EPFL

TÉCNICO LISBOA

Thèse n° 11681

Non-verbal Communication between Humans and Robots: Imitation, Mutual Understanding and Inferring Object Properties

Présentée le 6 avril 2023

École Polytechnique Fédérale de Lausanne
Faculté des sciences et techniques de l'ingénieur
Laboratoire d'algorithmes et systèmes d'apprentissage

À l'Instituto Superior Técnico (IST) da Universidade de Lisboa
Programme doctoral en robotique, contrôle et systèmes intelligents
Doutoramento em Engenharia Informática e de Computadores

pour l'obtention du grade de Docteur ès Sciences

par

Nuno Ricardo FERREIRA DUARTE

Acceptée sur proposition du jury

Prof. J. M. L. M. Lemos, président du jury
Prof. A. Billard, Prof. J. Santos-Victor, directeurs de thèse
Prof. M. Lopes, rapporteur
Dr S. Ivaldi, rapporteuse
Dr A. Sciutti, rapporteuse
Prof. A. J. M. Bernardino, rapporteur

École polytechnique fédérale de Lausanne

2023

Abstract

Humans can express their actions and intentions, resorting to verbal and/or non-verbal communication. In verbal communication, humans use language to express, in structured linguistic terms, the desired action they wish to perform. Non-verbal communication refers to the expressiveness of the human body movements during the interaction with other humans, while manipulating objects, or simply navigating in the world. In a sense, all actions require moving our musculoskeletal system which in return contribute to expressing the intention concerning the completion of that action. Moreover, considering that all humans share a common motor-repertoire, i.e. the degrees of freedom and joint limits, excluding cultural or society-based influences, all humans express action intentions using a common non-verbal language. From walking along a corridor, to pointing to a painting on a wall, or handing over a cup to someone, communication is provided in the form of non-verbal "cues", that express action intentions.

This thesis objective is hence threefold: (i) improve robot imitation of human actions by incorporating human-inspired non-verbal cues onto robots; (ii) explore how humans communicate their goals and intention non-verbally and how robots can use the same non-verbal cues to also communicate its goals and intentions to humans; and (iii) extract latent properties of objects that are revealed by human non-verbal cues during manipulation and incorporate them onto the robot non-verbal cue system in order to express those properties.

One of the contributions is the creation of multiple publicly available datasets of synchronized videos, gaze, and body motion data. We conducted several Human-human interaction experiments with three objectives in mind: (i) study the motion behaviors of both perspectives in human-human interactions, (ii) understand how the participants manage to predict the observed actions of the other; (iii) use the collected data to model the human eye-gaze behavior and arm behavior.

The second contribution is an extension to the legibility concept to include eye-gaze cues. This extension proved that humans can correctly predict the robot action as early, and with the same cues, as if it were a human doing it.

The third contribution is developing a human-to-human synchronized non-verbal communication model, i.e. the *Gaze Dialogue*, which shows the inter-personal communication of motor and gaze cues that occur during action execution and observation, and apply it to a human-to-robot experiment. During the interaction, the robot can: (i) adequately infer the human action from gaze cues, (ii) adjust its gaze fixation according to the human eye-gaze behavior, and (iii) signal non-verbal cues that correlate with the robot's own action intentions.

The fourth and final contribution is to demonstrate that non-verbal cues information extracted from human can be used by robots in recognizing the types of actions (individual or action-in-interactions), the types of intentions (to polish or to handover), and the types of manipulations (careful or careless).

Overall, the communication tools developed in this thesis contribute to enhance of human robot interaction experience, by incorporating the non-verbal communication "protocols" used when humans interact with each other.

Keywords: Non-verbal Cues, Human-Human Interaction, Eyes and Body Tracking, Mutual Understanding, Human-Robot Interaction

Resumo

Os seres humanos são capazes de expressar as suas acções e intenções recorrendo à comunicação verbal e/ou não verbal. Na comunicação verbal, os seres humanos utilizam a linguagem para expressar, em termos linguísticos, a acção que querem realizar. A comunicação não verbal refere-se à expressividade dos movimentos do corpo humano durante a interacção com outros seres humanos, quer seja durante a manipulação de objectos, quer seja simplesmente a deambular. De certa forma, todas as acções requerem o movimento do nosso sistema músculo-esquelético que, em troca, contribuem para expressar a nossa intenção relativa à conclusão dessa acção. Além disso, tendo em conta que todos os seres humanos partilham um repertório comum para movimentos corporais, ou seja, agimos de forma semelhante, excluindo influências culturais ou sociais, conclui-se que todos os seres humanos expressam intenções de acção utilizando uma linguagem não verbal comum. Desde o caminhar ao longo de um corredor, até apontar para uma pintura numa parede, ou entregar uma chávena a alguém, a comunicação é fornecida sob a forma de "sinais" não-verbais, que expressam intenções de acção.

Esta tese tem três objetivos: (i) melhorar a capacidade de imitação das ações dos humanos por parte dos robôs ao incorporar nos robôs a comunicação não verbal inspirada na dos humanos quando executam ações; (ii) explorar como os humanos comunicam aos outros as suas ações ou intenções de agir e como os robôs podem tirar partido dessa comunicação para expressar a sua ação ou intenção de agir; e (iii) extrair as propriedades latentes dos objetos que são reveladas pela comunicação não verbal dos humanos durante a manipulação dos mesmos e a sua incorporação no sistema não verbal dos robôs de maneira a expressar essas propriedades.

Uma das contribuições foi a criação de múltiplas bases de dados, gratuitamente disponíveis, de vídeos e um conjunto de dados sincronizados dos movimentos dos olhos e do corpo. Realizámos várias experiências com três objetivos em mente: (i) estudar os movimentos de ambas as pessoas nas interacções humano-humano, (ii) compreender como as pessoas conseguem prever as acções observadas dos outros; (iii) utilizar os dados recolhidos para modelar o comportamento do braço humano e o comportamento do olhos.

A segunda contribuição é uma extensão do conceito de legibilidade ao incluir os movimentos dos olhos baseado nos humanos. Esta extensão provou que os humanos podem prever correctamente a acção do robô tão cedo e com as mesmos sinais como se fosse um humano a fazê-lo.

A terceira contribuição é desenvolver um modelo de comunicação não verbal sincronizado entre humanos, intitulado de o *Diálogo dos olhos*, que mostra a comunicação interpessoal de sinais motores e visuais que ocorrem durante a execução e observação de uma acção, e aplicá-lo a uma experiência humano-robô. Durante a interacção, o robô é capaz de: (i) inferir adequadamente a acção humana a partir de sinais do olhar, (ii) ajustar a sua fixação do olhar de acordo com o comportamento do olhar humano, e (iii) sinalizar sinais não-verbais que se correlacionam com as intenções de acção do robô.

A quarta e última contribuição é uma demonstração que sinais não-verbais extraidos dos humanos pode ser usado por robôs para reconhecimento dos tipos de acções (individuais ou de acção-interacções), os tipos de intenções (polir a mesa ou entregar um objecto), e os tipos de manipulações (cuidadosas ou descuidadas).

Em resumo, os instrumentos de comunicação desenvolvidos nesta tese contribuem para o melhoramento das interações humano-robôs ao incorporarem os "protocolos" de comunicação não verbal usados pelos humanos entre sí.

Palavras-Chave: Sinais Não Verbais, Interação Humano-Humano, Deteção do Movimentos dos Olhos e do Corpo, Reconhecimento Mutuo, Interação Humano-Robo

Resumé

Les êtres humains sont capables d'exprimer leurs actions et leurs intentions par la communication verbale et/ou non verbale. Dans la communication verbale, les êtres humains utilisent le langage pour exprimer, en termes linguistiques, l'action qu'ils veulent réaliser. La communication non verbale fait référence à l'expressivité des mouvements du corps humain lors de l'interaction avec d'autres êtres humains, que ce soit lors de la manipulation d'objets ou simplement en marchant. En un sens, toutes les actions nécessitent le mouvement de notre système musculo-squelettique qui, en retour, contribue à exprimer notre intention concernant l'accomplissement de cette action. En outre, étant donné que tous les êtres humains partagent un répertoire commun de mouvements corporels, c'est-à-dire que nous agissons de manière similaire, à l'exclusion des influences culturelles ou sociales, il s'ensuit que tous les êtres humains expriment des intentions d'action en utilisant un langage non verbal commun. Qu'il s'agisse de marcher dans un couloir, de montrer du doigt un tableau sur un mur ou de tendre une tasse à quelqu'un, la communication se fait sous la forme de "signaux" non verbaux qui expriment des intentions d'action.

Cette thèse a trois objectifs : (i) améliorer la capacité des robots à imiter les actions des humains en incorporant dans les robots la communication non verbale inspirée de celle des humains lorsqu'ils effectuent des actions ; (ii) explorer comment les humains communiquent leurs actions ou leurs intentions d'agir aux autres et comment les robots peuvent tirer parti de cette communication pour exprimer leur action ou leur intention d'agir ; et (iii) extraire les propriétés latentes des objets qui sont révélées par la communication non verbale des humains lors de leur manipulation et leur incorporation dans le système non verbal des robots afin d'exprimer ces propriétés.

L'une des contributions a été la création de plusieurs bases de données de vidéos, librement accessibles, et d'un ensemble de données synchronisées sur les mouvements des yeux et du corps. Nous avons mené plusieurs expériences avec trois objectifs en tête : (i) étudier les mouvements des deux personnes dans les interactions homme-homme, (ii) comprendre comment les personnes peuvent prédire les actions observées des autres ; (iii) utiliser les données collectées pour modéliser le comportement des bras et des yeux humains.

La deuxième contribution est une extension du concept de lisibilité en incluant les mouvements oculaires humains. Cette extension a prouvé que les humains peuvent prédire correctement l'action du robot aussi tôt et avec les mêmes indices que si un humain le faisait.

La troisième contribution consiste à développer un modèle de communication non-verbale synchronisée entre humains, intitulé le *Gaze Dialogue*, qui montre la communication interpersonnelle des signaux moteurs et visuels qui se produisent pendant l'exécution et l'observation d'une action, et à l'appliquer à une expérience homme-robot. Pendant l'interaction, le robot est capable : (i) de déduire de manière appropriée l'action humaine à partir des signaux du regard, (ii) d'ajuster la fixation de son regard en fonction du comportement du regard humain, et (iii) de signaler des signaux non verbaux en corrélation avec les intentions d'action du robot.

La quatrième et dernière contribution est une démonstration que les signaux non verbaux extraits des humains peuvent être utilisés par les robots pour reconnaître les types d'actions (individuelles ou actions-interactions), les types d'intentions (polir la table ou livrer un objet), et les types de manipulations (soigneuses ou négligentes).

En résumé, les outils de communication développés dans cette thèse contribuent à l'amélioration des interactions homme-robot en intégrant les "protocoles" de communication non-verbale utilisés par les humains entre eux.

Mots-clés: Indices non verbaux, interaction homme-homme, suivi des yeux et du corps, compréhension mutuelle, interaction homme-robot.

Acknowledgments

I would to begin by expressing my gratitude to Prof. José Santos-Victor, my IST thesis supervisor. I'm truly grateful for his never ending patience throughout the course of my PhD endeavours. His feedback, advices and comments have helped me tremendously during my struggles and thanks to his presence and support I grew as a person and as a researcher. I am a much more confident researcher thanks to Jose and I will be forever in debt. I would like to also express my gratitude to Prof. Aude Billard, my EPFL thesis supervisor. My stay in LASA was wonderful and I could have not asked for a better environment to continue my research. You gave me liberty to continue the works that I had already done in Vislab, IST, while at the same time explore on new challenges that contributed to not only my interest but yours. I am grateful to both of my supervisors for pushing me to be better, finding time to review any of my papers, and for guiding me in this incredible journey.

This thesis would not have been possible without the effort, guidance, mentoring and comradery of Mirko Raković. The first year of the PhD was challenging because I did not know what I was supposed to do. After Mirko and I started collaborating, my confidence grew, motivation and productivity, as I began to build the path of this thesis. You shaped this thesis just as much as I did and I guarantee it that I would not have been as prepared now if we had not worked together. A tremendous Hvala.

There are other people who influenced my journey and this thesis which I would like to thank. To Konstantinos Chatzilygeroudis (Kostas) which throughout my stay at LASA was a collaborator, a teacher, and a mentor. To Lorenzo Jamone, for providing several insightful advices and suggestions whenever we had the chance to converse. And a special thanks to Prof. Alexandre Bernardino and Prof. José Gaspar, for their valuable feedback in many of my presentations to the VisLab group.

This work was funded by a Fundação para a Ciência e a Tecnologia (FCT) doctoral grant under the IST-EPFL Joint Doctoral Program (PD/BD/135116/2017). My PhD research was also funded through European projects (Secondhands and Corsmal). I would like to thank the European Commission for the support and the opportunities created through their funding.

I would like to praise some people who I met during my PhD journey. The people at LASA that were extremely friendly and made me appreciate my time spent there even more (Farshad, Gustav, Illaria, Michael, Michäel, Bernardo, Walid, Saurav, and Carolina). The people at VisLab whom I shared laughs, jokes, and fruitful conversations (Hugo, Ana Luisa, Borrego, Pedro, Avelino, Nuno, Carlos) and the two technicians (Ricardo and Weverton) that were always available to repaired the robots whenever I broke it, needed something, or just to share some thoughts about random topics.

To my family, in particular to my parents, sister, and grandfather for always being there, the four people whom have been present during my entire lifetime (and my sister since I can remember), they shaped me into who I am, they are the most important people in my life, the house where I grew up with the people that I grew up is all I need to recharge batteries, relax, and shutdown from the problems that I might have. I worked my but off to make you proud, if it wasn't for your support I would not have achieved a fifth of what I achieved.

To my short list of friends which have stuck with me from the time we have crossed paths (Rafa, Manel, Chico and Tomás) and to those that I've met over my many journeys and unfortunately have lost touch for various reasons I just want to say that you meant a lot to me at some point in my life, I'll forever keep the memories of our friendship and I just hope that in the future we can meet again.

Finally, the last paragraph is reserved to my dearest life partner Maria. We met during our PhDs and it was the best thing to combine learning about new things while growing as a human being by having you in my life and by my side. I could have not found a better half to complete me.

Anyone who has never made a mistake has never tried anything new.

Albert Einstein

Contents

List of Figures

List of Tables

Acronyms

CDS Coupled Dynamical System. 90, 91, 95, 98

dof Degrees of Freedom. 3

DS Dynamical System. 43–45, 48, 49, 51, 90–92, 95, 96, 98–102, 125–129

ESN Echo State Network. 111–116

FPS Frames per Second. 75

GMM Gaussian Mixture Model. 31, 35, 36, 45, 91, 93, 126, 127, 144

GMR Gaussian Mixture Regression. 31, 36, 37, 93, 127, 128

HHI Human-Human Interaction. 12, 22, 24, 31, 34, 37, 58, 63, 65, 66, 68, 69, 71–73, 75, 77, 79, 82, 83, 87, 90, 91, 93, 95, 99, 122, 131, 138, 143, 144, 148, 149

HMM Hidden Markov Model. 58, 63, 64, 66, 67, 75, 77, 78

HRI Human-Robot Interaction. 9, 10, 12, 24, 25, 27, 31, 42, 58, 71–75, 77, 79, 85, 87, 90, 98, 115, 116, 120, 127, 137–141, 145–149

Mocap Motion Capture System. 26, 32, 43, 51, 52, 58, 99, 106, 122, 135, 136, 138, 155, 156, 159, 161, 167

MTS Multivariate Time Series. 111–113

PCA Principal Component Analysis. 113, 114

RNN Recurrent Neural Network. 111–114

1

Introduction

> *If I have seen further it is by standing on the shoulders of Giants.*
>
> Isaac Newton,

Contents

1.1 Motivation

Humans, in the first few months of existence, begin perceiving the actions of others by observation of the surrounding world [Woodward et al., 2009]. It was believed that the ability of human infants to read the intentions of others was an innate trait, nowadays, thanks to studies in early cognitive development it is comprehended that early experiences contribute to social understanding.

In what follows is a road map on the understanding of other's goals and intentions from observation. The first example given is the seventh art, specifically renowned feature films, and how they convey so much information without the use of words ("Show, do not tell"). Then it continues with the examination of the human body posture which from a single instant/frame it is possible to extract the intended goal given the appropriate context. This capacity is accentuated given an example of proficient observers that can detect minute details in the small variations in the body configuration that disambiguate an action as successful or not. This remarkable understanding can be traced back in our brain by the activation of neural mechanisms during observation. These neural mechanisms, i.e. the mirror neurons, are also activated when we use our musculoskeletal system which generates the desired eye-gaze and body movements to perform the action. The motivation of this thesis ends with an overview of current robot scenarios and the advantages of integrating human understanding from observations onto robots.

Cinema and exposition

Movies can be a great source of entertainment yet, for the viewer to comprehend what a character is thinking or feeling, they must interpret their actions, body language, and facial expressions. There usually exists a narrator and dialogues between actors during scenes, but the film is a visual medium and those are add-ons to an already rich storyline. The expression "a picture is worth a thousand words" perfectly applies to cinema. Movie enthusiasts praise directors, cinematographers, and screenwriters who do focus on delivering elaborate, thought provocative, well-devised exposition which entertains and lets the audience decipher its meaning. What cinematographers call exposition involves explaining things to the audience with visuals, sounds, and clever ways to describe information while engaging in the story [Bell, 2004]. Bad exposition is when there are monotonous scenes of extensive speech, either narration or dialogue, to explain a piece of the plot, which makes for an unpleasant tedious experience that may result in disengagement. One example of poorly executed exposition is present in the 2016 feature-length film Suicide Squad (Figure 1.1 on the right). A ubiquitously poorly rated movie which, among other things, introduces characters in an overly descriptive, powerpoint-like, slides which crams all the information in one swift move instead of letting the viewer uncover over the course of the film. On the other hand, the left picture in Figure 1.1 shows one of the iconic moments from The Matrix film where Morpheus gives Neo two options

to continue his journey. This freeze-frame encapsulates so brilliantly the daring decision Neo had to make in order to uncover the truth.

Figure 1.1: Examples of good and bad exposition in movies. On the left: *The Matrix*. Dirs. Lana Wachowski and Lilly Wachowski. Warner Brothers, 1999. DVD. On the right: *Suicide Squad*. Dir. David Ayer. Warner Brothers, 2016. DVD. The first is a critically acclaimed masterpiece while the second is the infamous 2016 version of a failed superhero movie. The images are frames taken directly from each film.

The concept of "Show, do not tell" [Yarmolinsky, 1954] encapsulates the before mentioned differences in exposition. Once a technique used exclusively by writers and novelists to reveal information not through long descriptions but by invoking the reader's thoughts, senses, and feelings. It was soon adopted to cinematography given that movies work primarily by conveying visual cues during storytelling. The technique envisions an audience capable of intellectual reasoning to decode the information portrait in the visual exposition. "Showing" important information during movies scenes can be focusing the audience attentions on places, people, or things. Even a scene without dialogue can express emotion, reveal (i.e. the "telling" of) intention, or desire. It is not fun or enthusiastic hearing actors explain every important plot scene and the same can be applied to robots. It will not be pleasant if in order to interact with a robot, verbal instructions are required for every step of the collaboration. In that sense, this thesis is inspired by the "Show, do not tell" technique, where information is revealed by visual cues of the human and the robot.

Body posture, motion and gestures (the showing)

reveals a lot about someone's goals or intentions (the telling)

The First Human Language

The human body was the first communication language before any speech appeared ([Cummings, 2011]), from facial expressions, arm and leg movements, posture, head orientation, to eye contact. This capacity is referred to as non-verbal communication and involves all the Degrees of Freedom (dof) in our bodies. Our primate ancestors deduced, from observing the other's body and facial expressions, whether they are to be trusted or not [Kuiper, 2008]. In our daily lives, we befriend people that makes us feel safe and secure - which we do not think is a threat. Body language still takes an important part in assessing someone's personality [Cummings, 2011]. In an anatomical sense, each human being has a very similar body config-

uration providing similar responses when executing the same action. Consequently, we can recognize actions by recognizing the motions associated with said action. This repertoire is learned at a very young age when we imitate what our parents and relatives are doing with their hands, arms, face, and body [Revel and Andry, 2009]. As we grow we acquire knowledge throughout our life of how people behave and what is the intrinsic action associated with such behavior.

Figure 1.2: A person reaching for a book in a bookshelf. On the left it is a stock image taken from the web and, on the right, is the same image with the background removed after editing to keep only the subject in focus. Image taken from the web https: //www.pinterest.com/pin/398709373261271920/. Accessed 2 Dec. 2021.

As a practical exercise let us examine the left image in Figure 1.2. From a purely empirical analysis, we can detect a non-neutral and complex body configuration. The person is standing on top of books, the body is rotated to the side and tilted forward, the head is faced slightly upwards and to the right, the right arm is stretched to reach a specific object, and the hand is opened, while the left hand is probably holding the bookcase for support. Finally, one of the legs is raised and tilted backwards which we can assume is providing extra stability to counterbalance the weight of the body that is pushing forward. This is what can be seen from basic inspection and deduction. From the context described above, we can also infer that the eyes, although hard to see, are fixating a specific object, in this case, probably one of the books from the shelf. Additionally, from the context of the scenario (the bookcase) and given the described body posture we can assume that the lady is trying to reach a book on the top shelf. These are evidently decoded from the cues of the stretched arm with the head and eyes focusing a region close to the hand. All of this information can be extracted by simply inspecting the human body without any text description. This is quite a remarkable feat considering it (apparently) did not take much time, effort, and energy for the brain to process this information from a single image. Now, context is extremely important, and it is thanks to it that much of the deduction comes from (specific goal and intention). However, if we were to remove the context, illustrated by Figure 1.2 on the right, we could still argue that the person is reaching for something. The context allows us to decode the specificity of the object (a book), nonetheless, the action (reaching) was clear without the context.

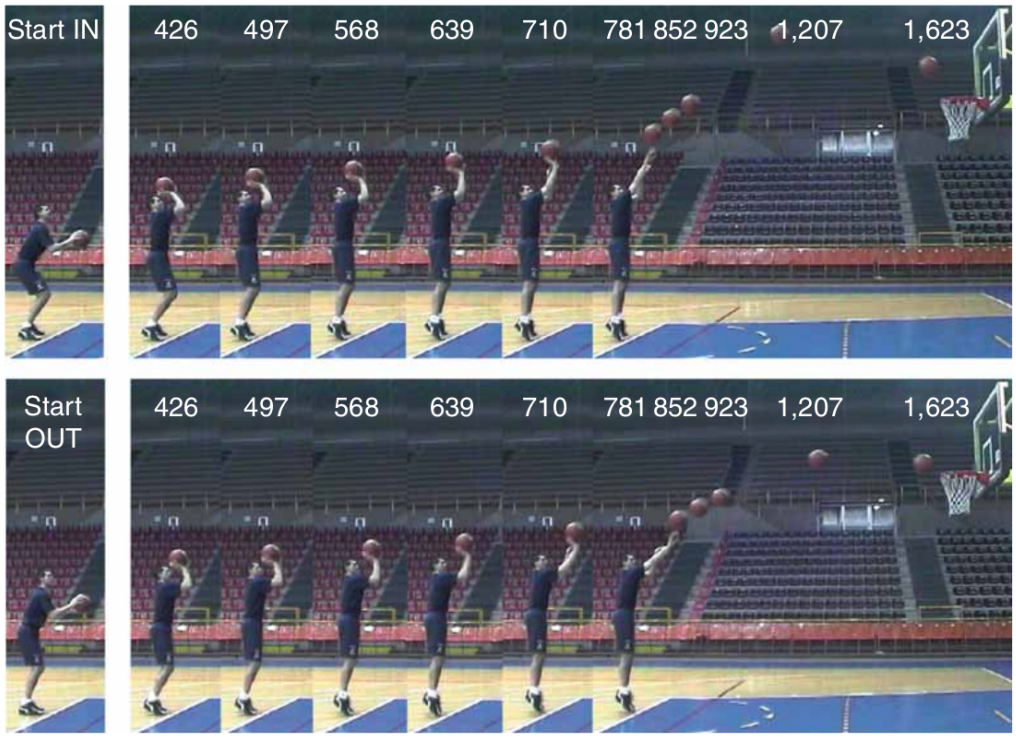


Figure 1.3: A subject attempting to make free-throws in basketball. The number on top reflects the frame where the video sequence was cut. Participants were then asked to answer whether they think the ball is going through the basket or not. The top sequence is of a successful attempt while the bottom is a failed attempt at making a free-throw. Reprinted from "Action anticipation and motor resonance in elite basketball players." by Salvatore Aglioti et al., 2008, Nature Neuroscience, page 2.

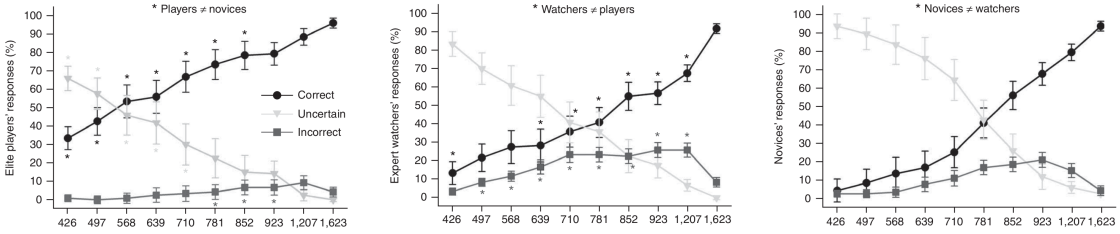


Figure 1.4: Percentage of successful responses to the video outcomes for the three groups of participants. The percentages are given to the 10 predefined video sequence cuts. Reprinted from "Action anticipation and motor resonance in elite basketball players." by Salvatore Aglioti et al., 2008, Nature Neuroscience, page 2.

The Minute Details

The impressive ability of human observation does not end here. The following work examines the aptitude for expert users to detect small variations in the arms. [Aglioti et al., 2008] performs an interesting exercise involving basketball free throws. Participants were asked to correctly guess whether the person in the video was going to successfully make a free throw or not. There are several video samples which include attempts where the person would make and others where it would miss. The participants include current professional basketball players (experts' group), ex-players turned coaches and journalists (watchers' group), and people with no knowledge of the sport (novices' group). The experiments involved video segments where the action would be cut at specific frames as shown in Figure 1.3. The participants were given 3 possible answers at each frame: "make", "miss", and "do not know".

The study revealed that, with no surprise, professional basketball players predicted the outcome of the action earlier and more accurately than the coaches and journalists, and even more than the novices viewers. This makes sense as they have the expertise of watching and playing on a regular basis compared to retired players or novices. The surprising results are that basketball players were capable of correctly guessing, with some precision, the outcome of the action before the ball leaving the person's hand (frame 710 and before). This suggests that the players are examining the body configuration (motion kinematics) to determine the ball's trajectory. As for novices and expert watchers, the predictive abilities mainly relied on the trajectory of the ball. Figure 1.4 shows that basketball players correctly guess up to 60% of the actions at the 710 frame mark (before leaving the hand), where the watchers and novices are below 40%. This is a remarkable ability, and although not accessible to everyone, it shows the tremendous information the human body can reveal if someone is sufficiently trained to look for them. Robots might be far away to detect these subtle differences but are more than ready to examine the human body and recognize the inherent actions.

Findings in Neuroscience

The experiments in [Aglioti et al., 2008] also analysed brain activity by looking at the corticospinal activity of people viewing the free throws video sequences. A transcranial magnetic stimulation (TMS) device was placed on the left primary motor cortex of each participant. The analysis found that the expert's and watchers' group corticospinal excitability is increased when they observe basketball actions but not when observing other sports actions, like kicking a football. This is in line with work in neuroscience where actions that are familiar to us tend to increase the motor-evoked potentials (MEP) higher than non-familiar ones [Cross et al., 2009]. Nonetheless, TMS analysis shows that observing other people performing actions induces activity in the MEPs from the muscles that would be active if those observed actions were performed by ourselves. This neuron activation is part of a bigger picture called the "mirror neuron" system (MNS).

The "mirror neuron" is a neuroscience theory on the neurophysiological mechanisms of action understanding in humans and other primates. There is a curious story surrounding this famous discovery. It is said that the first time these neurons were detected was by chance when studying the prefrontal cortex in macaques, a species of primates commonly used in experiments due to its resemblance to the human central nervous system. The scientists were performing other unrelated experiments with electrodes implanted in the macaque's brain when, during a lunch break, they noticed that some macaque's brain regions were actively flashing. They were confused given that those brain regions are only flashing during active task participation and the monkey was only seated, and ogling at the sandwich the scientist was eating. That was when they first discovered that there are regions of the brain where neurons are active during the execution of an action and active during observation of the same action

executed by someone else ([Rizzolatti and Craighero, 2004, Rizzolatti et al., 2001a]). Figure 1.5 shows an illustration of how the "mirror neurons" response occurs in the macaque's brain. Later experiments have discovered that the same neural response is present in the human brain [Mukamel et al., 2010]. Although the experiments did not use invasive electrodes but instead noninvasive indirect measurements, such as electroencephalography (EEG) or functional magnetic resonance imaging (fMRI), in principle, the results argue that similar "mirror neurons" exist in humans. The "mirror neuron" system delivers a plausible, although still subject to much debate theory that humans, as well as monkeys, recognize through observation the other's actions. The motor cortex is the region in the cerebral cortex responsible for the control and planning of motor movements. The first discovery of mirror neurons was in the F5 region in the premotor cortex. The premotor cortex is responsible for movement preparation and mirror-like neurons were found to be active during observation of goals, such as grasping an object [Gallese et al., 1996]. The primary motor cortex M1 which is responsible for controlling muscles and thereby movements is also active during action observation [Press et al., 2011]. Additionally, if there are lesions in those brain regions a decrease in action

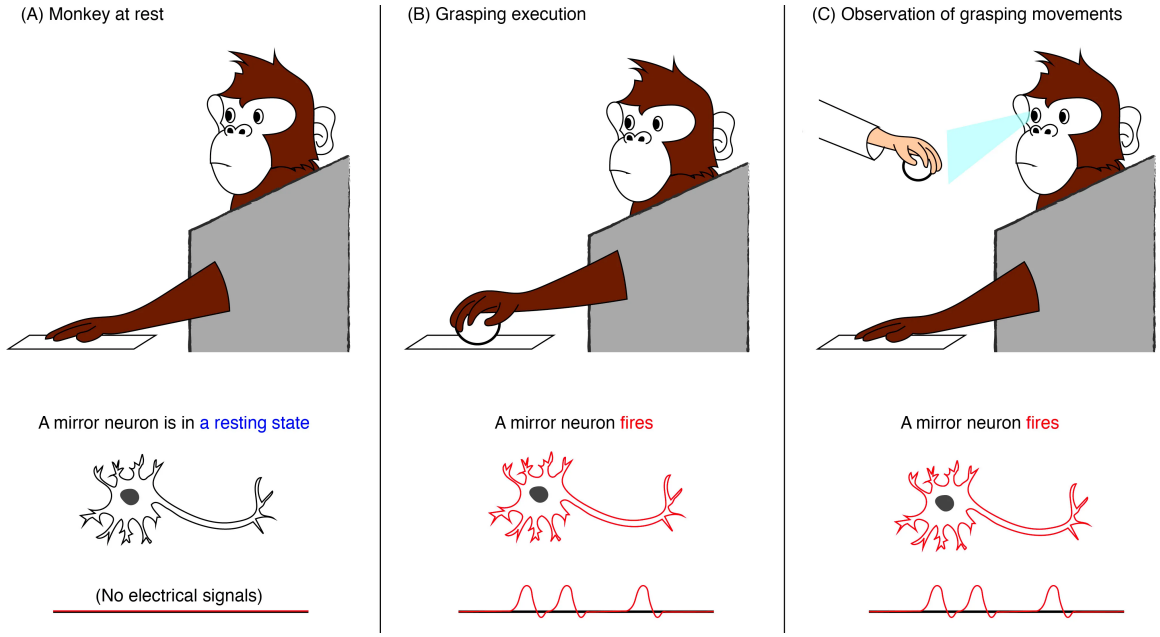


Figure 1.5: A "mirror neuron" fires an electrical pulse, or action potential, when the monkey either observes or executes a specific action. In this case, the "mirror neuron" responds to grasping action. The graph at the bottom shows what the action potentials (each depicted as a hump) would look like when measured with an electrode, as used by the researchers. Image and caption from the Harvard University website, "Mirror Neurons After a Quarter Century: New light, new cracks", by John Taylor and figures by Youngeun Choi, 2016, https://sitn.hms.harvard.edu/flash/2016/ mirror-neurons-quarter-century-new-light-new-cracks/ Accessed 2 Dec. 2021.

recognition is noticed [Keysers et al., 2018]. More, it has been confirmed that mirror neurons are not only responsible for action recognition but also capable of understanding the action intention [Iacoboni et al., 2005]. These neurophysiological studies, and many more, bring light to our inherent capability of decoding human activity from inspection of the visuomotor coordination, i.e. non-verbal communication. Our understanding of the human body posture, orientation, head-eyes fixation, object-oriented goal, etc, are reinforced by the activation of

our brain's inner neurological visuomotor mapping of those same actions [Rizzolatti et al., 2001b].

