we exploit a two-step training process. At first, we train an autoencoder to learn embedded representations of the original time-series. Then, such embeddings are used to train our cGAN, which generates new synthetic embeddings. The generator can be conditioned to output embeddings belonging to one of the two classes by giving as input a vector of labels and a random time-series noise (see Figure 7.1). Finally, a decoder is employed to reconstruct the time-series from the generated embedded representations. The main advantage of a generator that directly outputs in the embedded space is that in this way we encourage the model to focus on the relevant features of the dataset rather than on the samples themselves [177].

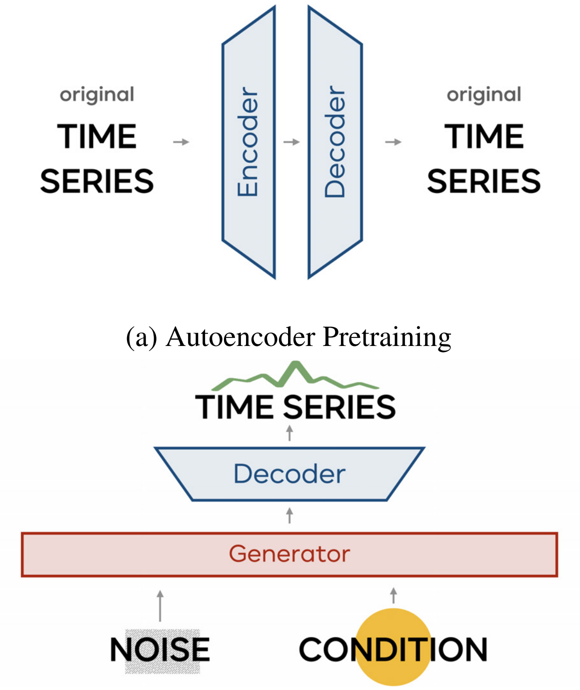
7.1 Methods

In this section I will first present the generative problem statement and the details of the network design and training. Then, the implementation on the two robots of the movements following the desired velocity profiles, with a communicative intent, will be addressed.

7.1.1 cGAN design and implementation

Generative problem statement

Suppose to have a distribution p_A of time series that describes the process A and another distribution p_B that describes the process B. We want to learn a Generator distribution $p_g(z|y)$, where z is random noise from a normal distribution, that is conditioned by some extra information y, for example a numerical label describing the target class we want to generate. Then $p_g(z|y)$ can be learned by a generator $G(z, y; \theta_g)$ whose output x should lie in



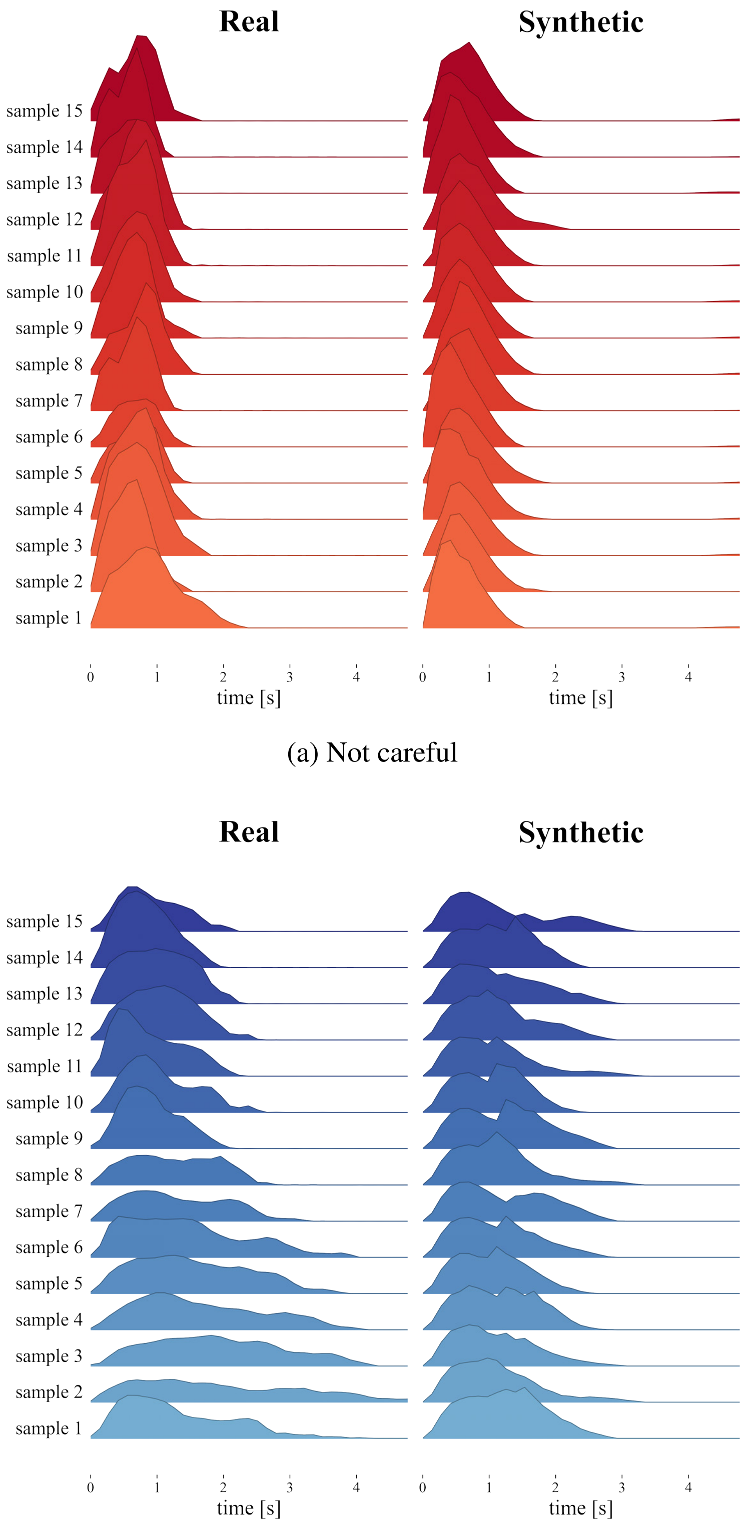
(a) Autoencoder Pretraining

(b) Inference

Figure 7.1 **Proposed architecture:** in (7.1a) is represented how is performed the training of the encoder/decoder network, to represent the time series in the embedded space. In (7.1b), are shown the steps for synthesizing a new velocity profile: our model takes as input random noise and a conditioning label, then the generator outputs in an embedding space that represents the characteristics of the signal we want to generate. Finally, the previously trained decoder translates the desired embedded properties into a time series

p_A or p_B based on the condition y. This learning process is supervised by a Discriminator $D(x,y;\theta_d)$, which outputs a single scalar representing the probability that x came from training data rather than $p_g(z|y)$.

We propose a conditional model for the generation of time series. Our model consists of four network components: encoder, decoder, sequence generator, and sequence discriminator (see Figure 7.1). The training of the first two components allows learning encoded representations of the time series in the dataset. The adversarial network (*generator* + *discriminator*) is exploited to learn to generate embedded representations in the latent space. These synthetic embeddings can then be used by the decoder to reconstruct plausible time series. To speed up the training, all velocity profiles were downsampled and eventually zero-padded to one-third of the original duration.



(a) Not careful

(b) Careful

Figure 7.2 **Comparison between real and synthetic data:** representation of a subset of the original velocity profiles that constitute the training data (real) and of the time series synthesized by the GAN, for both the not careful (7.2a) and careful (7.2b) classes. Please note that there is no one-to-one correspondence between each Real and Synthetic sample: they are randomly selected to give a qualitative overview of the original and synthetic dataset.

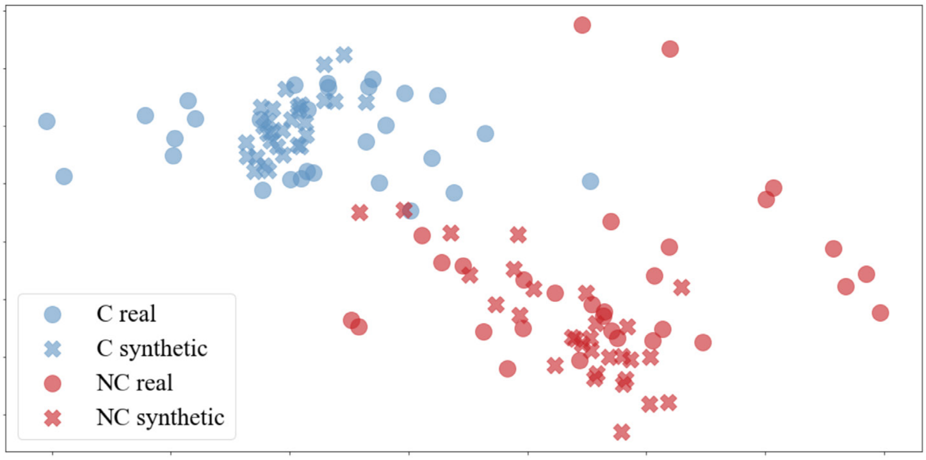


Figure 7.3 **PCA**: principal component analysis of the original and synthetic data distributions. The synthetic samples are close to the target classes but without overlaps, meaning that they are not copying the originals.

Encoder/Decoder networks

The *encoder/decoder* networks are jointly trained as an *autoencoder* before starting the adversarial training of *generator* and the *discriminator*. The *encoder/decoder* (Figure 7.1a) provide mappings between original time series and latent space, enabling then the adversarial network to learn about the data underlying temporal dynamics through lower-dimensional learning. The goal is to learn an embedded representation of the original data. This compressed representation, together with the output of the discriminator, is used as a supervision signal during training. In our work we implemented a simple LSTM with one encoder layer and one decoder layer. The latent dimension of the LSTM encoder was 16.

Generator/Discriminator networks

After the training of the *encoder/decoder* networks, the *generator* and the *discriminator* are trained. In the generator the prior input noise $p_g(z)$ and the conditioning label y (either 0 or 1) are combined in a joint hidden representation. The label is embedded into a vector of the same size as z, by using a Keras embedding layer, and then multiplied by the z vector, by using a multiplication layer. The result is fed to the generator which is composed of 1 LSTM layer and returns an embedding. In our case the dimension of z was equal to the downsampled velocity profiles lengths (43) and the desired embedding dimension was 16. This compressed representation is exploited for computing the supervised loss as shown in Eq.7.1.

$$\arg\min_{G}\max_{D} V(D,G) = \mathscr{L}_{GAN}(G,D) + \lambda \mathscr{L}_{L2}(G) \tag{7.1}$$

where:

$$\mathscr{L}_{GAN}(G,D) = \mathbb{E}_{x,y}[\log D(x,y)] + \mathbb{E}_{x,z}[\log(1 - D(x,G(z,y))] \tag{7.2}$$

and:

$$\mathscr{L}_{L2}(G) = \mathbb{E}_{x,y,z}[||x - G(z,y)||_2] \tag{7.3}$$

The loss function is composed of two terms, *(i)* the output of the discriminator, that needs to be fooled, and *(ii)* the l2 norm between the embeddings generated and those obtained by encoding the original data through the pretrained encoder. For stability reasons the L2 loss can be modulated by a constant factor lambda (λ), in our experiments equal to 100.

The *discriminator* takes as input the time-series and the associated label. The label is again turned into a dense vector of size z (43) by using a Keras embedding layer. The embedded label is concatenated with the time-series and passed to one GRU layer of size 32 and finally to a densely connected NN layer. This latter outputs the probability that a given velocity profile belongs to the target class defined by the associated label. We adopted a learning rate of 0.01 and Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ for both generator and discriminator. The batch size was 32 and the training lasted 60 epochs. These parameters might have to be fine-tuned appropriately depending on the considered dataset and domain of application. At inference time the embeddings produced by the *generator* can be finally reconstructed by the *decoder*. See Figure 7.1b for reference. Some of the velocity profiles generated by our model are shown in Figure 7.2, while Figure 7.3 offers a 2D visualization of how the synthetic samples match the original data distributions.

Dataset

We focused on the generation of velocity profiles associated with the transportation of objects with different properties, exploiting the same dataset presented in Chapter 3, involving four glasses, identical in shape and appearance, but differing in weight (light: 167 gr; heavy: 667 gr) and content: two of them were completely filled with water, necessitating extra caution to be moved. The hand kinematics during the transport of the glasses was recorded with a sampling rate of 22 Hz with active infrared markers. In Chapter 6 we showed how we trained separate networks for each of the four classes of motion, considering 1001 total transport movements. In this case, for training the GAN, we considered as feature the carefulness required and, as dataset, we used the subset of the 502 transport motions involving the light

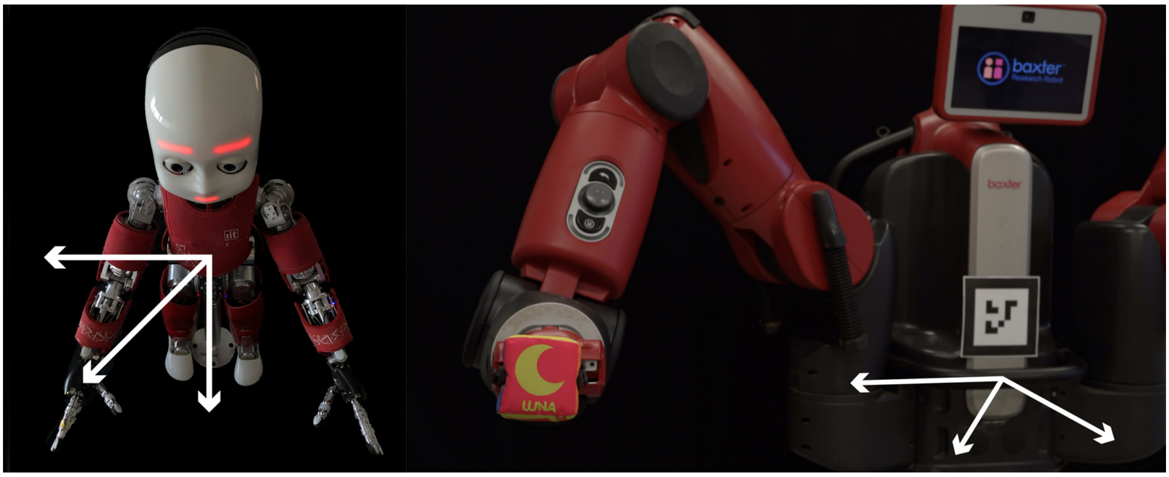


Figure 7.4 Robots' movements direction: the three planes along which the motions were generated, namely frontal, sagittal and oblique with respect to the robots' reference frame, are marked with white arrows.

glasses. This choice was made to avoid extra variability in the data attributed to the weight difference, while preserving instead the variance associated to multiple movement directions.

7.1.2 Robots' movements generation

The second part of this work focuses on the possibility of replicating the desired velocity profiles with the robots end-effector. Independently from the nature of the velocity profile (either synthetic or recorded from human examples), we want the robots to accurately reproduce the desired velocity profiles at the end-effector, following distinct spatial trajectories and maintaining the most communicative characteristics of the movement. We considered two robots, the humanoid robot iCub [109] and the collaborative robot Baxter [52], in order to perform rigorous analysis independently of the kinematics and control of the robots. Although these two robotic platforms present significant differences in terms of appearance and kinematics (e.g. spatial configuration of the Degrees Of Freedom) they both share the ability to interact with humans and to manipulate objects. To assess the ability of the robots to follow a generic velocity profile, we randomly selected three examples generated among the careful class (*ProfileC1, ProfileC2, ProfileC3*), and three among the not careful class (*ProfileNC1, ProfileNC2, ProfileNC3*). We also decided to evaluate the movement executions along three different planes: frontal, sagittal and oblique (intermediate between the other two). For a graphical intuition, see Figure 7.4. Due to the different kinematic configurations of the robots, their range of motion and workspace differed. All trajectories generated for

Baxter covered a path of 60 *cm*, while those performed by iCub were about 30 *cm*, depending on the plane on which the movement occurred. Finally, for each of the six selected velocity profiles and of the three trajectories, we executed the same movement ten times, to measure the robots' consistency in repeating the actions. See robotMovements-videos[1] for a clear view of the generated movements. To avoid any risk of damaging the robots' electronics with water, in the demonstrative videos we opted for showing a soft cube to represent a not careful transport motion and a pair of glasses as an example of a delicate object to move with caution. For both robots, we saved the end-effector Cartesian position during the trajectory execution for subsequent analyses. In particular, we computed the tangential velocity magnitude of the end-effector to compare it with the one planned by the controllers.

Baxter robot

We decided to control the Baxter Robot with MoveIt!, the standard ROS motion planning framework [31]. This framework is familiar to all those who use Baxter without necessarily installing additional software or possessing advanced knowledge of control theory. We first remapped the selected velocity profiles to fit the desired movement length, we kept constant the temporal duration of the action while modulating the magnitude of the velocity. This operational choice is supported by the isochrony principle. In human actions, the average speed of point-to-point movements rises with distance, and movement duration is only slightly dependent on movement range [166]; indeed, there are multiple proofs of time-invariance in the upper limb movements [4, 10].

Since our conditional network produced velocity profile norms, they were independent of the spatial trajectory. Fixed the desired start and end point of the trajectory, in Cartesian coordinates, we used Moveit! to plan the intermediate spatial points that the end-effector had to go through. Such desired trajectory consisted of *n* Cartesian points equally spaced by 0.01 m. The choice of using a fixed spatial step, while intervening on the temporal execution was driven by the MoveIt! controller implementation. MoveIt! was asked to plan a joint trajectory defined by these Cartesian points, as a sequence of poses in the joints space. Then, knowing the number of samples generated by Moveit and the velocity profile that we wanted the end-effector to follow, we computed the timestep by which each trajectory point had to be reached. By doing this, it was possible to define the Cartesian velocity of each spatial step and the desired velocity profile was replicated. Finally, we gave as input to the ROS

[1] Movements execution on Baxter and iCub robots
https://www.youtube.com/playlist?list=PL9sy5y8WKC4kX4AMGCdXYneK6-VqaR4M5

Joint Trajectory controller[2] the sequence of joints to be reached, defined by Moveit!, and the vector of timesteps.

iCub robot

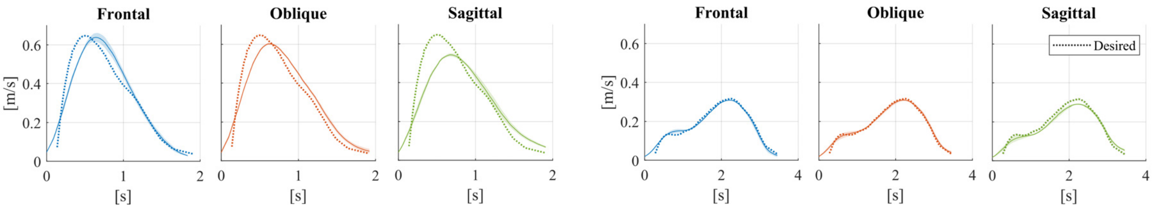
Similarly to Baxter, we controlled the iCub robot with its default “Cartesian Controller" developed by Pattacini et al. [124]. Also on iCub we have implemented movements that preserved the duration of the synthesized velocity profile; for this reason, the number of samples and the magnitude of the velocity norm had to be scaled to adapt to the different Cartesian trajectories. On iCub, the end-effector controller required to maintain a constant timestep between adjacent trajectory points: in order to replicate the desired velocity profile, we then varied the Euclidean distance between the spatial Cartesian points.

7.2 Results

7.2.1 Robots’ Movements Assessment

To assess the performances of the controller in replicating our reference velocity profiles, we compared the executed movements with the desired ones. First, in Sec. 7.2.1, we provide a qualitative overview of the performed motions. Then, we compare the maximum velocity amplitude reached with the desired one; this, to investigate whether the robots can reach the same range of velocity that humans apply when manipulating objects, delicate or not (see Sec. 7.2.1). Finally, we compute the Pearson correlation between each executed velocity profile and the planned one (see Sec. 7.2.1).

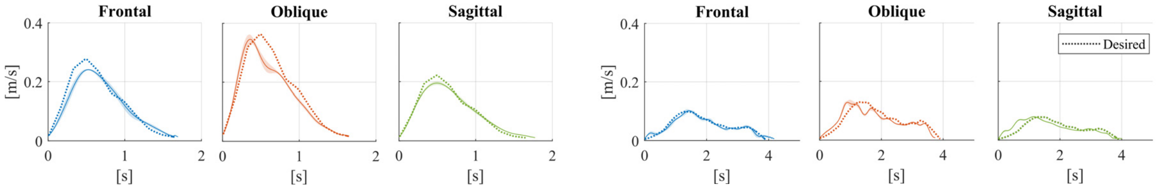
As previously explained in Sec. 7.1.2, we randomly selected from our dataset three velocity profiles to exemplify the C class and three for the NC one. We then considered three possible directions for the robots’ end-effectors to move: along the frontal, sagittal, and transverse planes. Each velocity profile was rescaled according to the selected trajectory, and each movement was repeated ten times, to assess the robot’s repeatability. This led us to have, for each one of the three directions, ten repetitions of the same movement for each of the considered velocity profiles.



(a) Not careful - ProfileNC2

(b) Careful - ProfileC1

Figure 7.5 **Baxter executed velocity profiles:** The represented profiles are respectively ProfileNC2 for the not careful class (7.5a) and ProfileC1 for the careful one (7.5b). Ten repetitions of such profiles are executed along three different movement directions: frontal, sagittal and oblique. The mean of the velocity executions is a solid line, while the standard deviation is a more transparent area of the same color. The desired velocity profile is dashed and it is identical in the three directions, since the distance covered by the Baxter robot was the same.