Neuroscientist António Damásio and his group have been trying to define the necessary requirements for a living being to be considered a conscious being. They have come up with the idea that in order to have a system that simulates the actions of others (a MNS) there must be first a system that simulates our actions before we execute them. This system is what they call *as-if* system. The brain can "simulate" activations in the somatosensorial regions, such as sensations of touch, temperature, or even body motion, *as-if* it were really occurring [Damasio et al., 1991]. The advantage of having a system that can simulate a state without in fact realizing it is to reduce time processing the information and save energy. For example, when we have to fixate an object in our periphery vision, the occipital lobe (the visual cortex of the brain) is alerted to the imminent movement of the eyes and prepares to smooth the transition of fixation to the object without vision blur. The prove of "mirror neurons" is the validation of the *as-if* system, a network of neurons capable of simulating visuomotor coordination of action that is not actually happening in the self.

It Starts at a Very Young Age

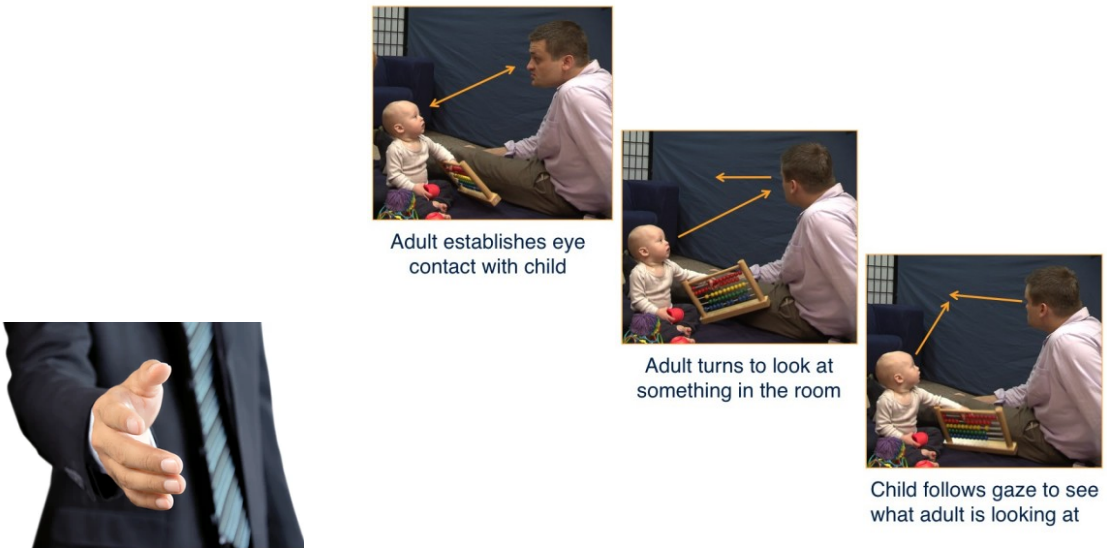


Figure 1.6: A person going for a handshake on the left, and on the right, it is an infant-parent interaction where the infant is following the caretaker's gaze. The left image was taken from the web (https://www.pinterest.com /pin/398709373261271920/). The part on the right is from the University of Washington, the Institute for Learning & Brain Sciences website, "The Importance of Early Interactions", 2016, https://doi.org/10.6069/trxn-kx52. Accessed 2 Dec. 2021.

Newborn begin their life as a clean slate (figuratively), without any knowledge, of how to walk, talk, move or perform other daily tasks. Although it is debated whether newborns have already innate skills Slater and Kirby [1998]. Infants start building their motor repertoire by observing their parents and imitating. Infants learn a lot by moving Nagai and Rohlfing [2007], watching and repeating what they do. Andrew Meltzoff a developmental psychologist has defined this as the "like me" theory of child development [Meltzoff, 2007]. Stating that

infants are constantly evaluating other people's actions to validate their belief that others are "like me". Andrew Meltzoff's "like me" theory corroborates the same principles of António Damásio's *as-if* theory.

"The eyes are the windows to the soul", and in a neurophysiology sense, it indicates the focus of attention and goal-direction to achieve or execute an action. [Mangold, 2015] estimated that humans receive about 85-90% of the information through their visual system. Figure 1.6 on the right shows the influence that eye gaze (accompanied with head-gaze) has on an infant. Showing that, starting at a very young gaze, we humans use gaze to understand others' focus and intentions. Children can participate in social interactions long before they can speak. At around 10 months children understand that eye-gaze can be used to communicate [Brooks and Meltzoff, 2002]. It is not just human eye-gaze that infants follow, [Meltzoff, 1999] has also found that infants are more likely to imitate an action by a human than if that action was performed by an innate device. Hence, body posture, gestures, mannerisms are all picked up by children and coded to a significant meaning, whether it is an emotion or an action.

Robots can Learn it Too

[Dragan et al., 2013] state that for the intention of others to be understood, they need to make their goal location unambiguous to us. Figure 1.6 on the left shows a hand gesture that is ubiquitous to everyone, and no verbal description is needed to understand the intention. Our body movements translate how we move around in the world, manipulate objects, and collaborate with others. In order to achieve this in Human-Robot Interaction (HRI), robots should, in the same manner as humans, perform understandable movements.

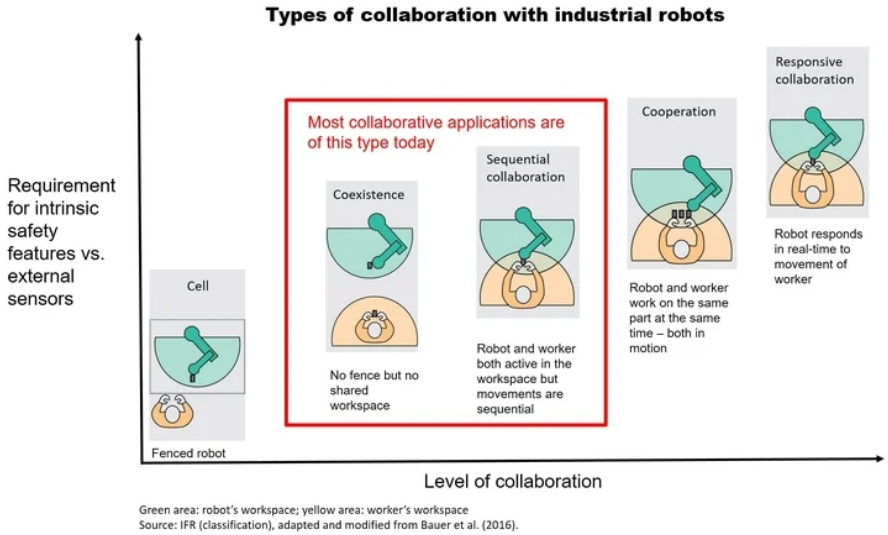


Figure 1.7: The different types of human-robot interactions and highlighted in red are the level of collaboration which are more frequent in todays research. This is a stock image taken from the web (https://ifr.org/news/top-trends-robotics-2020/). Accessed 2 Dec. 2021.

Robotics has been advancing to less restricted environments, i.e. cell-like situations with

little or no human interaction, to a more shared workspace with humans. Unfortunately, most applications developed today have little collaboration between a robot and a human, as seen in Figure 1.7. Industrial manufacturing is going through a major reform toward flexible and intelligent manufacturing (Industry 5.0). Breaking with the established safety procedures as the separation of workspaces between robot and human are removed. Some examples of proposals are by [Robla-Gomez et al., 2017] that introduces a viscoelastic covering on robots to absorb impacts when a collision can not be avoided or make the robot mechanically compliant for safe and intentional human-robot contacts. [Zanchettin et al., 2016] can adapt the robot's velocity by knowing where the human will be to avoid task interruption. All of these approaches are last resort solutions: either minimize damages when accidents occur or reduce the robot's capabilities (torque, velocity) to prevent serious accidents. We believe in solving the problem of shared space by providing robots with human observation techniques to infer intention and action. Few robots, nowadays, are cognizant of the implicit meaning of their actions resulting in random, unclear, signals to humans [Chatila et al., 2018, Avelino et al., 2021]. Most often than not, humans can not interpret robot actions in the purely functional manner that they are intended. I argue that the robotics research community should explore, even further, techniques to efficiently generate and understand implicit communication. It would promote understandable robot actions and avoid major accidents. I believe that robots with human-like motions will make action intentions more salient to humans.

1.2 Thesis Objectives

The importance of studying human non-verbal communication is to adhere these properties to humanoids and robots. We humans, without realizing it, are using a lot of subtle non-verbal cues from others humans to understand the action. The simple act of observing the human locomotion tell us the direction, an estimation of speed (fast, slow, etc), and possibly, from the context, where it is going. When working in a shared space, humans utilize eye gaze and body movements as cues to understand the actions of their workmates. By inferring the actions of others, we can efficiently adapt our movements to appropriately coordinate the interaction. At the same time, we can learn how humans behave and build robots that behave in a human-like manner. It is our belief that robots should possess this non verbalized vocabulary not only to understand humans quickly and effectively as well as to express their goals. This facilitates human understanding of the robot's actions and intentions. Figure 1.8 shows a diagram of our proposal for human-robot communication during HRI. The human, as stated previously, emits their non-verbal cues (eyes and body), and the robot should also possess that same capability. Our objective is to work on developing the communication system (the red region). First, a communication system that takes advantage of the gaze information provided by humans to recognize action intention and, at the same time, generate human-inspired eye-gaze cues that expresses robot's action intention. Secondly, a communication system that uses the

body cues (our focus is on the arm/hand) to recognize the action intention from the motion and, at the same time, generate human-inspired robot motion that expresses robot's action intention. These two communication systems need to process the non-verbal cues to extract useful features for understanding human action and reproduce robot non-verbal cues.

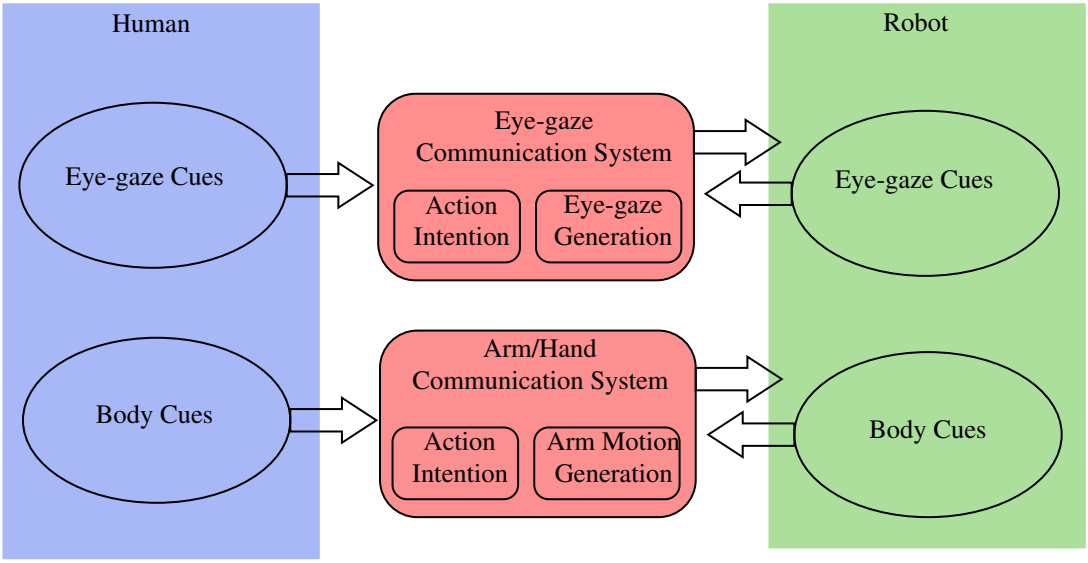


Figure 1.8: Diagram of human-robot non-verbal communication system.

Hence this thesis aims at:

- studying human non-verbal cues during object manipulation and human-human interaction
- developing human-inspired models of non-verbal cues that express human intention when manipulating objects and when interacting with other humans
- implement those models in robot controllers that (i) generate human-inspired non-verbal cues that express robot action intention, and (ii) recognize from human non-verbal cues the desired human action intention

Imitation learning is the most widely used form of learning during development. It reduces effort and time consumption in learning dexterous abilities that would otherwise take much longer to acquire in a trial-and-error learning technique. Imitation is also central to the development of fundamental social skills such as reading faces, body gestures [Meltzoff and Moore, 1977], and understanding the intentions and desires of other people [Over and Gattis, 2010]. Gaze helps us understand if the person is engaged in the action and what is the next step in the interaction. [Patla and Vickers, 2003, Bambach, 2013] have shown that people fixate points on which they will step approximately one second before reaching them, a.k.a. footprint fixation, hence navigational intent can be inferred from gaze patterns and I argue that action intent can also be inferred from gaze patterns.

Robots should conform to human expectations hence we aspire to have robots that read human non-verbal cues and express their own non-verbal cues. If the robot can understand human intention through our eye-gaze and arm-hand cues, in real-time, this allows for faster and reliable reactions. If the robot can generate its own eye-gaze and arm-hand cues, inspired on human non-verbal cues, then humans will have a more natural and fluid understanding of what are the robots goals and intentions. This combined provides a system that in trying to mimic Human-Human Interaction (HHI) progresses towards a future where humans and robots (HRI) do not need any external communication tool (a screen or remote controller) to work together seamlessly.

1.3 Contributions

The main contributions of this thesis are: (i) multiple publicly available datasets of HHI with synchronized videos, gaze and body motion data; (ii) proof that robot legibility can be a robot with predictable non-verbal cues; (iii) develop a non-verbal communication model (the *Gaze Dialogue*) which gives the robot the ability to check! infer the human action from gaze cues, adjust its gaze fixation according to the human, and signal non-verbal cues that correlate with the robot's own action intentions; and (iv) extract unknown object properties from human non-verbal cues during manipulation.

1.4 Thesis Outline

This thesis is organized in three different parts. Part I explores the imitation capabilities of robots to express human-like actions from non-verbal cues. Part II explores how humans communicate and how robots can use the same cues to communicate its goals and intentions to humans. Part III explores further the communication by exploring the intricate details in non-verbal cues that reveal object latent properties during manipulation.

In Part I, Chapter 3 addresses the human *gaze cues* during interpersonal interactions and the legibility of robot actions when using human-like *gaze cues*. Chapter 4 is on understanding polishing motions from human *motion cues* and robots reproducing the *motion cues* for accurate and legible polishing strategies. In Part II, Chapter 5 presents the *The Gaze Dialogue Model* encoding the interplay of the eye *gaze cues* during the dyadic interaction between a human and a robot. Chapter 6 proposes a motor resonance model that couples the robot and human *motion cues* during handover actions. In Part III, Chapter 7 is on inferring levels of liquid inside a cup from eye *gaze cues* of human-human and human-to-robot handovers. Chapter 8 studies the human *motion cues* during manipulation of cups empty or full with liquid and how robots can take advantage of such information to adapt its interaction when manipulating the cup.

The thesis concludes with an overall conclusion for each of the previous chapters, detailed answers to the Research Questions presented in chapter 2, and outlining limitations and future research directions. The final chapters A:D are appendixes which include information on the publicly available datasets and additional results from chapter D.

1.5 Publications

references, 02.biblio

Journals:

1. ferreiraduarte$_{a}ction_{2}018ferreiraduarte_{2}020_{b}enchmark$

2. ferreiraduarte$_{g}aze_{2}022ferreiraduarte_{r}ole_{2}022$

Conferences:

3. ferreiraduarte$_{d}ataset_{2}018ferreiraduarte_{c}oupling_{2}019$

2. ferreiraduarte$_{h}uman_{2}020ferreiraduarte_{2}021_{l}earning$

3. ferreiraduarte$_{r}obot_{2}022$

Workshops:

1. ferreiraduarte$_{a}ction_{2}019ferreiraduarte_{2}019_{b}ehave$

2

Literature Review

> *2. Nothing in life is to be feared, it is only to be understood. Now is the time to understand more, so that we may fear less.*
>
> Marie Curie,

Contents

This section is reserved to present the current state of the art in the field of robotics particularly centered in the human-robot collaboration approaches which take advantage of human non-verbal information. Figure 2.1 illustrates the structure of the section. It begins with motivating further the thesis by discussing relevant works in the fields of neuroscience and human psychology. It proceeds with the works centered on the analysis of non-verbal gaze and body cues with a particular focus on action and motion understanding. It finishes with the inclusion of current proposals on human non-verbal cues integration in robot experiments.

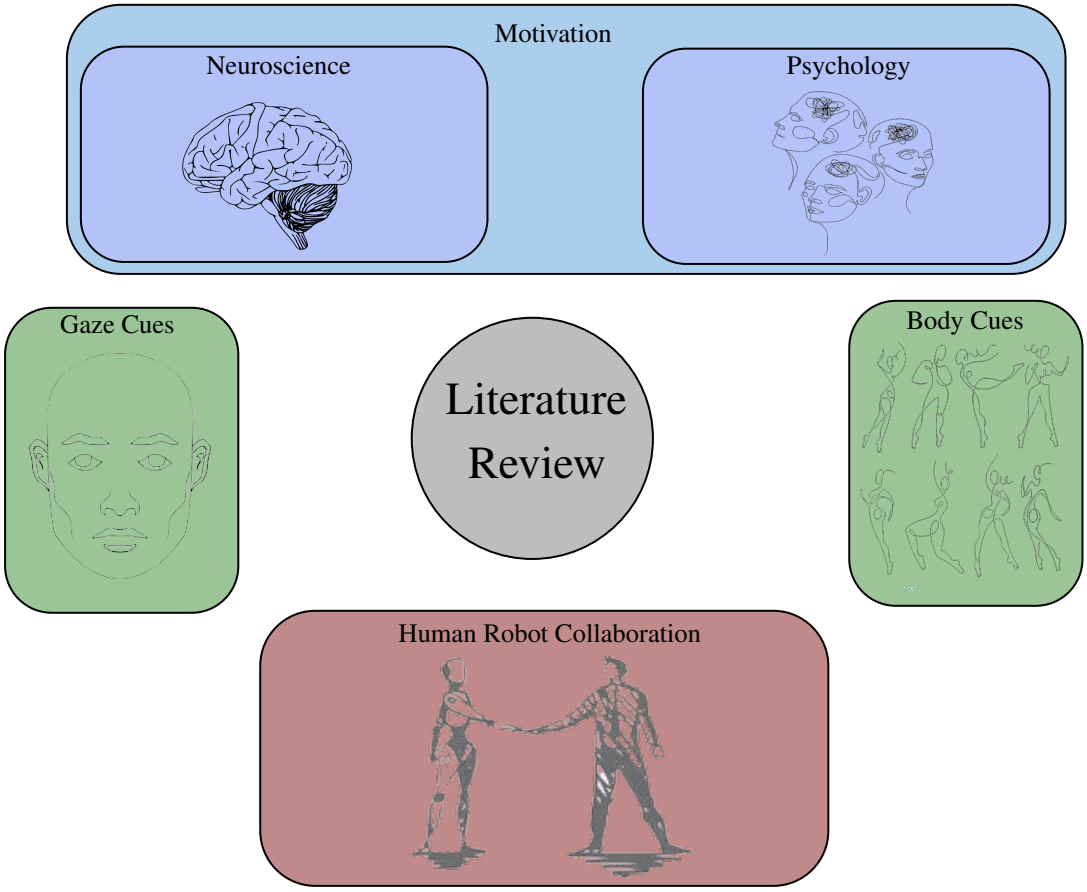


Figure 2.1: Diagram of the most important topics present in this chapter.

2.1 Neuroscience Background

Neuroscientists [Rizzolatti et al., 2001a] discuss that "mirror neurons" serve as a mapping function, both in primates and in humans, to explore the implications for understanding and imitating actions. First of all, it has been shown that in macaques mirror neurons only respond to actions that have physical results on objects, e.g. such as grasping an object. However, it has been observed that homologous human brain regions fire up by intransitive (meaningless) movements. This means that humans have a greater propensity for copying actions details, i.e. imitating, as for chimpanzees, they have a greater propensity for copying action outcomes, i.e.

emulating. Furthermore, when monitoring their own behaviour, humans have a bias toward monitoring kinematics, chimpanzees, on the other hand, have a bias toward monitoring goals ([Hecht et al., 2013]). The MNS is also important in anticipating future actions that are not yet fully visible ([Kokal et al., 2009, Welsh et al., 2005]) either other's or own actions ([Vernon et al., 2015]). [Kourtis et al., 2013] found evidence that people engaged in joint tasks represent in advance each other's actions to facilitate coordination.

Interestingly, mirror neurons are also active when the meaning of action can be inferred from sounds or other hints ([Wykowska et al., 2016]). [Olivier et al., 2007] have shown that the same cortical network (responsible for grasping) may contribute to language and number processing, supporting the existence of tight interactions between processes involved in cognition and action. [Schuch and Tipper, 2007] found that the simulation of another person's behavior goes beyond pure visuomotor processes. Observing another person successfully inhibit action or make an error evokes processes similar to those that occur when the action is completed. An example can be given when watching someone slip on ice. It provides a powerful error signal, enabling us to avoid the same hazard, even if we had not been explicitly monitoring this person's behaviour. Moreover, listening to others' actions can also be a triggering factor. A fraction of the monkey mirror neurons, in addition to their visual response, also become active when the monkey listens to an action-related sound (e.g. breaking of a peanut).

It has been suggested by [Iacoboni and Dapretto, 2006] that mirror neurons could result from evolutionary pressure to greatly facilitate action understanding. [Rizzolatti et al., 2001b] hypothesise two options that might explain how the brain implements action understandings: the visual hypothesis and the direct-matching hypothesis. The visual hypothesis states that action understanding is based on a visual analysis of the different elements that form an action, with no motor involvement. The direct-matching hypothesis holds that we understand actions when we map the visual representation of the observed actions onto our motor representation of the same action. [Jellema et al., 2000] support the visual hypothesis by showing that neurons activate with directed gaze to perform an action but do not activate when there is no directed gaze to the action. However, this does not indicate how the "validation" of the meaning of observed actions is achieved. [Umiltà et al., 2001] showed the mirror neurons in the F5 region in macaques verify the direct-matching hypothesis by showing that more than half of the recorded mirror neurons also discharged in the hidden condition. e.g. macaques knew there was food even when they saw only the hand. Monkeys, like humans, can infer the goal of an action, even when the visual information about it is incomplete. From a motor point of view, the gaze/reach association makes sense, e.g. infants realize that you can reach better something you gaze at. False-belief understanding is of particular interest because it requires recognition that others actions are driven not by reality but by beliefs about reality, even when those beliefs are false. [Krupenye et al., 2016] performed trials with great apes which proves the hypothesis, at least in principle, that non-humans also acquire a theory of mind when it

comes to interpreting the other's understanding.

Imitation is primarily directed toward reproducing the outcome of an observed action sequence rather than reproducing the exact action means ([Hilt et al., 2020]). [Erlhagen et al., 2006] perform experiments of grasping-placing sequence to show how the mapping from perception to action may contribute to the inference of the action goal. Consequently, greater uncertainty about others' actions will call for a greater need for trustful predictions and greater sensorimotor recruitment ([Wang et al., 2015, Sciutti et al., 2013]). We predict others' behavior via adopting the intentional stance. The intentional stance refers to the other's mental states such as beliefs, desires and intentions. [Wykowska et al., 2016] found that participants observing the iCub transporting an object, anticipated the action patterns similarly to when they observed a human. The robot evoked automatic 'motor matching' and 'goal reading' mechanisms in the observers. Emulating human-like behavior in artificial agents might lead to social cognitive mechanisms being invoked to the same extent as other human interaction partners would do. [Lopes and Santos-Victor, 2005] came up with a general architecture for action imitation based on visual mapping of the action observed, motor conversion of the action visualized, and visual transformation from the observed point to the robot's point of view. [Breazeal and Scassellati, 2002] proposed designs for building robots that can become sociable beings.

Synchronization, which is a physiological mechanism present in mammals ([Feldman, 2007]), occurs at the neurological level during human dyad interactions ([Hu et al., 2017]), where others refer to it as inter-brain synchronization ([Dumas et al., 2010, Ikegami and Iizuka, 2007]). [Konvalinka and Roepstorff, 2012] review the two-brain methodological approaches. The reviewed studies employ either fMRI, EEG, or NIRS recordings in both intra- and inter-personally, and integrate various conceptual frameworks. Many of the two-brain studies have identified functional similarities between brains in interaction. Two assumptions have been made: (i) that the brains of two interacting members are coupled via their behaviour, or (ii) that there is a brain-to-brain coupling mechanism between interacting partners that cannot be merely explained by the measured behavior of the two members. Whichever the assumption, synchrony is certainly an indispensable trait of social interaction. [Sisbot et al., 2010, Mörtl et al., 2014] show that humans tend to synchronize with interactive partners, either by adapting the speed to the partner, following the same gaze direction ([Lachat et al., 2012]), or manipulating the objects the same way. Moreover, when applied to human-robot relations, humans tended to synchronize better to humanoid robots than to non-humanoid robots [Hasnain et al., 2012, 2013]. This gives an advantage to humanoid robots when aiming for safe and efficient interaction. The advantage that humanoid robots have when it comes to the motor repertoire, is linked to their overall body structure when compared to the human body. Humans are predisposed to understand human movement [Simion et al., 2008]. As such, robots that reproduce movements that humans recognize the intention of, are ideal for sharing common spaces with them. The robot's intention would be understood quickly, preventing any

collision and harm during the interaction.

Inspired by the MNS hypothesis, we argue that for safe and reliable interactions between humans and robots, a robot must possess the ability to map human behaviour. Hence, our objective is to analyse humans during the manipulation of objects, either individually or in collaborative tasks. Mirror neurons might be involved in understanding the actions of others and might, therefore, be crucial in *action-in-interaction* communication ([Fadiga et al., 2005]). More, human manual interactions with objects are part of what sets us apart from the rest of the animal kingdom. Create and use tools, understand/learn from, or imitate each others' object-related actions in ways that other species do not ([Hecht et al., 2013]).

2.2 Psychology Background

According to psychologists, when two humans interact or communicate with each other they do not only use speech to convey the content of a message but also employ a large variety of non-verbal behaviors [Burgoon and Kendon, 1992]. Young infants, even at preverbal stages, take turn switching roles while maintaining bidirectional interaction without the use of explicit codes or declarative "procedures". Among the implicit but important signals of interaction, synchrony and rhythm appear to be fundamental mechanisms in early communication among humans [Andry et al., 2011]. [Nagai and Rohlfing, 2007] studied the *motionese*, that can help infants and robots to detect the meaningful structure of actions. Parents tend to modify their infant-directed actions, e.g., put longer pauses between actions and exaggerate actions, which are assumed to help infants understand the meaning and the structure of the actions. Scenes captured by parents and toddlers have different properties, and that toddlers scenes lead to models that learn more robust visual representations and drastically outperform parent-trained models in many conditions [Bambach et al., 2017].

Neuroimaging, developmental psychology, and social psychology suggest the existence of motor patterns that reflect the intention to act in a social context. [Georgiou et al., 2007] show that kinematic differences in reach-to-grasp actions are due to prior intentions, i.e. why one has decided to grasp the object. Moreover, reach-to-grasp can also be executed in cooperative and competitive setups. The cooperative task requires a slow careful action while the competitive task a fast movement (natural), which it is referred in this thesis as non-careful. However, the difference between cooperative and competitive could be because of the social and non-social conditions arising from the need to coordinate behavior with external timing signals in the social conditions. It is plausible that reaching and grasping an object with a cooperative intent leads to a motor strategy that is different from the motor strategy used to reach towards and grasp an object with a competitive intent. Social context shapes action planning in such a way that, although the to-be-grasped object remains the same, different kinematical patterns are observed [Becchio et al., 2010]. [Gaussier and Pitti, 2017] mention that the grasping trajectory of an object is impacted by the social environments. [Gupta et al., 2009] stated that

humans are capable of recognizing an object and its purpose by watching the grasping of other humans. Hence our brain is perhaps not planning the grasping as a sequence of elementary and independent subtasks. In social tasks, final-goals have also been reported as having an effect on reach-to-grasp kinematics such as giving vs placing an object [Becchio et al., 2008a], cooperative vs competitive actions [Becchio et al., 2008b], and even verbal vs non-verbal communicative intentions [Sartori et al., 2009]. [Iacoboni et al., 2005] showed that when observing people, we are often able to say not only what they are doing but also the why i.e. the prior intention motivating their action. The very same action is executed differently depending on whether it carries a communicative or a purely individual intent [Sartori et al., 2009].

[Imre et al., 2019] argue that altruistic behavior is not necessarily a consequence of deliberate cognitive processing but may emerge through basic sensorimotor processes such as error minimization. Affordances also play a key role by constraining the possible set of actions that an observed actor might be engaged in, enabling a fast and accurate intention inference [Jamone et al., 2018]. Some evidence in experimental psychology has suggested that imagery ability is crucial for the correct understanding of social intention. [Lewkowicz et al., 2013] called motor imagery as an internal representation of a given motor act without overt motor output. [Lewkowicz et al., 2013, Aglioti et al., 2008] showed that human agents can use small changes in gaze position and/or hand kinematics to anticipate above the chance level the end result before seeing the second-half of the sequence.

In collaboration, there is not a priori role distribution, but a spontaneous role distribution depending on the interaction history and mutual "online adaptation" [Jarrassé et al., 2014]. In contrast, cooperation occurs when different roles are ascribed to both people (leader and follower). [Fairhurst et al., 2014] suggest that leading is related not only to the implementation of a task-specific strategy but also to the adjustments of this strategy depending on the nature of the partner. From everyday experience, we know that it is generally easier to interact with someone who adapts to our behavior. Beyond this, achieving a common goal will very much depend on who adapts to whom and to what degree. [Sacheli et al., 2013] concluded that people when grasping bottle-shaped objects make their movements more "communicative" even when not explicitly instructed to do so. Concluding that being the leader of an interaction implies the (intentional) recruitment of communicative behaviors to convey essential information to others; acting as a follower implies adaptation to a partner not only on the basis of good predictive abilities but also depending on the ability to inhibit automatic resonance in order to focus on the partner's (and on the joint) goal. This supports the notion that joint actions imply a form of communication during which smooth coordination is achieved only when partners send motor signals effectively and are prompt to interpret them. It is the visual interaction and not the verbal that helps achieve unintentional coordination [Richardson et al., 2005].

Metacognition concerns the processes by which we monitor and control our own cognitive processes. Implicit metacognition enables us to adopt a "we-mode", through which we

automatically take account of the knowledge and intentions of others [Frith, 2012]. Explicit metacognition enables us to discuss with others the reasons for our actions and perceptions and overcome our lack of direct access to the underlying cognitive processes. [Frith, 2012] suggest that explicit metacognition is a uniquely human ability that has evolved through its enhancement of collaborative decision-making and this thesis's goal is to progress towards including it in robots as well.

Understanding nonverbal communication is crucial for building adaptive and interactive robots. We need to take seriously the role of turn-taking and role switching from imitating to being imitated if we intend to avoid the confusion between social embeddedness and *action-in-interaction* exchange and sharing. Imitating corrects the discrepancy between movement seen and movement done [Nadel et al., 2004]. From this can emerge both the capacity to learn new actions and the capacity to link one's perception to the others' action rather than to one's own action.

2.3 Non-verbal Gaze Cues

In non-humans, such as apes, the colour of the sclera is rather similar to that of the skin around the eyes. This might perhaps have evolved to deceive predators or even fellow primates who might compete for scarce resources. We humans may have evolved eyes with greater contrast between iris and sclera precisely because the risk of predators is minimal, and the benefits of an enhanced gaze signal in terms of communication [Hedge et al., 1978] and cooperation [Amati and Brennan, 2018] far outweigh the cost of an inability to deceive.

Saccadic eye movements reflect not only the action, but in addition, appear to have the purpose of obtaining very specific information [Hayhoe and Ballard, 2005]. The authors give an interesting example found in cricket players that fixate the bounce point of the ball just ahead of its impact, as the location and time of the bounce provide information for the desired contact point with the bat. Eyes movements are pro-active, i.e. saccades are often made to a location in a scene in advance of expected behavior [Johansson et al., 2001]. In other words, gaze precedes movements. When watching an actor manipulate objects, observers naturally direct their gaze to the object as the hand approaches and typically maintain gaze on the object until the hand departs [Flanagan et al., 2013, Parks et al., 2015]. Roughly a third of all fixations on objects could be identified as one of four: (i) locating objects used later in the action, (ii) directing the hand or object in hand to a new location, (iii) guiding the approach of one object to another, (iv) checking the state of some variable.

The structure of the eyes is such that it provides us with a particularly powerful non-verbal cue to the direction of another person's gaze [Langton et al., 2000]. This belief is expressed in the eye-mind hypothesis, stated by [Just and Carpenter, 1976], which posits that eye gaze is tightly linked to attention and cognitive processes. However, understanding where someone is directing their attention involves more than simply analyzing their gaze direction, head

orientation, and pointing gestures. It should be the ability to make sense of another individual's actions and crucially, to predict what they are about to do next. The "Theory of Mind" in humans [Humphrey, 1984] says the we humans are able to do this because we have evolved the ability to read the behaviors of others in terms of mental states such as knowing and believing. Socially relevant cues such as eye-gaze direction and head orientation trigger reflexive shifts of the other's visual attention. Nonetheless, non-biological directional cues such as arrows do not trigger reflexive shifts of attention [Green et al., 2013]. In turn, any information provided by the head can override directional signals from the body. If the face is viewed at a distance, or if the eyes are obscured by a shadow, the system defaults to signalling the direction of attention from the orientation of the head, or if this too is obscured, from the orientation of the body [Langton et al., 2000]. Head orientations operates to influence eye direction at a very early stage in processing [Langton et al., 2004], and the judged direction of gaze from isolated eyes is improved by adding a head as background [Kluttz et al., 2009]. [Castelhano et al., 2007] concluded that during real-world scene perception, observers are sensitive to the other's direction of gaze and use it to help guide their own eye movements. One central finding is that the ability to derive the location to which an observed actor is attending (gaze following) develops earlier than the ability to relate one's own and other's perceptions [Knoblich and Sebanz, 2008]. Gaze following has also been shown in behavioural studies on goats, dogs, and chimpanzees. However, the ability to relate to one's own and others' perceptions seems to be present only in humans emerging from 12 months onwards [Tomasello and Carpenter, 2007]. Infants start acquiring some rudimentary sensorimotor skills by the time they start to learn joint attention. Very young infants (2-3 years old) do not infer the mental state of "seeing" from another gaze direction but 4-6 years old do [Montgomery et al., 1998]. Infant's gaze is influenced by action familiarity [Elsner et al., 2014]. Authors in [Doniec et al., 2006] observed that infants gaze behavior followed the robot's arm movement, and before the robot fully completes the movement it knows which object it is going to grab.

[Nowak et al., 2017] studies whether temporal adaptation usually present in HHI also occurs during human-robot cooperation. [Sciutti et al., 2018, Sciutti and Noceti, 2018] developed social interactions between humans and robots to study human adaptation to robot behavior. One reason for non-adaptability could be the non-natural gaze behavior of the robot, in particular the lack of mutual gaze. Other works by [Kshirsagar et al., 2020] are lacking true human-robot eye contact given that the robot head is a screen with eyes connected to a robot arm and not a humanoid. [Kompatsiari et al., 2018] showed that eye contact is more engaging for humans and that reflects on the gaze following. [Just and Carpenter, 1980, Langton et al., 2000, Ristic et al., 2002] show that people reflexively follow human gaze cues suggesting that artificial agents are treated similarly to human agents.

[Grigore et al., 2013, Zheng et al., 2015] study robot-to-human handovers and found that using head and eye gazes respectively allowed for fewer failures and participants reached sooner for the object. [Anzalone et al., 2015] have used robot-to-human eye-gaze movements

while [Fan et al., 2017] used for human-to-robot eye-gaze movements to study whether the human is engaged. [Huang et al., 2015a] use a head-mounted eye-tracker to predict the ingredients chosen for making a sandwich. In [Khoramshahi et al., 2016] gaze cues significantly improved participants' reaction times to an avatar's movements. Participants in [Kompatsiari et al., 2021] experiments felt more engaged with the robot when it established eye contact. However, [Perugia et al., 2021] stated that gaze toward an object rather than gaze toward a robot is the most meaningful prediction of engagement. Aversion of gaze in a social chat is an indication of a robot's uncanniness and that the more people gaze at the robot in a joint task, the worse they perform. One thing is certain, tasks are done faster with eye-gaze than head-gaze [Palinko et al., 2016]. Head-based gaze estimation is good enough for detecting head turns [Yucel et al., 2013], but it is not good enough for object recognition. [Admoni, 2016] was focused on studying the influence of robot eye-gaze cues on human reaction, and decision making and in [Admoni et al., 2016] studied intentional delays, e.g. slowing down the release of the object, and it lead to increased participants' awareness of the robot's non-verbal gaze cues. [Ivaldi et al., 2017] noticed that if participants had a negative attitude towards the robot, they would look less at the robot's face and more at the location of the interaction, the robot's hand.

Having access to a partner's referential gaze behavior has been shown to be particularly important in achieving collaborative outcomes [Collier et al., 2015], but the process in which people's gaze behaviors unfold over the course of an interaction and become tightly coordinated is not well understood. [Collier et al., 2015] mention that tracking the gaze of a collaborating dyad could be used "in situ" to track their progression through a reference-action sequence. Which is what this thesis aims to do. Findings clearly show how hand, body, and eye gaze position, together with the agent's goals, can suggest an action-in-interaction or an individual action. By synthesizing gaze behaviors appropriately in coordination with the detected gaze of a human interlocutor, the robot could attempt to produce gaze behaviors that follow the same pattern of natural human-like gaze coordination [Collier et al., 2015].

2.4 Non-verbal Kinesics Cues

The human body is our physical interaction with the surrounding world, as such our movements reflect our intentions and actions. The simple act of walking communicates where we intend to go [Knapp et al., 2013]. [Cummings, 2011] provided an extensively detailed analysis of the subjects body posture, mannerisms, and ticks, during group conversations. The author was capable of categorizing people's personality from the conspicuous but subtle non-verbal signals. Humans seem able to effortlessly align their actions, goals and intentions with other humans during social interactions [Newman-Norlund et al., 2007]. Social cognition concerns the process of alignment of individual minds, even in the absence of a shared goal [Gallotti et al., 2017]. For instance, [Issartel et al., 2007] noticed that when people walk

together, they synchronize unconsciously their footsteps by steadily regulating their step size or frequency. We can deduce that immediate unconscious motor coordination can not be avoided when the subjects share visual information.

[Mörtl et al., 2012, Lorenz et al., 2011] found that human dyads synchronize their arm movements in goal-directed action tasks. This somewhat proves that the partner is not a "neutral" stimulus each agent needs to adapt to [Sacheli et al., 2012]. [Shen, 2012] studied motor interference and motor coordination and concluded that participants tended to synchronize better with humanoids compared to non-biological entities (e.g. pendulums or moving dots). [Nair et al., 2020] developed a model, inspired by the newborn ability to detect biological from non-biological motions [Simion et al., 2008], and concluded that both the model and human participants can reliably identify whether two actions are the same or not. [Vignolo et al., 2016, 2017] developed a similar model which enabled the robot to detect the presence of humans in its surroundings to provide the appropriate social behavior.

[Dragan et al., 2013] argue that for others to understand the intention of an action, the movement should be legible. The term legibility describes how quickly the trajectory unveils its end goal before the movement is complete. The word legible, traditionally an attribute of written text, refers to the ease of readability of handwriting. Predictability, on the other hand, refers to the quality of matching expectations.

[Butepage et al., 2018] worked on predicting a window of future human motion given a window of past frames from skeleton data. Although the approach is limited to simple movements (e.g. moving the arm from A to B) it predicted better legible motions. Other works [Busch et al., 2017, Pérez-D'Arpino and Shah, 2016, Stulp et al., 2015, Pfeiffer et al., 2016] explore the advantages of legible behaviors, but it is important to note the impact of trying to make all robot motions legible. [Bodden et al., 2016] states that motion synthesis methods have focused on functional objectives and that simple and straight-line motion can be as expressive as state-of-the-art legible motions synthesis.

[Rasch et al., 2018] studied the human arm motion when handing over an object. They experimented with two robots, one humanoid and one non-humanoid, and concluded that humans prefer the humanoid movement because of its biological motion. [Yamane et al., 2013] also, from a human-human database, analysed the human giver motion in order to provide a robot with correct hand poses to receive the object. The authors mention that the robot reacts while the human hand is moving as observed in the human-human condition. Fluent meshing in human-robot collaboration requires the robot to make its intentions clear to the human collaborator. [Kwon et al., 2018] enabled robots to express their incapability by non-verbally communicating what they are trying to accomplish and why they are unable to accomplish it.

In HHI, individuals naturally achieve fluency by anticipating the partner's actions. [Moreau et al., 2016] also found that users tend to place objects in such a way that it facilitates the action of the other user. This predictive ability is largely lacking in collaborative robots, leading to inefficient HRI.

2.5 Human-Robot Collaboration

When a robot is learning it needs to explore its environment and how its environment responds to its actions [Senft et al., 2017, Chandrasekaran and Conrad, 2015]. With the increase of autonomous capabilities of robots, the role of humans in the interaction is not reduced, on the contrary, human gains more high-level responsibilities. Therefore, it is important to consider the human in the control loop as a decision-making dynamical system [Musić and Hirche, 2017]. [Baraglia et al., 2016] came up with the conclusion that people collaborate best with a proactive robot, but still prefer having control of when the robot should help. [Huang et al., 2015b] developed two types of robot behaviors: (i) proactive, i.e. lead the action, and (ii) reactive, i.e. wait for the human to make a move. The authors concluded that the proactive behavior led to the greatest levels of team performance, but with the poorest user experience, while the reactive robot led to the poorest performance and greatest user experience. [Beckerle et al., 2017] surveyed HRI perspectives, current issues and opportunities in the field. They conclude that sensory feedback is a possibility to close human-machine control loops and could rely on models of basic sensory dimensions. Evaluation metrics going beyond questionnaires are scarce and functional assessment protocols that consider real-world task complexity and training progress are required, especially for learning devices. To tackle HRI systematically in design, human-oriented methods and human-in-the-loop experiments are promising topics for future research. This is exactly what I try to accomplish in this thesis.

[Bansal et al., 2019] show preliminary results that robots acting in consideration of the goals and interaction-awareness of others achieve higher efficiency as well as improved safety. However, [Gombolay et al., 2017] state that human participants' awareness of their team's actions decreased as the degree of robot autonomy increased. Researchers must be aware of increased autonomy and reduce situational awareness. [Li et al., 2016] developed a framework where the human needs to take an online corrective action to move the robot arm when there are uncertainties. After the human releases, the robot should continue to follow the path. This is an approach, although limited to only coordination of trajectories (not actions), which allows for robot adaptation to human action.

Motor resonance is based on the finding that executing and perceiving an action relies on the same substrates illustrated by behaviors such as motor coordination and motor interference Gallese [2003]. This has been validated even in neurological studies Chaminade et al. [2008]. Motor coordination is a hallmark of human interactions, and motor interference has been used to demonstrate the validity of the motor resonance framework to understand human perception of humanoid robots. In human-humanoid interaction, both agents are anthropomorphic, humans should rely on motor resonance the same way they do when facing a fellow human. [Issartel, 2009] propose that it should be reproduced in the behavior of humanoid robots the unintentional motor coordination described in human interactions. So that in order to optimize humanoid's social competence, robots and humans should mutually influence each other.

[Tsarouchi et al., 2016] state that sensors such as vision systems, or laser scanners, enable a more natural and direct human-like motion. [Wang et al., 2019] developed a controller which recognizes human handover intention from an electromyography sensor. Although it works well for intended scenarios there are some limitations in a multi-action setup where the handover is not the only expected outcome. The same can be remarked for [Nemlekar et al., 2019, Strabala et al., 2012]. [Chan et al., 2015] prepared a scenario where humans handed over twenty common objects and the experiments were recorded with a Motion Capture System (Mocap) system. The human givers use a different handover orientation depending on whether they are focusing on their own comfort or the receiver's comfort. This suggests that some household objects do not have a strong enough affordance characteristic that would prompt givers to hand them over in any particular orientation unless asked explicitly to consider the object's function. [Hansen et al., 2017] another scenario with ten participants recorded with a Mocap system showed that the handover strongly depends on the interpersonal distance between the giver and receiver, while object mass relates only to handover duration. The mass may change but the shape size and centre of mass were kept constant. [Vogt et al., 2018] develop another approach to seamlessly retrieve and pass objects to and from human users. Yet the system is time-dependent and not generalized to any type of handover since it requires the location of objects, robot and users at all times. [Medina et al., 2016] study the human-to-human handover and measured the load share during handover for a more robust robot release of the object. [Parastegari et al., 2017] focused on predicting the preferred object transfer position for humans, i.e. height of the object with respect to the table.

One finds more research about how to speed up a given technique than the comprehensive characterization of that technique, how it complements other approaches, and how it can be integrated with them [Ingrand and Ghallab, 2017]. The focus is on making it faster, more accurate, than actually understanding if it is useful or not. Acting cannot be reduced to the reactive triggering of sensory-motor commands mapped from planned actions or observed events. There needs to be a significant deliberation between what is planned and the commands achieving it. To collaborate effectively the autonomous system must know the user's goal. As such, most prior works follow a predict-then-act model, which first predicts the user's goal with high confidence, then assists given that goal. [Shervin Javdani et al., 2019] apply their framework to both shared-control teleoperation and human-robot teaming. In their studies, user's tended to prefer a predict-then-act approach for the simpler grasping scenarios, though not significantly so, as users varied greatly in their preferences and desires. [Hoffman, 2019] state that there is no systematic discussion on how to measure fluency and proposed metrics that are useful for social robotic interactions but do not define the importance of the role relationship between human and robot. The coordination between visual search and motion control has not been investigated. [Tseng and Mettler, 2019] proposes an approach to analyze the coordination between visual attention via gaze patterns and motion control. The human experimental data demonstrates that fixation is used primarily to look at the target, and humans

coordinate their motion and gaze to scan new areas.

The notions of safe and compliant hardware are not enough. [Dehais et al., 2011] compared three different robot motions: (i) most legible, safe, comfortable, (ii) most unsafe, and (iii) least legible. When analyzing the eye-gaze motion (ii) and (iii) led to statistically higher mean fixations time than (i). Suggesting that motion (ii) and (iii) were more complex since longer mean fixation duration are generally believed to be an indication of a participants difficulty extracting or interpreting information. Additionally, for motion (ii) participants exhibited the lowest mean of saccades while observing, representing an excessive focusing due to the unsafeness of the trajectory, as for motion (iii) had the highest number of saccades, meaning there was a lot of searching due to being a trajectory difficult to understand. A robot should plan its motion so that it is both safe and efficient. [Hayne et al., 2016] achieves it by avoiding the workspace previously occupied by the human. A similar work is by [Claudia and Shah, 2015], which developed a real-time target prediction of human reaching motion. In this case, it predicts the action of the human and plans a trajectory for the robot to reach its goal without colliding with the human.

[Lecun et al., 2015] expects unsupervised learning to become far more important in the long term. On recognition and prediction of human reaching motions, [Luo and Berenson, 2015] used an unsupervised learning method that outperform supervised methods. Humans and animals learning is largely unsupervised and the structure of the world is discovered by observing, not by being told the name of every object.

2.6 Research Questions

Before a robot can imitate a human, it needs to decode human behavior. Human behavior is predicated on the intention, and the intention can be interpreted from its verbal and non-verbal communication [Mavridis, 2015]. Verbal communication will not be addressed in this thesis since it is slow and unpractical for action intention. Non-verbal communication cues, on the other hand, are one, if not, the source of action understanding for humans. This thesis argues that robots can extract valuable information from non-verbal cues to use in HRI scenarios.

From reviewing the literature it is clear that there are plenty of works on non-verbal cues but none that explore the full potential in HRI. There is a lack of research on human eye-gaze cues for recognizing and reproducing actions. As for the hand-arm cues, although there is extensive research in human-robot collaboration, there is little research on synchronization between humans and robots as well as extracting in real-time latent information from non-verbal cues. [Takayama et al., 2011] claim that a robot showing its intention reassures the humans of their interpretations of robot behavior, thus making the robot more appealing and approachable. This thesis is also focusing on enabling robots to express their own human-like non-verbal cues. I argue that this is a communication tool necessary for robots to express understanding of human action and for humans to understand the robot reaction, i.e. mutual understanding.

As such this thesis attempts to answer the following research questions:

- Research Question 1 (RQ1) - Can robots execute actions and be successfully understood just by imitating human non-verbal cues?
- Research Question 2 (RQ2) - Can humans and robots mutually understand each other during interaction simply through non-verbal cues?
- Research Question 3 (RQ3) - Do human non-verbal cues reveal object properties and can it be detected by robots?
- Research Question 4 (RQ4) - Can robots use human-like eye-gaze and arm-hand cues to express actions, intentions, and motion profiles?

The idea is to aim for a world where robots behave like humans so that interacting with a robot would have no need for training or practicing since it will be identical to another human.

Part I

Imitating Human Actions

3

Using human-like *gaze cues* to imitate in pick-and-place and handover actions

> *One, remember to look up at the stars and not down at your feet. Two, never give up work. Work gives you meaning and purpose and life is empty without it. Three, if you are lucky enough to find love, remember it is there and don't throw it away.*
>
> Stephen Hawking,

Contents

Humans have the fascinating capacity of processing non-verbal visual cues to understand and anticipate the actions of other humans. This "intention reading" ability is underpinned by shared motor-repertoires and action-models, which we use to interpret the intentions of others as if they were our own. By inferring the actions of others, we can efficiently adapt our movements and appropriately coordinate the interaction. According to [Dragan et al., 2013], the intention of others can only be understood if and when the end-goal location becomes unambiguous to us. For that same reason, to improve HRI, robots should perform coordinated movements of all body parts, so that their actions and goals can be "legible" to humans.

3.1 Introduction

We start by defining a scenario of HHI, detailed in Section 3.2, to study non-verbal communication cues between humans, in a quantitative manner. The experiment consists of an actor performing goal-oriented actions in front of three humans sitting at a round table (Figure 3.1). The actor picks up a ball and either (i) places the ball on the table (*placing*) or (ii) gives the ball to one of the subjects (*giving*). The experiment is recorded and the actor's 3D body movements and eye-gaze information during the interaction are tracked.

The recordings were taken during the entire experiment, and used to design a human study. This is with the purpose of analysing three different cues: eye gaze, head orientation, and arm movement towards the goal position (Section 3.2.1). For this study, we prepared a gated experiment, using a set of video segments of increasing temporal duration, of each action performed by the actor. The video fractions are shown to the participants, and they are asked to predict the actor's intended action: *giving* the ball to one of the persons or *placing* the ball at one of three assigned markers on the table (6 possibilities in total).

The 3D body movements and eye-gaze information are used to develop a computational model of the human actions (Section 3.3). The arm movement was modelled with Gaussian Mixture Model (GMM), and Gaussian Mixture Regression (GMR) is used to generate the arm trajectory. The eye gaze behavior depends on the type of action: (i) for the *placing* the eye fixates the initial ball position and then it aims at the goal position (i.e. marker on the table); (ii) for the *giving* action, the eye gaze switches between the face of the human and end-goal position (i.e. the handover location). The computational model is incorporated in a controller for the iCub humanoid robot, with the purpose of investigating whether humans can "read" the robot actions as they read the actions of humans. A second human study using a robot as the actor performing the same set of actions. The video fractions of the robot-actor are then presented to another group of participants, who are asked to anticipate the robot's action intention (Section 3.4).

Our results show that we can model the non-verbal communication cues during HHI and transfer that model to a robot executing *placing* actions or *giving* a ball to a human.

3.2 The First Experimental Setup

The scenario can be seen in Figure 3.1. The actor picks the object from the initial position and executes one of these 6 preselected action-configurations (2 actions and 3 spatial directions):

Figure 3.1: Human-Human Interaction: an experiment involving one actor *giving* and *placing* objects and three subjects reading the intentions of the actor.

- ***placing*** on the table to the actor's **left** (P_L), **middle** (P_M), or **right** (P_R),
- ***giving*** the ball to the person on actor's **left** (G_L), **middle** (G_M), or **right** (G_R).

The OptiTrack Mocap system consists of 12 cameras all around the environment and 3 to 4 markers placed on the glasses and wrist making up rigid bodies for each of the relevant body parts. The Mocap system provides position and orientation data of all relevant body parts (head, torso, right-arm, left-arm) recorded at 120 Hz. The actor movements were recorded with an Mocap suit with 25 markers, placed on the upper torso, arms, and head. The eye gaze was recorded with the binocular Pupil Labs eye tracker [Kassner et al., 2014] at 60 Hz. Pupil-Labs Capture also provides additional information related to pupil detection, the two eye cameras frames, the external (world) camera frames, among other information. In Appendix A there are more details on the experimental setup, sensors, and data collection, as well as descriptions on the questionnaires.

3.2.1 Human Study

The study includes a questionnaire pertaining to the actions performed by an actor. 55 participants (40 male, and 15 female), age 31.9±13 (mean±SD) were presented with videos of *giving* or *placing* actions in the different spatial directions, and were asked to predict the action. The videos were fractioned into four types: eye gaze shift, head gaze shift, and arm movement. This can be understood as a gated experiment in which fractions of video segments are shown to subjects beginning when the actor grabs the object and ending when:

- there is a saccadic eye movement towards the goal - G
- 'G' plus the head rotates to the same goal - G+H
- 'G+H' plus the arm starts moving to the goal - G+H+A
- 'G+H+A' plus the arm finishes the trajectory to the goal - G+H+A+.

Figure 3.2 gives the overall success rate for all the different fractioned videos. The more temporal information is available to subjects, the better the decision is, the higher the success rate and the lower the variance, $F(2,5560)=1396.76$, $p<0.0001$. Gaze alone is responsible for a 50% success rate (3 times chance level of 1/6 = 16.7%).

The analysis is further refined by considering two variations: (i) how well can the subjects predict spatial orientation, irrespective of the *giving* vs *placing* action? and (ii) how can the subjects predict the action (*giving*, or *placing*) irrespective of the orientation (**left**, **middle**, or **right**)?

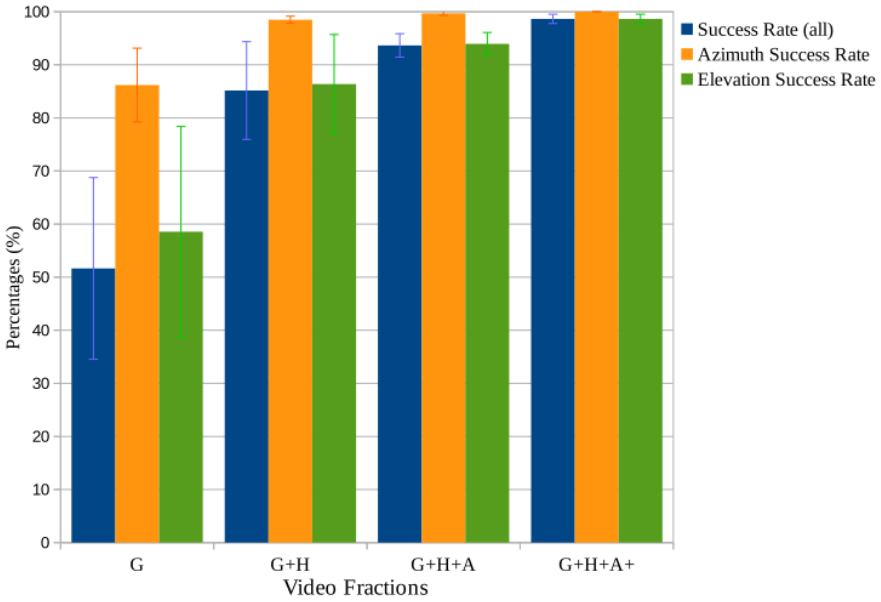


Figure 3.2: The average success of the participants identifying the correct action: overall success rate and success rate in identifying the direction of the action. The error is the standard deviation.

The prediction of spatial orientation does not depend strongly on the amount of temporal information. Subjects were only capable of understanding the action-type 60% (chance level of

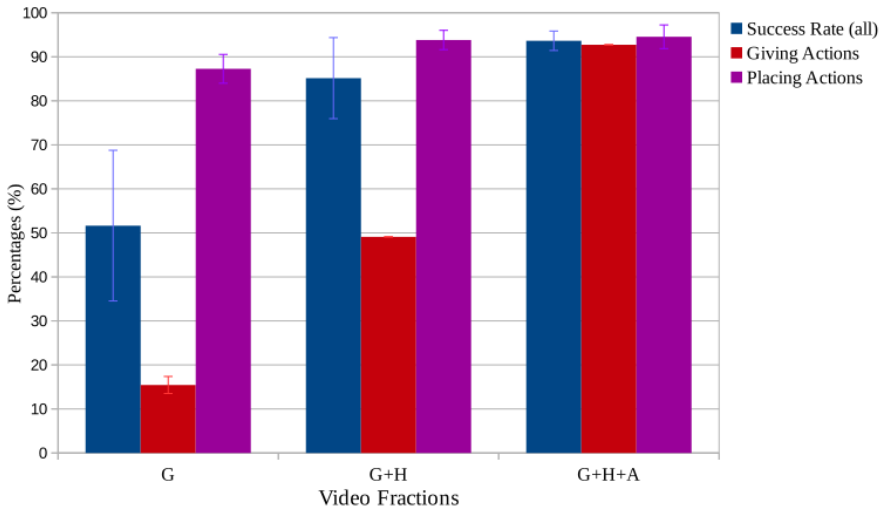


Figure 3.3: The average success of the participants identifying the correct action success rate in identifying the *giving* and *placing* actions. The error is the standard deviation.

50%) of the time for the first video fraction. When analysing Figure 3.3, for the *placing* action we have a success rate of 85% (chance level 50%) with gaze alone. However, we observe that for the *giving* action we get a success rate lower than chance level, F(1,5560)=2306.78, $p<0.0001$, indicating a bias towards *placing* in our HHI scenario. The reason for this can be observed in Figure 3.4 b-e. For *giving* actions there are different gaze trajectories. According to [Moon et al., 2014], humans prefer a *giving* action when the actor performs this switching behavior [Zheng et al., 2015] observed in Figure 3.4 d-e. This switching behavior can be seen as a confirmation routine to acknowledge to the other person that an interaction is taking place. For low information situations, which is the case for 'G' video fractions, the logical choice is to infer that the actor is not trying to communicate with us, which justifies the preference for the *placing* action.

These experiments clearly demonstrate, quantitatively, the importance of gaze in a dyadic action. This analysis shows that human eye-gaze provides key information to read the action correctly, and justifies the need to include human-like, eye-gaze control, in order to improve action-legibility and anticipation as required for efficient human-robot interaction.

3.3 Human Gaze Behavior and Kinesic Movement of Action Execution

3.3.1 Analysis of Gaze Behavior

Figure 3.4 shows five different cases of the spatio-temporal distribution of the fixation point marked with a green circle. Figure 3.4a shows the spatio-temporal distribution of fixation points for the P_M *placing* action in which the green circle is concentrated around the goal position of the red ball.

Figure 3.4b-3.4e show the spatio-temporal distributions of the fixation points during G_M

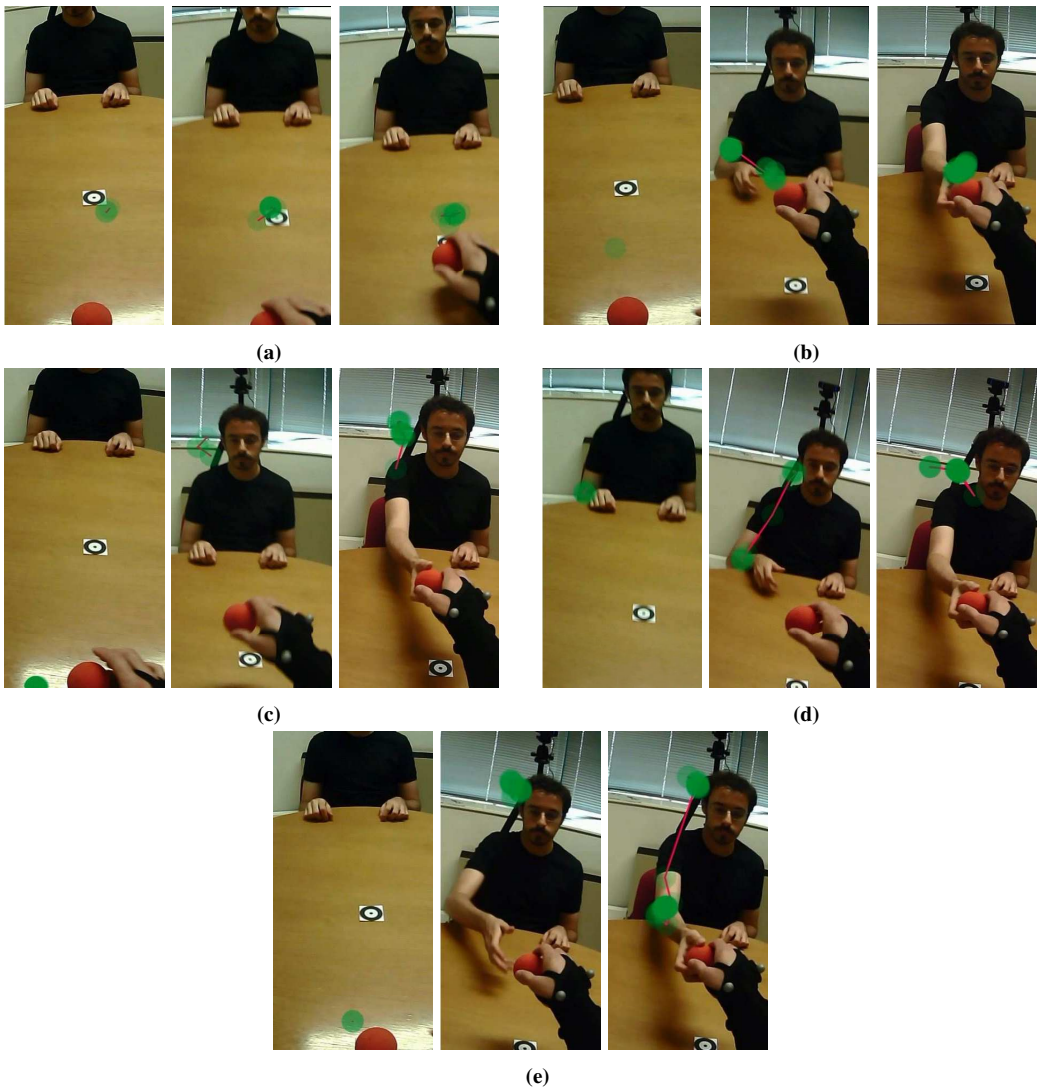


Figure 3.4: Sequence of images of spatiotemporal distribution of fixation point for *placing* and *giving* actions. Subgroup (a) is related to action P_M. The actor only fixates the center marker which is the end-goal point for the action. Subgroups (b)-(e) correspond to action G_M. The actor changes fixation point in 4 different patterns: (b) actor's only fixates the hand of the subject in front; (c) only fixating the subject in front; (d) the actor begins by fixating the subject's hand and it ends by fixating the subject's eyes; (e) the actor fixates the subject's eyes in the beginning and it ends the fixation by looking at the subject's hand.

giving action when the actor was fixating: (i) only the hand of the person, (ii) only the face of the person, (iii) first the hand and then the face, and (iv) first the face and then the hand. From this observed behavior, we designed a controller that will generate an equivalent switching behavior of the fixation point, i.e. a qualitatively similar eye-gaze behavior.