(a) Not careful - ProfileNC3

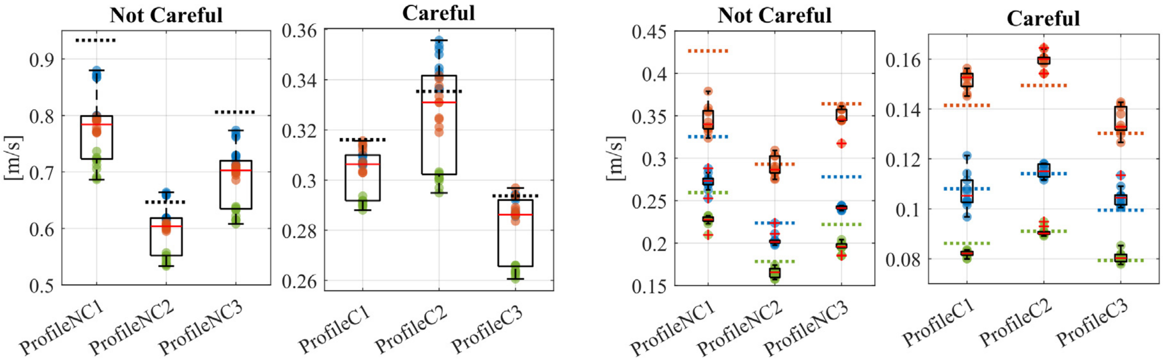
(b) Careful - ProfileC3

Figure 7.6 **iCub executed velocity profiles:** The represented profiles are respectively ProfileNC3 for the not careful class (7.6a) and ProfileC3 for the careful one (7.6b). Ten repetitions of such profiles are executed along three different movement directions: frontal, sagittal and oblique and the graphical conventions are the same of Figure 7.5, with the desired velocity profiles dashed.

Movement execution

In Figure 7.5 is shown how the Baxter robot executed, along the three different planes, one among the three selected not careful (7.5a) and careful (7.5b) velocity profiles. Even though the movement repetitions along the same direction are comparable, with a very low intra-class variability, some trajectories appear more suitable than others in reproducing the velocity at the end-effector; it can be noticed how, for the not careful movements (Figure 7.5a), along the frontal plane the velocity peak reaches the desired one, while in the sagittal one it remains lower. The execution of the careful movements (Figure 7.5b) shows no delay with respect to the planned velocity, with the profile peak and shape preserved for all the movement directions.

[2] http://wiki.ros.org/joint_trajectory_controller



(a) Baxter robot

(b) iCub robot

Figure 7.7 **Amplitude of the velocity peaks**: the desired amplitude is marked with a black dashed horizontal line for Baxter (7.7a) robot, while it follows the color code for iCub (7.7b): blue for frontal movements, orange for oblique and green for the sagittal ones. The red lines in the boxplots represent the medians, the black rectangles limit the 25th and 75th percentiles of the executed velocities. Since iCub covered paths of varying length, depending on the direction, we report in Figure 7.7b the boxplot for each plane.

Regarding iCub, in Figure 7.6 are represented two other velocity profiles among the selected ones, different from those represented for Baxter in Figure 7.5, one for the NC (7.6a) and one for the C (7.6b) class. On this robot, the shape of the not careful profiles is well aligned with the planned movement, and the velocity amplitude almost meets the desired one. Considering Figure 7.6b, a small delay between the planned and executed movements can be noticed for the Oblique and Sagittal directions.

Maximum velocity

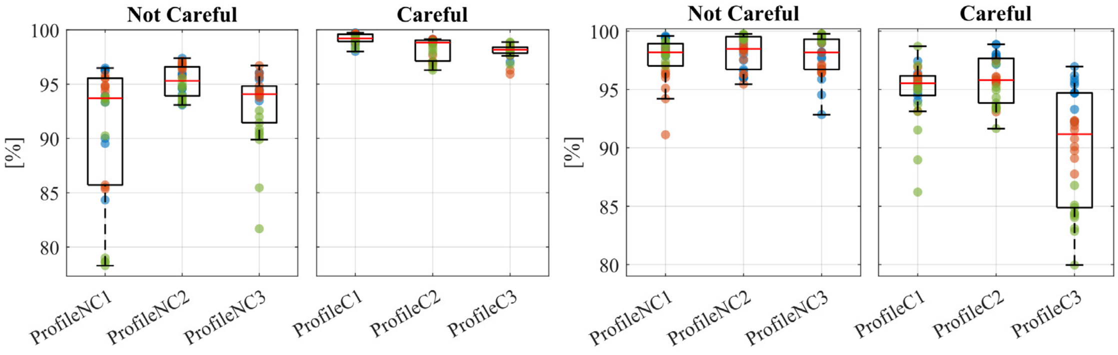
In Figure 7.7 is represented as a boxplot the maximum speed reached by the two end-effectors for the three careful and not careful selected velocities among all the repetitions. The planned maximum velocity is represented as a dashed line. The color code distinguishes the planes of the action execution: green along the sagittal plane, orange for the oblique and blue in the frontal.

For the Baxter robot (Figure 7.7a), the desired maximum speed is indicated by a single black horizontal bar; since the trajectories along the three planes covered the same distance of 60 *cm*, the associated velocity profiles were rescaled in the same way by the controller. Therefore, the Baxter robot was required to reach the same maximum speed independently from the movement direction. While the distribution of NC movements do not reach the desired speed, for most of the trials, the careful amplitudes are generally comparable with the planned one, with some variability depending on the trial considered. As anticipated by Figure 7.5, some trajectories are more suitable than others in reproducing the desired speed, and this reflects also in the maximum peak distributions. Indeed, Baxter reached different maximum speeds depending on the direction: for both the NC and C classes, and independently from the selected velocity profiles, the movements along the sagittal plane, in green, have the lowest maximum speed, followed by the ones in the oblique plane and lastly by the ones in the frontal plane, in blue, which are closest to the desired maximum speed. This reveals a trajectory dependency in the Baxter performances, that should be taken into account when generating new movements.

Considering iCub, the path covered along the three planes differed, due to the range of motion of each joint in the kinematic chain: the workspace along the three planes was different. Therefore, the same original velocity profile was rescaled differently for the three movements (see Figure 7.6, where examples of the desired velocity profiles for the NC and C classes are mapped in the three directions). Figure 7.7b reports the speed distributions with colored dashed lines indicating the desired values; it can be appreciated how, in most of the cases, the distribution associated with each direction of motion is comparable with the corresponding planned values. In particular, the oblique direction in orange required faster movements, while the sagittal one in green was the slowest. This was deliberately planned in advance for the iCub robot, according to the trajectory, as confirmed by the dotted reference lines. However, it can be noticed how, for the NC class, iCub executed peaks are systematically lower than the desired ones, with the exception of the oblique trajectories with *ProfileNC2*, which are totally satisfying.

Pearson Correlation Coefficient

Finally, in Figure 7.8, we considered the Pearson correlation coefficient to evaluate how well the executed velocities are related to the planned ones. In Figure 7.8a are represented the correlation coefficients for every trial obtained with Baxter robot. For the not careful class, the median correlation settles in the $94-95\%$ range, while in the careful class is above the 98% in all the three trials. Baxter robot efficiently explored the trajectories, maintaining a good correlation with the desired velocity, especially in the careful movements, which required lower speeds. The same reasoning was applied for iCub robot, and the resulting correlation coefficients are represented in Figure 7.8b. Differently from Baxter, the best results were obtained in the not careful, faster movements, with a median correlation above 98%. The careful executed movements were however well correlated with the planned ones, even though ProfileC3 showed a wider distribution. This example is the same whose velocity profiles are reported in Figure 7.6b, where a delay between the planned and executed



(a) Baxter robot (b) iCub robot

Figure 7.8 **Pearson correlation** on Baxter (7.8a) and iCub (7.8b) robot. For every movement, the correlation has been calculated between the planned and the executed velocity profile. The graphical conventions are the same as in Figure 7.7, with blue dots for frontal movements, orange for oblique and green for the sagittal ones.

movements can be detected in the sagittal and oblique trajectories. This reflects in the Pearson correlation.

7.3 Discussion

We can identify two main goals in this Chapter: (i) to artificially generate velocity profiles associated to the manipulation of object properties, exploiting generative networks and (ii) to assess whether two different robots can replicate human-like velocity profiles.

Regarding (i), we proposed a novel conditional GAN architecture that generates new synthetics embeddings, belonging to the C and NC class, depending on the provided conditioning label. We showed that our network is able to generate meaningful synthetic profiles, belonging to the same distribution of the original ones without exactly replicating them. Having a network producing artificial samples, instead of simply copying human examples, grants an unlimited source of movements, always different and unique, yet consistent with the object's properties and as communicative as human ones. Moreover, by using a conditional network, we could potentially generate data intermediate to the two initial classes, therefore completely new for the velocity profiles used to train the model. However, the meaning of such intermediate profiles, also in the execution phase, should be properly investigated and assessed in future studies.

Concerning (ii), we controlled two quite different robots in performing movements following the selected velocity profiles, proving that it is possible to replicate the same kind of movements with their end-effectors. From the analyses presented in Sec. 7.2.1, we can generally conclude that both the robots successfully executed the trajectory with the desired velocities. The other considerations are tailored to the two robots, but are meant to discuss what kind of robot characteristics can impact the synthesis of communicative robotic movements. The amplitude of the executed movements was different between the two robots, due to their dimension, shape, and structural design. While all Baxter trajectories covered 60 *cm*, iCub (a smaller robot) trajectories ranged around 30 *cm*, with variability depending on the considered plane. For this reason, even though the original velocity profiles were the same, they were re-scaled differently on the two robots, to cover a different distance with the same duration. Considering the two classes, C and NC, we detected a difference in the robots' executions. Indeed, according to the Pearson correlation coefficient in Figure 7.8, Baxter performed better in replicating the careful profiles, with lower accelerations and intensity, while iCub correlation with respect to the desired profile was higher for the not careful class. Figure 7.7 helps better understand the distinction between the two robots. Baxter presents a strong variability in the maximum velocity reached depending on the three directions of movement and, especially for the NC class, it fails to meet the planned magnitude. iCub shows more uniform performance with respect to the movement directions, however in the NC class we notice again a disparity between executed and planned velocity peaks. This is an important result for future interaction experiments, where the movement directions should be chosen along the most appropriate plan.

The lower performance detected with Baxter may be attributed to a combination of some mechanical limits of its actuators and of MoveIt!. Indeed, Baxter is a low-cost manufacturing platform with well-known constraints, due to the choice of prioritizing safety over accuracy in its movements [33]. The discrepancy detected on the planes where the movement occurs can be explained by a different involvement of the joints during the action. Such difference is noticeable, especially for the NC movements, which required higher accelerations (see Figure 7.5a). Baxter, in the first part of the movement, is not able to keep up with the acceleration of the planned profile, causing a delay of the moment when the maximum peak speed is reached. According to the available information [1], Baxter is able to reach speeds up to 1 m/s, hence the generated NC profiles are in the correct range, however it may need a longer time to achieve such magnitudes.

We can conclude that our approach successfully generates meaningful velocity profiles, and that it is possible to reproduce with different robotic platforms movements belonging either to the C or NC class. However, robots limitations should be taken into account when pursuing a similar approach. The advantage of using MoveIt! to control Baxter is that our method can be easily exported to any other robotics platform based on a ROS framework, for further testing with robots with different characteristics.

Even though the tested movements were associated with the manipulation of objects with peculiar properties, our approach offers a much broader perspective on the generation of communicative movements. Indeed, the controllers we designed can potentially follow any kind of velocity profile, obviously staying within the motion limits of the robots. For this reason, given the availability of a proper dataset to generate the velocity input, the end-effector could be controlled to communicate other objects' properties, such as the weight, or even an emotional attitude in the gesture, such as a gentle or a rude one.

In this work, we chose to focus on the trajectory of the end-effector to extend the domain of applicability of our approach: a one-to-one mapping of the human motion would be feasible only for those robots with a strictly human-like kinematic chain. Instead, the modulation of the end-effector motion is possible on most robotic platforms, exploiting standard algorithms to solve the inverse kinematics and motion planning. The following Chapter will address how the movements generated with iCub and Baxter are perceived by humans. Indeed, it remains to be assessed whether is sufficient to modulate the end-effector adopted velocity to communicate the carefulness in the gestures or if other joints in the kinematic chain should be properly controlled. Given the implementation on two quite different robots (iCub, humanoid, and Baxter, collaborative but with an arm kinematic chain far from the human one), their movements, even if controlled with the same principle, could produce a very different effect on the interaction. Alongside the perception of the communicative intention of the robot, we will investigate its effect on how humans interact with the robot and carry out their own tasks.

Part III

The effect of implicit communication: towards the interaction

Chapter 8

Robots with different embodiments can express and influence carefulness in object manipulation

Elaboration of an article published in:
International Conference on Development and Learning (ICDL 2022) [91]

8.1 Introduction

The previous Chapter presented two main contributions: a novel generative network for producing velocity profiles associated with the transport of an object and the control of different robotic platforms to reproduce with their end-effector the desired time-series. Such an approach arose from the necessity of producing communicative robot motion associated with manipulating objects with certain characteristics, in particular, to show carefulness, or its absence, in the gesture. However, if, from a technical point of view, the strategy proved to be solid, it still has to be assessed its efficacy in communicating the desired attitude of the robot.

As a broad review by Venture and Kulić points out [162], robots can exploit their embodiment to be communicative while performing a task to transmit more information to their partner. In particular, research on implicit communication focuses mainly on two goals: conveying the robot's intentionality or affective/emotional attitude. Another context where robots' communicative potential has been explored is the co-verbal one: generating human-like gestures correlated with speech makes the robot more understandable and perceived positively [138, 164]. In this Chapter, we want to investigate implicit communication associated with object manipulations performed by two robots with different embodiments, designing their movements to express carefulness or not during the transportation of objects.

Research questions and objectives

To provide the same amount of information about the object they are carrying, robots should modulate their kinematics accordingly, improving the task's safety and efficiency. Although from a technical point of view, we showed how robots could replicate those motions preserving the original kinematic qualities, we have no evidence that this allows robots to communicate the carefulness associated with an object to a human. To the best of our knowledge, there are no works in literature where the perception of the robot arm kinematics modulation, associated with the transport of objects with specific characteristics, has been investigated. In this context, the research questions we aim to target are the following:

> 1$^{\text{st}}$ Research Question (**RQ1**): Are human-derived motions enough for the robot to communicate the carefulness required to transport an object?
>
> 2$^{\text{nd}}$ Research Question (**RQ2**): Can the carefulness demonstrated by a robot influence how humans handle objects?
>
> 3$^{\text{rd}}$ Research Question (**RQ3**): Do different robot embodiments affect the carefulness perceived by humans?

To explore these research questions, we deployed the solution proposed in the previous Chapter to generate transportation motions on two different robotic platforms: iCub and Baxter. Due to their different conformation, these two platforms are suitable for exploring the possible effect of the embodiment **RQ3**. Videos recorded with the two robots have been used to investigate **RQ1** through an online questionnaire and **RQ2** with an in-presence experiment. Even though we are aware that the physical presence of robots has a stronger effect on the interaction [93], using videos allowed us to test simultaneously, if and how the appearance of the robot would influence the same set of participants.

8.2 Methods

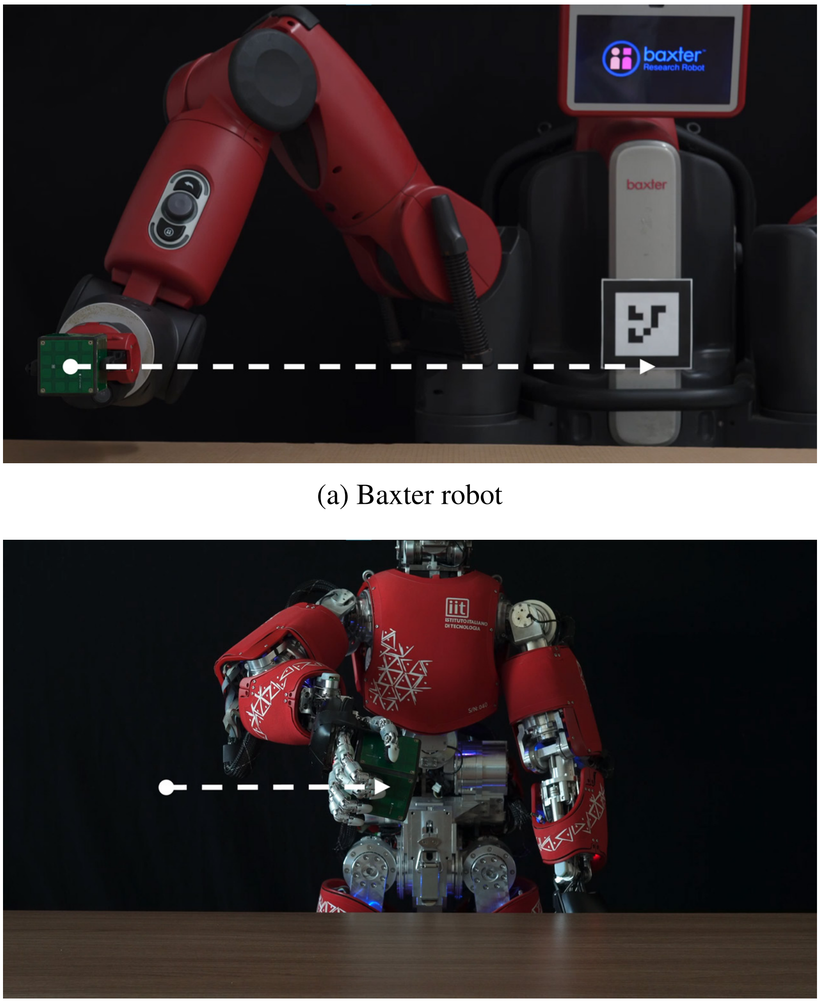
We decided to address **RQ1** by preparing a questionnaire including a few videos of Baxter and iCub transporting a blurred object, with what we programmed as a careful (C) or not careful (NC) attitude. Participants were asked to evaluate the movement's carefulness, on a five-points Likert scale from “Not careful" to “Careful" (carefulness score). Moreover, we presented a list of common-use objects and asked how careful we should be in transporting such items on the same scale. This is to clarify the concept of carefulness and to understand which attributes make an object more or less delicate to carry from the participant's perspective. Indeed, even though we generate robot motions focused on respecting the appropriate velocity profile, it could be necessary to consider other factors according to the object characteristic inducing carefulness. For example, while manipulating fragile objects, we want to minimize the risk of collisions.

Concerning **RQ2**, we opted for a within-subject in-presence experiment. In this case, participants transported a sensorized cube after watching videos of one of the robots carrying the same item, again either with a C or NC attitude. The experiment allowed us to investigate if participants, with no prior knowledge of the robots' attitude, would be influenced by their movement to the point of modulating their kinematics.

Both the experiments, relying on videos realized with different robotics platforms, will also allow quantifying **RQ3**. In the following part, we will give a short overview of how we generated iCub and Baxter movements. Then, we will describe the questionnaire. Finally, we will present the in-presence experiment setup.

Generating communicative movements on robots

To produce the movements of the two robots along the frontal plane, we followed the approach described in Chapter 7. We selected a subset of careful and not velocity profiles and controlled the robots' end-effectors to follow them. To prepare the stimuli for the questionnaire and the in-presence experiment, we recorded videos of iCub and Baxter moving the same object, a green cube, along the frontal plane. In particular, we used 2 distinct velocity profiles per condition (C/NC) per robot, resulting in 8 videos for the online questionnaire, while 6 videos for each combination for the experiment, resulting in 24 different videos. Baxter trajectories had a range of 60 cm. Instead, iCub ones were shorter, around 20 cm, due to its kinematic configuration and, since the movement's original duration was preserved, slower than Baxter ones. Indeed, the movements in the videos were characterized by a maximum velocity reached by the end-effector in the not careful condition of 0.281 ± 0.037 m/s for iCub, and of 0.871 ± 0.080 m/s for Baxter. In the careful videos instead, iCub reached a maximum velocity of 0.145 ± 0.061 m/s, while Baxter of 0.312 ± 0.049 m/s. Regarding the end-effector mean velocity, we report in the not careful condition for iCub 0.137 ± 0.023 m/s, while for Baxter 0.348 ± 0.129 m/s. In the careful videos, iCub had a mean velocity



(a) Baxter robot

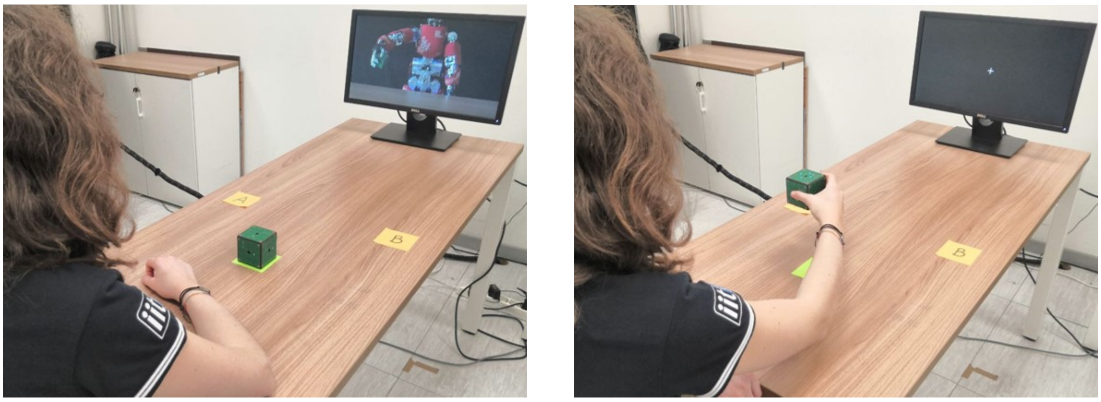
(b) iCub robot

Figure 8.1 Frames of the videos portraying the transport of a green cube along the frontal plane. Baxter (8.1a) is at the beginning of the movement, while iCub (8.1a) at its end. The dashed arrows represent the trajectory of the movement, which starts from the white dot and develops in the direction of the arrow.

of 0.076 ± 0.032 m/s, while Baxter of 0.132 ± 0.048 m/s. As an example, in Figure 8.1 are shown two images taken from the videos.