3.3.2 Analysis of Kinesics (Motor) Movement

GMM (inspired by [Calinon et al., 2007]) model the trajectories of the arm movement in a probabilistic framework. The motion is represented as a state variable $\{\xi_j\}_{j=1}^{N} \in \mathbb{R}^3$, where N is the total number of arm trajectories for all actions, and ξ_j are the Cartesian coordinates of the hand for *giving* or *placing* actions.

Figure 3.5 shows an example of the recorded trajectories of the actor's hand during execution of the P_R action. The middle column shows the recorded trajectories encoded in GMM, with covariances matrices represented by ellipses. We use four Gaussian distributions to model the behavior of the arm trajectory for each Cartesian coordinate. This is to take into account the minimum error and the increase of complexity of the problem. Then the signal is reconstructed using GMR. The new parameters, mean and covariance for each Cartesian coordinate, are defined as in [Calinon et al., 2007]. The right column represents the GMR output of the signals in bold and the covariance information as the envelope around the bold line.

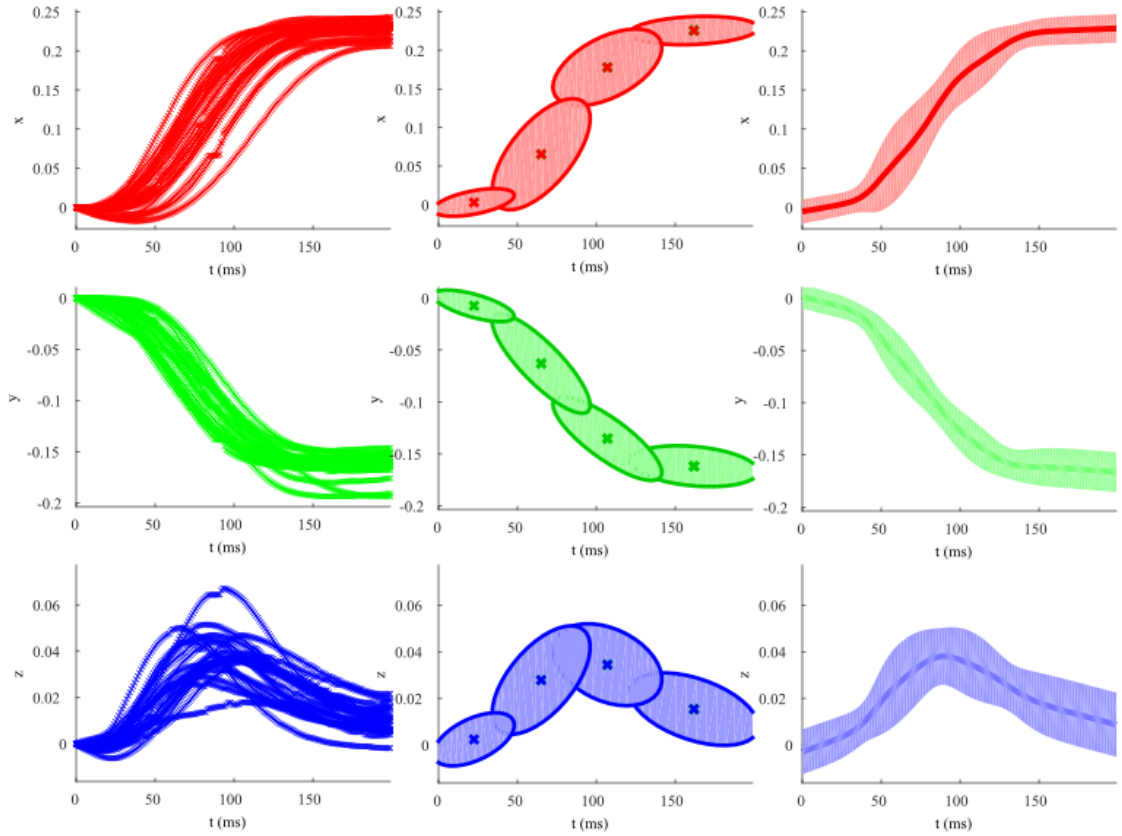


Figure 3.5: Recorded coordinates of human hand performing P_R action, representation of corresponding covariance matrices and output from GMR with covariance information.

Figure 3.6 shows the spatial distribution of the recorded data for all six actions represented by six different colours and the corresponding GMR.

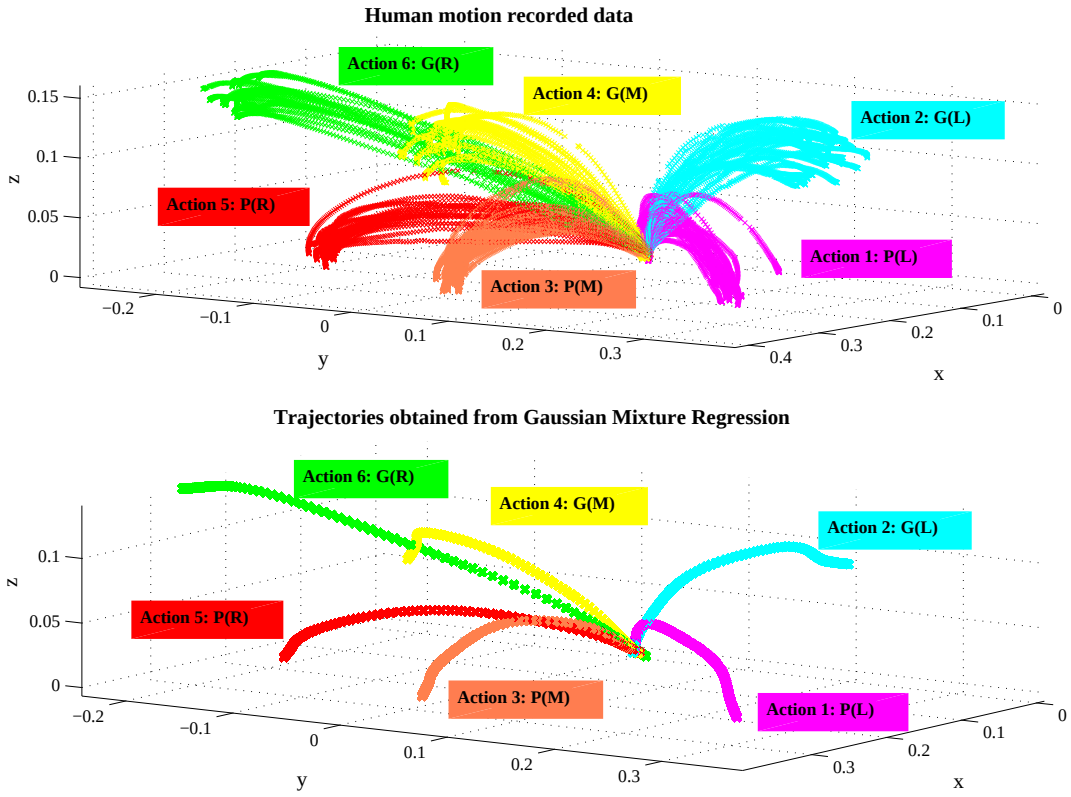


Figure 3.6: Spatial distribution of hand motion for all six actions (top) and corresponding output from GMR (bottom)

3.4 Robot Experiments

The model is embedded in a controller for a humanoid robot. The reference arm trajectory is generated with a GMR and the arm's joints are controlled with a minimum jerk Cartesian controller. The robot eye controller was based on the qualitative analysis of the human gaze behavior and the eye's and neck joints are simultaneously controlled using Cartesian 6-DOF gaze controller [Roncone et al., 2016].

The robot gaze controller was implemented as a state-machine that (qualitatively) replicates the gaze shift behavior observed during HHI. The controller's initial state is the starting location of the ball. Then, depending on the action, for (*placing*) it switches to the final location of the ball, for *giving* it switches between two states: (i) face of the person, (ii) handover location. Figure 3.7 shows the sequence of images, during the execution of the G_R action by the iCub robot and the corresponding images of the actor, when the actor looks first to hand of the other person and then switches to the face.

3.4.1 Human Subjective Analysis

To study the readability of robot's intention, we prepared a second questionnaire with the same set of actions performed by a robot. To assess the relative importance of the different non-verbal (eye, head, arm) cues we have added new conditions: (i) blurring the eyes in

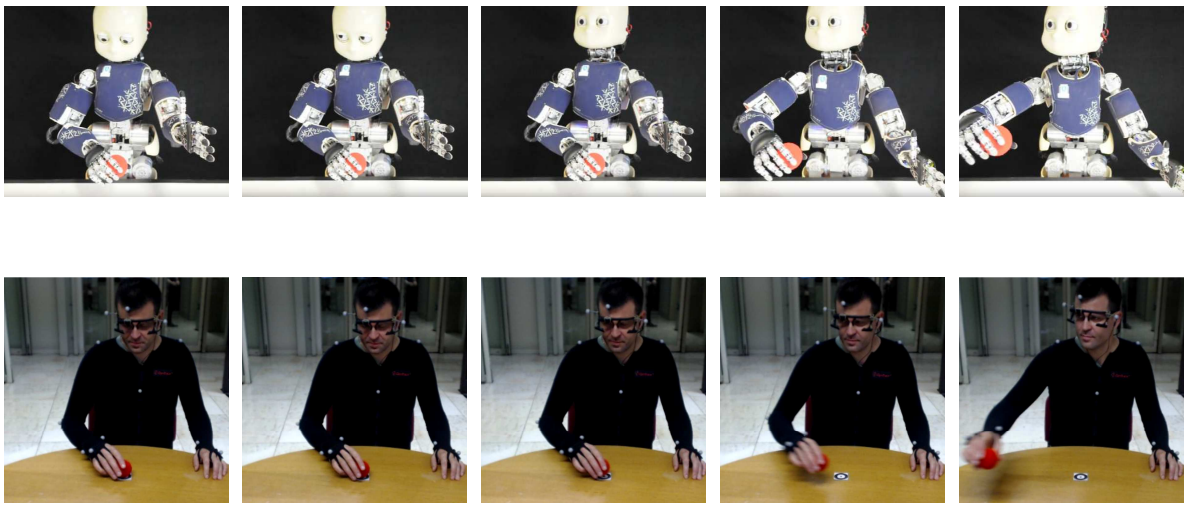


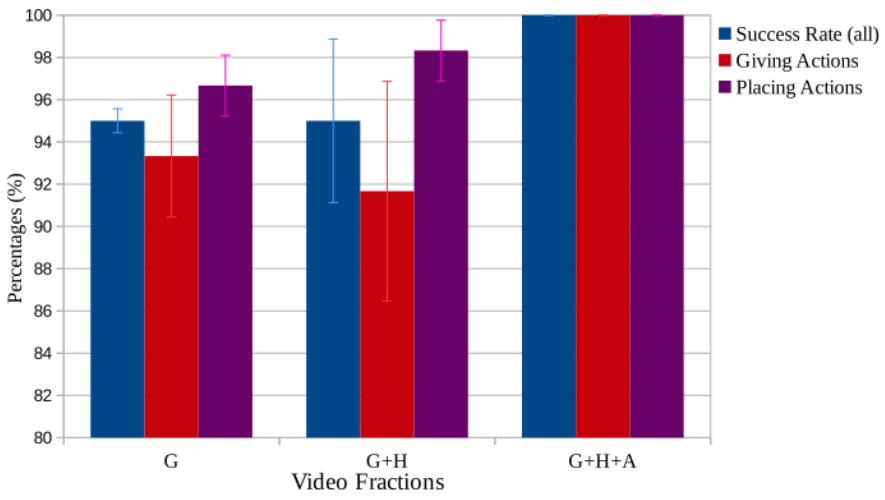
Figure 3.7: The sequence of images of a robot (top) and an actor (bottom) performing the G_R action. The first sequence is the initial point for both the actor and the robot. The second stage corresponds to when the short video stops at the video fraction 'G'. The third is at video fraction 'G+H'. Forth and fifth sequences are for the final two video fractions, corresponding to the arm motion.

the video, and (ii) blurring the entire head. Figure 3.8a shows the participants success rate in identifying the robot-action in the three cases: *giving* action, *placing* action or both. We analyse the effects of blurring on the success rate of *placing* and *giving* actions, Figure 3.8b. We can see that when blurring the eyes, and preserving only the head information ('(Blrd) G+H') the success rate drops around 5%. Since there is a clear distinction between the head orientation in *placing* and *giving* actions, for most people, this is enough information to predict the robot's intention. When blurring the whole head, the only information available is the motion of the arm. Here the participants take longer to understand the action and ambiguity rises. The videos in which only the arm is visible are comparable to experiments performed by [Dragan et al., 2013]. Following the author's terminology, our arm motion is not legible but predictable, and as such, they believe it will not give the most information to the user.

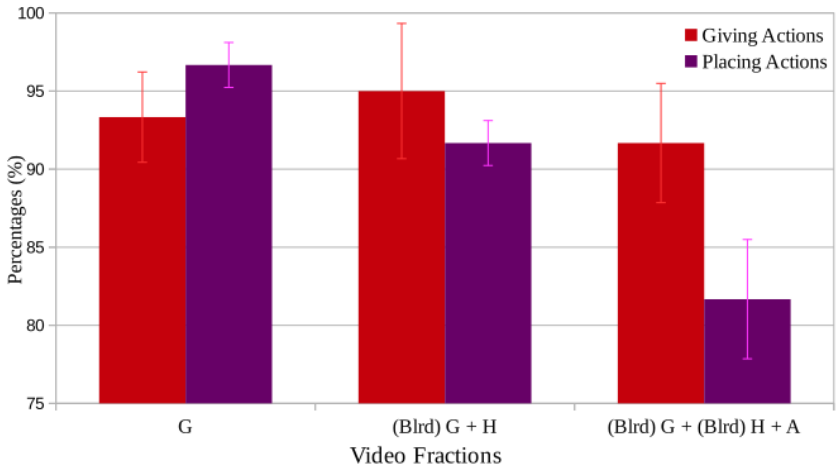
3.5 Final Remarks

[Dragan et al., 2013] proposed two types of arm movements (predictable and legible), and demonstrated that a legible arm movement, which is an overemphasised predictable motion of the human arm, can give more information about the action that the human or the robot is going to do. The experimental scenario involved two end-goals, close to each other. The participants were faster and more accurate to predict the end goal in the case of the overemphasised arm movement. However, there were only very few options in that scenario, and we argue that it would not generalize well if there were more end-goals (for example six as in our case).

We propose an alternative to embed action legibility with overemphasized arm motions, and extend the motion model to incorporate eye gaze information. Our approach improves



(a)



(b)

Figure 3.8: The success of the participants identifying the correct robot action: a) overall success, *giving* actions, and *placing* actions; b) The effects of blurring in the success rate.

legibility, by coordinating human-like eye-gaze behavior with natural arm movements. The resulting robot's behavior showed to be legible even for multiple sets of actions.

We validated these findings with a second human study, where subjects had to read/predict the intentions of a robot. In our experiments, it was much easier to read intentions of a robot than those of a human. We can explain this by looking at Figure3.7, that shows a side by side comparison of the action performed by the human and the robot. In the second pair of images, we see already a clear change in the eyes of the iCub, which is not yet visible in the case of the human actor. This can be due to the high contrast between the white face and black eyes of the iCub. A different perspective on these results will be addressed in the discussion of future work (Section 9.1). A link for the video is provided here to illustrate the different steps taken in this work - video.ACTICIPATE.ral-2018.

The final conclusion taken from the second human study is the importance of the robot's gaze for the overall readability of the coordinated motion. Figure3.8 shows that just by looking

at the arms without any gaze information the success rate drops below 85%. This also results in a slower prediction since the subjects have to wait for the arm of the robot to start moving which is slower than the movement of the eyes. Although 85% is a good result, it is only when we combine eyes and head movement that the results reach an almost perfect score. Our proposal combines the human gaze behavior with the human arm movement to achieve legible behavior to humans.

4

"Reading" human *motion cues* to imitate a polishing action

> *If you talk to a man in a language he understands, that goes to his head. If you talk to him in his language, that goes to his heart.*
>
> Nelson Mandela,

Contents

In this chapter the objective is to understand how humans perform different polishing strategies and develop an optimization approach to generate those polishing motions observed from humans. The generated polishing motion can be useful in HRI scenarios where a human instructs a robot to perform polishing actions, i.e. *follow along*, by demonstrating the type of behavior desired to complete the task.

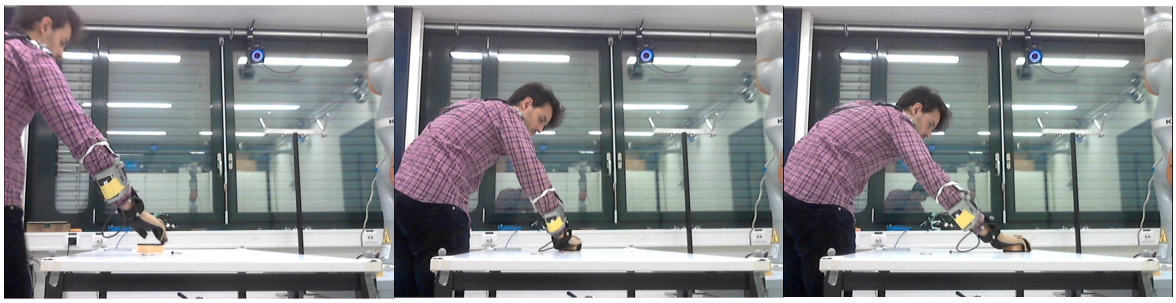


Figure 4.1: A Human performing a task of polishing a table.

4.1 The Second Experimental Setup

In order to acquire realistic polishing movements humans were instructed to grasp a polishing tool, i.e. a sponge, and proceed to polish a table. For each experiment, the human had full range of motion to perform any type of polishing motion. This resulted in very different motions as it is shown in the following section.

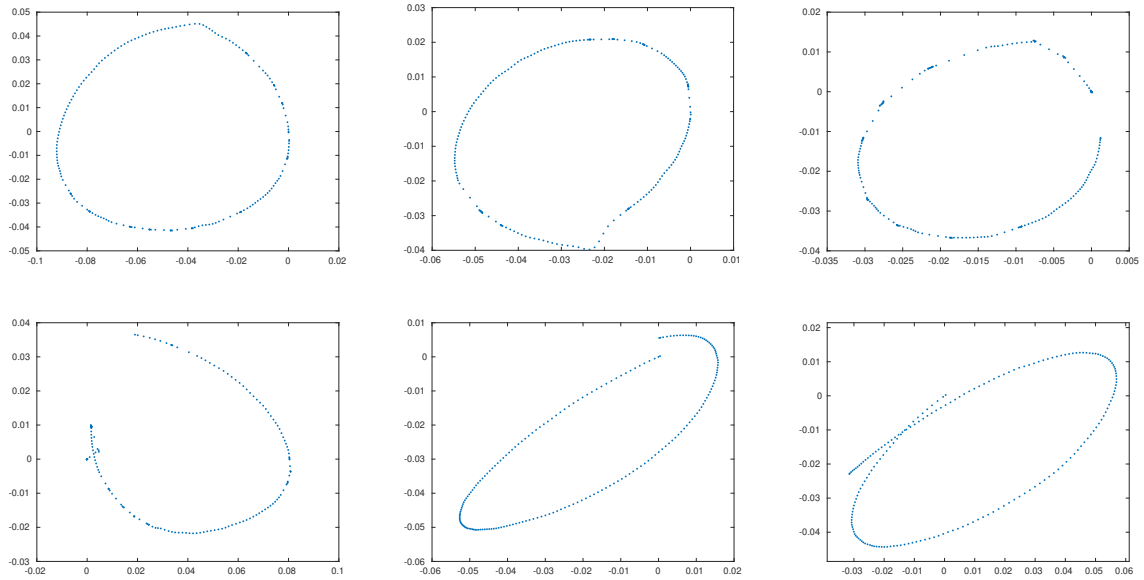


Figure 4.2: Illustration of some polishing motions extracted from the Human experiments.

The human experiments involved each human to wear infra-red markers on the wrist, elbow, and shoulder of the preferred arm for the polishing tasks so as to collect the Cartesian coordinates of the three arm-joints during the experiments. Figure 4.1 shows a standard

example of the experimental setup. The Mocap system was used to detect the three arm-joints markers as three separate rigid bodies and the data was recorded with a frequency of 50 Hz. Each human would perform the task of polishing a table for approximately 1 minute. Post-collection of the whole dataset, the wrist rigid body Cartesian coordinates was picked for the purpose of this work, as this is the joint that most accurately describes the behavior. Figure 4.2 represents some of the extracted polishing movements from the dataset. The z-axis coordinate was disregarded as this was reflecting merely the small variations of height of the wrist with respect to the table. As observed there are a large variety of polishing motions present in the dataset, which requires a representation of polishing motions with enough degrees of freedom to replicate all these different polishing styles. Next, it is introduced the methodology used and later the implementation to polishing motions.

4.2 Methodology

Dynamical System

The Dynamical System (DS) is widely used in this thesis. The materials presented here are adopted from [Khalil, 2002]. The general form of the DS is as follows

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, t, \mathbf{u}) \tag{4.1}$$

where $\mathbf{x} \in \mathbb{R}^d$ denote the state system vector of dimension $d \in \mathbb{N}$, t denotes time, and $\mathbf{u}$ is the control input. The input vector $\mathbf{u}$, in feedback control is defined as a function of the state, $\mathbf{u} = \mathbf{u}(\mathbf{x})$. In this case, the closed-loop dynamics become

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, t, \mathbf{u}(\mathbf{x})) = \mathbf{f'}(t, \mathbf{x}) \tag{4.2}$$

The general dynamics are time-dependent, however, for the purpose of this thesis, the focus is on state-dependent only systems which are formulated as

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}) \tag{4.3}$$

where $\mathbf{f} : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is a continuous and continuously differentiable function. Such DS are referred as *autonomous systems* since the evolution of the state $\dot{\mathbf{x}}$ only depends on the state $\mathbf{x}$. For a specific state x^* if $\mathbf{f}(x^*) = 0$ then state x^* is an equilibrium point of the DS. For an $autonomous$ DS it is necessary to define the equilibrium point as stable so as to prevent undesired behaviors.

The following is the definition of an equilibrium point:

Definition 4.2.1. (Equilibrium Point)
A point $x^* \in \mathbb{R}^d$ such that $\mathbf{f}(t, x^*) \equiv 0 \; \forall t > t_0$ is an equilibrium point of $\mathbf{f}$.

From this definition it is now possible to define a stable equilibrium point by defining stability or, also referred to, as Lyapunov stability.

Definition 4.2.2. (Stability)
An equilibrium point x^* is considered stable if for each $\epsilon > 0$ there exists $\delta = \delta(\epsilon, t_0) > 0$ such that:

$$\| \mathbf{x}(t_0) - x^* \| < \delta \;\Rightarrow\; \| \mathbf{x}(t) - x^* \| < \epsilon, \qquad \forall t > t_0$$

From the definition, an equilibrium point is stable if and only if nearby points remain nearby, or close enough to the point x^*. If nearby points converge to the equilibrium point then it is a asymptotically stable point x^*.

Definition 4.2.3. (Asymptotic Stability)
An equilibrium point x^* is considered asymptotically stable if it is stable and, if there exists $R(t_0) > 0$ such that:

$$\| \mathbf{x}(t_0) - x^* \| < R \;\Rightarrow\; \| \mathbf{x}(t) - x^* \| \to 0, \qquad \forall t \to \infty$$

To guarantee stability of the equilibrium point in an *autonomous* DS the Lyapunov's direct method is applied, which results in finding a scalar function, also known as a Lyapunov function, that follows the stability properties of a nonlinear system.

Theorem 4.2.1. (Stability of equilibrium point in autonomous nonlinear DS)
Let $\mathbf{x} = 0$ be an equilibrium point of a DS on the form $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$. Let $\mathcal{D} \subseteq \mathbb{R}^d$ be a region including the origin. Let $V(\mathbf{x})$ be a continuously differentiable function such that:

1. V is positive and definite in $\mathcal{D}$:
   $V(0) = 0$ and $V(\mathbf{x}) > 0 \; \forall \mathbf{x} \in \mathcal{D} \setminus \mathbf{0}$
2. $\dot{V}$ is negative semidefintive in $\mathcal{D}$:
   $\dot{V}(\mathbf{x}) \leq 0 \; \forall \mathbf{x} \in \mathcal{D} \setminus \mathbf{0}$

   Then $\mathbf{x} = 0$ is stable. If, additionally:
3. $\dot{V}$ is strictly negative definite in $\mathcal{D}$:
   $\dot{V}(\mathbf{x}) < 0 \; \forall \mathbf{x} \in \mathcal{D} \setminus \mathbf{0}$

   Then $\mathbf{x} = 0$ is *locally asymptotically stable*.

To note that Theorem 4.2.1 generalizes to any equilibrium point located in $\mathbb{R}^d$ by a simple change of variables. A DS which follows these properties means that every point nearby the equilibrium point, i.e. inside region $\mathcal{D}$, will converge to the equilibrium point. It was considered the state variable as the Cartesian coordinates of the wrist position $\xi(t)$. From the data it is collected *N* demonstrations of the interaction, yielding $\{{\xi^t}_n, \dot{\xi}^t_n\}$, $\forall t \in [0, T_n]$; $n \in [1, N]$, where ${\xi^t}_n$ and $\dot{\xi}^t_n$ are the state and derivative, respectively, for t time step, and the

n-th demonstration. T_n represents the number of samples in the n-th action. The collected data are a collection of instances of interactions which can be represented as first-order differential equations of the arm motion:

$$\dot{\xi} = f(\xi) + \in \tag{4.4}$$

where $f : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is a continuous and continuously differential function, with a single equilibrium point $\dot{\xi}^* = f(\xi^*) + \in$. $\in$ is a zero mean Gaussian noise which has the properties of dealing with errors and/or human motion variability. To take into account spatial perturbations, it is considered ξ in the reference frame of the target.

The DS is encoded using GMM which defines a joint distribution function

$$\mathcal{P}({\xi^t}_n, {\dot{\xi}^t}_n|\theta) = \sum_{k=1}^{K} \pi^k \mathcal{N}({\xi^t}_n, {\dot{\xi}^t}_n; \mu^k, \Sigma^k) \tag{4.5}$$

over the collected data as mixture of K Gaussian distributions [Khansari-Zadeh and Billard, 2011], where π^k, μ^k, and Σ^k are, respectively, the prior component, mean, and covariance matrix of the kth Gaussian. To compute the DS from Equation (4.4) the posterior mean of $\mathcal{P}({\dot{\xi}^t}_n|{\xi^t}_n)$ is estimated:

$$\hat{\dot{\xi}} = \sum_{n=1}^{K} h^k(\xi)(\Sigma^k_{\dot{\xi}\xi}(\Sigma^k_{\xi\xi})^{-1}(\xi - \mu^k_{\xi}) + \mu^k_{\dot{\xi}}) \tag{4.6}$$

where

$$h^k(\xi) = \frac{\pi^k \mathcal{N}({\xi^t}_n, {\dot{\xi}^t}_n, \mu^k, \Sigma^k)}{\sum_{i=1}^{K} \pi^k \mathcal{N}({\xi^t}_n, {\dot{\xi}^t}_n, \mu^i, \Sigma^i)}$$

$$h^k(\xi) > 0$$

and

$$\sum_{n=1}^{K} h^k(\xi) = 1$$

The GMM parameters can be initially guessed using Expectation Maximization and tuned further to minimize the error between the real (from the data) and generated velocities.

4.3 Formulation of Polishing Motions as Limit Cycles

Following the notions and properties for stable *autonomous* DS presented in Section 4.2 and assuming that the DS is locally asymptotically stable to a limit cycle, it is now possible to formulate the polishing dynamics. A polishing motion on a table can be approximated as limit cycles on a 2D surface. The coordinates are changed from Cartesian $(x, y)^T$ to polar coordinates $(r, \phi)^T$ as it is simpler to represent limit cycles than in Cartesian. Converting $x = r\cos\phi$ and $y = r\sin\phi$ the following DS results in:

$$\dot{\hat{\mathbf{x}}} = \begin{pmatrix} \dot{x} \\ \dot{y} \end{pmatrix} = \hat{\mathbf{f}}(\hat{\mathbf{x}}) = \begin{pmatrix} \dot{r}\cos\phi - r\dot{\phi}\sin\phi \\ \dot{r}\sin\phi + r\dot{\phi}\cos\phi \end{pmatrix} \tag{4.7}$$

as $r = \sqrt{x^2 + y^2}$, $\phi = \arctan(y, x)$, and the polar coordinates derivatives are $\dot{r} = -\alpha(r - r_0)$ and $\dot{\phi} = \omega$, respectively. The parameter $\alpha \in \mathbb{R}$ indicates the radial velocity, $r_0 \in \mathbb{R}$ stands for the radius of the limit cycle, and $\omega \in \mathbb{R}$ represents the angular velocity of the limit cycle. The $autonomous$ DS $\dot{\hat{\mathbf{x}}} = \hat{\mathbf{f}}(\hat{\mathbf{x}})$ approximates all polishing motions to circular limit cycles.

To generate any type of polishing motion the limit cycle is reshaped by applying a transformation matrix:

$$\hat{x} = M(x; \Theta) \tag{4.8}$$

where $\hat{x}$ is the canonical state which is a transformation of the real state x, and $M(, ; \Theta) : \mathbb{R}^m \rightarrow \mathbb{R}^m$ is the transformation matrix with p parameters ($\Theta = (\theta_1, ..., \theta_p)$). To get the generalized $autonomous$ DS for any polishing motion, Equation 4.8 is derived to obtain the expression in Equation 4.7:

$$\ddot{\hat{\mathbf{x}}} = \partial M_x(x; \Theta)\dot{x} + \partial M_\Theta(x; \Theta)\dot{\Theta} \tag{4.9}$$

where ∂ is the partial derivative of M over the state x or the parameters Θ. Given that $\partial M_x(x; \Theta) \in \mathbb{R}^{m \times m}$ is invertible and the second term is neglectable due to slow variation of the parameters, the equation is re-written as:

$$\dot{\mathbf{x}} = [\partial M_x(x; \Theta)]^{-1}\hat{\mathbf{f}}(M(x; \Theta)) \tag{4.10}$$

where $M(x; \Theta)$ is a transformation matrix with a translation, rotation, and scalar components:

$$M(x; \Theta) = HR(x + T) = \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix} \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} (x + \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}) \tag{4.11}$$

where H is the scalar diagonal matrix, with a and b the scaling coefficients in x and y axis, respectively, R is the rotation matrix, with θ the angle rotation, and T the translation vector with x_1 and x_2 respectively the components in the x and y axis. With $\partial M_x(x; \Theta) = HR$ the transformed DS is

$$\mathbf{f}(x, \Theta) = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} a^{-1} & 0 \\ 0 & b^{-1} \end{pmatrix} \hat{\mathbf{f}}(\begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix} \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} (x + \begin{pmatrix} x_1 \\ x_2 \end{pmatrix})) \tag{4.12}$$

With the DS formulation from Equation 4.12 any polishing motion is represented in 2D Cartesian space. For illustration purposes the next part is reserved for particularly interesting examples of limit cycles.

4.3.1 Examples of Limit Cycles

Figure 4.3 illustrates a few examples of the limit cycles that can be generated from changing the Θ parameters. The α, r_0, and ω are set to 10 rad/s, 0.5 meters, and $\frac{\pi}{2}$ rad/s, respectively. The first example is the simple circle limit cycle which is generated when ignoring the Θ parameters $[a, b, \theta, x_1, x_2] = [1, 1, 0, 0, 0]$. The remainder examples change one or several Θ

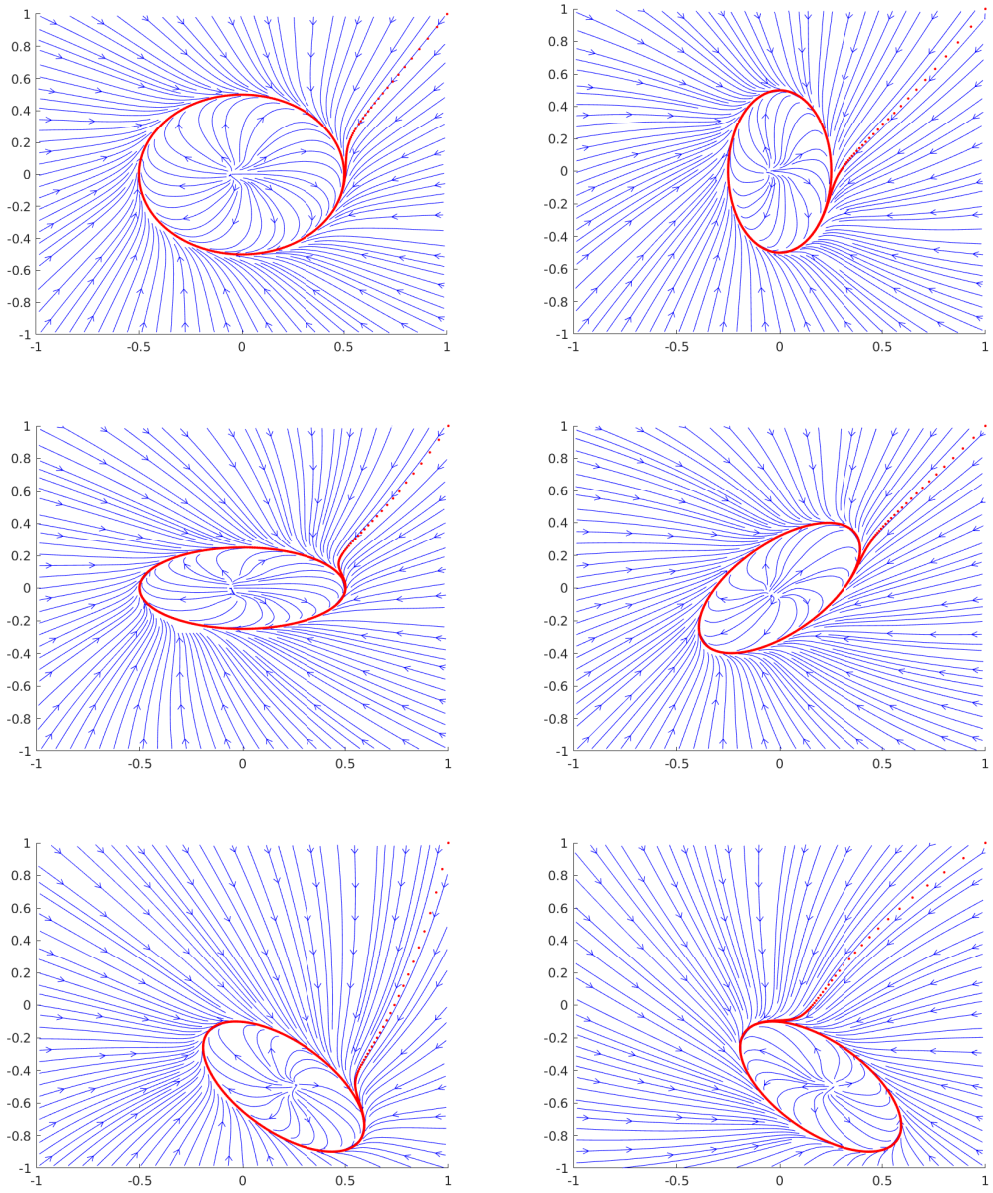


Figure 4.3: The examples on the top row have the respective parameters: (left) $[1, 1, 0, 0, 0]$, (right) $[2, 1, 0, 0, 0]$. The center row have the respective parameters: (left) $[1, 2, 0, 0, 0]$, (right) $[1, 2, 0, 0, 0.8]$. The bottom row have the respective parameters: (left) $[1, 2, 0.2, -0.5, 0.8]$, (right) $[1, 2, 0.2, -0.5, 0.8]$. The example on the bottom row has the ω value inverted.

parameters. For each limit cycle a trajectory is computed starting on the point $[1, 1]^T$ with the purpose to observe how a robot moves according to the dynamics.

4.4 Optimization Problem

The DS is defined to represent any ellipses in a 2D Cartesian space. The problem lies in finding the correct Θ parameters of the DS to match the different trajectories observed from the human dataset. To find the correct Θ parameters it is best to use an optimization solver which tries to find the optimal solution to the defined problem. The solver used is **fmincon** from MATLAB [Grant and Boyd, 2018] which is a gradient-based solver ideal to solve problems with continuous and continuously differentiable objective functions. As shown in the previous section the model needs to be fitted as a continuous and continuously differentiable DS. In order to find the best fit of the DS the main objective is to match the first derivative computed from the human data trajectories with the velocities generated from the optimal DS. Given that the parameters Θ of the described DS from the previous section are the only variables that are possible to optimize it seems feasible to find proper solutions. Before going for the full description of an ellipse with $\Theta = [\alpha, \text{radius}, \omega, a, b, \theta, x_1, x_2]$, a description of the simplest form of an ellipse, commonly known as a circle, just needs α, radius and ω. The definition of the minimization problem is the following:

$$\min J(\mathbf{x}) = \sum |\dot{\mathbf{x}}_r - \dot{\mathbf{x}}_d| \tag{4.13}$$

where

$$\dot{\mathbf{x}}_d = \begin{pmatrix} -\alpha(r - \text{radius})\cos\phi - r(\omega)\sin\phi \\ -\alpha(r - \text{radius})\sin\phi + r(\omega)\cos\phi \end{pmatrix}$$

and r and ϕ are computed from the Equation 4.7. The only thing that is provided to the optimization solver is the 2D (x,y) Cartesian coordinates of the human trajectories. The parameters Θ also have bounding conditions that is known a priori, such as a positive radius, and α is also strictly positive since to reach the limit cycle the radial velocity must be positive. Hence, the complete minimization problem is:

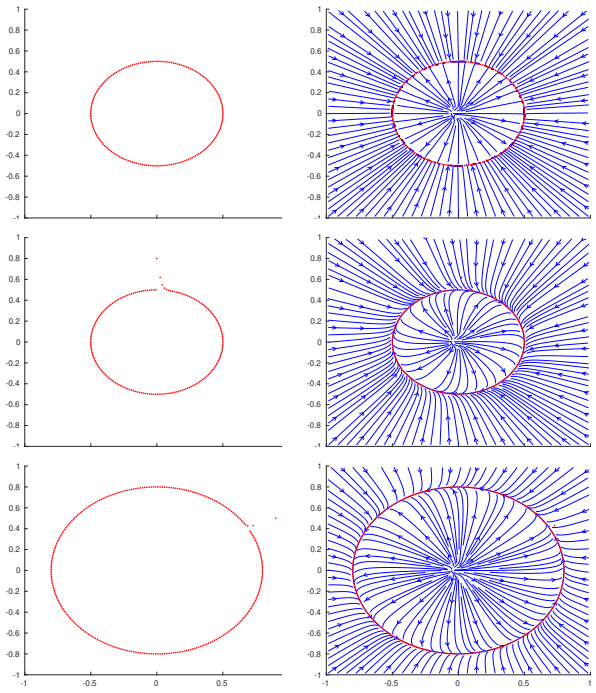
$$\min J(\mathbf{x}) = \sum |\dot{\mathbf{x}}_r - \dot{\mathbf{x}}_d| \tag{4.14}$$

subject to:

$$\alpha \geq 0 \qquad r \geq 0 \qquad \omega \in \mathbb{R}$$

As a proof of concept, the Θ parameters are hand-picked for simulated trajectories of circular motions. These motions are generated from the DS defined in Equation 4.7 and the results for the motions with corresponding optimal parameters are shown in Figure 4.4.

It is important to note that in order to get a usable value for α there are two suitable solutions, either to give some points that are outside the circle (as seen in the middle and bottom examples of Figure 4.4), or restrict the bounding conditions to more realistic values (set to $10 \geq r \geq 100$). As α is the radial velocity, if the data is only encapsulated in the limit



Final Objective	0.0094758
α	573.7266
radius	0.5003888
ω	-1.570796

Final Objective	3.0916e-11
α	30
radius	0.5
ω	-1.5708

Final Objective	2.6898e-11
α	40
radius	0.8
ω	1.0472

Figure 4.4: The top example is a DS with the $\Theta = [30, 0.5, -\pi/2]$, middle has $\Theta = [30, 0.5, -\pi/2]$, and bottom $\Theta = [40, 0.8, \pi/3]$. The second column represents the generated DS and the red trajectory is the simulated circle motion. The third column has a table for the optimal results from the solver.

cycle there is no way to discover the parameter α.

To represent ellipses more parameters need to be added as in Equation 4.12. It was then decided to have $\Theta = [\alpha, \text{radius}, \omega, a, b, \theta]$, the last two parameters (x_1, x_2) were ignored as to focus solely on origin centered ellipses. By adding more parameters to optimize, an increase from 3 to 6 degrees of freedom, the initial objective function did not find the real Θ parameters to build the limit cycles. Minimizing solely the error velocity of the problem didn't fulfill the requirements of the DS. As a result, it was added other constraints to the problem:

$$\min J(\mathbf{x}) = \sum \frac{\|\dot{\mathbf{x}}_r - \dot{\mathbf{x}}_d\|}{\|\dot{\mathbf{x}}_r\|} + \sum \frac{\|r_r - r\|}{\|r_r\|} + \sum \frac{\|\phi_r - \phi\|}{\|\phi_r\|} + \sum \frac{\|\dot{r}_r - \dot{r}\|}{\|\dot{r}_r\|} \tag{4.15}$$

where

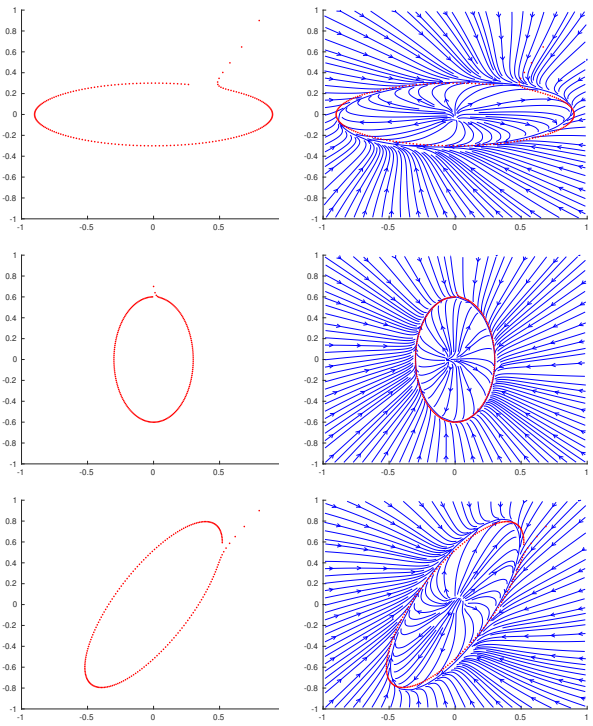
$$\dot{\mathbf{x}}_d = \begin{pmatrix} \cos\theta(a^{-1})(\dot{\hat{\mathbf{x}}}_d^x) - \sin\theta(b^{-1})(\dot{\hat{\mathbf{x}}}_d^y) \\ \sin\theta(a^{-1})(\dot{\hat{\mathbf{x}}}_d^x) + \cos\theta(b^{-1})(\dot{\hat{\mathbf{x}}}_d^y) \end{pmatrix} \qquad \dot{\hat{\mathbf{x}}}_d = \begin{pmatrix} \dot{\hat{\mathbf{x}}}_d^x \\ \dot{\hat{\mathbf{x}}}_d^y \end{pmatrix} = \begin{pmatrix} \dot{r}\cos\phi - r(\dot{\phi})\sin\phi \\ \dot{r}\sin\phi + r(\dot{\phi})\cos\phi \end{pmatrix}$$

$$\dot{r} = -\alpha(r - \text{radius}) \qquad r = \sqrt{\hat{x}^2 + \hat{y}^2} \qquad \hat{x} = a\cos(\theta)x + a\sin(\theta)y$$
$$\dot{\phi} = \omega \qquad \phi = \arctan(\hat{y}, \hat{x}) \qquad \hat{y} = -b\sin(\theta)x + b\cos(\theta)y$$

subject to:

$$\alpha \geq 0 \qquad r \geq 0 \qquad \omega \in \mathbb{R} \qquad a \geq 0 \qquad b \geq 0 \qquad \theta \in \mathbb{R}$$

Given that $r_r = \sqrt{x^2+y^2}$, $\phi_r = \arctan(y,x)$ and $\dot{r}$ is the first derivative of the polar coordinates of the input data computed as $\dot{r}_r = \frac{x\dot{x}+y\dot{y}}{\sqrt{x^2+y^2}}$. A normalization factor was applied to all the minimizer's variables to balance the weight of minimizing each error. The results for ellipse trajectories, as well as the optimal parameters outputted, are shown in Figure 4.5. In Appendix D.1 and D.2 are presented more ellipse motions and the respective limit cycle with the optimal Θ parameters.



Final Objective	1808.0671
α	10.7821
radius	0.890471
ω	-1.18782
a	2.88586
b	1.0007
θ	1.62252

Final Objective	27.3243
α	24.6717
radius	0.602065
ω	-1.38563
a	2.00191
b	1.00445
θ	0.0337262

Final Objective	297.8379
α	13.9532
radius	0.902118
ω	-1.25065
a	2.99054
b	1
θ	5.80337

Figure 4.5: The top example is a DS with the $\Theta = [20, 0.9, -\pi/2, 1, 3, 0]$, middle has $\Theta = [30, 0.6, -\pi/2, 2, 1, 0]$, and bottom $\Theta = [20, 0.9, -\pi/2, 1, 3, \pi/3]$.

The solutions to most ellipse trajectories are extremely accurate and provide usable parameters to recreate the limit cycles. With the exception of some particular trajectories, as seen in Appendix D.1 and D.2, the optimization problem found the correct Θ parameters. All generated simulated data was composed of 200 data points (except when specifically mentioned) with a frequency rate equal to the recorded human dataset (50 Hz), and the initialization of the optimization problem parameters is $\Theta = [1, 1, 0.1, 0.1, 0.1, 1]$ for all trajectories.

4.5 Solver Solution to Human Demonstrations

For the human trajectories of the dataset some preprocessing was necessary. A Smoothing filter was applied to the trajectories so the first derivatives of the Cartesian coordinates became less erratic. A 10-point moving average was applied to the 200 data points of each trajectory. Additionally, all the ellipses were recentered to the origin since our optimization problem does not account for the x_1, x_2 parameters. The limit cycle results of the human dataset examples exhibited in Figure 4.2 are illustrated in Figure 4.6.

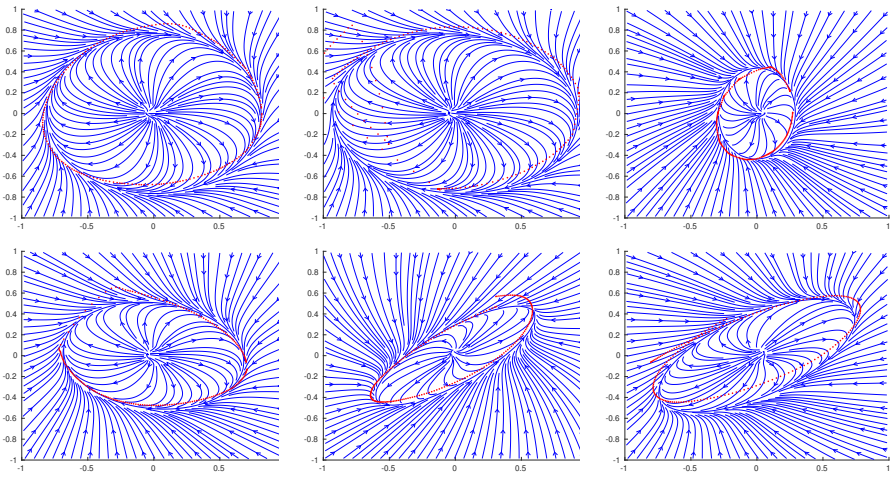


Figure 4.6: Generated DS's from the optimal parameters of the optimization function for different human ellipse trajectories.

The first idea of just minimizing $|\dot{\mathbf{x}}_r - \dot{\mathbf{x}}_d|$ made sense but proved to be insufficient, specially when adding the full Θ parameters. The problem might have to be with being highly unconvex with many local minimas, so the solver would find the first one even though it did not meet the desired Θ parameters. After adding newer constraints in order to respect the polar coordinates in addition to the initial objective, the optimal parameters got closer to the desired ones. In the end, the objective function from Equation 4.15 met all the Θ requirements although with a few exceptions (like the one in Appendix D.1). The overall results for the datasets of the human ellipses trajectories are good enough that can be replicated by a robotic platform in the context of polishing a table.

4.6 Robot Experiments

Figure 4.7 shows some examples of the generated limit cycles in an online human-robot scenario where the human is demonstrating to the Kinova robot the type of polishing motion it wants. It could have also been included the location of the polishing motion inside the table. This is easily computed by including two parameters x_1, x_2 corresponding to the Cartesian coordinates (X and Y axis) into the problem constraints. However, this two new constraints are not really required and it is simpler to extract the location directly from the Mocap system. This avoids extra parameters in the optimization function that would increase computational time which is undesirable for online systems.

The KUKA iiwa 7dof arm in these experiments is controlled in torque mode via the closed-loop DS-based impedance controller ([Kronander and Billard, 2016]) presented in Section 8.3.4. Due to the passivity provided by this control law, the robot can be actively perturbed while executing the commanded velocities from the learned DS. Appendix D.2 illustrates the compliant controller running in real-time while polishing a table. The Kinova gen3 arm is controlled using the kortex_ros[1] package for ROS and velocity commands were used to control

[1]The official repository to interact with the Kinova robot https://github.com/Kinovarobotics/ros_kortex

the end-effector in Cartesian coordinates at 40 Hz for linear (m/s) and angular (rad/s) velocities. For each trial the optimization solver received 800 data points of human polishing. The human hand position is tracked with a Mocap rigid body streaming data to ROS at 120 Hz. The data

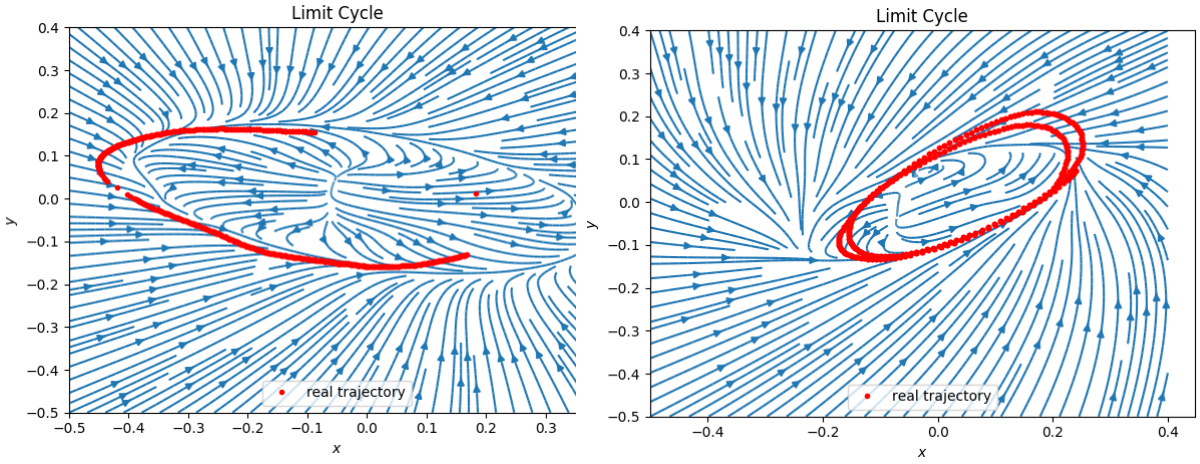


Figure 4.7: Example of generated DS's from the optimal parameters of the optimization function from real-time human demonstrations of polishing motions to the Kinova robot.

is sent to the solver when the human hand begins to move on the table. The data is processed without any filtering and the computed Θ parameters are used to generate the appropriate limit cycle DS. Figure 4.8 illustrates both robots performing the limit cycles DS with optimized Θ parameters by the optimization solver. Figure 4.8 on the first two rows shows the KUKA robot polishing a table after recognizing the polishing strategy from human demonstration. The last rows in Figure 4.8 shows the human demonstrating in real time the polishing motion and after the robot reproducing the same polishing motion in another location of the table. Most of the trials were successful in recognizing the approximate polishing strategy of the human. The main issue that may result in failure to recognize a limit cycle is missing to detect the marker's location due to occlusions of the rigid body. The capability of solving with no filtering applied to the position of the marker proves that the system is robust to noise and loss of data. The main assumption that limits the legibility of this approach is the constant speed in the elliptic trajectory. This approach did not model the velocity of the polishing motion only the trajectory so it does not take into account how fast the human was polishing. The speed during polishing could infer different strategies (e.g. scrubbing a stain or softly cleaning a delicate plate) and it is something that is intended to be explored in the future.

4.7 Final Remarks

This section shows yet another approach for utilizing the human non-verbal cues to recognize actions. In this case the problem focuses on extracting from the human motor movements the arm motion trajectory to detect the polishing pattern and apply it to a robotic platform. The developed system records the human non-verbal cues, computes the appropriate polishing dynamics and commands the robot to follow the recognized polishing motion all while the

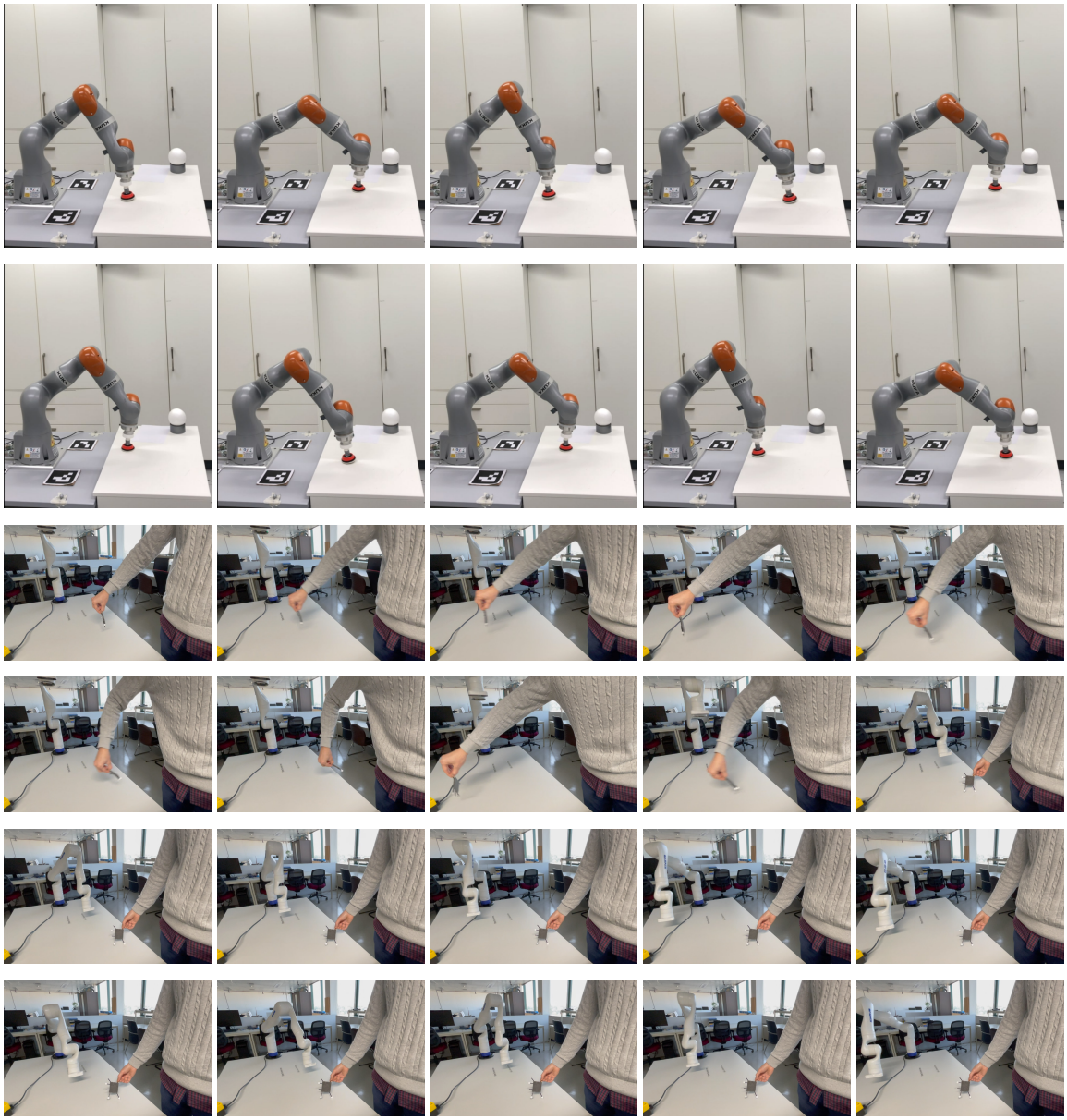


Figure 4.8: Limit cycle DS producing polishing motions on the KUKA robot (top) and Kinova robot (bottom). Human demonstrates a polishing strategy and robot imitates by generating a limit cycle that reflects the polishing motion.

human is polishing. When revisiting the state of the art, most "learning from demonstration techniques" where robots learn to perform tasks by human guidance ([Khoramshahi and Billard, 2019, Figueroa, 2019]) involve a human manually forcing the robot joints or end-effector to specific positions in order to record the desired robot kinematic configurations. This is not only time consuming but also tremendously inefficient since this requires multiple demonstrations to explain multiple tasks. Our approach, on the other hand, is inspired on the monkey-see-monkey-do style of learning. This is a more natural way of a subject explaining to a robot how to perform a task - just imitate what is happening - which requires less effort since there is no manual labour as in previous approaches.

Part II

Understanding Human Intention while Expressing Robot Goals

5

The Gaze Dialogue Model

> *Any fool can know. The point is to understand.*
>
> Albert Einstein,

Contents

Eye movements are particularly important [Sebanz et al., 2006] for joint-action coordination, and humans rely strongly on gaze perception to anticipate the intentions of others [Ricciardelli et al., 2002]. When working on a joint task, the human eye gaze alternates between looking at each other's eyes, seeking the confirmation and engagement of the counterpart, and fixating the goal position before and during reaching actions [Johansson et al., 2001]. Authors in [Sebanz and Knoblich, 2009] report that the ability to gaze at the right location in a timely manner substantially enhances coordination with other individuals. In infant-parent relationship the eye-gaze communication presents itself as a tool to study in depth the infant's development [Yamamoto et al., 2019, Kuboshita et al., 2020, Yamamoto et al., 2020]. Humans can routinely engage in joint actions, and coordinate their movements with others in very sophisticated manners. Such interactions occur in situations as diverse as cooking, cleaning, assembling complex structures, carrying heavy loads, or performing team sports. These tasks involve a collaborative process to coordinate attention, communication, and actions to achieve a common goal [Huang et al., 2015a]. During this process, humans observe the behavior of their partners to anticipate their actions, and to plan their own actions accordingly.

5.1 Introduction

This chapter addresses the *Gaze Dialogue* between two people working on a joint task. This is a collaborative task as seen in Figure 5.1, involving a series of actions where each person is either a leader or a follower. After one action is completed, the roles are changed, i.e. the leader becomes a follower, and vice-versa. The process is repeated until the whole task is finished. Two types of actions are considered: *individual action* and *action-in-interaction*. In individual actions, e.g. a *placing* action, the leader picks up an object and places it on a table. During *action-in-interaction*, e.g. a *giving* action, the leader picks up an object and hands it over to the follower.

Figure 5.1: On the left is a HHI experiment with two humans performing a task of assembling two towers, without any verbal communication. The experiment requires them to be *placing* objects on top of a tower or a *giving* the objects to the other person. On the right is a HRI experiment where a human is performing the same task as before, but interacting with a robot with human-like gaze behavior.

During the experiments, the eye movements of both actors are recorded using an eye-tracking system. The recorded data are labelled and analysed, as detailed in Section 5.2. The

recorded *Gaze Dialogue* data is then used to train a Hidden Markov Model (HMM), as described in Section 5.3. The *Gaze Dialogue* model incorporates the inter-dependency/coordination between the leader's and follower's gaze movements. The *Gaze Dialogue* model serves two key functionalities: (i) predicting the gaze fixations of others and planning one's own fixations; (ii) using the gaze fixations to predict the actions of others and to plan/generate one's own actions.

In order to validate the performance of the *Gaze Dialogue* model the results computed by the model are compared against the dataset acquired in the HHI experiments. The fixations and actions performed by one of the subjects in the HHI are used as input to the model, predicting the fixations and actions of the other human in the interaction. In the next step of the performance validation, the model is implemented in a humanoid robot controller, which drives the robot eye fixations, during the HRI experiments (Section 5.4). The robot controller, inspired on the *Gaze Dialogue* model, takes the human gaze fixations as the input to predict the human next gaze fixations and action performed, while at the same time, generating its own appropriate gaze fixations and planning its own actions.

The results show the robot successfully identifying the actions of the human partner, and acting in a manner that is consistent with the HHI scenario. The behavior of the robot is described quantitatively and it can be visualised in the supplementary material in Section 5.4.3. Finally, Section 5.5 has some conclusions and establish directions for future work. This approach contributes to a better computational modelling of the eye-gaze behavior during HHI scenarios, as well as to endow humanoid robots with similar non-verbal communication skills, thus enhancing HRI and collaboration.

5.2 The Third Experimental Setup

The experiment is a dyadic interaction task for constructing the two towers from a stack of three objects placed next to each participant, Figure 5.2. The description of the task and the stack of objects is occluded from the other participant. In order to complete the task, the actors are required to perform a series of simple actions:

- As a leader, ***placing*** an object on the tower
- As a leader, ***giving*** an object to the person to place on the other tower.
- As a follower, observer the person ***placing*** an object on the tower
- As a follower, receive the ball from the ***giving*** action by the leader

To capture such eye movements, in this experiment both participants were wearing Pupil-Labs binocular gaze trackers. During the performed actions, participants' head gaze, as well as wrist movement, was recorded using Optitrack Mocap system. In Appendix B there are more details on the experimental setup, sensors, and the dataset information.