8.2.1 Online questionnaire

At the beginning of the questionnaire, we asked participants their age, gender, and general knowledge about robotics, with possible answers going from 1 (“None - I have no idea what it is") to 5 (“Professional - I work, or worked, in the area of robotics"). In the first part, we asked them to observe the 8 C/NC videos of Baxter and iCub moving their end-effector, presented



(a) Participant watching one of the iCub’s videos where it transports the sensorized cube

(b) The participant places the cube in the indicated position, after watching the video

Figure 8.2 Example of the experimental setup where the participants were asked to manipulate the cube (8.2b) after watching a video of one of the robots transporting it (8.2a).

in random order. In the questionnaire, the videos of Baxter’s and iCub’s movements had the same setting as in Figure 8.1, but the end-effector was blurred. For each video, participants had to rate from 1, “Not careful" to 5 “Careful" the observed action, answering the question “How attentive was the robot in moving the object?". Indeed, they were told the robot was moving a cubic container, and since the end-effector was blurred, they could not see what was inside it. In the second part, we proposed a list of 12 items and asked to imagine they would have to be carried around. The instructions to this part stated ‘For each of them, evaluate how much attention and care should be associated with the manipulation", again on a scale from 1, “Not careful" to 5 “Careful". We chose the items among everyday use objects, and they were: glass full of water till the brim, rubber ball, pair of scissors, worn out pelouche, lit candle, wooden cube, origami, plastic bottle, crystal cup, USB charger, face mask and a pair of glasses. 49 people (age: 25.5 ± 5.7, 25 females, 23 males, 1 non-binary/genderfluid) freely chose to answer the questionnaire.

8.2.2 In-presence experimental setup

We ran an in-person experiment to see if, beyond understanding the carefulness in the robots’ movements when explicitly questioned about it, observing the robots’ actions could influence how a person carries an object. The experiment involved 11 healthy right-handed subjects that voluntarily agreed to participate in the data collection (8 females, 3 males, age: 26.5 ± 3.1). All volunteers are members of our organization, but none is implicated in the research.

For this study, we used 6 videos per robot per condition, for a total of 24 trials, reproduced in casual order and with casual direction of the robot movement: always on the frontal plane, but going either from right to left or vice versa, by mirroring the video.
During the experiment, participants were asked to look carefully at the video displayed on the monitor in front of them (see Figure 8.2a) and afterward, following the instruction given by a synthetic voice, to move the cube they would find in front of them either to position A or B, marked by a colored paper square on the table (see Figure 8.2b). The object they had to move was the iCube, a green sensorized cube of side 5 cm [152].

Each participant was asked to read the same instructions carefully designed so as not to introduce bias in the experiment. In particular, we explained that they would watch videos of the robots transporting the iCube. We used several prototypes of this cube, aesthetically indistinguishable but more or less delicate depending on their assembly state and internal components. At this point, we showed a box containing three iCubes to be used for the experiment. Even though we actually used only one of them, the purpose of having three was to avoid participants assuming a priori a binary interpretation of the videos (C or NC, even though this was exactly what we implemented in the robot actions). At the end of each trial, the participants were asked to close their eyes, for the experimenter to place a new cube in the starting position without influencing them with how the cube was moved. A sound informed the subject to open their eyes and prepare to watch a new video. We explained that the purpose of the experiment was to simulate a collaboration between the robot in the video and the participant. So, they were asked to imagine that, in every trial, the cube on the table was the one manipulated by the robot in the video, who virtually deposited it on the table for them to grab. To avoid a simple copy of the observed movement, as already proposed in other studies [98], while the robots' actions occurred in the frontal plane, the participants moved the iCube in the sagittal direction. Before starting with the video reproduction, we ran a baseline of 6 trials with no stimuli, for the participants to familiarize with moving the cube from the starting position to the ones in A or B. To precisely record the hand kinematic during the transport of the cube, we used active infrared markers from the Optotrak Certus®, NDI, motion capture (MoCap) system, which gave the 3D position of each marker with respect to a common reference frame, with a frequency of 100 Hz. We positioned four markers on the metacarpal bones of the right hand and considered for further analyses the most visible one in each trial. For the purpose of this study and the sake of space, we report the results of the analysis of the kinematics data coming from the motion capture system and not those recorded through the iCube.

8.3 Results

To analyse the acquired data, we used Jamovi software [2], in particular the GAMLj module [55] for mixed models.

8.3.1 Online questionnaire

To assess if the implicit information embedded in the robots' movements is communicative of their carefulness, we first evaluated the responses regarding the perceived carefulness in the videos observed by the participants. Considering that two videos were presented for each robot in the C and NC conditions, each participant evaluated 8 videos, resulting in a total of 392 carefulness scores, ranging from 1 (NC) to 5 (C). From these data, we ran a mixed model assuming the carefulness Score as dependent variable, the robot type, the condition and their interaction as factors, and the subjects as cluster variables. The resulting metrics are shown in Figure 8.3. The effect of condition resulted significant ($NC - C$, $estimate = -1.041$, $SE = 0.0878$, $t = -11.85$, $p < 0.001$) as well as the effect of the robot ($iCub - Baxter$, $estimate = 0.429$, $SE = 0.0878$, $t = 4.88$, $p < 0.001$). Moreover, also the interaction between robot and condition is significant ($careful \cdot Baxter$, $estimate = 0.531$, $SE = 0.1756$, $t = 3.02$, $p = 0.003$). These results indicated that the videos showing what we planned as careful movements were significantly perceived as more careful, getting a score higher by 1.041. Also, iCub was generally perceived as more careful than Baxter. The carefulness score attributed to Baxter videos covered a wider range, with a stronger difference between the C and NC conditions (see Figure 8.3). For both robots, the not careful movements were perceived as rather neutral on the Likert scale.

Regarding the carefulness score associated with the proposed 12 items, the results are shown in Figure 8.4. The answers allowed to create a scale going from the objects which were considered not critical to be transported, such as the rubber ball, to the one which requires the highest attention in the manipulation, that is the glass of water full till the brim. Bringing together the results of the carefulness scores attributed to the robots' videos, in Figure 8.3, and to the proposed objects, in Figure 8.4, we could evaluate for which items on the list the robots' movements might be considered suitable. On average, none of the robots' movements seem adequate to move the items evaluated as delicate as the *Lit Candle*, onwards. iCub NC movements obtained a carefulness Score which could match the *Scissors* and *Pair of Glasses*. Even though we did not ask specifically about the item, we can posit that iCub NC movements are suitable, in general, to manipulate objects with an equivalent fragility. Following the same reasoning, Baxter NC attitude seems suitable for those objects which obtained a neutral evaluation, as the *Origami*.

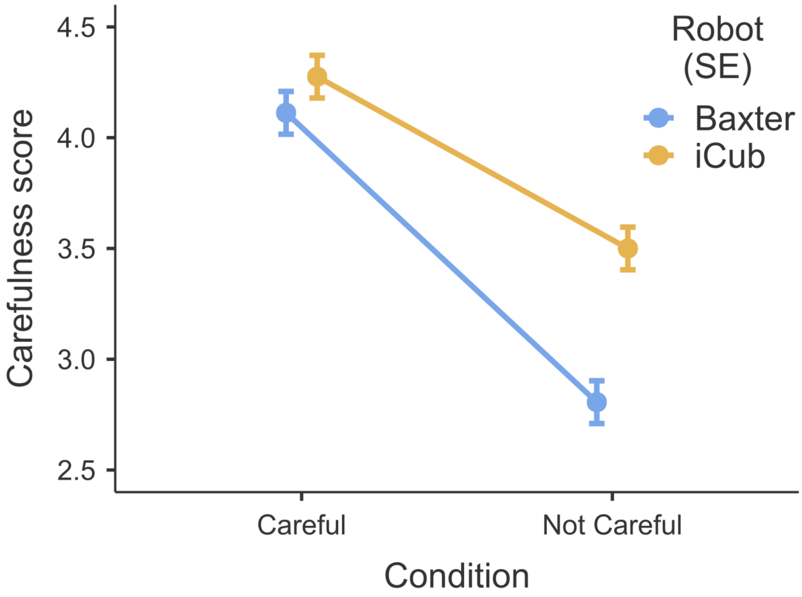


Figure 8.3 Carefulness perception of Baxter and iCub videos proposed in the questionnaires. Both robots were perceived as significantly more careful when executing the movements planned as careful. Moreover, there is a significant difference in the perception of the two robots in both conditions, more striking in the not careful one.

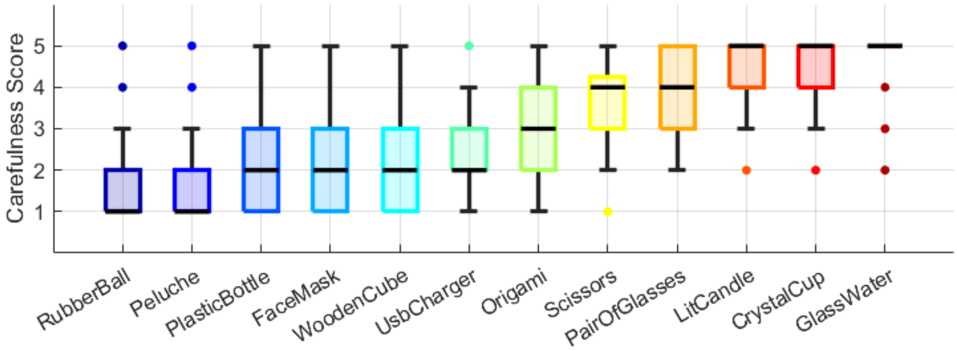


Figure 8.4 Carefulness score obtained by each item proposed in the questionnaire. The colored boxes represent the 25th and 75th percentiles, the black horizontal line the median. The single dots are outliers.

8.3.2 In-presence experiment

From the results presented in the previous part, we verified that the movements produced with iCub and Baxter were indeed perceived as more or less careful, and coherently with what we planned. The aim of the in-presence experiment was to go a step further and to investigate whether the carefulness modulation of robot motion was sufficient to affect how participants performed their own movements. To answer this question, we computed the hand velocity norm from the components of the velocity, obtained by deriving the three-dimensional position of the considered marker. As a metric to quantify the modification of human movement, we considered the maximum velocity reached and the duration of the transport movement performed after watching the video. Indeed, previous studies [45] and our findings in Chapter 6 showed that the velocity peak is a significant measure to distinguish between careful and not careful movements: in particular, careful movements are characterized by lower velocities and longer durations.

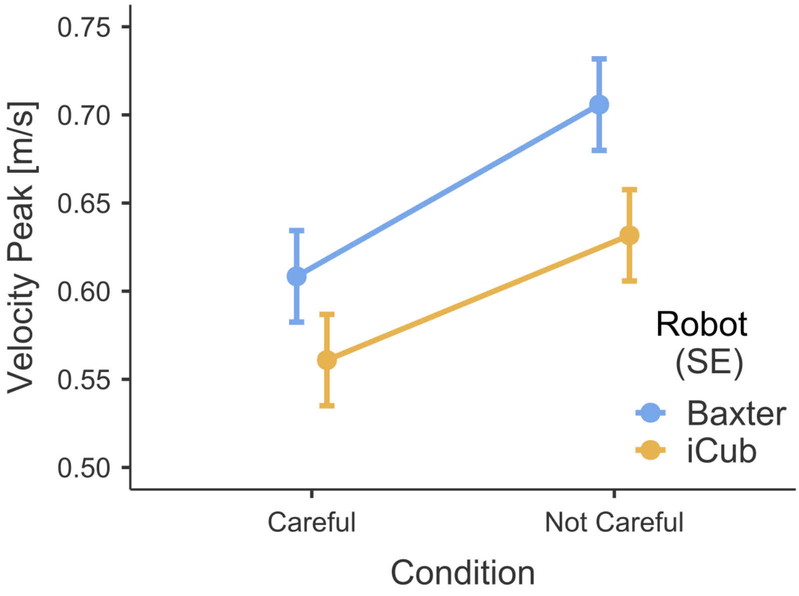


Figure 8.5 Kinematics modulation of the cube transport movements performed by participants. The hand velocity is significantly modulated by the condition of the video associated to the movement: when the robot performed a not careful movement, participants reached a significantly higher maximum velocity than when the robots showed a careful attitude. There is also a significant effect of the robot.

In the in-presence experiment, six videos were presented for each robot in the C and NC conditions, for a total of 24 trials for each of the 11 participants. We ran a mixed model assuming the magnitude of the velocity peak as the dependent variable and the robot type, the condition and their interaction as factors, and the subjects as the cluster variable.

The effect of condition resulted significant ($NC-C$, $estimate = 0.0841$, $SE = 0.0127$, $t = 6.60$, $p < 0.001$) as well as the effect of the robot ($iCub-Baxter$, $estimate = -0.0608$, $SE = 0.0128$, $t = -4.77$, $p < 0.001$). The interaction between robot and condition was instead not significant ($careful \cdot Baxter$, $estimate = -0.0266$, $SE = 0.0255$, $t = -1.04$, $p = 0.297$). The results are shown in Figure 8.5. It can be noticed that the transport movements performed after a careful video stimuli have indeed lower peaks of velocity, with an estimate of the velocity reduction equal to 0.841 cm/s. There is also an effect of the robot, with the videos of iCub soliciting slower movements, with an average reduction in the peak speed of -0.608 cm/s, with respect to Baxter. This is in agreement with what we found from the questionnaires, where iCub movements got a higher carefulness score (see Figure 8.3). For the duration of the movement, we ran the same mixed model using the temporal duration as dependent variable. Again, we found a significant effect of the condition ($NC-C$, $estimate = -0.1796$, $SE = 0.0233$, $t = -7.70$, $p < 0.001$) and of the robot ($iCub-Baxter$, $estimate = 0.0886$, $SE = 0.0233$, $t = 3.80$, $p < 0.001$), not of their interaction ($careful \cdot Baxter$, $estimate = 0.0758$, $SE = 0.0466$, $t = 1.63$, $p = 0.105$). When moving the cube from the starting position to the final one, the covered distance was always about 30 cm, so the time difference detected depends on how the participants modulated their movement speed. Confirming our hypothesis, after observing careful robot manipulations, participants moved the cube in a slower way, with an estimated increase of duration of 179.6 ms. Analogous considerations were verified as well for other kinematics metrics, such as the mean velocity of the transport or its maximum acceleration.

As an additional control, we also compared the movement duration and the velocity peak of the transport after the video stimuli with those of the cube manipulation in the initial baseline, when no videos were presented and participants performed a simple pick and place of the cube. We ran two mixed models, one for each kinematic metrics as dependent variable, the condition (baseline, C or NC) as factor, and the subjects as cluster variable. The results for the velocity peak metrics are shown in Figure 8.6. For both the metrics, we obtained a significant difference between the Baseline and the careful condition (Movement duration, $C-Baseline$: $estimate = 0.1296$, $SE = 0.0282$, $t = 4.59$, $p < 0.001$; Velocity peak, $C-Baseline$: $estimate = -0.0644$, $SE = 0.0163$, $t = -3.94$, $p < 0.001$). This shows that when the robots displayed an attentive attitude, the participants moved significantly slower than they would have naturally. Instead, the models did not evidence a significant difference between the Baseline and the NC condition, even though, as shown in Figure 8.6, participants produced slightly faster and shorter movements than during the baseline (Movement duration,

$NC - Baseline$: $estimate = -0.0497$, $SE = 0.0282$, $t = -1.76$, $p = 0.08$; Velocity peak, $NC - Baseline$: $estimate = 0.0196$, $SE = 0.0163$, $t = 1.20$, $p = 0.231$).

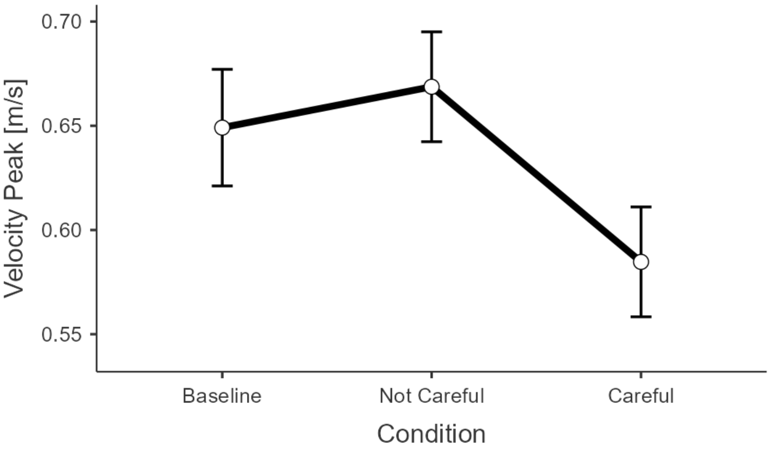


Figure 8.6 Comparison of the maximum velocity reached during the cube transportation in the different phases of the experiment: the baseline, when no stimuli was provided, and the main part where either careful or not careful videos were presented. There is a significant difference between the baseline and the careful condition.

8.4 Discussion

In this Chapter, we aimed to answer three main research questions: **RQ1** assess if the implicit information embedded in the robots' movements is perceived; **RQ2** evaluate the potential modulation produced by the robot movements on how humans handle objects and, **RQ3**, investigate whether the robots' embodiment affects how their actions are perceived.

Even though the idea of carefulness associated with object manipulation has already been explored, it is not easy to give an unequivocal definition of the term. For instance, unlike weight, carefulness cannot be measured with a defined scale. From the questionnaire answers in Figure 8.4, about grading the carefulness required for manipulating common-use objects, what emerges is that the concept of carefulness does exist and it is shared among people. Moreover, the emerging gradation correlates with distinct attributes of the items, which may contribute to making an object delicate to handle: the content (glass with water), the material (crystal), the preservation of the object state (lit candle), the value (pair of glasses) or the potential danger (scissors). The questionnaires outcomes positively confirmed that the movements we produced, with both the robotic platforms, were indeed communicative of the desired feature, that is the presence or absence of a careful attitude in the action (see

Figure 8.3). Moreover, the robots' gestures proved to be efficient in soliciting a motor change in the participants: the simple request to imagine a collaboration with the robot, such as virtually carrying the same object, was enough to bring out a motor adaptation, even along a different direction with respect to the plane where the robot action took place. In a real one-to-one human-robot collaboration, the modulation of human movements in response to the robot ones could become even stronger. Interestingly, we found an effect of the robot: iCub was perceived as generally more careful in its movements, while Baxter presented a more marked difference between the two conditions. The same result also emerged when comparing the kinematics metrics of the in-presence experiment, presented in Section 8.3.2: iCub videos solicited slower actions by the participants, i.e., more careful manipulation of the cube they found on the table. We can ascribe this to two main factors: how the movements were generated and the embodiment of the robots. As explained in Section 8.2 and in the previous Chapter with more detail, when controlling the robot end-effector to follow the desired trajectory with a velocity profile of choice, in our implementation, the duration of the original movement is preserved; this means that for longer trajectories the velocity of the movement needs to be higher, whereas for shorter paths the original velocity profile is stretched. The different sizes of the two robots cause the trajectories covered by iCub to be shorter and lower the speed of its movements. iCub's appearance is more human than Baxter's, not only in the hand shape but also in the kinematic configuration of its arm; this could lead people to identify themselves more with its movements and perceive them, regardless, as more delicate. However, the difference between the C and NC conditions in both robots was significant for the carefulness score attributed to the videos and for the kinematic measures of motor contagion. Hence, even with a platform not designed to have an anthropomorphic shape - such as Baxter - a bioinspired modulation of movement speed effectively communicates motion carefulness. The movements we implemented were rated as going from a rather careful attitude to a neutral one. Indeed, what we labeled as not careful movements, were perceived as neutral and Figure 8.6, allows to better understand better how a C or NC movement is modulated with respect to standard baseline actions. In everyday life, moving objects in a not careful manner seems to be the natural way of transporting them: the kinematics modulation emerges when we are challenged with a difficult task or if we want to show caution. This seems reasonable to optimize the actions' efficiency and should also be considered when programming robots' behavior.

In the next Chapter we will present how the same approach can be applied to a robotic manipulator, with kinematics far from the human one, studying an in-presence dyadic interaction.

Chapter 9

The effect of robot communicative motions in a joint task

Elaboration of two articles
Published in: International Conference on Social Robotics (ICSR 2022 [90])
Submitted to: International Conference on Intelligent Robots and Systems (IROS 2023)

The studies presented in this Chapter were conducted at the Institute for Systems and Robotics (ISR), Instituto Superior Técnico (Universidade de Lisboa), Lisbon, Portugal, with the supervision of Professor José Santos-Victor and the collaboration of Nuno Ferreira Duarte, whom I would like to thank sincerely.

9.1 Introduction

How we perform our actions naturally embeds information on the features of the object transported. Collaborative robots can transmit information in an analogous way by modulating the strategy used to transport objects with their end-effectors, as explored in Chapter 8. This possibility can promote spontaneous interactions by making an implicit yet effective communication channel available. In this Chapter, we investigate if humans correctly perceive the implicit information shared by a robotic manipulator through its movements during a dyadic collaboration task and the eventual advantages of using an expressive controller with respect to a neutral one. Hence, we move a step forwards by applying the generative methods presented in Chapters 6 and 7, to produce communicative motions with a robot that is not humanoid and whose arm configuration and kinematics are distant from the human one.

We designed two collaborative experiments.
In the first one, the robot was tasked with grasping a set of plastic cups and passing them to the human partner, who needed to sort the simulated characteristics of the glass based on the robot's behavior. This allowed us to investigate not only the perception of the robot's attitude but also to measure the effect, if any, of the communicative modulation on the human task. Indeed, with respect to the study presented in the previous Chapter, which was carried out through videos, having a real dyadic interaction grants a stronger assessment of the implicit communication going on.

In the second one, we propose a novel and realistic interaction between humans and robots where the parts have to collaborate in handling cups with different content: empty or filled with water almost to the brim. The robot receives the cups from the human and sorts them, adopting a neutral motion controller or our proposed expressive one that generates movements designed according to the cup content, depending on the experimental condition. In this case, we studied the human kinematics adaptation in the reaching and transport of the cups, the efficacy of the robot in completing such a challenging task, and participants' preferences between the two experimental conditions.

It must be mentioned that in the same work, it was also validated an online classifier to differentiate careful/not careful human movements, associated with the cups' content. Results showed that the carefulness during the handover of full cups can be reliably inferred online, well before action completion. However, the classifier design and development were defined independently from my thesis work [44, 45], hence its detail will not be reported.

Research questions and objectives

We explore the following research questions in the first scenario of robot-to-human handovers **S1**. First, **RQ1.1**, whether the attitude conveyed by the robot's movements is perceived as expected, i.e., if our controller correctly expresses the carefulness (or its absence); in this sense, we hypothesize **H1.1** that participants would correctly classify if the observed actions were careful or not, even if a not-humanoid robotic manipulator performed the communicative movement. We also verify **RQ1.2** if a robot transporting objects and expressing the appropriate human-like behavior can invoke motor adaptation in the human response. We assume **H1.2** that, if a contagion emerges, careful actions from the robot will elicit slower movements in the human, and vice versa. An insight in this direction was already presented in Chapter 8; however, a further investigation in this direction is worthwhile, given the direct interaction with a non-humanoid robot presented in this case.

In the second scenario **S2**, the two agents collaborate in handling cups empty or filled with water. With the first research question **RQ2.1**, we evaluate human kinematics before and during the object manipulation, to observe how the movement is modulated in the context of interaction with a robot and expand previous knowledge on careful handling [45, 89]. In this sense, we hypothesize **H2.1**: to measure motor adaptations analogous to those observed in Chapter 3, with full cups eliciting slower actions, which may be affected by the dyadic interaction context. A *within-subject* study design, comparing our proposed "expressive" controller against a neutral motion controller, was chosen to answer the second research question **RQ2.1** on the advantages of deploying communicative action generations in an HRI context. Hence, we test our hypotheses **H2.2** through questionnaires, assuming that participants would prefer the expressive condition and that it would also improve task efficiency.

In the following, I will present the two studies separately, going over their methods, results, and discussion.

9.2 Methods – S1

The objective of our study is to assess whether, in a one-to-one interaction, the generated robot's movements are informative of the properties of the transported object. Moreover, we evaluate if robot behavior affects how humans perform their tasks.
We will first explain how we synthesized the required velocity profiles and controlled a Kinova Gen3 robot with 7 degrees of freedom to execute them; then, we will describe the experimental setup and design.

9.2.1 Generating communicative movements on the robot

To have the robot communicate through its movements the object properties, the robot's end-effector followed the velocity profiles generated by the Generative Adversarial Networks (GANs) model. Our interest is in generating movements to convey whether the transported object requires caution and care to be transported (careful movement) or it is safe to move without any particular concern (not careful movement). Previous studies [45, 140] and Chapters 3-5 assessed this kind of object manipulation and showed a marked difference in the kinematics of the human hand associated with the two classes of motions, exploiting it to automatically detect the *carefulness* in the action.
To define meaningful movements associated with the properties of the carried object, we

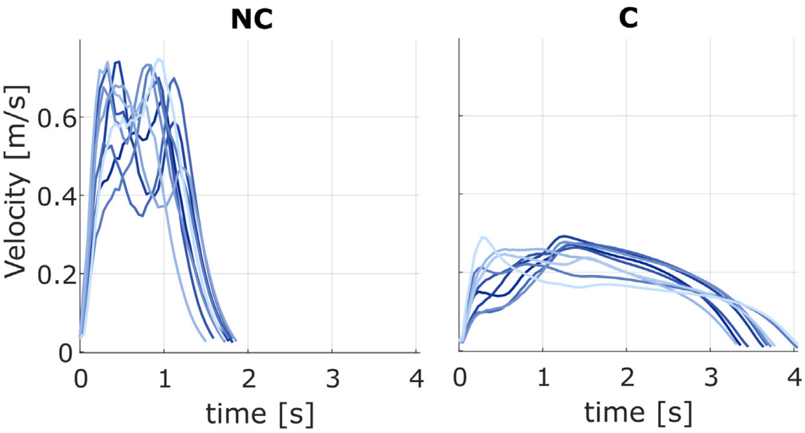


Figure 9.1 Velocity profiles generated by the GANs associated to not careful (NC) and careful (C) transportation of objects. These velocity profiles were used to control the robot during the human-robot experiment.

modulated the velocity profile adopted by the robot end-effector, following the approach explained in detail in the previous Chapters. The robot task consisted in grasping a cup from a table and handling it over to the participant, adopting the designed communicative motion in the porting action. For this specific study, from the trained GANs, we synthesized ten velocity profiles for each of the two classes to be replicated by the Kinova robot. A representation of the generated data is available in Figure 9.1.

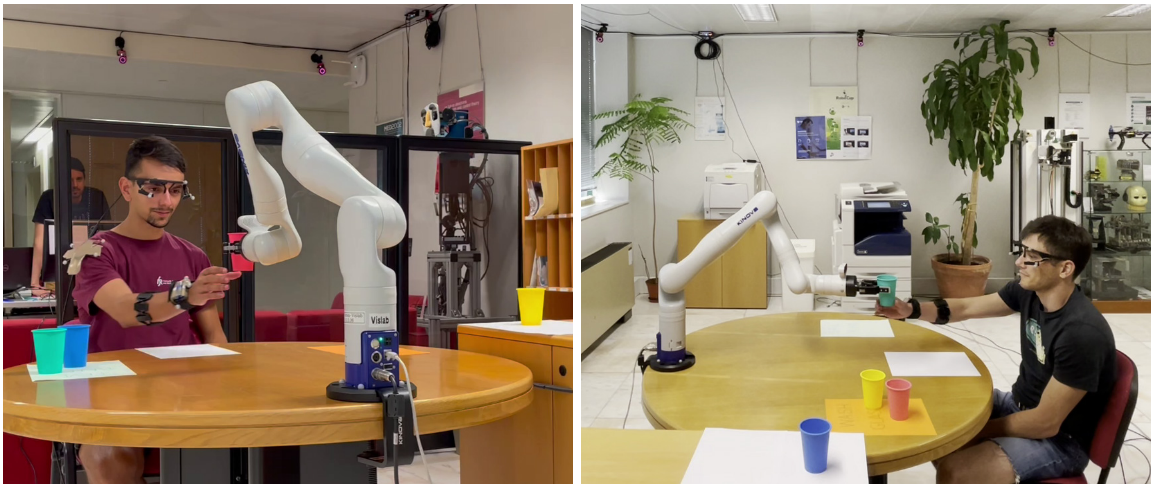
The Kinova Gen3 robot was controlled using ROS and the package kortex_ros[1]. Such package provides a velocity controller in Cartesian space, which moves the end-effector at 40 Hz in linear (m/s) and angular (rad/s) velocities. Attached to the end-effector is the Robotiq 85 two-finger gripper[2] used to grasp the cups. This work applies two high-level controllers: (i) a velocity PI controller and (ii) a velocity GAN controller. The first controller is responsible for picking the cups from the table, and the second is for transporting and handing over the cups to the participant. The former generates a constant velocity profile throughout the trials, while the latter follows one of the 20 GAN velocity profiles selected (10 careful and 10 not careful) during the experiment. For each GAN motion trajectory, the velocity profile is decomposed into the 3D Cartesian velocity coordinates by setting the current location and final location (handover point) at each time step. The handover location was fixed in advance to avoid any variability that could influence participants during the experiment. The position of the participant's wrist was tracked with the motion capture system, while the position of the gripper is computed by the robot's forward kinematics given

[1] Official repository of the Kinova Gen3 ROS package: https://github.com/Kinovarobotics/ros_kortex

[2] Official website of the gripper: https://robotiq.com/products/2f85-140-adaptive-robot-gripper

the known joint angles and the location of the robot base also tracked with the motion capture system. More details on the sensors are provided in the following Section. The handover release moment was obtained by applying a threshold: the robot opened the gripper to release the cup whenever the distance between the gripper and the participant's wrist was below a fixed value. This simple design was enough to grant a smooth and reactive handover required for our experiment.

9.2.2 Setup and experiment design



(a) Setup frontal view

(b) Setup lateral view

Figure 9.2 *Setup:* when interacting with the Kinova Gen3 robot, participants were seated at a table. Once grasped the cup from the robot gripper, they had to put it down on one of the three areas delimited on the table. The motion capture markers used to analyse the human kinematics are visible on the participants' right wrist.

Participants were asked to sort the items handled by the Kinova Gen3 robot by positioning them on the appropriate areas marked on the table where they were seated. We designed the experiment for the participants to focus on the robot's behavior and not on the characteristics of the items. For this reason, we used identical plastic cups: in the instructions, we explained that we were simulating a cocktail bar scenario, where the robot and the human had to collaborate in sorting the glasses between those full to be served to the clients, and the used and empty ones, to be washed; in such context, the cups were meant to be either full of a liquid or empty: however, we explained to the participants that due to the danger of having a robot transporting water, all the cups were empty. This granted that participants could not rely on any visual cue or the actual object features to decide where to place the cup.

In every trial, the Kinova robot grasped a cup from the table next to it (see Figure 9.2b) and transported it towards the participant, following either a careful (or not) velocity profile generated by the GANs (associated respectively, to the transport of a full or empty glass). The task for the participants was then to grasp the cup from the robot gripper and place it in the appropriate area on the table: on the “To be served” area, on the right, if they thought that the cup was actually meant to be full, or on the “To be washed” area, in case they assumed the cup was indeed empty. A third area, in the middle, was available to place the cups whose virtual content was not clear to the participant to avoid forcing them into making a decision. They were not informed about the modulation of the robot transport movements and, since the cups were all the same, they had to rely on the robot behavior to make their decision[3]. We used Optitrack[4] motion capture system (MoCap), with an acquisition frequency of 120 Hz, to track the position of the participant wrist and shoulder.

Twelve healthy participants, all members of Instituto Superior Técnico, voluntarily took part in the experiment. Each evaluated 20 robot movements, where the sequence of careful and not careful modulation was randomized once and then maintained for every participant. The interactions were organized as five blocks of a sequence of four trials. At the end of each block, the experimenter put the cups on the table next to the robot. This resulted in a total of 240 movements evaluated, equally balanced between careful or not robot behavior.

9.3 Results – S1

One of our aims was to verify whether modulating the robot end-effector velocity to express carefulness can inform participants about the virtual content of the manipulated glasses. Figure 9.3 shows the participants’ accuracy in evaluating, for each trial, if the observed transportation motion was meant to be associated with a delicate object, i.e., careful robot movement, or not. We represented with a dark bar the percentage of correct answers the participants gave, i.e., when they correctly interpreted the robot’s attitude. Considering the total number of evaluated trials (240), 189 were correctly classified with no indecision, resulting in an accuracy of 78.75%. The transparent colored bars represent the misclassified movements. For instance, when we generated a robot action modulated to communicate a not careful attitude, while participants associated it with the transport of a full cup. As can be noticed, misunderstanding a not careful action for a careful one was the most frequent occurrence, especially in the first trials. In detail, 90% of the careful robot movements were

[3] Sample video of the human-robot interaction: https://www.youtube.com/watch?v=HVahS-0tn6g

[4] Optitrack website: https://optitrack.com/cameras/flex-13/

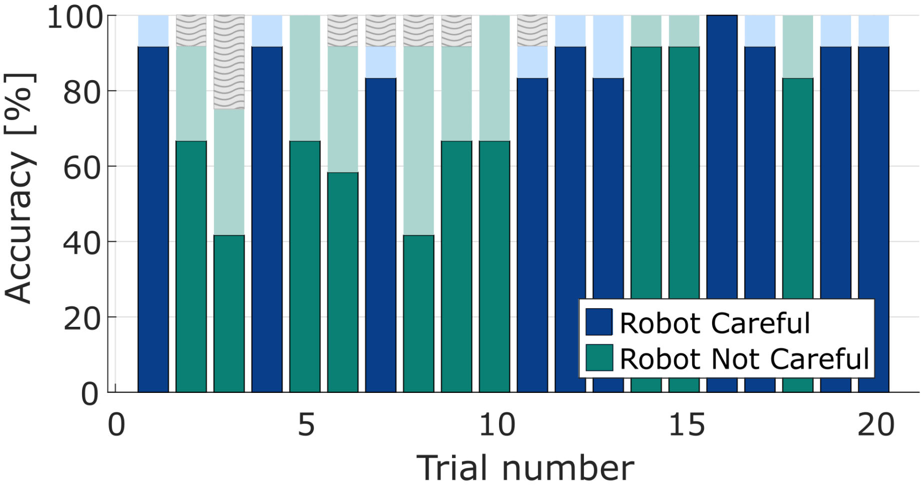


Figure 9.3 *Perception of robot's movements:* Percentage of correct interpretation of the robot's transportation movements during the experiment. When the robot performed careful movements, in blue, they were correctly perceived 90% of the times. NC motions required more trials to be consistently classified. The dark bars represent the percentage of correct classification of the movement from the participant, the transparent bars the percentage of wrong attribution; finally, the light gray bars with a wavy pattern, the percentage of "Unknown" answers in each trial.

perceived as such, whereas 75% of the not careful ones were correctly interpreted. Finally, the grey bars with a wavy pattern represent those trials where participants preferred not to make a choice and placed the cup on the neutral area on the table. Also, these occurrences, which happened in 9 trials out of 240, decrease as the experiment progresses.

Another aspect we were interested in investigating is whether the robot's two attitudes had any effect on how participants performed their tasks. An exploratory inspection of the hand velocity data encouraged us to deepen this intuition: Figure 9.4 reports an example of the velocity adopted by one participant when reaching for the cup in the robot gripper. A noticeable modulation in the participant's movements correlates with the attitude shown by the robot. When the robot handled the cup with a caring attitude, also the participant reached for it with slower and prolonged action compared to the not careful situation. To assess this modulation quantitatively in human actions, we considered the duration and median velocity of the movements as relevant features. It should be noted that the trajectories covered by participants with their hand to reach the robot gripper were comparable between the conditions, being the handover location fixed and the starting position on the table. Therefore, duration and velocity metrics are meaningful to compare the two conditions given the same traveled distance. To perform statistical analyses on the acquired data, we used

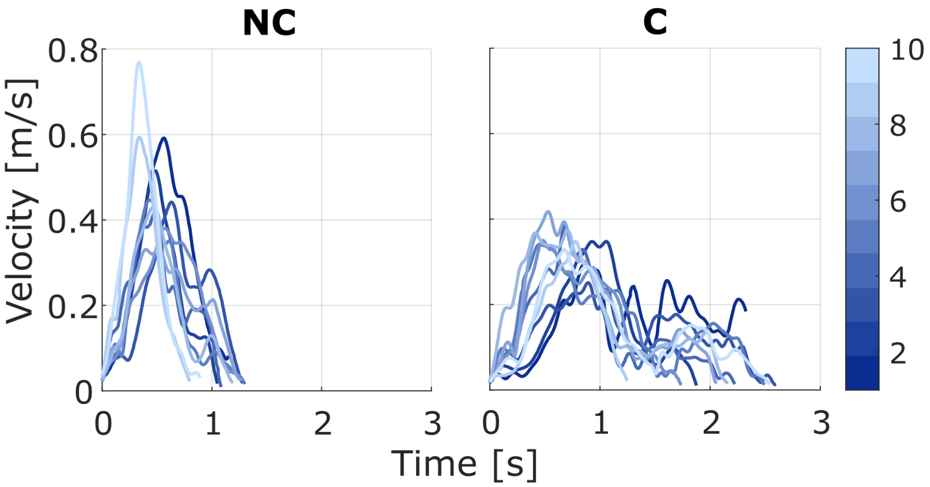
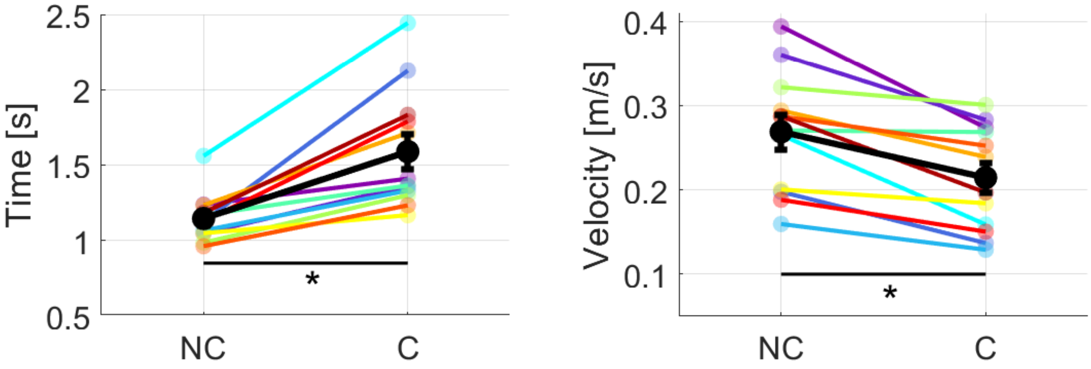


Figure 9.4 *Velocity reaching movement:* profiles adopted by one participant when reaching for the cup in the robot gripper. It is noticeable a modulation of both the duration and the maximum velocity depending on the style of the movement adopted by the robot: not careful (NC) or careful (C). The colormap is associated to the trial numbers, in order.



(a) Reaching duration

(b) Reaching median velocity

Figure 9.5 *Reaching movement:* in (9.5a) mean duration of the participants reaching movements towards the robot's gripper. When the robot performs a careful (C) transportation movement, participants are significantly slower in reaching for the cup. Also the median velocity adopted in the reaching movements (9.5b) is modulated by how the robot moved when transporting the cup. The mean values for each participant are represented in a different color. The thick black lines represent the mean over the twelve participants, with the standard error. The star indicates a significant difference with $p < 0.001$.

Jamovi software[5], in particular the GAMLj module[6] for mixed models. Figure 9.5a shows the mean durations of the participants reaching movements toward the robot gripper. We ran a mixed model assuming the duration of the participants' reaching movements as the dependent variable, the carefulness in the robot movement as a factor, and the subjects as cluster variables. The effect of condition resulted significant ($C-NC$, $estimate = 0.443$, $SE = 0.055$, $t = 8.00$, $p < 0.001$), indicating that when the robot end-effector was following a careful velocity profile, the subsequent human reaching action was longer, with an extended duration estimate of 0.443 seconds. A second mixed model was used to evaluate the median velocity adopted by the participants when reaching the robot gripper (see Figure 9.5b), using this time the median velocity as dependent variable: when the robot was careful, participants significantly diminished their median velocity, with an estimated reduction in speed of $0.055m/s$ ($C-NC$, $estimate = 0.055$, $SE = 0.012$, $t = -4.60$, $p < 0.001$). These findings prove that the modulation of the robot movements affected how participants moved to reach and take the cup from the robot. This happened even when there was no reason to adapt to the object properties since we consider a reaching movement and all the cups had exactly the same characteristics. We also verified, for both the duration and the median velocity of the reaching movements, if there was an interaction with the participants' accuracy in evaluating the robot's behavior in every trial. We used the accuracy in their classification as an additional factor in the mixed model, but we found no interaction with how they performed the reaching duration or velocity. The modulation in response to the robot attitude also occurred when participants did not recognize it explicitly.