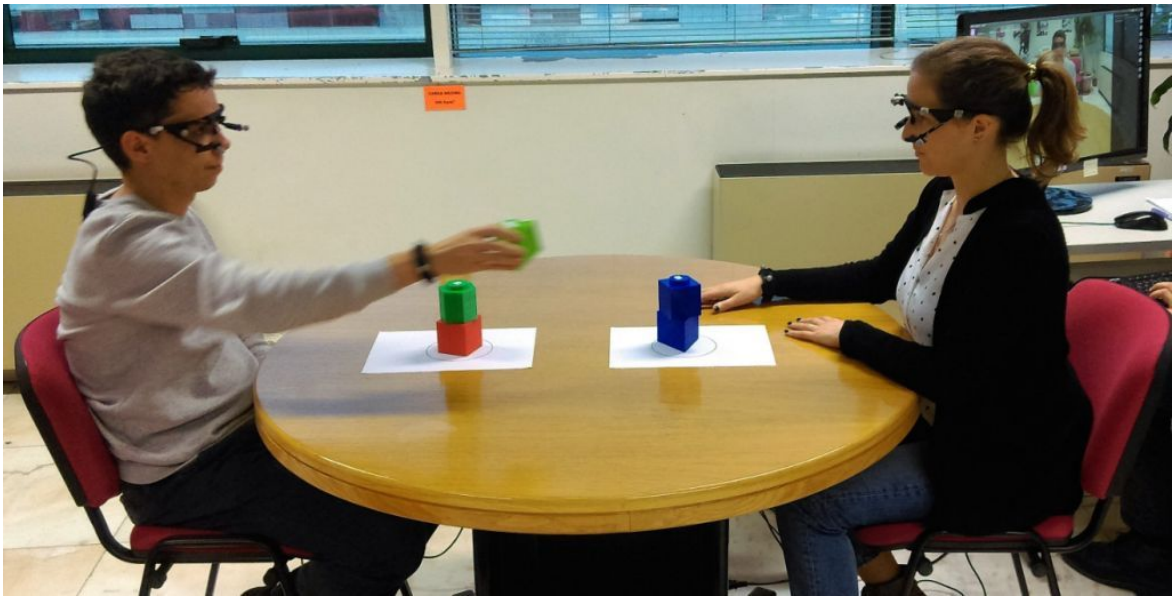


Figure 5.2: Human-human interaction experiment with two humans that are performing a task of assembling two towers, without any verbal communication.

Table 5.1: Examples of leader's gaze behavior for each action with total duration in video frames for each region of interest.

Giving	Labels	B	FH	FT	FH	LOH	FT	FF	FT
	Duration (frames)	143	7	23	7	21	6	38	29
Placing	Labels	B	LOT	FF					
	Duration (frames)	78	31	8					

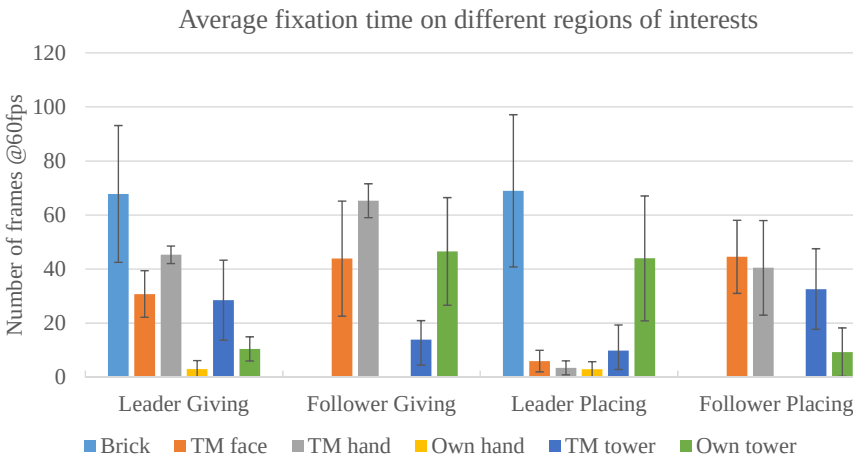
5.2.1 Gaze Behavior in a Collaborative Task

After the data is collected the significant regions of interest, i.e. gaze fixations, that are fixated most often are identified. For a *leader*, the following fixations are considered: brick (B), follower's face (FF), follower's hand (FH), leader's own hand (LOH), follower's tower (FT), leader's own tower (LOT). The fixations defined for a follower are: leader's face (LF), leader's hand (LH), leader's tower (LT), follower's own tower (FOT). Then, these fixations are used to manually label both the leader's and follower's gaze behavior. Table 5.1 shows one example of the leader's gaze fixation labelling process for one *giving* and one *placing* action.

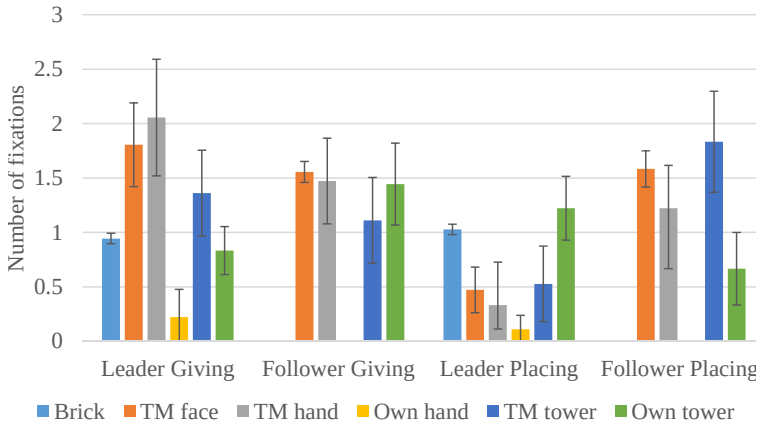
Besides the gaze behavior, the significant events of an action are also annotated: action start, object picked, object handed over[1], object placed, and end of action. The Figure 5.3 shows the plot of an average duration of fixation to different regions of interest across 72 actions for both roles, and an average number of fixations for identified regions of interest.

Conclusions can be drawn from the analysis of Figure 5.3, both from the leader's and the follower's perspectives. For the *giving* action, the leader has multiple gaze fixations, and the gaze fixation time is longer compared to the *placing* action. The leader fixates the brick equally for the two types of actions. In the case of the *placing* action, instead, the leader focuses mainly on his/her tower, whereas for the *giving* action, the leader switches several times between follower's face, hand and tower, and fixates those regions of interest for a significant amount

[1]This label exists only for the giving action.



(a) Number of frames (60 FPS video) for each region of interest.



(b) Number of times the region of interest is selected.

Figure 5.3: Quantitative measure of fixations in frames (a) and frequency (b) for the different regions of interest per action and perspective. TM stands for Teammate.

of time. The follower's regions of interest are different from the leader, i.e. fixating the brick does not exist, and looking at his/her hand after the brick is handed over was negligible. The follower's gaze fixation behaviors are comparable between the *giving* and *placing* action whilst there is a significant difference for the leader's gaze fixations. This is due to the follower's lack of information about the type of the action. As a consequence, the follower spends a significant amount of time consecutively fixating the leader's face and/or hand presumably attempting to "read" the action. The main difference is the number and duration of gaze fixations between the leader's and his/her tower. This occurs when the follower is already aware of the action and fixates the goal, that is, either his/her tower, for visually controlling the *giving* action, or the leader's tower, to monitor the execution of the *placing* action.

Figure 5.4 shows the computed probabilities of the leader's gaze fixations over time, for both *giving* and *placing* actions. These probabilities were calculated and averaged for all actions present in the dataset. The (empirical) probability was estimated by calculating the relative frequency of each gaze fixation over time, after normalising the time-duration of all actions. The plot shows that the leader starts by fixating the brick. For the *giving* action,

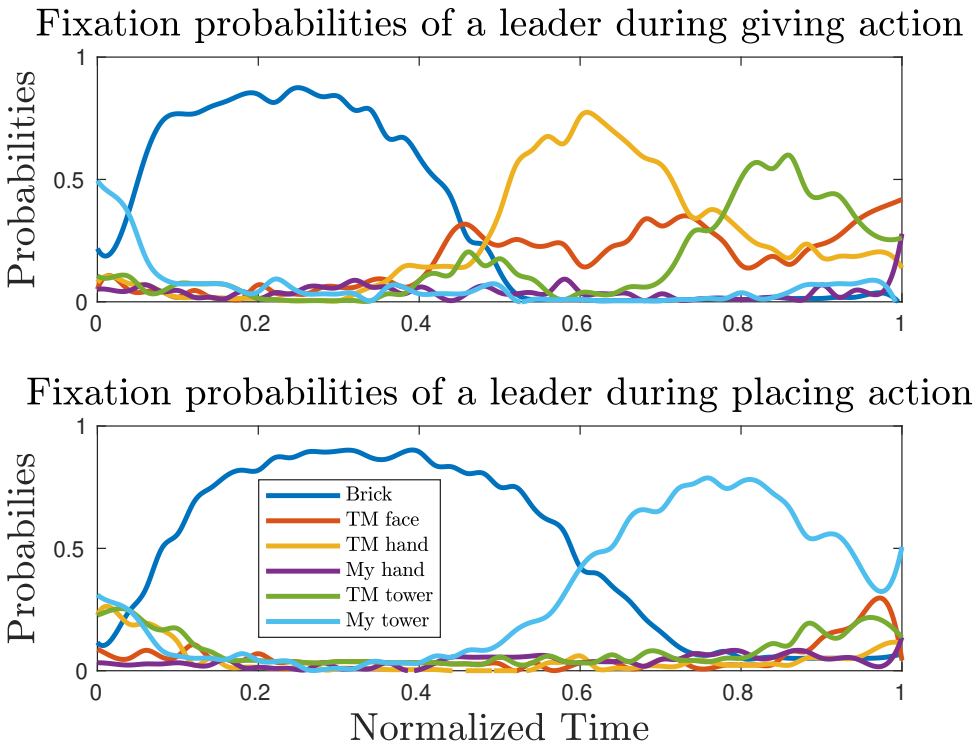


Figure 5.4: Fixation probabilities of Leader's gaze for giving (top figure) and placing (bottom figure) action.

the leader successively fixates the follower's hand, the follower's tower and, finally, the face. In the *placing* action, the leader fixates the brick first, and then his/her own tower, almost until the end of the action. At the very end, the probability to fixate the follower's face and tower increases. Figure 5.5 shows the follower's gaze fixations when observing the leader performing either a placing or a giving actions. The most notable fixations are TM Hand, and My Tower, which are predominant during a giving action. These occur when the goal of the follower is to grasp the object from the leader's hand and to place said object on his/hers tower. At the beginning until about 50% of the total time, the most probable fixation is TM Face, this indicates that the follower is trying to decode the leader's action intention, though it is not solely the face but the leader's hand and tower are also fixated at this time. Since a placing action is an *individual action* a follower is not participating, hence there is not one most probable but several, reflecting a more passive role. Although at the end of the action, the follower tends to fixate on the leader's tower as it becomes clear that a placing action is taking place.

The main conclusions from this analysis are: (i) it is possible to predict the leader's action from the gaze fixations; (ii) there is a clear distinction between the leader and the follower's gaze fixations; (iii) from the leader's perspective, there is a considerable focus on the brick, which is negligible in the follower's case. The reason has to do with the roles of each subject in the experiments: the leader needs to manipulate the brick to complete the action, whereas the follower only needs to follow the leader's behavior; (iv) when comparing the two types of

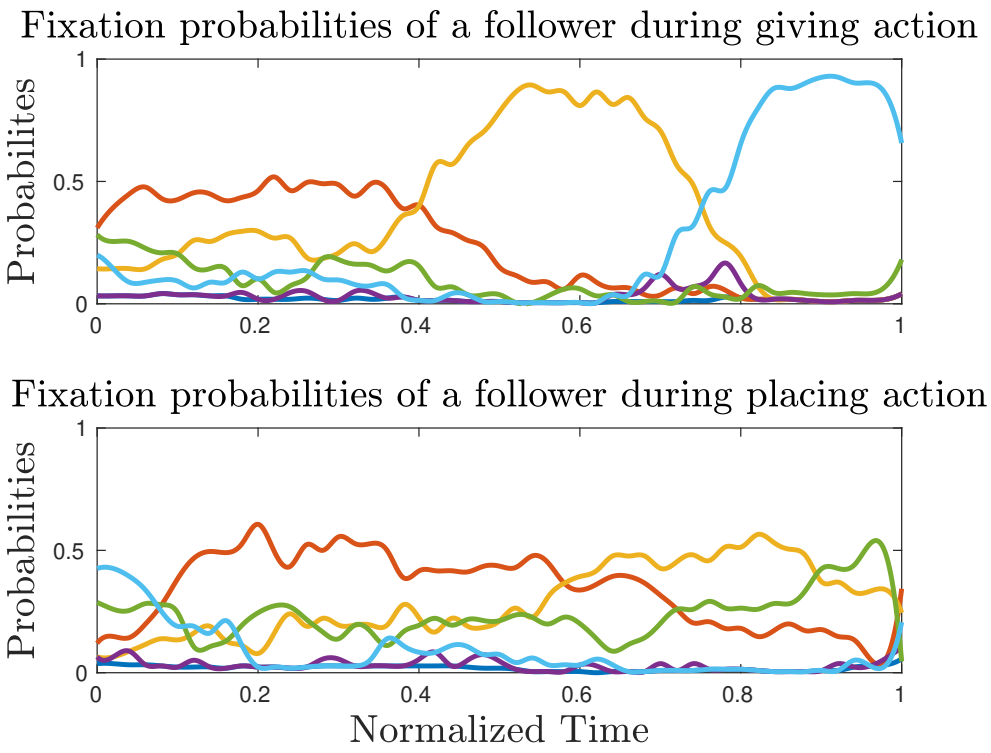


Figure 5.5: Fixation probabilities of Follower's gaze for giving (top figure) and placing (bottom figure) action.

actions, the follower had similar gaze fixations for both actions, whereas the leader changed the gaze behavior depending on the type of action. This has to do with nature of each action, one is an interaction which requires to communicate intent, while the other is an individual action of *placing* a brick on the tower. In the case of the follower, since the action was unknown in the beginning, it is expected that the behavior would be similar until the moment when the follower understands the action intention. After the follower decodes the action, the gaze behavior would change accordingly, which may indicate the slight change in the tower fixations for the two actions.

5.3 The Gaze Dialogue Model

The proposed *Gaze Dialogue Model* integrates the eye-gaze communication that occurs during an interaction with two humans, with the arm-motor actions which result from the interaction. It starts with a general model that represents each human as a separate system, with eye-gaze and arm-motor movements, and the interpersonal links of non-verbal communication. The eye-gaze communication, i.e. the gaze fixations, are used for predicting the fixations of others while, at the same time, generate one's fixations. The associated actions can be inferred from understanding the gaze fixations. The arm-motor cues represent the action of each actor. It predicts the actions of others, while, at the same time, it plans one's actions, generating the appropriate motor commands to complete the action.

The *Gaze Dialogue Model* can also be adapted to the present HHI experiment and to the leader-follower relation that is extracted from the data. The HHI *Gaze Dialogue Model* contains both the eye-gaze communication and the arm movements, as in the general model, with the difference that the leader's action is instructed and pre-defined, as the purpose of the follower is to understand the action, *giving* or *placing*, and act accordingly.

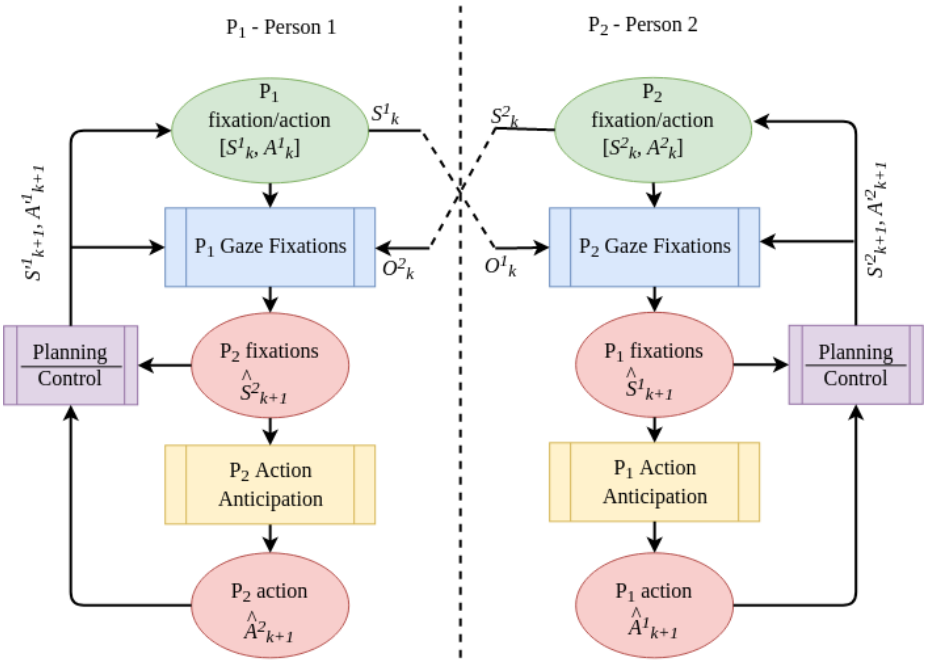


Figure 5.6: Block diagram of proposed general *Gaze Dialogue* model

The proposed *Gaze Dialogue Model* block diagram is shown in Figure 5.6. The states are defined as the gaze fixation S_k and type of action A_k, for each actor, at time instant k. The model is composed of the Gaze Fixations system, identified by the blue blocks, and the Action Anticipation system, the yellow blocks. The Gaze Fixations system is responsible for predicting the fixations of others, and generating one's own fixations. The role of the Action Anticipation system is to predict the actions of others, and to plan one's own action.

The *Gaze Dialogue Model* uses the history of a person's gaze fixations and actions, together with the observations O_k of the gaze fixations of the other person. The Gaze Fixations system predicts the gaze fixation of the other person at time $k + 1$, while the Action Anticipation system predicts the type of action performed by the second person. The predictions of the fixations and actions, together with one's fixations and actions, are eventually fed back to the Planning/Control system. This block is responsible for determining the person's next gaze fixation and which action to perform.

Different approaches can be applied for modelling the gaze behavior including Gaussian Mixture Modeling, Support Vector Machines, Neural Networks, among others. It was decided to use HMM as the modelling tool as it naturally encodes the interaction process by representing the eye-gaze fixations as states, and human actions as outcomes. In addition, HMMs are able to generate the robot eye-gaze movements in real-time while predicting, at the same time,

the human action entirely from human eye-gaze fixations. Finally, the number of parameters to estimate is compatible with the relatively small size of the data-sets collected during the human studies. The general approach of the *Gaze Dialogue Model* is described in Section 5.3 for the Gaze Fixations system and in Section 5.3 for Action Anticipation system.

Gaze Fixations

The *Gaze Dialogue* between two persons is modelled with a HMM. Each actor has an associated internal state variable:

$$S_k \in \left\{U_1, ..., U_N\right\}$$

where $U_1, ..., U_N$ are the admissible state values, i.e. fixations, and $k \in \left\{1, ..., T\right\}$ denotes the discrete time instants. The actor has access to an instantaneous observation:

$$O_k \in \left\{V_1, ..., V_M\right\}$$

where $V_1, ..., V_M$ are the fixations of the other actor.

The two sequences (state and observation sequence)

$$S = \left(S_1, ..., S_T\right), O = \left(O_1, ..., O_T\right)$$

are represented by the HMM $\lambda = (\pi, C, D)$ where π denotes the probability distribution of the state variable at time $k = 1$, $C = (c_{i,j})$ denotes the transition matrix and $D = (d_{i,j})$ the matrix of output probabilities [Rabiner, 1990].

Since there are two actors, denoted by P_1 and P_2, the above sequences are duplicated:

$$\begin{array}{ll} S_k^1 \in \left\{U_1^1, ..., U_N^1\right\} & O_k^1 \in \left\{V_1^1, ..., V_M^1\right\} \\ S_k^2 \in \left\{U_1^2, ..., U_N^2\right\} & O_k^2 \in \left\{V_1^2, ..., V_M^2\right\} \end{array}$$

and different HMMs are used to generate the state and observation sequences for each actor: $\lambda^1 = (\pi^1, C^1, D^1)$ and $\lambda^2 = (\pi^2, C^2, D^2)$.

The joint probabilities of the state and observation sequences for the two actors are given by:

$$P\left(S^1, O^1\right) = \prod_{k=1}^{T} c^1_{S^1_{k-1}, S^1_k} \cdot d^1_{S^1_k}\left(O^1_k\right)$$

$$P\left(S^2, O^2\right) = \prod_{k=1}^{T} c^2_{S^2_{k-1}, S^2_k} \cdot d^2_{S^2_k}\left(O^2_k\right)$$

In the perspective of the interaction actor P_1, it predicts the fixation at time $k+1$ of P_2, $\hat{S}^2_{k+1}$, and generated its next fixation, S'^1_{k+1}. In the perspective of P_2, it predicts the fixation of P_1, $\hat{S}^1_{k+1}$, and generate the next fixation, S'^2_{k+1}, of P_1.

Action Anticipation

For the Action Anticipation system, the purpose is to predict the actions of others and plan one's actions, based on the gaze fixations of the actors. For actor j where $j = [1, 2]$, to predict the action of others $\hat{A}^i_{k+1}$ its combined the information related to the other gaze fixations $\hat{S}^i_{k+1}$, where $i \neq j$ is the other actor. Based on the current gaze fixation of the other, the action probabilities from Table 5.3 are used to update an exponential moving average:

$$P^i_a(k+1) = (1-\alpha)P^i_a(k) + \alpha\delta(k) \tag{5.1}$$

where k refers to time, and α is a constant smoothing factor. $\delta(k)$ is the probability $P^i_a(k)$ of action a occurring when actor i is looking at gaze fixation S^i_k at time k. P^i_a is the probability of actor i performing action a. During the interaction, the *Gaze Dialogue Model* predicts the actions of others, and allows one to plan our own actions. The predicted action $\hat{A}^i_{k+1}$ of actor i is updated for every new gaze fixations $\hat{S}^i_{k+1}$ the actor i is gazing for each time $k+1$. Whilst at the same time $k+1$, the Action Anticipation system is planning, the associated action A'^j_{k+1} to actor j. The exponential moving average mechanism ensures a smooth evolution of the action probabilities, and filters out spurious noisy measurements.

The following section describes the *Gaze Dialogue Model* for the leader-follower scenario in the HHI experiments and Section 5.3.2 explains the corresponding Action Anticipation system.

5.3.1 Gaze Fixations for the Human-Human Interaction

To model the HHI experiments described in Section 5.2, the general *Gaze Dialogue Model* was adapted to a leader-follower relation. Figure 5.7 shows the block diagram of the model with a few modifications to reflect the leader-follower experiments from our scenario. From [Gallotti et al., 2017] in a leader-follower scenario, the leader leads the action, while the follower adapts its behavior to match the leader's intention. As such, in the *Gaze Dialogue* model, the leader's block system is not in closed-loop, since the leader is not influenced by the follower. Since the action is instructed to the leader, the leader's has to generate his/her own gaze fixations and action. The leader is thus not required to predict the gaze fixations of others nor to predict the actions of others. On the other hand, the follower "reads" the leader's non-verbal cues (arm movements and gaze fixations) to infer the leader's action and, consequently, prepare his/her own action and provide the appropriate non-verbal cues.

The State and Observation values match the labelled and leader/follower pair gaze fixations described in Section 5.2. The leader has six different states, and follower four.

For the leader/follower pair, it denotes the states of the leader and follower respectively by

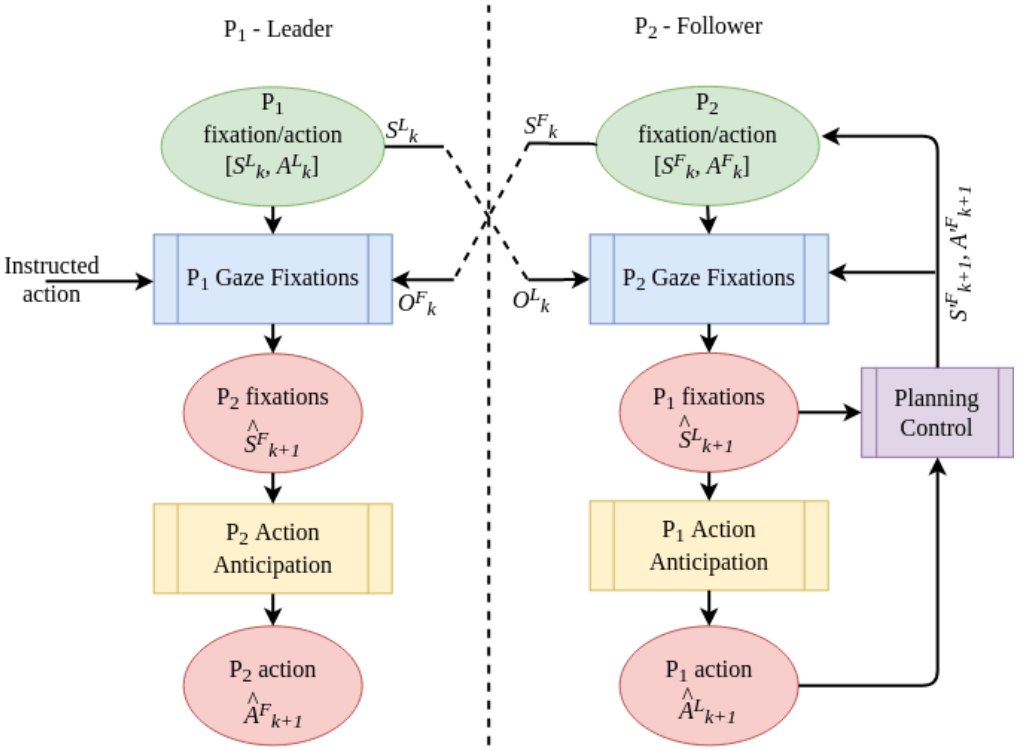


Figure 5.7: Block diagram of leader-follower *Gaze Dialogue* model

S^L and S^F. The states and observations of the leader-follower become:

$$S_k^L \in \left\{U_1^L, ..., U_N^L\right\} \quad O_k^L \in \left\{V_1^L, ..., V_M^L\right\}$$
$$S_k^F \in \left\{U_1^F, ..., U_N^F\right\} \quad O_k^F \in \left\{V_1^F, ..., V_M^F\right\}$$

the HMMs for the leader-follower relation has the following parameters, $\lambda^L = \left(\pi^L, C^L, D^L\right)$, and $\lambda^F = \left(\pi^F, C^F, D^F\right)$.

For the leader-follower relation of the HHI experiments, the state and observation sequences of both actors are related, as the state of the leader actor is the observation of the follower, and vice-versa:

$$S^F = O^L \quad \text{and} \quad S^L = O^F \tag{5.2}$$

The two sequences $S = \left(S_1, ..., S_T\right), O = \left(O_1, ..., O_T\right)$ and the leader has state sequence S and observation sequence O, while the follower has the opposite. Hence the two probabilistic models are:

$$P\left(S, O\right) = \prod_{k=1}^{T} c^L_{S_{k-1}, S_k} \cdot d^L_{S_k}\left(O_k\right)$$

$$P\left(O, S\right) = \prod_{k=1}^{T} c^F_{O_{k-1}, O_k} \cdot d^F_{O_k}\left(S_k\right)$$

with different C and D matrices. The two HMMs are learned from the *giving* and *placing* actions data, and the obtained transition and emission matrices are given in Table 5.2. Hence there are 4 HMMs in total in the *Gaze Dialogue Model*, one for each person and for each action.

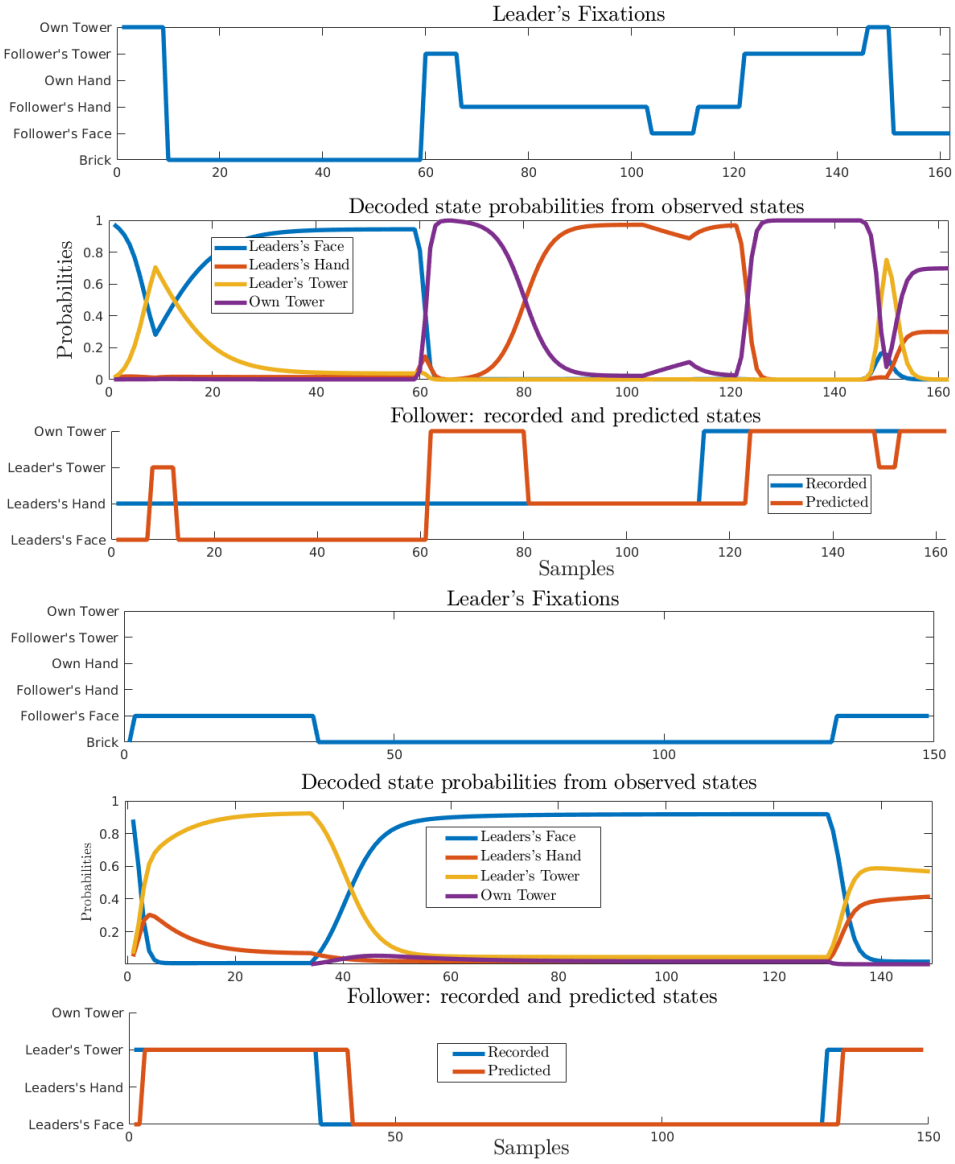


Figure 5.8: Simulations of Leader's and Follower's internal model in the case when the leader's behavior during a giving action (top) and placing action (bottom). The top plot shows the leader's recorded gaze fixations, the middle plot the follower's fixation probabilities, and the bottom shows the follower's recorded and most likely fixations.

The HMMs are used by the leader to predict the follower's next state $\hat{S}^F_{k+1}$ and, conversely, by the follower to predict the leader's next state, $\hat{S}^L_{k+1}$. More important, by using posterior decoding, the follower can plan its next fixation S'^F_{k+1} in response to the leader's behavior.

The model evaluation is performed by taking the leader's gaze fixations (blue line on Figure 5.8 top) as the input and estimate the predicted behavior of the follower using the posterior state probabilities (Figure 5.8 middle). The predicted gaze fixations of the follower (red line on Figure 5.8 bottom) are gaze fixations with the highest probability in each time instate. The follower's predicted gaze fixations are compared against one instance of actual (recorded) gaze fixations for the same action as the known leader's behavior (blue line on Figure 5.8 bottom).