9.4 Discussion – S1

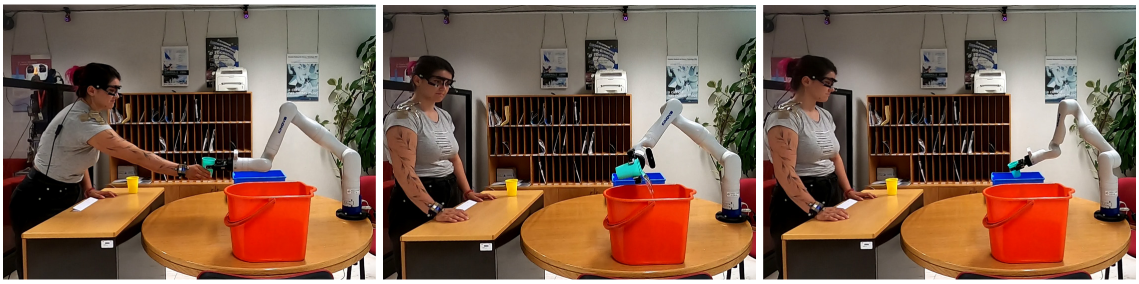
We exploited a generative approach to produce robot movements that could implicitly communicate if a handled object required or not carefulness to be transported. To avoid influencing the choice, all the items transported by the Kinova robot were identical (empty plastic cups). The participants had to decide if they were supposed to be virtually full or empty, without any particular hint or instruction on how to proceed. Firstly, we assessed **RQ1.1** whether our controller can express caution in the gestures or its absence. According to the results shown in Figure 9.3, our hypothesis **H1.1** was verified: we notice that the *careful* robot actions have been perceived as such since the first trials of the experiment. Regarding *not careful* actions, there is a learning trend in how they were perceived during

[5] Jamovi software website: https://www.jamovi.org

[6] General analyses for linear models Jamovi module: https://gamlj.github.io/

the experiment. In the first trials, they were sometimes mistaken for actions associated with transporting a full cup. As the experiment progressed, the difference between the two modulations became more evident, with an accuracy in the participants' choices above 80%. Reflecting on the original dataset (see Chapter 3) of human movements used to train the GANs associated with the transport of full and empty cups, we can observe that a *not careful* attitude is the standard in our actions. Indeed, when no particular circumstances are forcing us, for instance when picking and placing an ordinary object, we tend to move in a "neutral" way, and we can shortly describe our approach as not careful. This was also confirmed by the findings of Chapter 8, where we showed that the actions performed during the baseline and after watching an inattentive attitude were comparable. On the contrary, a strong kinematics modulation appears when we are paying attention to not spill the contents of a glass. This careful kinematics shaping is what we truly modeled in the communicative robot's movements, and it is rewarding that careful movements were perceived correctly from the beginning.

This study also allows us to evaluate **RQ2.1**, the effect that the implicit modulation of the robot actions has on the interaction. Even though participants knew from the beginning that the plastic cups were all the same and all empty, there was a modulation in how they approached the robot gripper, confirming our second hypothesis **H2.1**. We gave an overview of this phenomenon in Figure 9.4 and a quantitative assessment in Figure 9.5. If the robot manifested a careful attitude, adopting a lower magnitude in the velocity profile and a longer duration of the movement (see Figure 9.1 for reference), also the reaching movement of the humans was significantly slower. Interestingly, this also happened when participants had trouble explicitly recognizing the motion style and classifying the cup: the contagion in how they performed the reaching task was still present. This result emphasizes how important it is to modulate the actions of robots appropriately, with a view toward collaborative interaction. Indeed, we proved a motor contagion from the robot to the human, even if there was no need for the participants to directly associate with the task and adapt their motor strategy. We observed natural coordination emerging from such a simple task, where the pace of the human spontaneously adapted to the robot one, mimicking, even unconsciously, the attitude observed. Human-robot motor contagion on velocity was already observed whenever the robot velocity profile is biologically plausible [17, 161, 170]. In our approach, the reasonableness of the velocity profiles was granted using a generative network trained on human examples. The findings in our study extend the existing evidence of motor contagion in Human-Robot Interaction, proving that robotics arms can also leverage it to convey appropriate ways



(a) Handover (b) Pouring (c) Dropping

Figure 9.6 Sequence of actions by the human and robot. From left to right: human-to-robot handover of a full cup (9.6a); robot pouring the water content into the orange bucket (9.6b); robot placing the emptied cup in the blue box (9.6c). When manipulating an empty cup, after the handover (9.6a), the robot directly drops the cup in the blue container (9.6c). Human is wearing an eye-tracking device, infrared motion tracking markers, and an IMU sensor on the wrist.

of handling fragile objects. Through motion alone, it was possible to open a channel of communication between the two agents, with measurable effects on the interaction.

Finally, it should be noted that we obtained these results by modulating the movements of a 7 degrees of freedom robotic manipulator, not a humanoid robot. Nevertheless, even though its kinematics was far from the one of a human arm, it was possible to achieve the desired communication intent by simply modulating the end-effector control. This proves the power of the proposed approach and its potential scalability in other contexts and with other robots, also industrial ones, where implicit communication through motion could improve the efficiency and safety of a joint collaborative task. Given these findings and observations, we then planned to exploit the same controller and have the robot actually manipulate full and empty cups, to assess how the movement's modulation affects trust, perceived competency, and efficiency in a dyadic interaction while facing a challenging task. This is the core of our second study with Kinova Gen3 robot, which will be presented in the following.

9.5 Methods – S2

This section describes the human-robot interaction experiments, details the sensors used for the dataset acquisition, and presents the administered questionnaires.

9.5.1 Setup and experiment design

Participants stand in front of a table with four identical plastic cups placed in a row, equidistant from each other. These cups differ in content, being two empty and two filled with water almost to the brim, constituting two types of objects to be handover: empty or full. Participants faced a Kinova Gen3 robot fixed to a table with two distinct recipients at the robot side. As seen in Figure 9.6, on the left side of the robot, there is an orange bucket meant to contain water, while the blue drawer on the right stores the empty (or emptied) cups.

The experiment is presented as a collaborative task, where the human should help the robot clean the table by handing over the cups, from the rightmost cup to the leftmost, one at a time. The robot receives the cup (see Figure 9.6a); in case the cup contains water, it pours the content into the orange bucket (Figure 9.6b), and finally, it places the empty cup in the blue drawer (see Figure 9.6c).

We adopted a *within-subject* study design where participants are exposed, in a randomized order, to two conditions associated with the controller used by the robot to complete the task: a neutral motion and an expressive motion. In each condition, participants completed 12 handovers to the robot, divided into three blocks, with the experimenter resetting the setup and the four cups on the table at the end of each block. The sequence of empty and full cups to be handed over was balanced but changed in every block. The naturalness of the human-to-robot handover was made possible by constantly tracking the 3D wrist pose with a motion capture system, without requiring a pre-determined and fixed location. As explained in the previous study, the robot's gripper position was computed using forward kinematics, and a threshold was set to grasp the cup when the distance between the gripper and the participant's wrist was below a certain value.

Our study involved 15 right-handed participants (8 females, 7 males, 26.6 ± 6.2 years old) who provided written informed consent. They were all naive regarding the purpose of the experiments and not directly involved in our research. The self-reported level of knowledge in robotics was: 40.0% professional or advanced, 33.3% average, and 26.7% little or none.

Sensors and Data Description

Movements' kinematics and visual information were recorded throughout the experiment. Pupil Labs head-mounted glasses [80] were used to track eyes' movements providing 2D gaze fixation on a POV perspective and information on pupil dilation. For more information on the Pupil Labs data acquisition, consult the repository[7]. We used again the OptiTrack

[7] Pupil-Labs software repository: https://docs.pupil-labs.com/core/software/pupil-player

Table 9.1 Sensor specifications and size of recordings.

Sensor	Type of data	Frequency (Hz)	Total Size
OptiTrack	Motion Tracking	120	∼1.1 GB
Pupil Labs	Eye Tracking	{30, 120}	∼100 GB
LPMS-B	IMU	400	∼320 MB
GoPro	Video output	60@1080p	∼40 GB

Note: The PupilLabs streams 30 Hz@720p for the gaze fixation and 120 Hz@320p for the pupil detection system of each eye.

MoCap, consisting of 12 infrared cameras around the room, tracked the position of head, shoulder, and wrist, through reflective rigid bodies suitably designed. Additionally, an Inertial Measurement Unit (IMU), LPMS-B model[8], was placed on the participants' wrists. The acquisition with the three sensors was synchronized through ROS, providing a synchronized timestamp. Moreover, the main events in each trial were manually annotated during the experiment, providing labels for human grasping, handover, robot pouring, and object release. Finally, an external RGB camera recorded the experiments from the viewpoint pictured in Figure 9.6. Table 9.1 presents the sampling frequency for each sensor and the corresponding total amount of data[9]. 360 actions ($15 participants \times 12 handovers \times 2 conditions$) were recorded and performed successfully without dropping the cup or spilling the content. Due to ambiguities or missing data from the MoCap system, in the kinematic analyses of the current study, we considered a total of 310 trials, 76 and 82 handovers for empty cups, respectively, in neutral and expressive conditions, and 70 and 82 trials, involving full cups for neutral and communicative, respectively.

9.5.2 Questionnaires

Together with a quantitative kinematic description of the human motion, we were interested in investigating if and how the controller conditions would affect how participants *explicitly* perceived the task. For this reason, we conducted a pre-questionnaire to understand the general perception and propensity to robotics; we then administered the same set of scales after each of the two robot controller conditions and finally conducted a post-questionnaire

[8] IMU sensor datasheet: https://www.lp-research.com/wp-content/uploads/2013/06/LpmsBUsersGuide1.2.7.pdf

[9] The human-to-robot *handover* actions dataset is made publicly available on the institution's website https://vislab.isr.tecnico.ulisboa.pt/datasets_and_resources/#hcups_water

to assess the global perception of the experimental conditions. All the items were based on a 5-point Likert scale, except for the post-questionnaire. Indeed, at the end of the experiment, we asked participants to think about the two interactions experiences they had with the robot and express their preferences concerning a list of 9 items, such as “In which one of the two parts you were more comfortable in interacting with the robot?” or “Which one of the two parts of the interaction you enjoyed the most?”. We also included a speed-bump item in the questionnaires after each interaction to check the participants’ level of attention during the completion. See Section 9.6.2 and Table 9.2 for further details on the scales. We used Jamovi software to analyze the data collected in the survey by using Wilcoxon rank-sum tests, correlation analysis, and binomial proportion tests.

9.5.3 Robot Motion Controller

We designed two conditions, neutral and expressive, to compare the efficacy and the effect on the interaction of adding a communicative layer to the generated actions. We used again the 7 Degrees of Freedom (DOF) Kinova Gen3 robotic manipulator to accomplish the challenging task of handling cups that may be filled with water. The ROS architecture was the same already presented in Section 9.2.1.

Our primary interest is to test and deploy a controller that mimics what happens naturally in human manipulations, being communicative of the object properties while remaining task-oriented. Its expressive ability has already been tested in **S1** [90], where participants could grasp the implicit message conveyed by the robot’s gestures and even showed the effects of motor contagion, although the robot was not humanoid. In this case, participants act instead as givers and perfectly know the characteristics of the object beforehand. Hence, the impact of the robot’s implicit communication is reduced. However, even in a context where the manipulator’s role is passive, we hypothesize that human-inspired motions can benefit the interaction, making it more fluid, smooth, and desirable to the human partner.

Neutral condition (NEU)

In the neutral condition, the actions were always designed with the same approach, generating a constant velocity profile throughout the trials with a simple proportional controller. The value of the adopted velocity was empirically chosen to be suitable for transporting all the cups. The choice of always using the same constant velocity granted the successful transport of all the full cups without any content spilling but was not optimized for the transport of the empty ones, which posed no risk instead.

Expressive condition (EXP)

To produce the robot's actions adapted to the properties of the object involved - *careful* if full of water, *not careful* if empty - we modulated the end-effector velocity, using velocity profiles norms generated by the Generative Adversarial Networks (GANs) replicating the approach described in Section 9.2.1. However, in this specific study, to avoid introducing a not controlled variable that may impact the comparison with the neutral condition, we randomly selected one velocity profile for careful and one for the not careful movement to be replicated by the Kinova robot. We used a careful attitude to transport the full cup from the handover position to the orange bucket, where the robot emptied the content. The designed not careful velocity profile was used instead to take the empty cups from the handover position or the orange bucket to the blue drawer where the cup is released.

9.6 Results – S2

In this section, we will present our findings relative to the analysis of human kinematics during the task, the results from the questionnaires, and the metrics assessing task efficiency.

9.6.1 Human motion

Figure 9.7 represents participants' hand velocity when moving towards the cup to grasp it (first velocity peak) and then its transportation toward the robot gripper (second peak). This representation results from applying a second-order Butterworth low pass filter with a cut-off frequency of 8 Hz to the hand velocities derived from the 3D trajectories recorded with the MoCap system for each trial. The velocity profiles were resampled to the median duration for each class, and their mean was computed for every time instant, together with the standard deviation. A noticeable difference in the velocity adopted depends on the cup involved: empty cups, associated with not careful motions, elicit quicker motions with higher speed magnitude. Instead, full cups require careful actions to avoid spilling the content, resulting in slower movements with lower accelerations. The robot controller does not influence how participants took the cup and handed it over, as could be expected from their role as givers. In Figure 9.7, the two distributions are distinguished by the content of the glass and not by the controllers applied on the robot, which are quite overlapped instead. These observations from Figure 9.7 were confirmed by running a mixed model with Jamovi GAMLj module separately on the reaching and the transport actions. We assumed the maximum velocity of participants' movements as the dependent variable, the subjects as cluster variables, and the content of

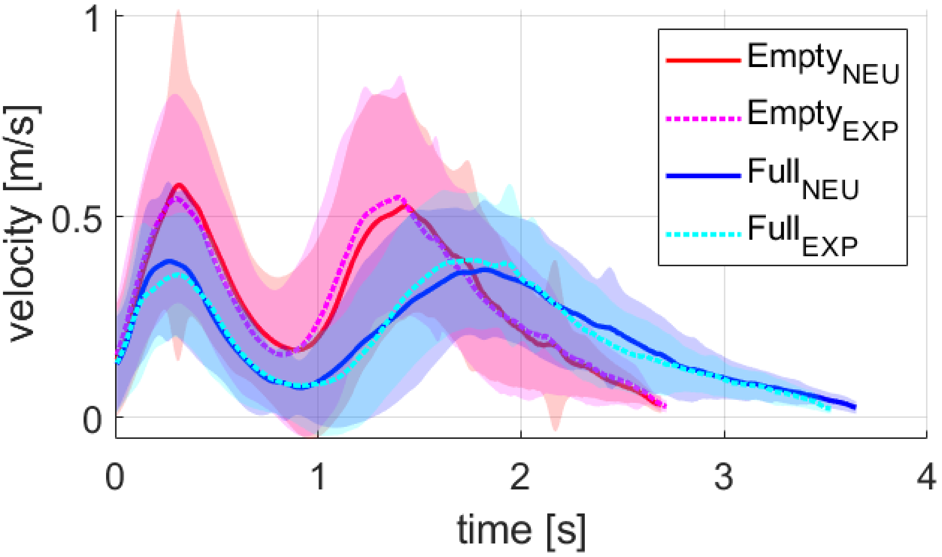


Figure 9.7 Hand mean velocities and standard deviation, in transparency, associated with the reaching (first peak) and transportation phase (second peak) of full and empty cups, in shades of blue and red, respectively. The robot controller used to transport the cup at a later time did not influence the human behavior as giver.

the cup, requiring or not care, the controller type, and their interaction as a factor. The effect of condition resulted significant for both the action phases (**Reaching:** $Full - Empty$, $estimate = -0.202\mathrm{m\,s^{-1}}$, $SE = 0.032$, $t = 56.6$, $p < 0.001$, **Transport:** $Full - Empty$, $estimate = -0.276\mathrm{m\,s^{-1}}$, $SE = 0.037$, $t = -7.49$, $p < 0.001$). This result shows that the modulation in human action appears not only when directly interacting with the object, when we observe a decrease of the velocity magnitude of $0.276\mathrm{m\,s^{-1}}$ if the cup content is full, but even before touching the object; indeed, participants adapt the motion according to the presence of water, with an estimate of $0.202\mathrm{m\,s^{-1}}$ of velocity reduction when preparing to grasp a full cup. These observations confirm previous results in the analysis of human kinematics associated with the carefulness feature [45, 89], but interestingly extend the effect of object properties also to the kinematics of reach-to-grasp actions. Observing the first velocity peak in Figure 9.7, the adjustment in the reach-to-grasp action is principally related to the magnitude of the velocity, as mentioned previously, whereas its duration is comparable for empty and full cups, differently to what happens during the transportation phase. It is the first time we have observed anticipated motor adaptation in this context; further investigations need to be conducted to understand why this anticipatory effect arose: it may be ascribed to the standing posture of participants or, even more interestingly, to the collaborative setting. Indeed, the intention to communicate and make predictable and readable gestures in a social context is well-described by the signaling theory [125]. For instance, it has been proved that the emerging kinematic pattern in grasping gestures differs between individual and social conditions [143].

Table 9.2 Questionnaire's scales, N=15

Scale	Item example	Condition	Cronbach's α	Mean±SD	p
Positive Attitudes About Robots (PARS) [12]	I think the use of robots can have a positive impact on society	Pre	.83	4.50 ± 0.57	-
Negative Attitudes About Robots (NARS) [121]	I would feel uneasy if I was given a job where I had to use robots	Pre	.02	-	-
Anxiety (Anx.) [120]	I would be anxious about the kind of movements the robot would make	Pre	.72	2.33 ± 0.73	-
Anxiety (Anx.) [67]	I was afraid to make mistakes with the robot	NEU	.63	1.47 ± 0.44	0.23
		EXP	.64	1.57 ± 1.50	
Competence (Comp.) [50]	I think that the robot was competent	NEU	.91	3.92 ± 0.84	0.90
		EXP	.88	3.87 ± 0.83	
Cognitive Trust in HRI (Cogn.) [13]	I would feel a need to monitor the robot's work	NEU	.71	3.76 ± 0.50	0.06
		EXP	.82	3.90 ± 0.56	
Affective Trust in HRI (Affect.) [13]	This robot would act cooperatively	NEU	.51	-	-
		EXP	.79	4.09 ± 0.64	
Evaluation of Robot Movements (Eval.) [14]	The robot's movements looked natural	NEU	.79	4.08 ± 0.39	0.26
		EXP	.83	4.13 ± 0.39	

Note: The p value, when reported, refers to a Wilcoxon rank sum test between Neutral and Expressive robot conditions

9.6.2 Questionnaires

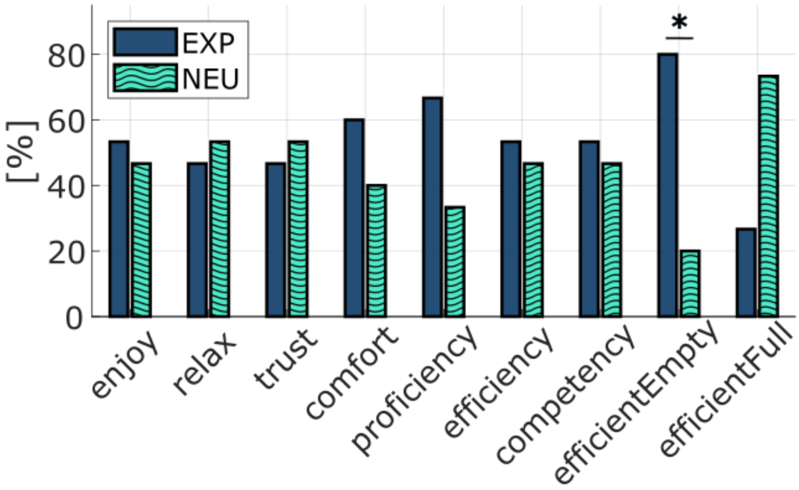


Figure 9.8 Post-questionnaire preferences between the Neutral and the Expressive GAN-based robot conditions.

The items used in the questionnaires were short versions adapted from validated scales. Cronbach's α was used to evaluate the internal consistency and reliability of each of them and is reported in Table 9.2. If the value of α was above 0.60, unit indices were produced by averaging the responses to the individual items included in each scale. The "NARS" and "Affective Trust in HRI" scales presented a low Cronbach's value ($\alpha = 0.02; 0.51$ respectively); hence they were not considered in the analyses. We also ran Shapiro-Wilk's tests to verify the normality of the samples. When the distribution resulted gaussian, we employed parametric tests; otherwise, we used the non-parametric versions. Before starting the experiment, we asked participants to answer a few items to characterize the population sample better. Their mean values are reported at the beginning of Table 9.2. Not surprisingly, we found a positive correlation between the self-reported knowledge of robotics and the PARS scale (Spearman's rho: $0.457, p < 0.05$). By administering a set of scales after each interaction with the robot, we wanted to assess if the participants perceived the two conditions differently. As previously mentioned, the order of the conditions was balanced and randomized to avoid introducing possible bias; however, as reported in Table 9.2, we found no significant difference. Only the Cognitive Trust in HRI scale reported a slight preference for the expressive condition, but still not statistically meaningful. Considering separately single items and using again a Wilcoxon rank sum test to compare the two conditions, we found two significant differences: "This robot would act consistently" (NEU: 3.73 ± 1.03, EXP: 4.20 ± 0.68, $p < .05$); "The robot moved too slowly" (NEU: 2.80 ± 1.20, *EXP*: 2.20 ± 0.78, $p < .05$). The percentages associated with participants' preferences in the post-questionnaire, after experiencing the two experimental conditions in random order, are reported in Figure 9.8. In this case, the distribution of the answers was binomial, so we ran a two outcomes proportion test. We found a significant difference in the item "In which one of the two parts was the robot more efficient in transporting the empty cups?" ($p < .05$). The robot was perceived as more efficient when controlled with the not careful GAN's velocity profile. We can also observe a tendency to choose the neutral controller instead when dealing with the full cup. This can be understood by referring to Section 9.5.3, where we explained that in the design of the neutral condition, we chose a velocity profile intermediate between careful and not that still allowed to complete the task successfully with no spilling; in this sense, the neutral robot was faster when transporting the cup with water. A trend is also noticeable in the Comfort and Proficiency items, where the participants well received the expressive controller.

The general questionnaire's outcome does not allow us to conclude about a strong preference of participants between the two conditions. The experiment design required the two agents to collaborate in any case, and the drive to comply and adapt may have overshadowed the differences in robot movements. At the end of the experiment, 12 participants out of 15 reported perceiving a distinction between the two blocks, but only 7 (46.7%) attributed it to a difference in the velocity; the others referred, for instance, to the pouring angle, the gripper strength, or the position of emptying or handover, which did not actually change. However, we can further reason that participants did not consciously recognize or, at least, did not explicitly appreciate the added value of communicative movements. In study **S1** [90], where the robot had the active role of giver, its human-inspired actions accomplished the goal of communicating carefulness associated with object properties and even elicited motor adaptation. In the current scenario, the participants' role as givers made them fully aware of the object's properties. In this sense, the robot had nothing to communicate except to accomplish the task without failures, which happened in both conditions. Therefore, in such a collaborative task, the robot must first function properly to be accepted, while communicative movement is not necessarily perceived overtly as better.

9.6.3 Global metrics

One of our hypotheses when designing a human-inspired robot controller was that it would improve the smoothness and efficiency of the interaction when adopting the appropriate level of carefulness. For this reason, we decided to examine a quantitative metric: the interaction

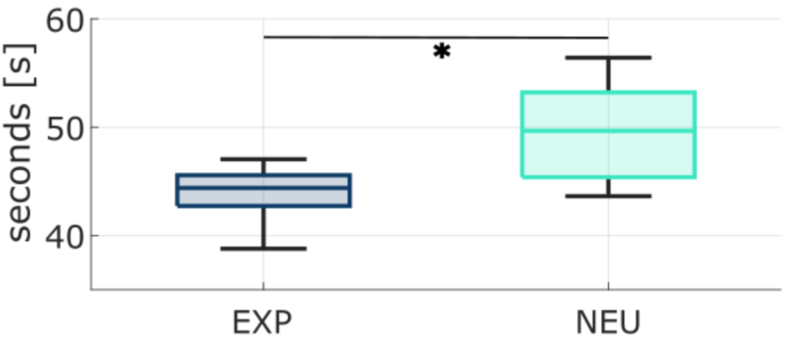


Figure 9.9 Difference in the block duration considering only the human contribution.

duration over the block of four consecutive cup handovers. In this case, we are interested in a fluency metric that, regardless of the time taken by the robot, which can vary depending on the condition and the cup handed, measures the human efficiency in the task. Hence, we annotated from videos the duration of the robot's action. We considered the net time by subtracting from the total duration of each block the time taken by the robot's movements from each completed handover to the end of the trial, i.e. when it released the empty glass into the blue container. Figure 9.9 illustrates the corresponding results, where we find a significant difference in the net duration of the blocks, in seconds, confirmed by a Wilcoxon rank sum Test (NEU: 49.5 ± 4.35, EXP: 44.4 ± 3.06, $p < .001$). This result is extremely interesting because it proves that, even if not consciously perceived by participants, the expressive modulation of the robot's actions makes the overall interaction smoother and more fluid, with a reduction of 5 seconds on the net duration. The impression we had observing the videos is that when the robot exhibited a human-inspired behavior, the latency time between each trial was reduced.

9.7 Discussion – S2

In this study, we presented a realistic scenario where a human and a robot successfully collaborate in handling full and empty cups, with the goal of understanding how to infer human carefulness during handover actions, as well as to generate similarly "expressive" robot movements.

We explore the natural adaptations that occur in human motion when handling cups with different filling levels which we observed, for the first time, not only during the transportation but also in the reach-to-grasp movements, confirming our first hypothesis **H2.1**. This result could allow the classifier to detect the care adopted and the related object properties in advance, allowing the robot to prepare the most appropriate motion strategy. We found interesting insights comparing a proportional neutral controller and our human-inspired expressive one.

When the robot was not expressive, keeping the same velocity for the whole experiment regardless of the cup content, participants perceived it as less consistent and found its actions generally too slow. Compared to an oblivious robot that makes no assumptions about the cup's properties, our approach mimics the human strategy: being careful when needed and speeding up otherwise. Although participants have not acknowledged this difference explicitly, we found a quantitative measure of the interaction efficiency in the experiment duration. Indeed, with the robot showing human-like behavior, humans reduce the time required to complete their tasks, thus improving the fluency and smoothness of the collaborative interaction. Our second hypothesis **H2.2** is therefore partially verified: participants did not broadly prefer our expressive motion controller, yet we found measurable facilitation from its deployment.

To conclude, this study validates a framework for action perception and generation in HRI. In future works, the complete architecture should be tested in a dyadic interaction where both the human and the robot have an active role in carrying objects. In such a context, the classification output (see [44, 45] for implementation details) could be directly used to plan adequate robot motions with GANs. Indeed, in this experiment, we wanted to avoid the risk of spilling water and compromising the interaction. To assess separately how participants perceived the two conditions and the detection performances, the classifier label was not used to decide the most proper attitude the robot should adopt.

Part IV

Conclusion

Chapter 10

Final discussion

10.1 Overview

In this thesis, we discussed the importance of implicit communication in supporting a successful exchange of information between two agents, human or artificial. Robots can greatly benefit from resorting to this additional means of communication; it is complementary to other forms of dialogue, and it grants immediate and natural reciprocal understanding by relying on the fundamental mechanisms of social interaction. Given these considerations, we aimed to tackle several conceptual and technological open issues targeting, in particular, the common but challenging task of manipulating objects and understanding their latent properties. The global objective of this work is to contribute to the design and development of robots able to interact with us more smoothly and seamlessly. This goal is in the interest of human-robot collaboration and, by exploiting some principles of social robotics, it can be suitable for HRI scenarios such as industry or assistive contexts. Indeed, with the increasing expansion of robot use in our society, it becomes paramount that they can communicate clearly, quickly, and efficiently with us, for a safe and fruitful collaboration. We believe that the easiest and most effective way of achieving so, is to take inspiration and mimic those mechanisms that regulate human-human communication: being automatic, they do not require any additional training from the human side. Indeed, if we want robots to become collaborators, the effort of the interaction cannot rely entirely on the human, who otherwise would only be a user.

This broad perspective has been applied in this thesis to the communication of object properties, and the philosophy behind the work has always been to take inspiration from human behavior. Indeed, if we observe another person dealing with an object, we can make some assumptions about their goal, intention, and also on hidden features of the item they are transporting. The same happens if we are the active agent: although we often do it unconsciously, the way we move is revealing and for instance helps our co-actors to plan for a handover appropriately. On this reasoning is based a key observation exploited in this research: observing a closed container on a table may not be informative enough to understand how to lift and transport it properly. Of course, we could estimate its weight based on the size and the material, and plan a grasp according to its shape and affordances; however, we could easily be misled by its external appearance and overlook the latent properties of the object. As humans, we tend to overcome this issue by relying on a huge amount of prior experience and quick adaptations to the context. However, for a robot, doing the same is particularly challenging and prone to errors and safety risks.

The solution can be found in human-robot interaction itself. It is precisely how we deal with an object that reveals its latent properties. Robots can exploit the cues that are naturally embedded in our motions to foster the inference process. Although their perceptive system differs from ours, they nonetheless have access to the surroundings through appropriate sensors and models. Hence, developing a framework for the automatic detection of latent features, such as the carefulness of human gestures, is in the scope of this work.

To promote reciprocal information exchange, also the robot's actions should account for the properties of the object they are transporting and, once again, mimicking human kinematics can be a winning strategy. Indeed, it accomplishes two main purposes: the generated actions are well-suited, safe, and efficient in dealing with a specific object, granting the successful execution of the task; moreover, an external observer can easily interpret the motion features and its associated meaning, by relying on the same neural circuits of motor resonance. In this way, as it happens in our actions, the two layers of executing goal-oriented and legible actions would automatically match.

The following sections will provide a summary of the main findings and contributions to knowledge relative to human actions understanding, robotic expressive motions generation, and effects on the interaction, linked to the manipulation of objects with different properties.

10.1.1 Perception

At the beginning of our work, we investigated the impact of manipulating objects with different characteristics on action kinematics. Indeed, as mentioned before, the common thread of our research is first to learn human strategies, then try to transfer them to robotic agents. By collecting a purposefully designed dataset, described in Chapter 3, we focused on two specific properties that affect object handling: the weight and the care required for the manipulation, obtained by filling plastic cups with water. Subsequent analyses confirmed not only that the different features had an impact on the transport actions, but that such effect was measurable with multiple sensors and particularly striking for the presence or absence of carefulness.

Given these considerations, we worked on possible algorithms to indirectly infer the characteristics of the various objects by looking at how participants moved them. Our innovative contribution relies exactly in this approach: instead of addressing the question by treating it as a classical object recognition problem, we opted to exploit the information coming from the manipulation and use it as a shortcut. By doing so, elements such as the exterior appearance of the objects were completely disregarded, granting flexibility to our method, which can thus be applied to unknown objects. Ultimately, we are not directly interested in understanding the object itself, but more in learning the impact its manipulation has on our kinematics. Of course, many technical challenges must be sorted out before achieving the final goal. First, it needed to be established which kinematics features could be used for the discrimination and how two retrieve them from the different sensors. In Chapter 4 we addressed them by focusing on the offline detection of observed carefulness and weight on specifically segmented transport actions. In a very promising way, we verified that kinematic information such as the adopted velocity profile was relevant in constructing a classifier. Moreover, such a feature was also easily extracted using only the robot's camera and calculating the optical flow of the scene. In Chapter 5 we defined a complete framework for online detection by having the robot observe the scene, with no additional sensors, in an ecological interaction scenario. We tested the architecture also with modified tasks with respect to those present in the original training set, achieving an accuracy up to 81.3%. However, we made the decision to disregard the weight of the cups, while keeping it as an element of variability in the experiments. Indeed, both the initial analysis of the human kinematics and the results with the offline classifiers showed that the weight difference in our dataset was somehow masked and overcome by the presence or absence of water in the cups. With this first part of the research, we collected evidence that the kinematics during the action is indeed a reliable source of information on the object manipulated by a human and that it is possible for the robot to automatically extract some motor descriptors to use in an inference process. However, in these experiments, the robot's role was limited to being a passive observer. There was still work to be done to get to a real interaction with the robot, where the exchange of implicit signals happens in a bidirectional way.

10.1.2 Action

When addressing the problem of generating expressive robot motions, we took inspiration once again from human behavior. We were interested in finding a method that could be as flexible as possible while allowing for capturing the essential kinematics features that make an action communicative. The studies on motion kinematics and carefulness detection allowed us to identify as a key feature of the transport action the hand velocity profile, considered as a time-series with its full dynamics. Therefore, following immediate parallelism, the idea was to modulate the speed of the robot's end-effector, working in the Cartesian space. This approach simplifies the problem by taking into account only the end-effector's dynamic, which needs to be attentively controlled to replicate the desired velocity profile, while the other joints of the kinematic chain move to follow without the need for a complex one-to-one mapping from the human arm. Hence, from a technical point of view, this approach can be easily applied to different robotic platforms, independently from the configuration of their chain, as presented in Chapter 7.

To implement this approach, it was necessary to identify velocity profiles that the robots could follow. For this task, I collaborated with another Ph.D. student, Luca Garello, who had greater expertise in artificial intelligence and learning methods. For our method to be general, we wanted to avoid a direct copy of the human motion and we decided to resort to Generative Adversarial Networks, as described in Chapter 6. These models, able to synthesize new data after appropriate training, have been extensively used for data augmentation and image generation; however, their application in robotics has been limited. We explored different implementations and we converged on using a single conditioned model which outputs a careful or not velocity profile according to the chosen label (see Chapter 7). Indeed, the network was trained with the hand velocity profiles during the transport actions included in the dataset, and we assessed its capability to learn the original distribution and produce novel but meaningful velocity profiles, with kinematics characteristics that respected the human ones. An interesting aspect of this approach is that it granted to capture not only the two classes of motions, but also their intrinsic variability. Repeated human actions, while being task-oriented and efficient, are nonetheless never exactly the same [61], and we wanted to transport the same natural fluctuations also in the robot's gestures. Variability plays such a key role in joint actions, to the point that is a predictor of how fast a motor task will be learned [136].

The second part of my Ph.D. journey allowed me to face the challenges of generating robotic actions following a human-inspired approach. From a technological standpoint, we succeeded in producing transport actions that matched the desired speed pattern on two different robots, humanoid and not. However, it remained to be assessed the communicative efficacy of our method: could the human co-actors perceive the intended robot attitude? The final part of this research aimed to address this question.

10.1.3 Interaction

To evaluate the efficacy and the impact of action modulations on the interaction, we conducted a series of experiments with different setups and robotic platforms. In Section 2.4, we introduced possible strategies to assess HRI and their pro and cons quantitatively. In our experiments, we combined videos and questionnaires with kinematics metrics to detect eventual motor contagion. Indeed, the first allows for an explicit measure of the interaction perception. At the same time, the latter gives insights into deeper adaptations happening during the collaboration, based on the fundamental resonance phenomena that regulate human perception-action mechanisms.

In Chapter 8, we presented a first experiment involving the two robots iCub and Baxter. From an online questionnaire showing videos of transport actions performed by the two robots, we confirmed that the two designed conditions careful and not careful were indeed distinguishable. We also observed how the iCub robot was in general perceived as more careful than Baxter. This could be linked to its human-like embodiment, which may be perceived as more delicate when moving. However, a precise choice in the movement design may have played a decisive role; indeed, the velocity profiles produced with the GANs are intended as norms, therefore detached from a specific trajectory. When controlling the robot end-effector to move from point A to B, the velocity profile needs to be re-scaled to cover the desired path appropriately. Inspired by the time-invariance and isochrony principles [4, 10, 166] we decided to maintain the original duration of the movement while scaling the velocity amplitude; this lead, for iCub shorter trajectories (given its child-like dimensions), to slower movements compared to Baxter's.

We also monitored participants' pick-and-place actions after watching videos of the two robots moving the same object. We found an impact of the robots' attitude on the velocity they adopted in their own task: if the robot was careful, they automatically slowed down the transport action. This indicates a contagion phenomenon between the robot's observed actions and the motor plan on the human side, which is an implicit measure of resonance and adaptation between the two agents [17, 27, 148]. We also confirmed the insight from the questionnaires, which displayed iCub as more careful. Another interesting finding from this experiment was that not careful actions were perceived as rather neutral. This was found both in the questionnaires and in the kinematics analysis of participants' transport actions: after observing a not careful motion, their motor response was comparable to the one in the baseline, when no stimuli were provided. In the original dataset, what we described as not careful movements were optimized standard actions, smooth and quick, for transporting empty plastic cups. It is with particular requirements and constraints in place, such as the risk of spilling the content of a full cup, that a strong kinematics modulation appears in our gestures.

Two other studies, presented in Chapter 9, were conducted at the Institute for Systems and Robotics of Instituto Superior Técnico during my research period in Lisbon.
In the first experiment, the underlying research questions remained the same, but we designed a dyadic robot-to-human handover interaction with a robotic manipulator. This was important not only to test the action generation approach on a completely different kinematic chain, but also to assess if the communicative effect was still in place with such robot embodiment. We found that motions designed to be careful were correctly perceived from the beginning, with an accuracy of 90%, whereas the classification error on not careful actions diminished as the experiment progressed, obtaining an overall accuracy of 75%. This still can be related to the previous discussion of not careful movements being actually standard in our motion planning and perception. We detected a form of motor adaptation again, in this case in the reaching actions towards the robot gripper: participants tended to match the velocity profile displayed by the robot, even if the objects involved in the handovers did not require any specific modulation, always being the same. This is a particularly relevant result. First, it was observed during the interaction with a robotic manipulator, with an embodiment far from the human one. Secondly, despite this difference in appearance, it was possible not only to open an implicit channel of communication with the human but this had also a measurable impact on the task. Previous studies explained that phenomena such as motor contagion are observable in human-robot interaction, preferably with humanoid-shaped robots, as long as the behavior displayed by the agent respects the laws that regulate biological motions [17, 27, 148]. In the proposed strategy, the robots' motion plausibility is granted by using human-derived velocity profiles to control their end-effectors.

In the second experiment, we studied the mirrored scenario of human-to-robot handovers of empty and completely full cups, where the manipulator carried and sorted them appropriately without ever spilling their contents. We implemented two controllers to be compared: a neutral one, which mimicked an oblivious attitude from the robot and transported all the cups with the same constant velocity; and the expressive controller based on GANs, already used in the first study, that replicates the human strategy: being careful when needed and speeding up otherwise. Analyzing human motion, we confirmed the usual modulation in the action kinematics to the content of the cup. However, for the first time, we observed an adaptation that anticipates the physical interaction with the object for what regards the maximum velocity of the reach-to-grasp motions. In this sense, it may be that the collaborative context, induced participants to be more communicative, as hypothesized by [125]. A more prudent explanation is that such preparatory motions respected the already mentioned principles of time invariance [4, 10, 166], being their duration comparable regardless from the cup, but were slowed down to grasp the full cup more precisely. Participants did not perceive a marked difference between the two robot conditions. It must be considered that in this task the robot's role was exclusively passive, and participants, as givers, perfectly knew the cups' content beforehand. In this sense, the impact of implicit communication in the robot gestures seems reduced, and the collaboration with the robot was generally well perceived in any case. However, from a quantitative point of view, we proved that when interacting with the robot showing an adaptive behavior, the overall task duration diminished, improving the efficiency of human actions and the smoothness of the collaboration.

10.2 Final considerations, limitations, and future works

The work presented in this thesis aims to contribute to the knowledge of the use of implicit communication for object manipulation in HRI context. It focuses on a specific topic, which is the manipulation of objects and, in particular, the effect their properties have on the motion. In this regard, the number of features considered was limited. We analyzed the effects of the weight and the content of cups on the manipulation, later focusing only on the latter to study the attitude we defined as carefulness, required to avoid spilling the cup's content. This study lays the foundation for an approach that is grounded on human behaviors and proposes methods to transfer the same mechanisms to artificial agents, but it would greatly benefit from an extension of the properties examined and their interaction; for instance, how do features such as the temperature or the shape of an object affect in-hand manipulation? Is their effect impactful enough to be detectable from the arm kinematics, or can other motion phases, such as the reaching, be taken into account for proper recognition? It has also been observed how other factors, such as emotional attitudes and vitality forms, affect how we handle objects [38, 160], and it may be interesting to follow the same approach we proposed, to allow the robot to recognize them and eventually plan an appropriate reaction.

Our strategy to recognize the carefulness in the motion is non-invasive and quite easy to deploy, requiring a simple RGB camera to observe the scene; moreover, it completely disregards the object's external appearance and is quite robust to obstructions as well. It performs a continuous classification of the movements observed through the optical flow. Still, it is also based on the assumption that the main source of movement in the scenario is the object manipulation itself. Additional algorithms should be developed to address those cases when the human is walking around the room, other users are present, or the robot is moving around (ego-motion filtering). A possible solution would be to integrate into the framework an additional sensor, such as a wrist-worn inertial sensor, which has already proved to be a valuable source of information for detecting carefulness; while making the setup a bit more invasive, it would still be tolerable even in industrial work environments.

Most of our work dealt with the concept of *carefulness*. We started by defining it as the attitude displayed when moving a cup full of water, supported by other studies [45, 46, 140, 111]. However, we are aware that this definition is limited. Other factors may induce a person to perform a careful manipulation. In Chapter 8, we started to shed some light on the matter, but further work is needed to come to a shared and unambiguous definition of carefulness. A starting point could be, once again, to study human behavior and observe how different elements, such as the potential danger, fragility, assembly state, or material of an object play a role. Regarding the generation of expressive movements, we resorted to generative networks and focused on controlling the robots' end-effectors to follow velocity profiles belonging to the desired distribution. Indeed, previous findings suggested that most of the communicative power of the motion is contained in its velocity dynamic. This strategy grants both a manipulation suitable for the specific object, which will be handled properly, and the expression of the communicative intent. Such an approach also has the advantage of being independent of a specific kinematic chain and, therefore, suitable to be applied on various robotics platforms. However, depending on the specific robot and task, it may arise the need to optimize the trajectory of secondary joints to avoid uncanny motions that could be detrimental to the interaction and the conveyed implicit message. Moreover, generative methods are data-driven, and this represents an advantage since the underlying learning process is automated; however, they require new datasets to be collected whenever a new class of motion needs to be added. A conditional approach, as the one we described in Chapter 7, could be however exploited to produce intermediate classes. In our attempt to design generic robotic actions communicative of carefulness, we privileged the biological plausibility and followed an approach as human-inspired as possible by replicating the complete dynamics of the velocity profile. However, it may be sufficient to transfer on the robot only specific key factors, such as the average speed of the motion, or the duration of the action, to obtain the same communicative effect. Is it the full dynamic of the motion velocity that matters, or can it be simplified by modulating a single factor? In future studies, it could be interesting to verify it by adjusting, rather than the complete motion profile, only the speed (absolute or relative between trials) with the standard robot controllers of the different robots. In the last experiment, we observed that the standard neutral controller was successful in completing the task; yet, the collaboration was smoother when the robot showed flexibility and adopted biologically inspired velocity profiles. This is a preliminary result that suggests how following the human-inspired road, however challenging, still adds something to the interaction.

To conclude, in this thesis we presented implicit communication applied to action perception and generation. The two aspects are obviously strictly linked. A comprehensive experiment where the robot perceives online the carefulness in the partner's gestures and uses the classification output to promptly react, by planning an adequate handover and transport action, needs to be realized. By deploying the complete framework we could gather definitive insights into how implicit communication fosters and promotes safe, efficient, and smooth interactions.