The last plot of Figure 5.8 shows that the predicted gaze fixations follow the leader's gaze fixation change, which means that the follower's predicted gaze fixations are "aligned" with

	Leader	Follower
Giving	$C_G^L=\begin{bmatrix}0.9861 & 0.0016 & 0.0045 & 0.0016 & 0.0041 & 0.0020\\ 0.0018 & 0.9567 & 0.0316 & 0.0009 & 0.0081 & 0.0009\\ 0.0012 & 0.0241 & 0.9630 & 0.0006 & 0.0105 & 0.0006\\ 0 & 0 & 0.0541 & 0.9279 & 0.0180 & 0\\ 0.0019 & 0.0183 & 0.0144 & 0.0010 & 0.9587 & 0.0058\\ 0.0477 & 0.0053 & 0.0053 & 0.0186 & 0 & 0.9231\end{bmatrix}$ $D_G^L=\begin{bmatrix}0.4702 & 0.2838 & 0.1310 & 0.1151\\ 0.0799 & 0.5245 & 0.0477 & 0.3479\\ 0.1381 & 0.5705 & 0.0317 & 0.2597\\ 0.3091 & 0.4273 & 0.0818 & 0.1818\\ 0.0650 & 0.2367 & 0.0260 & 0.6723\\ 0.3837 & 0.2820 & 0.2238 & 0.1105\end{bmatrix}$	$C_G^F=\begin{bmatrix}0.9683 & 0.0209 & 0.0070 & 0.0038\\ 0.0064 & 0.9796 & 0.0030 & 0.0111\\ 0.0437 & 0.0198 & 0.9226 & 0.0139\\ 0.0030 & 0.0012 & 0.0012 & 0.9947\end{bmatrix}$ $D_G^F=\begin{bmatrix}0.6997 & 0.0396 & 0.1227 & 0.0217 & 0.0319 & 0.0843\\ 0.3022 & 0.1861 & 0.3626 & 0.0215 & 0.0832 & 0.0444\\ 0.6199 & 0.0752 & 0.0894 & 0.0183 & 0.0407 & 0.1565\\ 0.1818 & 0.1832 & 0.2449 & 0.0136 & 0.3507 & 0.0258\end{bmatrix}$
Placing	$C_P^L=\begin{bmatrix}0.9867 & 0.0008 & 0.0004 & 0.0008 & 0.0004 & 0.0109\\ 0.0098 & 0.9755 & 0.0098 & 0 & 0 & 0.0049\\ 0.0569 & 0.0081 & 0.9268 & 0 & 0.0081 & 0\\ 0 & 0 & 0 & 0.9722 & 0 & 0.0278\\ 0.0200 & 0.0114 & 0 & 0 & 0.9629 & 0.0057\\ 0.0064 & 0.0051 & 0.0013 & 0.0006 & 0.0064 & 0.9803\end{bmatrix}$ $D_P^L=\begin{bmatrix}0.4543 & 0.2213 & 0.2284 & 0.0960\\ 0.1387 & 0.4380 & 0.4161 & 0.0073\\ 0.2857 & 0.0408 & 0.1735 & 0.5000\\ 0.3299 & 0.4639 & 0.1856 & 0.0206\\ 0.1869 & 0.6920 & 0.0588 & 0.0623\\ 0.2286 & 0.4163 & 0.3367 & 0.0184\end{bmatrix}$	$C_P^F=\begin{bmatrix}0.9682 & 0.0143 & 0.0162 & 0.0012\\ 0.0090 & 0.9793 & 0.0104 & 0.0014\\ 0.0259 & 0.0121 & 0.9586 & 0.0035\\ 0.0180 & 0.0060 & 0.0419 & 0.9341\end{bmatrix}$ $D_P^F=\begin{bmatrix}0.7546 & 0.0131 & 0.0192 & 0.0220 & 0.0371 & 0.1540\\ 0.4273 & 0.0479 & 0.0032 & 0.0359 & 0.1597 & 0.3259\\ 0.5570 & 0.0575 & 0.0172 & 0.0182 & 0.0172 & 0.3330\\ 0.7250 & 0.0031 & 0.1531 & 0.0063 & 0.0563 & 0.0563\end{bmatrix}$

Table 5.2: HMM parameters for the leader (L) and follower (F) defined by transition matrix C and emission matrix D for *(G)iving* and *(P)lacing* actions.

the leader's gaze fixations. As expected, the recorded follower's gaze fixations for a single instance/specific action may differ from the predicted (probabilistic) gaze fixations, however, most of the time, the predictions match the observed fixations.

Table 5.3: Probabilities for giving and placing action with respect to the leader's gaze fixation

	Giving	Placing
Brick	0.496	0.504
Follower's face	0.841	0.159
Follower's hand	0.931	0.069
Own hand	0.520	0.480
Follower's tower	0.748	0.252
Own tower	0.186	0.814

5.3.2 Action Anticipation for the Human-Human Interaction

From the HHI experiments, only two actions are possible for the leader, *giving* and *placing*, or the follower, receiving and not-receiving. Taking into account the leader-follower relation, a receiving action is associated with a *giving* action and not-receiving to a *placing* action.

The prediction of a certain action combines the information related to the follower's current fixations, with the past probability of the same action. These probability signals are denoted as P_G and P_P, respectively for the *giving* and *placing* actions. The action probabilities are the following:

$$P_G(k+1) = (1-\alpha)P_G(k) + \alpha\delta(k) \tag{5.3a}$$

$$P_P(k+1) = (1-\alpha)P_P(k) + \alpha\delta(k) \tag{5.3b}$$

where P_G and P_P denote the probability of *giving* and *placing* action, respectively, for each time step k, and $\alpha = 0.05$. The update $\delta(k)$ depends on the values of Table 5.3, evaluated for each gaze fixation of the leader at time k. In the HHI experiment, the leader is "instructed" which action to perform (*giving* or *placing*). The action is unknown to the follower who needs to understand it from the non-verbal communication cues. The *Gaze Dialogue Model* infers the leader's action from the leader's gaze fixations which are used to generate the follower's gaze fixations, and planning for the follower's own action. The follower's Action Anticipation

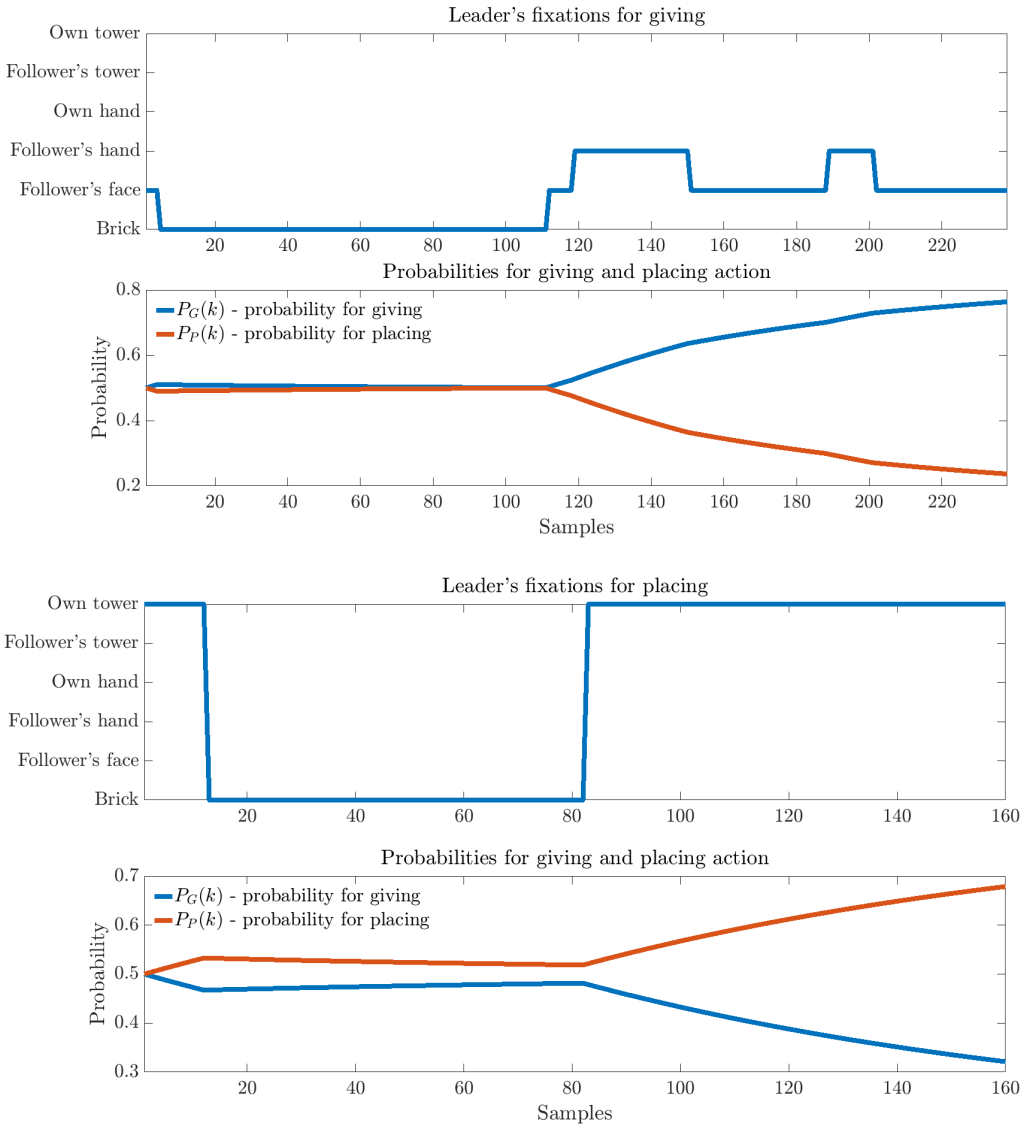


Figure 5.9: Change of the signals $P_G(k)$ (blue line) and $P_P(k)$ (red line) with respect to the leader's gaze fixations for *giving* (two top figures) and *placing* action (two bottom figures)

system uses the leader's observed gaze fixations. Each gaze fixation is associated with the probability to choose between two actions as given in Table 5.6. The probabilities were derived from the duration of each gaze fixation for each action, as given in Table 5.1, divided by the total duration of gaze fixation throughout the HHI experiments. Table 5.6 shows that the leader fixations at the brick or at his/her own hand are negligible for both actions, as

the probabilities are close to 50%. Instead, other gaze fixations provide stronger gaze cues towards one of the two actions. The leader's gaze fixation at the follower's face, hand or tower clearly communicate the intention of *giving* the brick, whereas gaze fixations at his own tower, presumably to visually guide the arm to properly place the brick, become strong cues for the *placing* action.

The Action Anticipation system is composed of two signals that represent the probabilities for the *giving* ($P_G(k)$) and *placing* ($P_P(k)$) actions, over time, with the initial values set to 50%. These signals are updated in each iteration as follows. First, the action is selected based on the leader's current gaze fixation and the probabilities shown in Table 5.3. For example, if the leader's fixates the follower's face there is a 84.1% chance to select a *giving* action and a 15.9% to select a *placing* action. Based on the selected action, the δ values of Equations 5.3a and 5.3b is updated to calculate the value of the signals $P_G(k)$ and $P_P(k)$ for the next time k. In case the leader gaze fixates the follower's face, and the *placing* action is selected, the δ of 0.159 is used for the signal $P_P(k)$ and -0.159 for the signal $P_G(k)$. The output signals $P_G(k)$ and $P_P(k)$ are smoothed with a moving average, and normalized with respect to the number of samples (i.e. the number of gaze fixations observed) collected up to time k. This approach is similar to a Markov Reward Process [Li, 2010] that adds a reward signal to each state. The reward is determined in the same manner, but in our case, the purpose is to decide which type of action, in order to prevent oscillatory behavior of the action prediction.

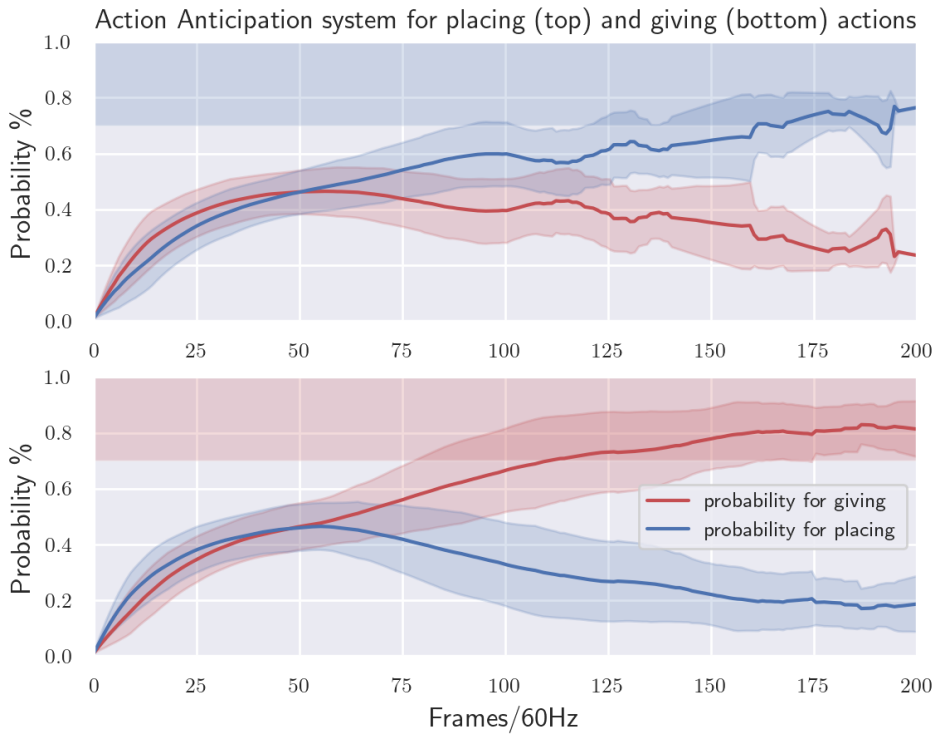


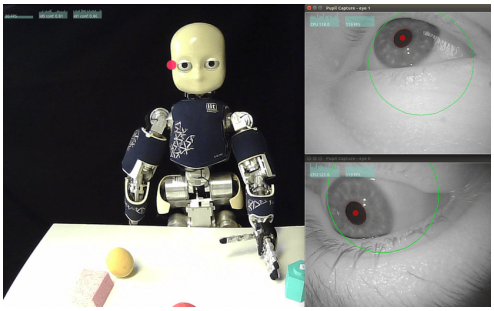
Figure 5.10: The Action Anticipation results on the entire HHI dataset on classifying the actions as a placing or giving action, respectively. Shadow area is the standard deviation while the shadow rectangular area on the top reflects the threshold of 75% accuracy.

The results on the accuracy of the Action Anticipation system can be viewed in Figure 5.10. The action is classified either *placing* or *giving* when the prediction reaches above 70% (the region is marked with a shaded color). A *giving* action can be correctly classified at around 60% completion which for an action that takes on average 2 seconds to finish, corresponds to a reaction time of 1.12 seconds. As for a *placing* action, it takes longer to predict, around 80% completion, which puts the reaction time at 1.36 seconds. The slower prediction could be caused by a prolonged period of time fixating the brick, as illustrated in Figure 5.4, which brings ambiguity to the system. Figure 5.9 exemplifies a scenario where the leader starts by first fixating the brick. The Action Anticipation system cannot predict which of the actions the leader is performing. When the leader gaze fixation switches to the follower's face or hand, the probability for *giving* increases, and when the leader gaze fixations switch to his own tower, the probability for *placing* increases. The relation between the P_G and P_P signals is used to predict the leader's action, A^L_{k+1}, that is set to be equal to the planned action of the follower A'^F_{k+1}. In this work, a simple comparison to decide between *giving* and *placing* actions is applied. If the difference between signals P_G and P_P is greater than a predefined threshold, the inferred action is *giving*, otherwise it is *placing*.

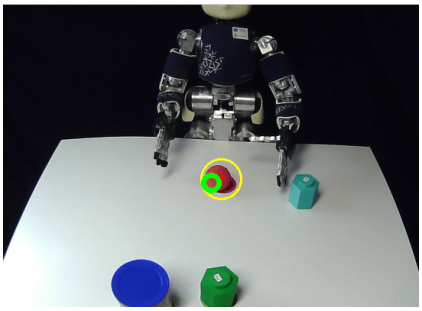
The main conclusions that can be drawn from the modelling are: (i) the *Gaze Dialogue Model* can generate gaze fixations for the *giving* and *placing* actions that are similar to the ones observed in the HHI data; (ii) the *Gaze Dialogue Model* can predict accurately the follower's next gaze fixations, when provided with the leader's real gaze fixations, from the HHI dataset, (iii) it is possible to predict the correct actions from the gaze fixations using our *Gaze Dialogue Model*.

5.4 Robot Experiments

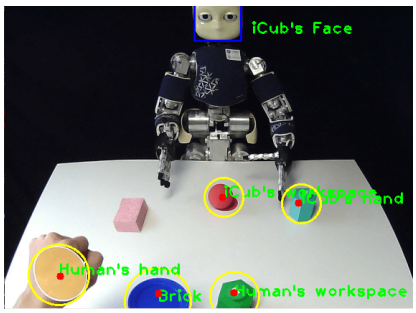
The *Gaze Dialogue Model* is able to represent actions using eye-gaze communication and motor information. The eye-gaze communication allows to predict the gaze fixations of others, while at the same time, generate one's own gaze fixations. Additionally, from those same gaze fixations, it can be used to predict the actions of others, while at the same time, plan one's own corresponding actions, which are represented as arm movements. The following section addresses the validation of the *Gaze Dialogue Model* in a human-robot scenario. The section begins by describing the two types of HRI: (i) robot-as-a-leader, and (ii) robot-as-a-follower; however, the focus of this work will be on robot-as-a-follower, for reasons that will become clearer below. Secondly, it describes the human-in-the-loop system, which is important for the interaction between robot-as-a-leader and robot-as-a-follower. The section finishes with a discussion on the results of HRI experiments and an analysis of the interaction comparing to the HHI experiments.



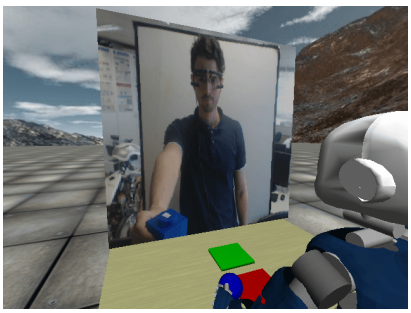
(a) Egocentric view of the human from the head mounted eye-tracker. The infra-red cameras recording the subject's eyes are depicted on the right. The human gaze fixation is depicted by a red marker in the robot's face.



(b) Human gaze fixation is on the red ball. The red ball detection is marked by the yellow circle, and the human gaze fixation is the green hallow circle.



(c) All the important regions, and correct labels, for the gaze fixation are identified. Additional objects which are not relevant are considered outliers.



(d) Experimental setup for the HRI scenario. Human subject is wearing a head mounted eye-tracker and the relevant objects are present.

Figure 5.11: Different perspectives of the Human-Robot Experimental setup: (a) is the view-point of the human; (b) illustrates the human gaze fixation and one of the objects identified; (c) all the labels from the HHI experiments are identified; (d) the perspective of the robot when interacting with a human. A video showing the interactions is available in video.GazeDialogue.ieee-2022

5.4.1 Robot Setup in the Leader-Follower Scenario

HRI experiments were carried out with the iCub robotic platform [Metta et al., 2010]. As a humanoid robot, the iCub has a body structure that is similar to the human body, so that humans can more easily understand the robot's motor behavior and, hence, its intentions [Kelley et al., 2010]. It has 2 cameras, on the head of the robot, that are capable of vergence and version, in a way similar to the human oculomotor control system.

In both HRI experimental scenarios there are things in common. Firstly, the human actor wears the Pupil Labs eye tracker, introduced in Section 5.2.1. The objective here is to track the human gaze fixations while (s)he interacts with the robot. The software and the gaze fixation point are shown in Figure 5.11a. Secondly, concerning the low-level controllers, the eye-gaze saccadic movements in the iCub is driven by the Cartesian 6-DOF gaze controller described in [Roncone et al., 2016]. As for the arm movements, a minimum jerk Cartesian controller is applied to control the iCub's arm and torso [Pattacini et al., 2010]. Finally, the arm movements are synchronised with the eye-gaze movements, specifically when a switch between gaze fixations occurs. The validation was made with the HHI experiments data, and applied to the HRI experiments for reproducibility of the arm movements. The most common examples occur when the human switches from fixating the brick to another region-of-interest.

It is usually associated with the beginning of either the *giving* or the *placing* action.

The robot-as-a-leader scenario does not require the robot to use any sensor data from the human. As illustrated in the block diagram of Section 5.3, our approach focuses on the non-verbal communication from the leader to the follower. Since in the robot-as-a-leader, the leader is always aware of the action, it does not require any feedback from the follower. It is assumed that once the robot takes the leader's role, the *Gaze Dialogue Model* generates the robot's gaze fixations and plans its action to execute either a *giving* or a *placing* action. The eye-gaze communication and arm movements are assumed to be communicated and 'read' by the human follower. The leader's eye-gaze communication for *giving* or *placing* actions is determined as the most likely gaze fixations observed in the HHI dataset. Figure 5.4 shows the probabilities of each gaze fixation during *giving* and *placing* actions, respectively, over time. The robot-as-a-leader interacting with a human can be seen in the supplementary video material video.GazeDialogue.ieee-2022.

In the case of the robot-as-a-follower, the human wears the eye-tracking system during *placing* and *giving* actions. As the robot has to follow the interaction, it has to interpret the relevant gaze fixations from the human. In this scenario, the human is part of the control loop, by providing feedback to the robot controller through his/her gaze fixations. This information is streamed, in real-time, to allow the robot to predict the human gaze fixations and the human action while, also in real-time, generating the robot's own gaze fixations and plan the robot's actions. The diagram in Figure 5.12 illustrates the human-in-the-loop modules involved in the HRI.

5.4.2 Human-in-the-Loop System

To provide the robot with the human gaze fixations, i.e. Human-in-the-loop, labelling and segmenting of the eye-tracker data is necessary to convert 2D images of the eye-tracker, to human gaze fixations. The algorithm is called Visual Focus of Attention (VFOA), inspired on the work from [Sheikhi and Odobez, 2012] which involves tracking the region where the human is fixating in a 2D image. The following subsections explain the necessary steps to segment, label, and communicate all the important eye-gaze fixations to the robot.

Gaze Fixation Point

The first step of the implementation of the block diagram of Figure 5.12 involves synchronizing the gaze fixation point provided by the LSL network [Kothe], and the video frame received directly by the Capture software of the Pupils Labs. The gaze fixation point is marked by a green hollow circle, seen in Figure 5.11b, and it is recorded at 120 Hz. The world camera, i.e. the egocentric view of the human, is published at 30 frames per second (30 Hz). Since the frequency of the gaze fixations stream is 4 times faster than the stream of the world camera, the process runs every 4 gaze fixation points from the buffer sent by the LSL network.

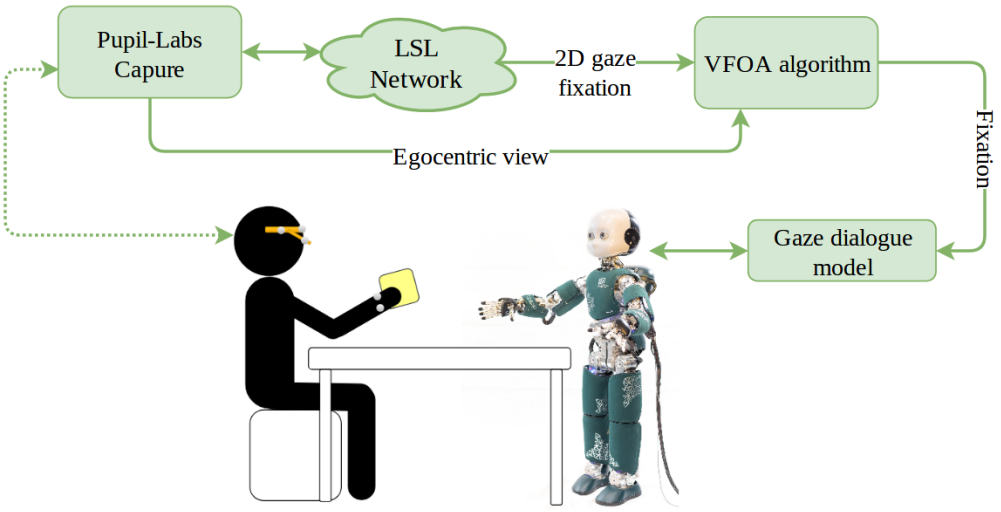


Figure 5.12: Diagram illustrating the connections between the different modules that make up the communication of the human eye gaze to the robot fixations. The first module is related to the software that acquires the data from the eye tracker - Capture by Pupil Labs [Kassner et al., 2014]. From this module the 2D fixation point of the subject's gaze projected onto the world view camera on the eye tracker is gathered. The stream of the world view camera, together with 2D gaze fixations through LSL network [Kothe], is sent to the Visual Focus of Attention algorithm module to track the relevant fixations. The final module is the implementation of the gaze dialogue model described in Section 5.3

Whenever the green hollow circle, i.e. the human gaze fixation, is inside a region of interest, the VFOA algorithm classifies the gaze fixation point as a valid state S_k, and it is sent to the *Gaze Dialogue* model.

Object Detection

The VFOA algorithm classifies as important eye-gaze fixations the states S_k and, correspondingly, the observations O_k. To classify the valid gaze fixation points, a color-based algorithm which extracts the relevant colors as the relevant objects to the HRI setup is used. Table 5.4 identifies the objects, with the corresponding colors, extracted in the HRI experiments and the associated label given to the *Gaze Dialogue Model*. An example of a HRI setup with the VFOA algorithm classifying objects of different colors with its corresponding label is in Figure 5.11c.

Label	color	RGB	Object
Brick	blue		Cylinder
iCub's Workspace	red		Sphere Shape
Human's Workspace	green		Hexagonal Shape
iCub's Hand	cyan		Cyan Sticker
Human's Hand	yellow		Yellow Sticker

Table 5.4: Associated label to the colored object in the HRI setup.

Face Detection

A Haar cascade classifier algorithm detects the iCub's face. The cascade is trained with real images of the iCub's face. This classifier can detect the iCub's face in the HRI scenario quite accurately with very few false positives during the trials. Figure 5.11c) shows all the regions of interest, including the iCub's face, detected from the VFOA algorithm output.

The *Gaze Dialogue Model* was implemented in the Human-in-the-loop system as follows. Firstly, the human gaze fixations are used as observations O_k and the robot's gaze fixations as the current state S_k. Secondly, the robot can predict the leader's gaze fixation $\hat{S}^L_{k+1}$ and action $\hat{A}^L_{k+1}$, using the appropriate HMM described in Table 5.2 and the Action Anticipation algorithm explained in Section 5.3.2. Thirdly, the predictions are fed into the Planning/Control block. Fourthly, the posterior decoding executes to generate the follower's gaze fixations S'^F_{k+1}. Finally, the leader's predicted action is used to plan the follower's action SA'^F_{k+1}. From this information it is determined which HMM model is used in the iteration $k+1$ to generate the next eye-gaze communication and arm movement of both leader and follower. The follower's gaze fixations are given as input to the robot eye controller [Roncone et al., 2016] to drive the eyes towards the correct 3D space gaze fixation point. The Action Anticipation system is used to decide whether the robot starts its arm movement toward the hand-over location, in the case of *giving*, or stands still, in the case of a *placing* action.

5.4.3 Results of the Human-Robot Interaction Experiments

As discussed in Section 5.2, the robot-as-a-leader validates the leader's correct gaze fixations for *giving* and *placing* actions, as shown in Figure 5.4.

Concerning the robot-as-a-follower validation, the human is instructed to perform the two types of actions plus an additional one: (i) giving, (ii) placing and (iii) fooling. The first two actions are the same as present in the HHI experiment, hence the subject interacting with the robot, albeit naive to the previous experiment, acted naturally without any further instructions. A total of 40 trials with one participant are performed, 20 trials performing both placing and giving actions. The Human-in-the-Loop system with the *Gaze Dialogue Model* ran online at 20 Frames per Second (FPS) and the steps shown are the iterations where new human gaze fixations O_{k+1} are received. It is not guaranteed that the Human-in-the-Loop system will output meaningful gaze fixations (1 out of the 6 fixations) for every single frame (for all 20 FPS). This is reasonable since sometimes humans divert their gaze to unrelated locations, considered in the human analysis as outliers, and in the HRI setting the occurrence did not affect the *Gaze Dialogue Model* or Action Anticipation system. Figure 5.13 shows the mean and standard deviation of the Action Anticipation systems for all of the trials in the HRI. Most of the interactions are correctly classified (average of 75 % or above) with 40 iterations which correspond to around 4 to 5 seconds of real-time human gaze fixations sequence. As for the third action, the subject is instructed to cause a perturbation during the execution of a handover

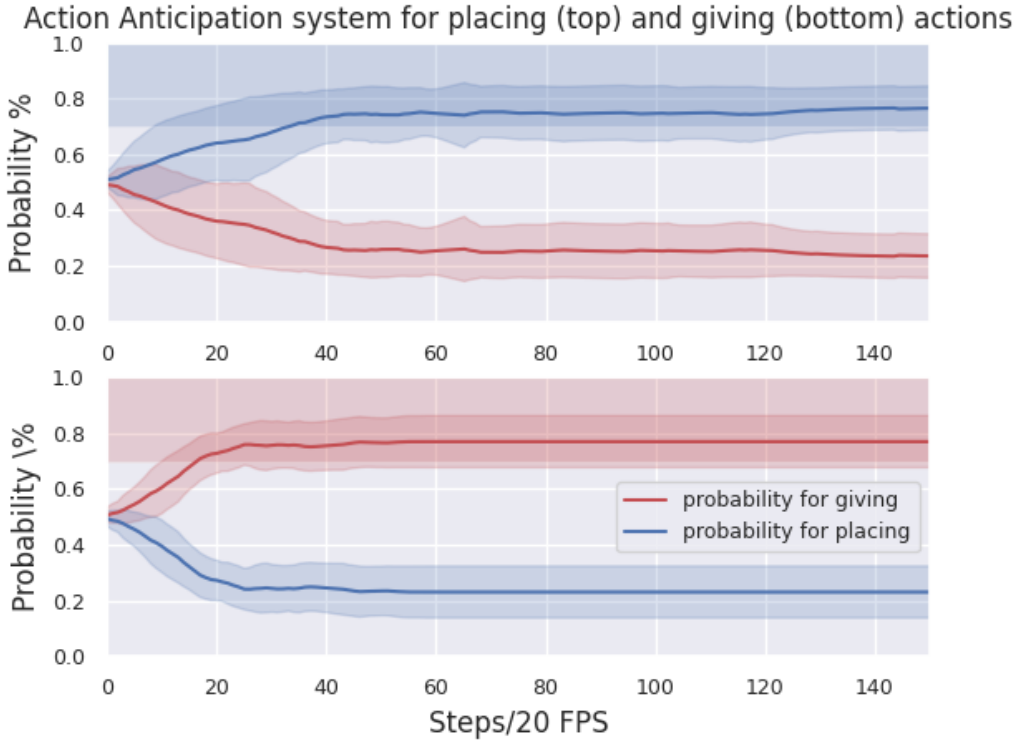


Figure 5.13: The Action Anticipation results for HRI trials on classifying the actions as a placing or giving action, respectively.

by switching to a *placing* action. The purpose of creating a fooling action is to show the active adaptation of the *Gaze Dialogue Model* to the different gaze fixations of the human. Figure 5.14 shows the human gaze fixations. During the fooling action, the human is instructed to execute a *giving* action, and before handing over the object, the human is instructed to place the brick in his workspace. During the first 200 iterations the human gaze fixations are mostly on the follower's tower, which correlates with a *giving* action. After the 200^{th} iteration, the human fixates his/her own tower, which correlates with a *placing* action. The robot generates its gaze fixations, driven by the model, and fixates its tower at the beginning, before successively updating its gaze fixations to the leader's face and to the leader's tower. In short, the results of the gaze fixations for the fooling action illustrates a fast reaction to the non-verbal gaze communication cues exhibited at run-time.