Publications

A list of publications, either published or in the publication process, that relate to the thesis's contribution or were made during the Ph.D. period is given below. The * stands for equal contribution and shared first name.

Published

- **Lastrico, L.**, Carfí, A., Rea, F., Sciutti, A., Mastrogiovanni, F. (2023) Toward Implicit Communication of Object Properties for Human-Robot Interaction. In VIII Congress of the National Group of Bioengineering (GNB 2023)

- **Lastrico, L.**, Ferreira Duarte, N., Carfí, A., Rea, F., Mastrogiovanni, F., Sciutti, A., Santos-Victor, J. (2022) If You Are Careful, So Am I! How Robot Communicative Motions Can Influence Human Approach in a Joint Task. Conference Paper. In 14th International Conference on Social Robotics (ICSR 2022), vol. Lecture Notes in Computer Science, (no. 13817), DOI 10.1007/978-3-031-24667-8_24

- Pusceddu, G., Cocchella, F., Bogliolo, M., Belgiovine, G., **Lastrico, L.**, Casadio, M., Rea, F., Sciutti, A. (2022). Training School Teachers to Use Robots as an Educational Tool: The Impact on Robotics Perception. Conference Paper. In 14th International Conference on Social Robotics (ICSR 2022), vol. Lecture Notes in Computer Science, (no. 13817), DOI 10.1007/978-3-031-24670-8_10

- **Lastrico, L.**, Garello, L., Rea, F., Noceti, N., Mastrogiovanni, F., Sciutti, A., Carfí, A. (2022) Robots with Different Embodiments Can Express and Influence Carefulness in Object Manipulation. Conference Paper. In IEEE International Conference on Development and Learning (ICDL 2022), pp. 280-286, DOI 10.1109/ICDL53763.2022.9962196

- Cocchella, F.*, Pusceddu, G.*, Belgiovine, G., **Lastrico, L.**, Rea, F., Sciutti, A. (2022) “iCub, We Forgive You!” Investigating Trust in a Game Scenario with

Kids. Workshop Paper. In IEEE Trust, Acceptance and Social Cues in Human-Robot Interaction (SCRITA Workshop 2022), in RO-MAN 2022 Conference, DOI 10.48550/arXiv.2209.01694

- **Lastrico, L.**, Carfí, A., Rea, F., Sciutti, A., Mastrogiovanni, F. (2021) From Movement Kinematics to Object Properties: Online Recognition of Human Carefulness. Conference Paper. In 13th International Conference on Social Robotics (ICSR 2021), vol. Lecture Notes in Computer Science, (no. 13086), pp. 61-72, DOI 10.1007/978-3-030-90525-5_6

- Garello, L.*, **Lastrico, L.***, Mastrogiovanni, F., Sciutti, A., Noceti N., Rea, F.(2021) A Generative Model Towards Conditioned Robotic Object Manipulation. Abstract report at the 3rd Italian Conference in Robotics and Intelligent machines (I-RIM 2021). DOI 10.5281/zenodo.5900621

- Antonj, M., Zonca, J., **Lastrico, L.**, Casadio, M., Sciutti, A. (2021) A pilot study towards the implementation of perceptual and motor adaptation in robots. Abstract report at the 3rd Italian Conference in Robotics and Intelligent machines (I-RIM 2021). DOI 10.5281/zenodo.6367971

- Dicenzi, M., **Lastrico, L.**, Carfí, A., Sciutti, A., Mastrogiovanni, F., Rea, F., (2021) Recognizing Italian Gestures with Wearable Sensors. Abstract report at the 3rd Italian Conference in Robotics and Intelligent machines (I-RIM 2021). DOI 10.5281/zenodo.5900573

- Garello, L.*, **Lastrico, L.***, Rea, F., Mastrogiovanni, F., Noceti, N., Sciutti, A. (2021) Property-Aware Robot Object Manipulation: a Generative Approach. Conference Paper. In IEEE International Conference on Development and Learning (ICDL 2021), pp. 1-7, DOI 10.1109/ICDL49984.2021.9515667. **Best paper nominee**

- **Lastrico, L.**, Carfí, A., Rea, F., Mastrogiovanni, F., Sciutti, A. (2020) Towards the Estimation of Object Characteristics by Observing Human Manipulation. Abstract report at the 2nd Italian Conference in Robotics and Intelligent machines (I-RIM 2020). DOI 10.5281/zenodo.4781580

- **Lastrico, L.**, Carfí, A., Vignolo, A., Sciutti, A., Mastrogiovanni, F., Rea, F. (2020). Careful with That! Observation of Human Movements to Estimate Objects Properties. Conference Paper. In 13th International Workshop on Human Friendly Robotics

(HFR 2020), vol. 18, pp. 127-141, DOI 10.1007/978-3-030-71356-0_10. **Best paper nominee**

Submitted or in preparation

- **Lastrico, L.**, Ferreira Duarte, N., Carfí, A., Rea, F., Mastrogiovanni, F., Santos-Victor, J., Sciutti, A. (2023) Like Robots, Like Humans: Pupil Dilation during Collaborative Object Manipulation. **Submitted to:** the 32nd IEEE International Conference on Robot and Human Interactive Communication (IEEE RO-MAN 2023)

- **Lastrico, L.***, Ferreira Duarte, N.*, Carfí, A., Rea, F., Sciutti, A., Mastrogiovanni, F., Santos-Victor, J. (2023) Expressing and Inferring Action Carefulness in Human-to-Robot Handovers. **Submitted to:** IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2023)

- **Lastrico, L.**, Carfí, A., Rea, F., Sciutti, A., Mastrogiovanni, F. (2023) On the Effects of Object Characteristics on Human Manipulations: a Multimodal Dataset. **Submitted to:** International Journal of Robotics Research (IJRR)

- Garello, L.*, **Lastrico, L.***, Sciutti, A., Noceti, N., Mastrogiovanni, F., Rea, F. (2022) Synthesis and Execution of Communicative Robotic Movements with Generative Adversarial Networks. **Manuscript in preparation**, for the Special Issue in IEEE Transactions on Cognitive and Developmental Systems (TCDS) DOI 10.48550/arXiv.2203.15640

References

[1] (2015). *Baxter Research Robot - Technical Specification Datasheet & Hardware Architecture Overview*. Rethink Robotics.

[2] (2021). The Jamovi Project. Jamovi (Version 1.6).

[3] Admoni, H., Weng, T., Hayes, B., and Scassellati, B. (2016). Robot nonverbal behavior improves task performance in difficult collaborations. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 51–58.

[4] Ambike, S. and Schmiedeler, J. P. (2008). *Time-Invariant Strategies in Coordination of Human Reaching*, pages 205–214. Springer Netherlands, Dordrecht.

[5] Ambrosini, E., Pezzulo, G., and Costantini, M. (2015). The eye in hand: predicting others' behavior by integrating multiple sources of information. *Journal of Neurophysiology*, 113(7):2271–2279.

[6] Ansuini, C., Cavallo, A., Campus, C., Quarona, D., Koul, A., and Becchio, C. (2016). Are we real when we fake? attunement to object weight in natural and pantomimed grasping movements. *Frontiers in Human Neuroscience*, 10.

[7] Antonsson, E. K. and Mann, R. W. (1985). The frequency content of gait. *Journal of Biomechanics*, 18(1):39–47.

[8] Anzalone, S., Boucenna, S., Ivaldi, S., and Chetouani, M. (2015). Evaluating the engagement with social robots. *International Journal of Social Robotics*, 7.

[9] Apicella, T., Slavic, G., Ragusa, E., Gastaldo, P., and Marcenaro, L. (2022). Container localisation and mass estimation with an RGB-D camera. In *the 32nd IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 9152–9155, Singapore.

[10] Averta, G., Valenza, G., Catrambone, V., Barontini, F., Scilingo, E., Bicchi, A., and Bianchi, M. (2019). On the time-invariance properties of upper limb synergies. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, PP:1–1.

[11] Bartneck, C., Kulić, D., Croft, E., and Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics*, 1:71–81.

[12] Bernotat, J. (2021). *Keep an Eye on Stereotypes–The Impact of Gender Stereotypes (Toward Humans and Robots) on Language Processing*. PhD thesis, Department of Psychology, Bielefeld University.

[13] Bernotat, J., Eyssel, F., and Sachse, J. (2017). Shape it – the influence of robot body shape on gender perception in robots. In *Social Robotics*, pages 75–84, Cham. Springer International Publishing.

[14] Bernotat, J., Eyssel, F., and Sachse, J. (2021). The (fe)male robot: How robot body shape impacts first impressions and trust towards robots. *International Journal of Social Robotics*, 13.

[15] Bicho, E., Louro, L., and Erlhagen, W. (2010). Integrating verbal and nonverbal communication in a dynamic neural field architecture for human–robot interaction. *Frontiers in neurorobotics*, 4.

[16] Bingham, G. (1987). Kinematic form and scaling: further investigations on the visual perception of lifted weight. *Journal of experimental psychology. Human perception and performance*, 13(2):155—177.

[17] Bisio, A., Sciutti, A., Nori, F., Metta, G., Fadiga, L., Sandini, G., and Pozzo, T. (2014). Motor contagion during human-human and human-robot interaction. *PLOS ONE*, 9(8):1–10.

[18] Bliek, A., Bensch, S., and Hellström, T. (2020). How can a robot trigger human backchanneling? In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 96–103. IEEE.

[19] Brand, R. J., Baldwin, D. A., and Ashburn, L. A. (2002). Evidence for 'motionese': modifications in mothers' infant-directed action. *Developmental Science*, 5(1):72–83.

[20] Brophy, E., Wang, Z., She, Q., and Ward, T. (2022). Generative adversarial networks in time series: A systematic literature review. *ACM Comput. Surv.*

[21] Bütepage, J., Ghadirzadeh, A., Öztimur Karadağ, Ö., Björkman, M., and Kragic, D. (2020). Imitating by generating: Deep generative models for imitation of interactive tasks. *Frontiers in Robotics and AI*, 7:47.

[22] Campanella, F., Sandini, G., and Morrone, M. C. (2011). Visual information gleaned by observing grasping movement in allocentric and egocentric perspectives. *Proceedings of the Royal Society B: Biological Sciences*, 278(1715):2142–2149.

[23] Cangelosi, A. and Ogata, T. (2019). *Speech and Language in Humanoid Robots*, pages 2261–2292. Springer Netherlands, Dordrecht.

[24] Canigueral, R. and Hamilton, A. F. D. C. (2019). The role of eye gaze during natural social interactions in typical and autistic people. *Frontiers in Psychology*, 10.

[25] Carfì, A., Foglino, F., Bruno, B., and Mastrogiovanni, F. (2019). A multi-sensor dataset of human-human handover. *Data in Brief*, 22:109–117.

[26] Cavallo, A., Koul, A., Ansuini, C., Capozzi, F., and Becchio, C. (2016). Decoding intentions from movement kinematics. *Scientific Reports*, 6:37036.

[27] Chaminade, T., Franklin, D., Oztop, E., and Cheng, G. (2005). Motor interference between humans and humanoid robots: Effect of biological and artificial motion. In *Proceedings. The 4th International Conference on Development and Learning, 2005*, pages 96–101.

[28] Chan, W. P., Nagahama, K., Yaguchi, H., Kakiuchi, Y., Okada, K., and Inaba, M. (2015). Implementation of a framework for learning handover grasp configurations through observation during human-robot object handovers. In *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*, pages 1115–1120.

[29] Chang, J.-Y., Tejero-de Pablos, A., and Harada, T. (2019). Improved optical flow for gesture-based human-robot interaction. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 7983–7989.

[30] Chu, F.-J., Xu, R., Seguin, L., and Vela, P. A. (2019). Toward affordance detection and ranking on novel objects for real-world robotic manipulation. *IEEE Robotics and Automation Letters*, 4(4):4070–4077.

[31] Coleman, D., Sucan, I., Chitta, S., and Correll, N. (2014). Reducing the barrier to entry of complex robotic software: a MoveIt! case study. *arXiv*.

[32] Constable, M. D., Kritikos, A., and Bayliss, A. P. (2011). Grasping the concept of personal property. *Cognition*, 119(3):430–437.

[33] Cremer, S., Mastromoro, L., and Popa, D. O. (2016). On the performance of the Baxter research robot. In *2016 IEEE International Symposium on Assembly and Manufacturing (ISAM)*, pages 106–111.

[34] Curioni, A., Knoblich, G., and Sebanz, N. (2019). *Joint Action in Humans: A Model for Human-Robot Interaction*, pages 2149–2167. Springer Netherlands, Dordrecht.

[35] Curioni, A., Voinov, P., Allritz, M., Wolf, T., Call, J., and Knoblich, G. (2022). Human adults prefer to cooperate even when it is costly. *Proceedings of the Royal Society B: Biological Sciences*, 289(1973):20220128.

[36] Dehais, F., Sisbot, E. A., Alami, R., and Causse, M. (2011). Physiological and subjective evaluation of a human–robot object hand-over task. *Applied Ergonomics*, 42(6):785–791.

[37] Di Cesare, G., Errante, A., Marchi, M., and Cuccio, V. (2017). Language for action: Motor resonance during the processing of human and robotic voices. *Brain and Cognition*, 118:118–127.

[38] Di Cesare, G., Vannucci, F., Rea, F., Sciutti, A., and Sandini, G. (2020). How attitudes generated by humanoid robots shape human brain activity. *Scientific Reports*, 10(1):16928.

[39] Dimiccoli, M., Patni, S., Hoffmann, M., and Moreno-Noguer, F. (2022). Recognizing object surface material from impact sounds for robot manipulation. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9280–9287.

[40] Donahue, J., Hendricks, L. A., Rohrbach, M., Venugopalan, S., Guadarrama, S., Saenko, K., and Darrell, T. (2017). Long-term recurrent convolutional networks for visual recognition and description. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):677–691.

[41] Donnarumma, F., Dindo, H., and Pezzulo, G. (2018). Sensorimotor communication for humans and robots: Improving interactive skills by sending coordination signals. *IEEE Transactions on Cognitive and Developmental Systems*, 10(4):903–917.

[42] Dragan, A. D., Bauman, S., Forlizzi, J., and Srinivasa, S. S. (2015). Effects of robot motion on human-robot collaboration. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, HRI '15, page 51–58.

[43] Dragan, A. D., Lee, K. C. T., and Srinivasa, S. S. (2013). Legibility and predictability of robot motion. In *Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction*, pages 301–308, Tokyo, Japan.

[44] Duarte, N. F., Billard, A., and Santos-Victor, J. (2022a). The Role of Object Physical Properties in Human Handover Actions: Applications in Robotics. *IEEE Transactions on Cognitive and Developmental Systems*, pages 1–1.

[45] Duarte, N. F., Chatzilygeroudis, K., Santos-Victor, J., and Billard, A. (2020). From human action understanding to robot action execution: how the physical properties of handled objects modulate non-verbal cues. In *2020 Joint IEEE 10th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, pages 1–6.

[46] Duarte, N. F., Raković, M., and Santos-Victor, J. (2022b). Robot learning physical object properties from human visual cues: A novel approach to infer the fullness level in containers. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 10375–10381.

[47] Dwarampudi, M. and Reddy, N. V. S. (2019). Effects of padding on LSTMs and CNNs. *ArXiv*, abs/1903.07288.

[48] Esteban, C., Hyland, S. L., and Rätsch, G. (2017). Real-valued (medical) time series generation with recurrent conditional GANs. *arXiv*.

[49] Farnebäck, G. (2003). Two-frame motion estimation based on polynomial expansion. In *In: Proceedings of the 13th Scandinavian Conference on Image Analysis*, LNCS 2749, pages 363–370, Gothenburg, Sweden.

[50] Fiske, S. T., Cuddy, A. J., and Glick, P. (2007). Universal dimensions of social cognition: warmth and competence. *Trends in Cognitive Sciences*, 11(2):77–83.

[51] Fitts, P. M. (1954). The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology*, 47(6):381.

[52] Fitzgerald, C. (2013). Developing Baxter. In *2013 IEEE Conference on Technologies for Practical Robot Applications (TePRA)*, pages 1–6. IEEE.

[53] Flash, T. and Hogan, N. (1985). The coordination of arm movements: an experimentally confirmed mathematical model. *Journal of Neuroscience*, 5(7):1688–1703.

[54] Fulton, M., Edge, C., and Sattar, J. (2022). Robot communication via motion: A study on modalities for robot-to-human communication in the field. *J. Hum.-Robot Interact.*, 11(2).

[55] Gallucci, M. (2019). GAMLj: General analyses for linear models.

[56] Garcia, D. (2010). Robust smoothing of gridded data in one and higher dimensions with missing values. *Computational Statistics and Data Analysis*, 54:1167–1178.

[57] Garello, L., Lastrico, L., Rea, F., Mastrogiovanni, F., Noceti, N., and Sciutti, A. (2021). Property-aware robot object manipulation: a generative approach. In *2021 IEEE International Conference on Development and Learning (ICDL)*, pages 1–7. Garello and Lastrico equally share the authorship.

[58] Gehrig, D., Kuehne, H., Woerner, A., and Schultz, T. (2009). Hmm-based human motion recognition with optical flow data. In *2009 9th IEEE-RAS International Conference on Humanoid Robots*, pages 425–430.

[59] Gentili, R., Cahouet, V., and Papaxanthis, C. (2007). Motor planning of arm movements is direction-dependent in the gravity field. *Neuroscience*, 145(1):20–32.

[60] Gibson, J. J. (1977). The theory of affordances. *Hilldale, USA*, 1(2):67–82.

[61] Gielniak, M. J., Liu, C. K., and Thomaz, A. L. (2013). Generating human-like motion for robots. *The International Journal of Robotics Research*, 32(11):1275–1301.

[62] Gulletta, G., e Silva, E. C., Erlhagen, W., Meulenbroek, R., Costa, M. F. P., and Bicho, E. (2021). A human-like upper-limb motion planner: Generating naturalistic movements for humanoid robots. *International Journal of Advanced Robotic Systems*, 18(2):1729881421998585.

[63] Hall, J. A., Horgan, T. G., and Murphy, N. A. (2019). Nonverbal communication. *Annual Review of Psychology*, 70(1):271–294.

[64] Hamilton, A., Joyce, D., Flanagan, J., Frith, C., and Wolpert, D. (2007). Kinematic cues in perceptual weight judgment and their origins in box lifting. *Psychological research*, 71:13–21.

[65] Harada, K., Hauser, K., Bretl, T., and Latombe, J.-c. (2006). Natural motion generation for humanoid robots. In *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 833–839.

[66] Hassanin, M., Khan, S., and Tahtali, M. (2021). Visual Affordance and Function Understanding: A Survey. *ACM Computing Surveys*, 54(3):1–35.

[67] Heerink, M., Krose, B., Evers, V., and Wielinga, B. (2010). Assessing acceptance of assistive social agent technology by older adults: the almere model. *I. J. Social Robotics*, 2:361–375.

[68] Hemeren, P. and Rybarczyk, Y. (2020). *The Visual Perception of Biological Motion in Adults*, pages 53–71. Springer International Publishing, Cham.

[69] Hochreiter, S. and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

[70] Hoffman, G. (2019). Evaluating fluency in human–robot collaboration. *IEEE Transactions on Human-Machine Systems*, 49(3):209–218.

[71] Hoffman, G. and Ju, W. (2014). Designing robots with movement in mind. *J. Hum.-Robot Interact.*, 3(1):91–122.

[72] Huang, C.-M., Cakmak, M., and Mutlu, B. (2015). Adaptive coordination strategies for human-robot handovers. In *Robotics: Science and Systems*.

[73] Huang, Y., Bianchi, M., Liarokapis, M., and Sun, Y. (2016). Recent data sets on object manipulation: A survey. *Big Data*, 4(4):197–216.

[74] Huang, Y. and Sun, Y. (2019). A dataset of daily interactive manipulation. *The International Journal of Robotics Research*, 38(8):879–886.

[75] Huber, M., Rickert, M., Knoll, A., Brandt, T., and Glasauer, S. (2008). Human-robot interaction in handing-over tasks. In *RO-MAN 2008 - The 17th IEEE International Symposium on Robot and Human Interactive Communication*, pages 107–112.

[76] Iashin, V., Palermo, F., Solak, G., and Coppola, C. (2021). Top-1 corsmal challenge 2020 submission: Filling mass estimation using multi-modal observations of human-robot handovers. In *Pattern Recognition. ICPR International Workshops and Challenges*, pages 423–436.

[77] Ishikawa, R., Nagao, Y., Hachiuma, R., and Saito, H. (2021). Audio-visual hybrid approach for filling mass estimation. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10-15, 2021, Proceedings, Part VIII*, pages 437–450. Springer.

[78] Jamone, L., Ugur, E., Cangelosi, A., Fadiga, L., Bernardino, A., Piater, J., and Santos-Victor, J. (2018). Affordances in Psychology, Neuroscience, and Robotics: A Survey. *IEEE Transactions on Cognitive and Developmental Systems*, 10(1):4–25.

[79] Johansson, G. (1975). Visual motion perception. *Scientific American*, 232(6):76–89.

[80] Kassner, M., Patera, W., and Bulling, A. (2014). Pupil: An Open Source Platform for Pervasive Eye Tracking and Mobile Gaze-based Interaction. *arXiv:1405.0006 [cs]*. arXiv: 1405.0006.

[81] Khusainov, R., Azzi, D., Achumba, I. E., and Bersch, S. D. (2013). Real-time human ambulation, activity, and physiological monitoring: Taxonomy of issues, techniques, applications, challenges and limitations. *Sensors*, 13(10):12852–12902.

[82] Knepper, R. A., Mavrogiannis, C. I., Proft, J., and Liang, C. (2017). Implicit communication in a joint action. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '17, page 283–292.

[83] Knoblich, G., Butterfill, Stephen, and Sebanz, N. (2011). Psychological research on joint action: Theory and data. In *Advances in Research and Theory*, volume 54 of *Psychology of Learning and Motivation*, pages 59–101. Academic Press.

[84] Kokic, M., Stork, J. A., Haustein, J. A., and Kragic, D. (2017). Affordance detection for task-specific grasping using deep learning. In *2017 IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids)*, pages 91–98.

[85] Koul, A., Cavallo, A., Ansuini, C., and Becchio, C. (2016). Doing it your way: How individual movement styles affect action prediction. *PLOS ONE*, 11(10):1–14.

[86] Kucherenko, T. (2018). Data driven non-verbal behavior generation for humanoid robots. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, ICMI '18, page 520–523.

[87] Lacquaniti, F., Terzuolo, C., and Viviani, P. (1983). The law relating the kinematic and figural aspects of drawing movements. *Acta Psychologica*, 54(1):115–130.

[88] Lastrico, L., Carfì, A., Rea, F., Sciutti, A., and Mastrogiovanni, F. (2021a). From movement kinematics to object properties: Online recognition of human carefulness. In *Social Robotics*, volume 13086, pages 61–72. Springer International Publishing.

[89] Lastrico, L., Carfì, A., Vignolo, A., Sciutti, A., Mastrogiovanni, F., and Rea, F. (2021b). Careful with that! observation of human movements to estimate objects properties. In *Human-Friendly Robotics 2020*, pages 127–141. Springer International Publishing.

[90] Lastrico, L., Duarte, N. F., Carfí, A., Rea, F., Mastrogiovanni, F., Sciutti, A., and Santos-Victor, J. (2022a). If you are careful, so am I! How robot communicative motions can influence human approach in a joint task. In *Social Robotics*, volume 13087, pages 267–279. Springer International Publishing.

[91] Lastrico, L., Garello, L., Rea, F., Noceti, N., Mastrogiovanni, F., Sciutti, A., and Carfì, A. (2022b). Robots with different embodiments can express and influence carefulness in object manipulation. In *2022 IEEE International Conference on Development and Learning (ICDL)*, pages 280–286.

[92] Law, T., Leeuw, J., and Long, Jr, J. (2021). How movements of a non-humanoid robot affect emotional perceptions and trust. *International Journal of Social Robotics*, 13.

[93] Li, J. (2015). The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents. *International Journal of Human-Computer Studies*, 77:23–37.

[94] Li, S. and Zhang, X. (2017). Implicit intention communication in human–robot interaction through visual behavior studies. *IEEE Transactions on Human-Machine Systems*, 47(4):437–448.

[95] Liu, H. and Wang, L. (2018). Gesture recognition for human-robot collaboration: A review. *International Journal of Industrial Ergonomics*, 68:355–367.

[96] Lohan, K., Ahmad, M. I., Dondrup, C., Ardón, P., Pairet, È., and Vinciarelli, A. (2020). *Adapting Movements and Behaviour to Favour Communication in Human-Robot Interaction*, pages 271–297. Springer International Publishing, Cham.

[97] Lohan, K. S., Lehmann, H., Dondrup, C., Broz, F., and Kose, H. (2019). *Enriching the Human-Robot Interaction Loop with Natural, Semantic, and Symbolic Gestures*, pages 2199–2219. Springer Netherlands, Dordrecht.

[98] Lombardi, G., Zenzeri, J., Belgiovine, G., Vannucci, F., Rea, F., Sciutti, A., and Di Cesare, G. (2021). The influence of vitality forms on action perception and motor response. *Scientific Reports*, 11.

[99] Lorenz, T., Weiss, A., and Hirche, S. (2016). Synchrony and reciprocity: Key mechanisms for social companion robots in therapy and care. *International Journal of Social Robotics*, 8:125–143.

[100] Lyu, Y., Yang, Y., and Ru, J. (2015). Gesture motion detection algorithm based on optical flow method. In *2015 IEEE International Conference on Computer and Communications (ICCC)*, pages 128–132.

[101] Macciò, S., Carfì, A., and Mastrogiovanni, F. (2022). Mixed reality as communication medium for human-robot collaboration. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2796–2802.

[102] Mandal, F. B. (2014). Nonverbal communication in humans. *Journal of Human Behavior in the Social Environment*, 24(4):417–421.

[103] Mandikal, P. and Grauman, K. (2021). Learning dexterous grasping with object-centric visual affordances. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6169–6176.

[104] Manera, V., Becchio, C., Cavallo, A., Sartori, L., and Castiello, U. (2011). Cooperation or competition? discriminating between social intentions by observing prehensile movements. *Experimental brain research*, 211:547–556.

[105] Maurice, P., Huber, M. E., Hogan, N., and Sternad, D. (2018). Velocity-curvature patterns limit human–robot physical interaction. *IEEE Robotics and Automation Letters*, 3(1):249–256.

[106] Mavridis, N. (2015). A review of verbal and non-verbal human–robot interactive communication. *Robotics and Autonomous Systems*, 63:22–35.

[107] McEllin, L., Sebanz, N., and Knoblich, G. (2018). Identifying others' informative intentions from movement kinematics. *Cognition*, 180:246–258.

[108] Metta, G., Fitzpatrick, P., and Natale, L. (2006). YARP: Yet another robot platform. *International Journal of Advanced Robotic Systems*, 3:43–48.

[109] Metta, G., Natale, L., Nori, F., Sandini, G., Vernon, D., Fadiga, L., von Hofsten, C., Rosander, K., Lopes, M., Santos-Victor, J., Bernardino, A., and Montesano, L. (2010). The iCub humanoid robot: An open-systems platform for research in cognitive development. *Neural Networks*, 23(8):1125–1134. Social Cognition: From Babies to Robots.

[110] Micelli, V., Strabala, K., and Srinivasa, S. (2011). Perception and control challenges for effective human-robot handoffs. In *Proceedings of RSS '11 RGB-D Workshop*.

[111] Modas, A., Xompero, A., Sanchez-Matilla, R., Frossard, P., and Cavallaro, A. (2021). Improving filling level classification with adversarial training. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 829–833. IEEE.

[112] Mogren, O. (2016). C-RNN-GAN: Continuous recurrent neural networks with adversarial training. In *Proceedings of the Constructive Machine Learning Workshop (CML) at NeurIPS*, Barcelona, Spain.

[113] Mottaghi, R., Schenck, C., Fox, D., and Farhadi, A. (2017). See the Glass Half Full: Reasoning About Liquid Containers, Their Volume and Content. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1889–1898, Venice, Italy. IEEE.

[114] Mousavian, A., Eppner, C., and Fox, D. (2019). 6-DOF Graspnet: Variational grasp generation for object manipulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2901–2910.

[115] Murray, M., Walker, N., Nanavati, A., Alves-Oliveira, P., Filippov, N., Sauppe, A., Mutlu, B., and Cakmak, M. (2022). Learning backchanneling behaviors for a social robot via data augmentation from human-human conversations. In *Conference on Robot Learning*, pages 513–525. PMLR.

[116] Nagasaki, H. (2004). Asymmetric velocity and acceleration profiles of human arm movements. *Experimental Brain Research*, 74:319–326.

[117] Neverova, N., Wolf, C., Lacey, G., Fridman, L., Chandra, D., Barbello, B., and Taylor, G. (2016). Learning human identity from motion patterns. *IEEE Access*, 4:1810–1820.

[118] 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.

[119] Nishimura, Y., Nakamura, Y., and Ishiguro, H. (2020). Long-term motion generation for interactive humanoid robots using gan with convolutional network. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, page 375–377, Cambridge, United Kingdom.

[120] Nomura, T., Suzuki, T., Kanda, T., and Kato, K. (2006a). Measurement of anxiety toward robots. In *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*, pages 372 – 377.

[121] Nomura, T., Suzuki, T., Kanda, T., and Kato, K. (2006b). Measurement of negative attitudes toward robots. *Interaction Studies*, 7:437–454.

[122] Noseworthy, M., Paul, R., Roy, S., Park, D., and Roy, N. (2020). Task-conditioned variational autoencoders for learning movement primitives. In *Conference on robot learning*, pages 933–944. PMLR.

[123] Pang, Y. L., Xompero, A., Oh, C., and Cavallaro, A. (2021). Towards safe human-to-robot handovers of unknown containers. In *30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*, pages 51–58, Virtual.

[124] Pattacini, U., Nori, F., Natale, L., Metta, G., and Sandini, G. (2010). An experimental evaluation of a novel minimum-jerk cartesian controller for humanoid robots. In *2010 IEEE/RSJ international conference on intelligent robots and systems*, pages 1668–1674. IEEE.

[125] Pezzulo, G., Donnarumma, F., and Dindo, H. (2013). Human sensorimotor communication: A theory of signaling in online social interactions. *PloS One*, 8:1–11.

[126] Plamondon, R. and Alimi, A. M. (1997). Speed/accuracy trade-offs in target-directed movements. *Behavioral and Brain Sciences*, 20(2):279–303.

[127] Raković, M., Duarte, N. F., Marques, J., Billard, A., and Santos-Victor, J. (2022). The gaze dialogue model: Nonverbal communication in hhi and hri. *IEEE Transactions on Cybernetics*, pages 1–0.

[128] Rasch, R., Wachsmuth, S., and König, M. (2018). A joint motion model for human-like robot-human handover. In *2018 IEEE-RAS 18th International Conference on Humanoid Robots (Humanoids)*, pages 180–187.

[129] Ravichandar, H., Polydoros, A. S., Chernova, S., and Billard, A. (2020). Recent advances in robot learning from demonstration. *Annual Review of Control, Robotics, and Autonomous Systems*, 3(1):297–330.

[130] Rea, F., Vignolo, A., Sciutti, A., and Noceti, N. (2019). Human motion understanding for selecting action timing in collaborative human-robot interaction. *Frontiers in Robotics and AI*, 6:58.

[131] Reinhardt, J., Pereira, A., Beckert, D., and Bengler, K. (2017). Dominance and movement cues of robot motion: A user study on trust and predictability. In *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1493–1498.

[132] Rizzolatti, G., Fadiga, L., Fogassi, L., and Gallese, V. (1999). Resonance behaviors and mirror neurons. *Archives italiennes de biologie*, 137 2-3:85–100.

[133] Rosen, E., Whitney, D., Phillips, E., Chien, G., Tompkin, J., Konidaris, G., and Tellex, S. (2019). Communicating and controlling robot arm motion intent through mixed-reality head-mounted displays. *The International Journal of Robotics Research*, 38(12-13):1513–1526.

[134] Runeson, S. and Frykholm, G. (1981). Visual-perception of lifted weight. *Journal of experimental psychology. Human perception and performance*, 7:733–40.

[135] Runeson, S. and Frykholm, G. (1983). Kinematic specification of dynamics as an informational basis for person-and-action perception: Expectation, gender recognition, and deceptive intention. *Journal of Experimental Psychology: General*, 112:585–615.

[136] Sabu, S., Curioni, A., Vesper, C., Sebanz, N., and Knoblich, G. (2020). How does a partner's motor variability affect joint action? *PLOS ONE*, 15(10):1–24.

[137] Sainath, T. N., Vinyals, O., Senior, A., and Sak, H. (2015). Convolutional, long short-term memory, fully connected deep neural networks. In *Proceedings of the 2015 IEEE ICASSP*, pages 4580–4584, Brisbane, Australia.

[138] Salem, M., Kopp, S., Wachsmuth, I., Rohlfing, K., and Joublin, F. (2012). Generation and evaluation of communicative robot gesture. *International Journal of Social Robotics*, 4:201–217.

[139] Sanchez, J., Corrales, J.-A., Bouzgarrou, B.-C., and Mezouar, Y. (2018). Robotic manipulation and sensing of deformable objects in domestic and industrial applications: a survey. *The International Journal of Robotics Research*, 37(7):688–716.

[140] Sanchez-Matilla, R., Chatzilygeroudis, K., Modas, A., Duarte, N. F., Xompero, A., Frossard, P., Billard, A., and Cavallaro, A. (2020). Benchmark for human-to-robot handovers of unseen containers with unknown filling. *IEEE Robotics and Automation Letters*, 5(2):1642–1649.

[141] Sandini, G., Sciutti, A., and Rea, F. (2018). *Movement-Based Communication for Humanoid-Human Interaction*, pages 1–29. Springer Netherlands, Dordrecht.

[142] Saponaro, G., Jamone, L., Bernardino, A., and Salvi, G. (2017). Interactive Robot Learning of Gestures, Language and Affordances. In *GLU 2017 International Workshop on Grounding Language Understanding*, pages 83–87. ISCA.

[143] Sartori, L., Becchio, C., Bara, B. G., and Castiello, U. (2009). Does the intention to communicate affect action kinematics? *Consciousness and Cognition*, 18(3):766–772.

[144] Scassellati, B. (1999). Imitation and mechanisms of joint attention: A developmental structure for building social skills on a humanoid robot. *Lecture Notes in Computer Science*, pages 176–195.

[145] Schaal, S. (2006). *Dynamic Movement Primitives -A Framework for Motor Control in Humans and Humanoid Robotics*, pages 261–280. Springer Tokyo, Tokyo.

[146] Schaefer, K. E. (2016). *Measuring Trust in Human Robot Interactions: Development of the "Trust Perception Scale-HRI"*, pages 191–218. Springer US, Boston, MA.

[147] Schulz, T., Torresen, J., and Herstad, J. (2019). Animation techniques in human-robot interaction user studies: A systematic literature review. *J. Hum.-Robot Interact.*, 8(2).

[148] Sciutti, A., Bisio, A., Nori, F., Metta, G., Fadiga, L., Pozzo, T., and Sandini, G. (2012). Measuring human-robot interaction through motor resonance. *International Journal of Social Robotics*, 4:223–234.

[149] Sciutti, A., Mara, M., Tagliasco, V., and Sandini, G. (2018). Humanizing human-robot interaction: On the importance of mutual understanding. *IEEE Technology and Society Magazine*, 37(1):22–29.

[150] Sciutti, A., Patane, L., Nori, F., and Sandini, G. (2014). Understanding object weight from human and humanoid lifting actions. *Autonomous Mental Development, IEEE Transactions*, 6:80–92.

[151] Sciutti, A., Patanè, L., and Sandini, G. (2019). Development of visual perception of others' actions: Children's judgment of lifted weight. *PLOS ONE*, 14(11):1–15.

[152] Sciutti, A. and Sandini, G. (2019). The role of object motion in visuo-haptic exploration during development. In *2019 Joint IEEE 9th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, pages 123–128.

[153] Sisbot, E. A. and Alami, R. (2012). A human-aware manipulation planner. *IEEE Transactions on Robotics*, 28(5):1045–1057.

[154] Stein, S. and Mckenna, S. (2013). Combining embedded accelerometers with computer vision for recognizing food preparation activities. *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*.

[155] Strabala, K., Lee, M. K., Dragan, A., Forlizzi, J., and Srinivasa, S. S. (2012). Learning the communication of intent prior to physical collaboration. In *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*, pages 968–973.

[156] Strabala, K., Lee, M. K., Dragan, A., Forlizzi, J., Srinivasa, S. S., Cakmak, M., and Micelli, V. (2013). Toward seamless human-robot handovers. *J. Hum.-Robot Interact.*, 2(1):112–132.

[157] Su, Z., Hausman, K., Chebotar, Y., Molchanov, A., Loeb, G. E., Sukhatme, G. S., and Schaal, S. (2015). Force estimation and slip detection/classification for grip control using a biomimetic tactile sensor. In *Proceedings of 15th IEEE-RAS International Conference on Humanoid Robots*, pages 297–303, Seoul, Korea.

[158] Tenorth, M., Bandouch, J., and Beetz, M. (2009). The tum kitchen data set of everyday manipulation activities for motion tracking and action recognition. *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, pages 1089–1096.

[159] Ugur, E., Nagai, Y., Sahin, E., and Oztop, E. (2015). Staged development of robot skills: Behavior formation, affordance learning and imitation with motionese. *IEEE Transactions on Autonomous Mental Development*, 7(2):119–139.

[160] Vannucci, F., Di Cesare, G., Rea, F., Sandini, G., and Sciutti, A. (2018). A robot with style: Can robotic attitudes influence human actions? In *2018 IEEE-RAS 18th International Conference on Humanoid Robots (Humanoids)*, pages 1–6.

[161] Vasalya, A., Ganesh, G., and Kheddar, A. (2018). More than just co-workers: Presence of humanoid robot co-worker influences human performance. *PLOS ONE*, 13(11):1–19.

[162] Venture, G. and Kulić, D. (2019). Robot expressive motions: A survey of generation and evaluation methods. *J. Hum.-Robot Interact.*, 8(4).

[163] Vesper, C., Schmitz, L., and Knoblich, G. (2017). Modulating action duration to establish non-conventional communication. *Journal of Experimental Psychology: General*, 146(12):1722–1737.

[164] Viet Tuyen, N. T., Elibol, A., and Chong, N. Y. (2020). Learning from humans to generate communicative gestures for social robots. In *2020 17th International Conference on Ubiquitous Robots (UR)*, pages 284–289.

[165] Vignolo, A., Noceti, N., Rea, F., Sciutti, A., Odone, F., and Sandini, G. (2017). Detecting biological motion for human–robot interaction: A link between perception and action. *Frontiers in Robotics and AI*, 4:14.

[166] Viviani, P. and Flash, T. (1995). Minimum-jerk, two-thirds power law, and isochrony: converging approaches to movement planning. *Journal of experimental psychology. Human perception and performance*, 21 1:32–53.

[167] Viviani, P. and Schneider, R. (1991). A developmental study of the relationship between geometry and kinematics in drawing movements. *Journal of experimental psychology. Human perception and performance*, 17 1:198–218.

[168] Wang, H., Zhu, C., Ma, Z., and Oh, C. (2022). Improving generalization of deep networks for estimating physical properties of containers and fillings. In *the 32nd IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 9147–9151, Singapore.

[169] Wang, Z., She, Q., and Ward, T. E. (2021). Generative adversarial networks in computer vision: A survey and taxonomy. *ACM Comput. Surv.*, 54(2).

[170] Watanabe, K. (2008). Behavioral speed contagion: Automatic modulation of movement timing by observation of body movements. *Cognition*, 106(3):1514–1524.

[171] Wu, X., Sahoo, D., and Hoi, S. C. (2020). Recent advances in deep learning for object detection. *Neurocomputing*, 396:39–64.

[172] Xiong, Y. and Quek, F. K. H. (2006). Hand motion gesture frequency properties and multimodal discourse analysis. *International Journal of Computer Vision*, 69:353–371.

[173] Xompero, A., Donaher, S., Iashin, V., Palermo, F., Solak, G., Coppola, C., Ishikawa, R., Nagao, Y., Hachiuma, R., Liu, Q., et al. (2022). The corsmal benchmark for the prediction of the properties of containers. *IEEE Access*.

[174] Xu, J., Broekens, J., Hindriks, K., and Neerincx, M. (2015). Mood contagion of robot body language in human robot interaction. *Autonomous Agents and Multi-Agent Systems*, 29:1216–1248.

[175] Yang, F. and Peters, C. (2019). Appgan: Generative adversarial networks for generating robot approach behaviors into small groups of people. In *Proceedings of the 28th IEEE International Conference on Robot and Human Interactive Communication*, pages 1–8, New Delhi, India.

[176] Yang, W., Paxton, C., Mousavian, A., Chao, Y.-W., Cakmak, M., and Fox, D. (2021). Reactive human-to-robot handovers of arbitrary objects. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3118–3124.

[177] Yoon, J., Jarrett, D., and van der Schaar, M. (2019). Time-series generative adversarial networks. In *Proceedings of 33rd Conference on Neural Information Processing Systems*, Vancouver, Canada.

[178] Yu, L.-F., Duncan, N., and Yeung, S.-K. (2015). Fill and Transfer: A Simple Physics-Based Approach for Containability Reasoning. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 711–719, Santiago, Chile. IEEE.

[179] Zeng, A., Song, S., Yu, K.-T., Donlon, E., Hogan, F. R., Bauza, M., Ma, D., Taylor, O., Liu, M., Romo, E., Fazeli, N., Alet, F., Dafle, N. C., Holladay, R., Morona, I., Nair, P. Q., Green, D., Taylor, I., Liu, W., Funkhouser, T., and Rodriguez, A. (2022). Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching. *The International Journal of Robotics Research*, 41(7):690–705.

[180] Zhao, Z.-Q., Zheng, P., Xu, S.-T., and Wu, X. (2019). Object detection with deep learning: A review. *IEEE Transactions on Neural Networks and Learning Systems*, 30(11):3212–3232.

[181] Zheng, Q., Wu, W., Pan, H., Mitra, N., Cohen-Or, D., and Huang, H. (2021). Inferring object properties from human interaction and transferring them to new motions. *Computational Visual Media*, 7.

[182] Zhou, A., Hadfield-Menell, D., Nagabandi, A., and Dragan, A. D. (2017). Expressive robot motion timing. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '17, page 22–31.

[183] Özkan, E. E., Gurion, T., Hough, J., Healey, P. G., and Jamone, L. (2021). Specific hand motion patterns correlate to miscommunications during dyadic conversations. In *2021 IEEE International Conference on Development and Learning (ICDL)*, pages 1–6.