In addition to the recorded gaze fixations probabilities, Figure 5.15 shows the output of the Action Anticipation system and the predicted action of the human at each iteration. As the interaction starts, the *Gaze Dialogue Model* and, more specifically, the Action Anticipation system, predicts a *giving* action. The decision concerning the *giving* action was made when the difference between the signals $P_G(k)$ and $P_P(k)$ exceeds a pre-defined threshold. The threshold is empirically determined and it influences how fast the *Gaze Dialogue Model* reacts to non-verbal communication cues. This decision was used by the robot to decide whether the action is *giving*, as well as to start its arm movement, i.e. arm reaching towards the handover location, or a *placing* action, to move the arm back to the rest position and continue observing.

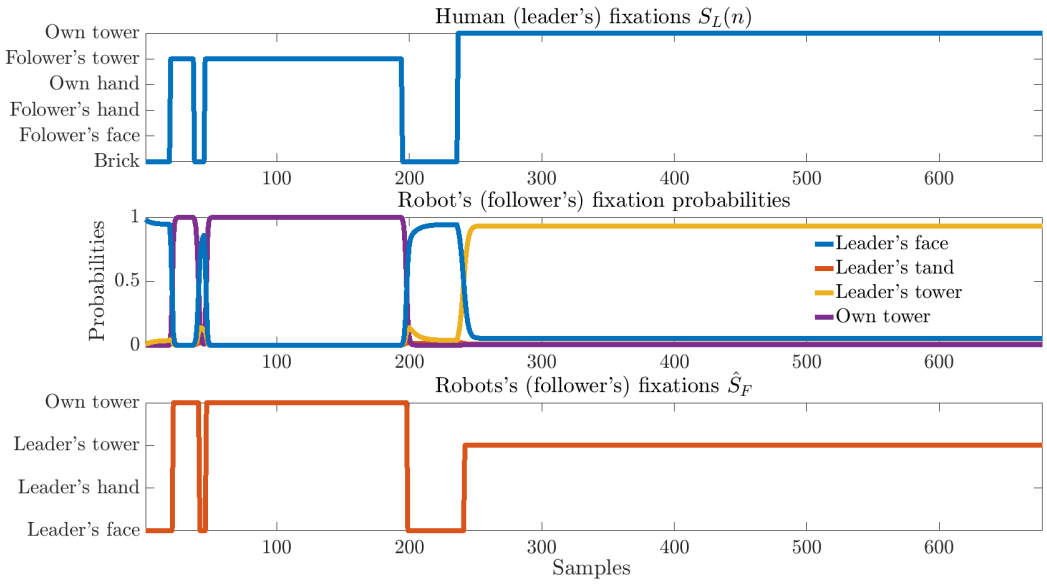


Figure 5.14: Human and iCub's fixations when human as a leader is fooling a robot (starts with giving and after some time switches to placing action): leader's gaze fixations (top); probabilities of follower's fixations (middle); follower's decoded most likely gaze fixations (bottom).

Once the leader fixates his/her tower, the probability for a *placing* action increases. As a result, the robot returns to its rest position while observing the human performing a *placing* action. This experiment validates the capability of the *Gaze Dialogue Model* to: (i) adapt to human gaze fixations, (ii) update the action observed, (iii) generate correct coupling robot-as-a-follower gaze fixations, and (iv) plan the according action. All of this simultaneously and in real-time.

These tests lead us to the following conclusions: (i) the developed *Gaze Dialogue Model* controller is capable of generating gaze fixations, in the robot-as-a-follower case, from real-time gaze fixations of the human subject; (ii) the gaze fixations generated by the controller are similar to the one's observed in the HHI and modelling section; (iii) the controller can predict the human action from the HHI dataset or from real-time gaze fixations in a HRI scenario; (iv) the Human-in-the-loop system can translate online the human VFOA into relevant gaze fixations during the HRI experiments; (v) the *Gaze Dialogue Model* successfully adapts to the action intention of the human during the interaction.

5.5 Remarks

The *Gaze Dialogue Model* emerges during dyadic interactions involving individual (*placing*) actions and actions-in-interaction (*giving*) actions. The implemented model uses the data collected during HHI experiments. The data consisted of paired, synchronised gaze fixations of people involved in the collaborative task. The *Gaze Dialogue Model* combines four HMMs that are selected based on the role of the person, leader or follower and two types of action: *giving* and *placing* for each role. After completing the statistical analysis, with the results

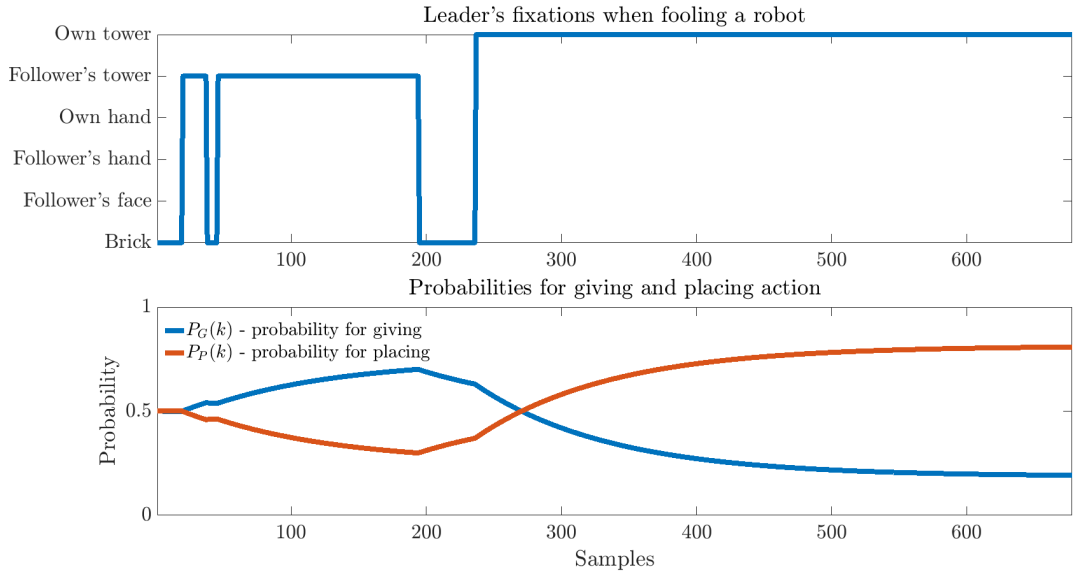


Figure 5.15: Robot's action prediction when human is fooling a robot (starts with giving and switches to placing action)

shown in Figure 5.3, it was clear that the gaze movements are a stochastic process that can be described with associated probabilities. The leader-follower experimental setup asks of the follower to acquire predictive capabilities and adjust to the other interaction partner. The leader's goal is solely to concentrate on the task at hand without taking into account to the follower. Therefore, the leader's action was considered to be known, and the model was deterministic, corresponding to the most likely gaze fixations of a human as a leader. The HMMs are used to predict the gaze fixations of the leader and the leader's action is inferred from the leader's gaze fixations. The posterior decoding is used to plan the follower's gaze fixations, based on the follower's previous fixations and the observed leader's gaze fixations. The inferred leader's action is used in HMMs for both (i) predicting the leader's gaze behavior and (ii) posterior decoding of follower's gaze fixations.

The model was implemented in the iCub robot controller and tested in HRI scenarios. For the robot-as-a-leader scenario, the iCub produces non-verbal gaze communication signals that correlate with the instructed action and may thus be interpreted by the human. For the experiments in the robot-as-a-follower case, a Human-in-the-loop approach is used, and the human gaze fixations are fed back to the model running in the robot controller. The Human-in-the-loop allows the robot to (i) infer the human action, and (ii) to adjust it's gaze fixations according to the human action.

In the case where humans try to fool the robot, by first performing a *giving* action and then suddenly switching to a *placing* action, the *Gaze Dialogue Model* proved to be robust to changes of intention. The robot first began by inferring the *giving* action and starting to move the arm towards the handover location. As soon as the human switches the gaze to his tower and starts a *placing* action, the robot quickly "understands" the change, moves the arm back to the rest position, and continues to observe the human. The accompanying video

material video.GazeDialogue.ieee-2022 shows the aforementioned cases, robot-as-a-leader and robot-as-a-follower.

The iCub eye-gaze saccadic controller performed gaze fixations at a speed approximate to a human, which allows for an accurate representation on the robot to the output of the *Gaze Dialogue Model*. On the other hand, due to hardware restrictions the arm-motor movements were slower on average to a human. This delay between the eye-gaze fixations and the arm movements when performing actions resulted in longer execution times when compared to the HHI experiments. Since the modelling of the robot's behavior is based on the HHI experiment data, if we have a robot with similar human arm-movements speed a more natural behavior of the robot would be achieved.

Before concluding this chapter, an extension that focuses on the alignment of the leader's behavior during dyadic interactions is proposed. The recorded gaze movements of dyads are used to build a model of the leader's gaze behavior (Section 5.6.1). The follower's gaze behavior data is used for two purposes: (i) to determine whether the follower is involved in the interaction, and (ii) if the follower's gaze behavior correlates to the type of action under execution. Information from (ii) is used to plan the leader's actions in order to sustain the leader/follower alignment in the social interaction. The model of the leader's gaze behavior and the alignment of the intentions is evaluated in a HRI scenario (Section 5.6.3), with the robot acting as a leader and the human as a follower. During the interaction, the robot (i) emits non-verbal cues consistent with the action performed, (ii) predicts the human actions, and (iii) aligns its motion according to the human behavior. Section 5.7 is reserved for conclusions future work.

5.6 Extending the Gaze Dialogue: proposal for modelling the leader's non-verbal cues

In the previous approach the leader's gaze behavior was pre-defined as the average, most likely behavior observed from the HHI scenario. Although this behavior may work on average for most interactions, an HRI is never deterministic since humans are naturally unpredictable and stochastic.

The terminology of [Gallotti et al., 2017] is adopted concerning the interaction roles, where one agent can be viewed as the leader and the other one as the follower, in the sense that the follower adapts his/her behavior to the leader, but not the other way around. Hence, in a HRI scenario, a robotic follower will adapt to a human leader. However, when the robot is the leader, the model behaves deterministically and it does not adapt to the behavior of the human follower. In this case, the robot (leader) does not take the speed of the human participant into account, and it is not concerned with the human's understanding of the action. The contribution of the current section is on tackling this issue.

As such, a reliable model for the leader's behavior needs to take the feedback of the follower's behavior into account. In this way, it becomes possible to achieve the third level of interaction [Gallotti et al., 2017], where both agents, the leader and the follower, adapt to each other in order to achieve a mutual alignment. The focus of this work is on closing the loop of the mutual alignment, by adapting the behavior of the actor performing the action (leader), to the behavior of the actor observing and eventually participating in the interaction (follower).

5.6.1 Analysis of the Leader's Gaze Behavior

The dyad interaction experiment is the one described in Section 5.3.1. The two actors have to perform a turn-taking task of *placing* an object on the table, or *giving* the object to the other person. Out of 72 actions, a total of 36 actions were *giving* and 36 were *placing*. The gaze behavior of all 144 actions are labeled with identified relevant fixations and events throughout the action (the labelling is identical to the previous section). The fixations are object (i.e. brick), team-mates' face (TM face), team-mates' hand (TM hand), own hand, team-mates' tower (TM tower), and own tower; and the events are object picked, object handed over, and object placed. Object handed over exists only in the *giving* action. The focus of this work is two-fold: (i) the gaze behavior of the leader during the *giving* action, more specifically on how he/she behaves before and after the handover, and (ii) follower's gaze fixation behavior when the action is *giving* or *placing*.

Figure 5.16 shows the time spent on each of these gaze fixation states, throughout the whole action, and for the two perspectives. In addition to the total amount of time spent on each state, there is a distinction between the gaze behavior before and after the handover. For these experiments, the handover time is defined as the moment when the leader's hand releases the object, and it is identified by the change in the fingers acceleration with respect to the brick.

Figure 5.16 (top image) shows how the leader is mainly focused at the object, and the TM face and hand, right before the handover. The brick is fixated when the leader is visual searching and/or grasping the object - the gaze assisting the motor control function. After the object is grasped, the leader looks mainly at the TM face, hand, and towers - the non-verbal cues to communicate the intention - the gaze engaged in communication purposes. Before the handover, Figure 5.16 (bottom image), the follower fixates the TM's face and hand, aiming at reading the action intention of the leader - communicative gaze. After the handover, the non-verbal cues serve purely functional goals. As the object is already in the follower's possession, the remainder of the action requires the follower to fixate his own tower and controlling the arm towards the goal - the functional role of gaze to assist the motor control.

5.6.2 Modeling of the Leader's Gaze behavior

In order to align the leader to the behavior of the follower a new model was built. Figure 5.17 shows the block diagram for modeling the gaze behavior and aligned motion planning of

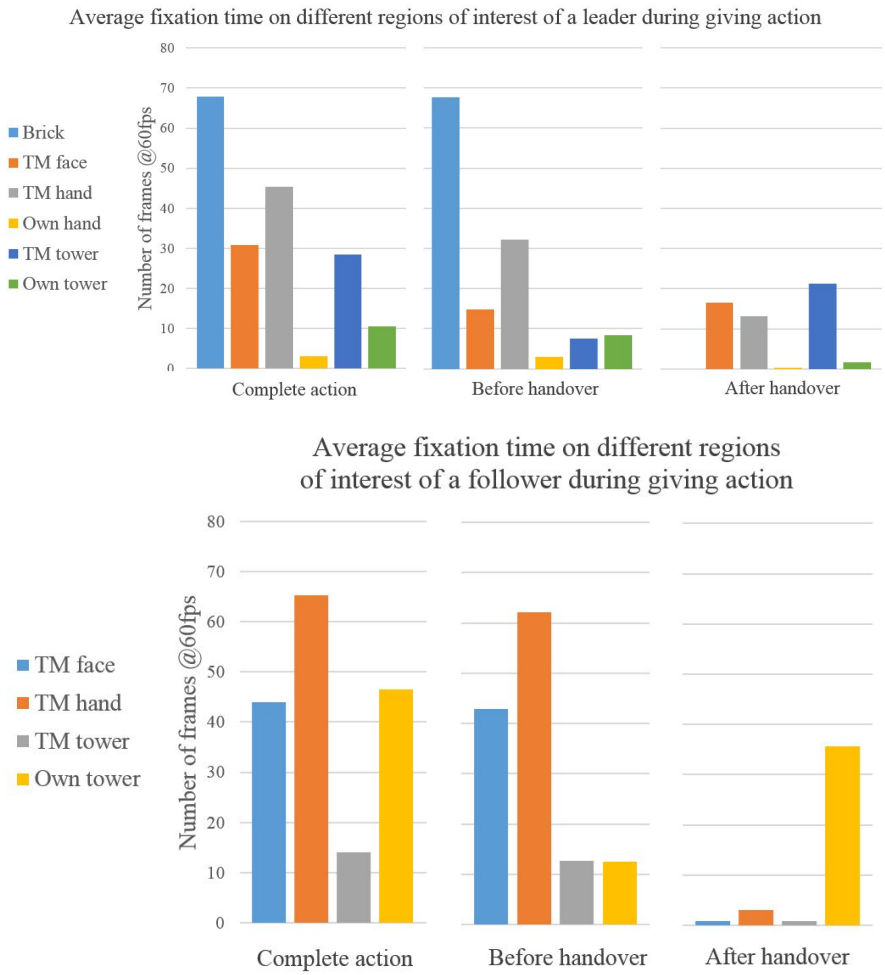


Figure 5.16: Cumulative analysis of the gaze behavior during the HHI experiment for the complete action, before and after handover, showing the leader's (top) and the follower's fixations (bottom). TM stands for Teammate.

agents P_1 and P_2. This required the adaptation of the block diagram from the general gaze dialogue model from Section 5.3 to Figure 5.17.

The state of each agent is defined as the gaze fixation S_k and type of action A_k. The fixations $[S_1(k), S_1(k-1), ...]$ are emitted by agent P_1, which are from the perspective of agent P_2, represented as observations $[O_1(k), O_1(k-1), ...]$. Simultaneously, fixations $[S_2(k), S_2(k-1), ...]$ are emitted by agent P_2, and represented as observations $[O_2(k), O_2(k-1), ...]$ of agent P_1.

The 'Gaze behavior models' encodes the leader's gaze stochastic behavior, that depends on the type of action. Action understanding uses the gaze fixation of the human to estimate the probabilities of *giving* versus *placing* action. This is fed back to the 'Planning/Control' block for the motion planning of the agent and selection of appropriate gaze behavior model.

Human Gaze Behavior

The leader's gaze behavior is modeled with Discrete-Time Markov Chains (DTMC) [Biagini and Campanino, 2016]. A DTMC represents the evolution of a system that stochastically

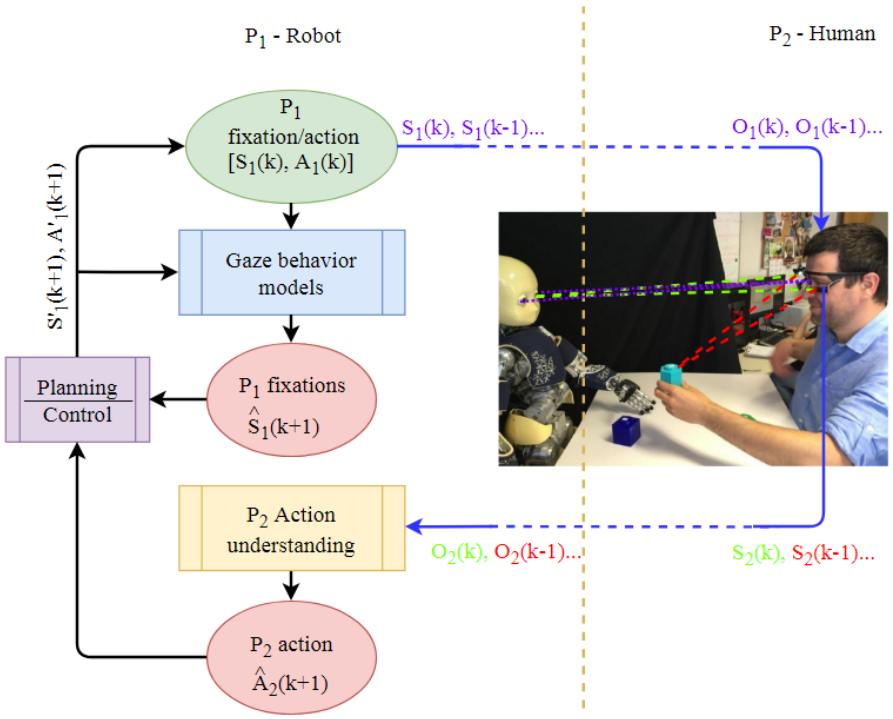


Figure 5.17: Block diagram of the proposed leader's gaze behavior and alignment model.

switches from one state to another, at discrete time instances. The model has an associated internal state variable: $S_k \in \{U_1, ..., U_N\}$ where $U_1, ..., U_N$ denotes admissible state values, i.e. fixations, and $k \in \{1, ..., T\}$ denotes the discrete time instants. In the case of a *giving* action, the leader has six admissible states before the handover, and four states after (Fig. 5.18). This corresponds to the top image from Fig. 5.16 with six fixations before handover. After the handover, the brick is never fixated and the fixation of one's own hand is negligibly small.

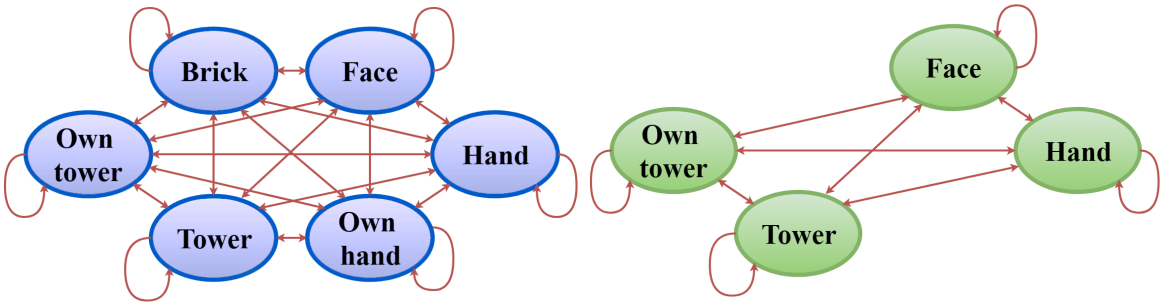


Figure 5.18: DTMC for the behavior of a leader: (left) before the brick handover; (right) after the brick handover.

The two DTMCs (for the period before and after the handover) are represented by transition matrices learned from the HHI data, which has labeled fixations of the dyad throughout all the actions. Transitions of the fixations for *giving* before and after handover are counted, and the obtained transition matrices are given in Table 5.5.

The admissible states that correspond to the indexes of the rows and columns of the transition matrices are: 1 - Brick, 2 - TM Face, 3 - TM Hand, 4 - Own hand, 5 - TM tower and

Table 5.5: Transition matrix before handover A^L_{bhon} and after handover A^L_{ahon} for the *giving* action

Handover	Leader
Before	$A^L_{bhon}=\begin{bmatrix} 0.9861 & 0.0016 & 0.0045 & 0.0016 & 0.0041 & 0.0020 \\ 0.0038 & 0.9505 & 0.0438 & 0.0019 & 0 & 0 \\ 0.0018 & 0.0211 & 0.9718 & 8.81e^{-04} & 0.0044 & 0 \\ 0 & 0 & 0.0571 & 0.933 & 0.0095 & 0 \\ 0.0072 & 0.0145 & 0.0435 & 0.0036 & 0.9239 & 0.0072 \\ 0.0566 & 0.0031 & 0.0031 & 0.0126 & 0 & 0.9245 \end{bmatrix}$
After	$A^L_{ahon}=\begin{bmatrix} 0.9623 & 0.0205 & 0.0154 & 0.0017 \\ 0.0309 & 0.9423 & 0.0247 & 0.0021 \\ 0.0196 & 0.0039 & 0.9712 & 0.0052 \\ 0.0179 & 0.0179 & 0 & 0.9643 \end{bmatrix}$

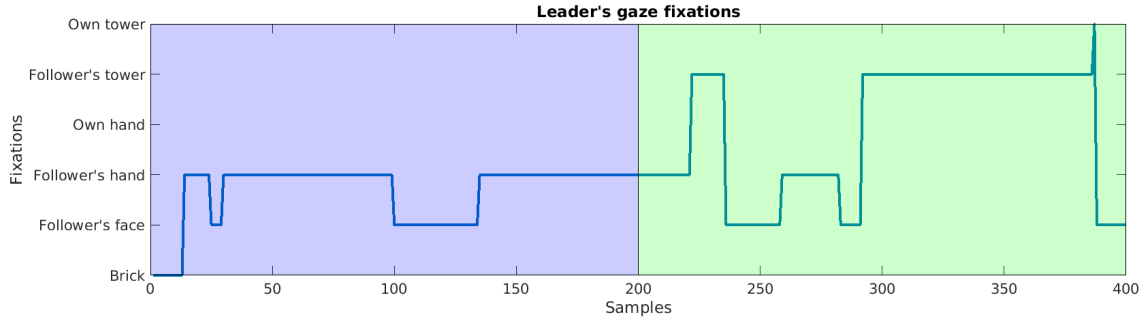


Figure 5.19: Leader's fixations when is applied the DTMC before handover (blue section) and DTMC after handover (green section).

6 - Own tower, before handover; and 1 - TM Face, 2 - TM Hand, 3 - TM tower and 4 - Own tower, after handover. To illustrate the output behavior that can be obtained with the DTMCs, a fixation sequence of 400 samples is generated (Figure 5.19), the first 200 samples using the DTMC before handover and 200 samples using the DTMC after handover. Figure 5.19 show that the fixations before handover are the brick, follower's face, and hand. After the handover, the fixations are the follower's face, hand, and tower, with very short fixation of the own tower. The leader's fixation are given in the top image of Figure 5.16.

Human Action Understanding

Referring to Figure 5.17, the robot (agent P_1) has access to the fixations of the human (agent P_2) which are represented as observations $O_2(k) \in \{V_1, ..., V_M\}$. The admissible fixations of the human are denoted by $V_1, ..., V_M$. The type of action is inferred from the HHI data of the follower's gaze fixations, by calculating the (average) empirical probabilities for *giving* versus *placing* conditioned to the follower's fixation, see Table 5.6.

When the follower looks at the leader's face, the probabilities for *giving* and *placing* are respectively 49.5% and 50.5%, meaning that it is not a strong cue for the action. Instead, when the follower looks at the leader's hand or at his own tower, it signals that the follower

Table 5.6: Average probabilities for the *giving* and *placing* actions, with respect to the follower's gaze fixations

	Giving	Placing
Leader's face	0.495	0.505
Leader's hand	0.617	0.383
Leader's tower	0.294	0.706
Own tower	0.844	0.156

understood that the leader intends to give him the brick. Finally, if the follower fixates the leader's tower, this is a strong signal that the follower understood that the leader will perform a *placing* action.

To select which action is being performed, the action probability is estimated by combining the information related to the instantaneous follower's fixations, with the past history of that probability. These probability signals are denoted as P_G and P_P, respectively for the *giving* and *placing* actions.

Based on the current instantaneous follower's fixation, the action probabilities from Table 5.6 is used to update P_G and P_P with same approach from Eqs. 5.3a and 5.3b:

$$P_G(k+1) = (1-\alpha)P_G(k) + \alpha\delta(k)$$

where k refers to time, and $\alpha = 0.05$. The update $\delta(k)$ depends on the values of Table 5.6, evaluated with the instantaneous follower's fixations. If the follower is currently fixating the leader's hand, and the *giving* action is selected, P_G is updated with $\delta(k) = 0.617$, and P_P is updated with $\delta(k) = -0.617$. If the *placing* action is selected, P_G is updated with $\delta(k) = -0.383$, and P_P is updated with $\delta(k) = 0.383$. This mechanism ensures a smooth evolution of the action probabilities and filters out spurious noisy measurements.

An example of human fixation, and the output of action understanding block are given in Figs. 5.21 and 5.22. In Figure 5.21, the human is engaged in the action and the probability of *giving* is always higher than the probability for *placing*. However, in the second example, during a certain period of time, the human fixates the leader's tower, communicating that he is understanding that the agent will perform a *placing* action. In this period, the probability for *placing* grows, until the human switches the fixations to the agent's hand or its own tower. The second example will illustrate on-line alignment of the leader's action planning from the follower's gaze cues.

5.6.3 Robot Experiments

The iCub robotic platform [Metta et al., 2010] is the robot chosen for these experiments. It is the same experimental scheme presented in Section 5.12 with the objective of tracking the gaze fixations of the human as a follower, while (s)he interacts with the robot. A Cartesian-based gaze controller [Roncone et al., 2016] was used to control the robot's eyes when fixating 3D coordinate points. The motor control of the torso, arm, hand, and fingers was done with a

minimum jerk Cartesian controller [Pattacini et al., 2010], which is responsible for guiding the movement of the robot to grasp the object, as well as to move the object to the handover location, and return to the resting position.

The human gaze fixations are used as an observation and the robot's fixations as the current state. Using DTMC and the prediction algorithm explained in Section 5.6.1, it is possible to predict the follower's gaze fixation and the executed action. The predictions are fed into a planning and control block. The posterior decoding algorithm is executed in order to control the alignment of the leader's fixations. The follower's predicted action is used to plan the leader's action. This information is also used to determine which DTMC model will be used in the next iteration for both the follower's action and fixation predictions, and for calculating the posterior decoding for the leader's fixation.

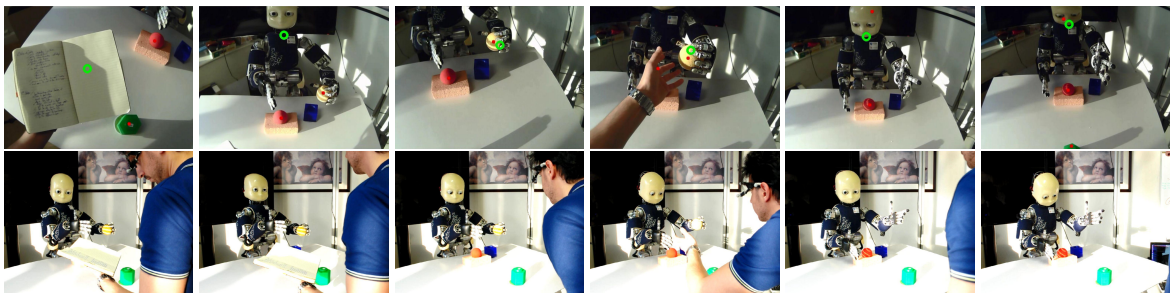


Figure 5.20: A robot interacting with a human initially disengaged from the interaction. The green hallow circle is the human gaze fixation. The gaze of the human can be classified as looking at relevant cues or outliers otherwise.

Figure 5.20 shows a robot performing a *giving* action. The HRI experiment starts with the human not attending to the robot, and looking at his notebook, i.e. outlier.

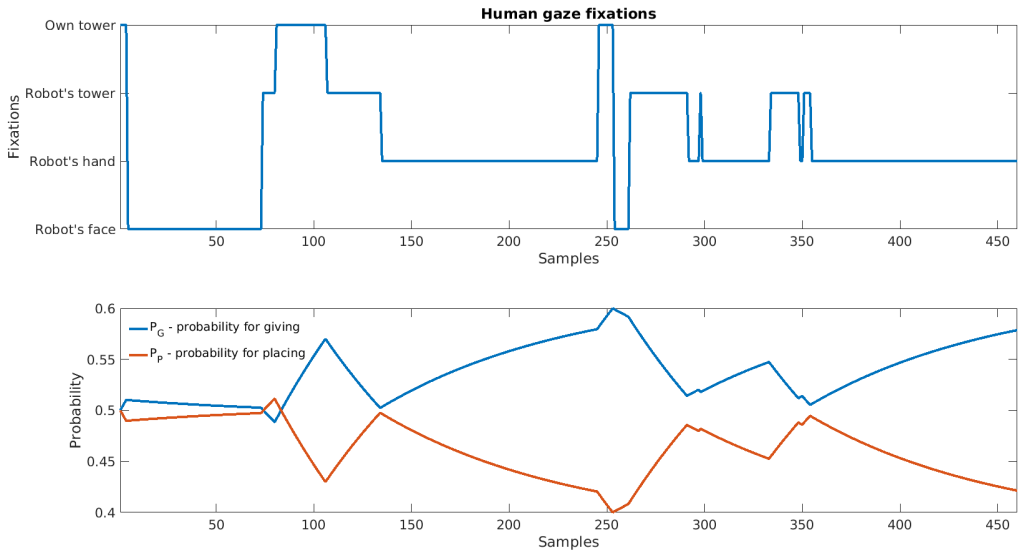


Figure 5.21: On the top is human gaze fixations during the HRI experiment. On the bottom is the prediction of the understood action.

During that time, the robot is continuing the non-verbal communication described in Section 5.6.2. This is an attempt of reaching action alignment with the human through the robot's gaze behavior. Since the robot does not get any information from the human, i.e. no

important cue provided by the eye tracker, the robot assumes the human did not yet understand the interaction intention, and will not complete the *giving* action. After the robot manages to catch the attention of the human, i.e. the human is looking at important cues of the interaction - states S_2 of the gaze behavior - the robot realizes the human understood the interaction intent, and proceeds to complete the handover action, see Fig. 5.21. The top image in Fig. 5.21 shows the human gaze fixations the moment the human starts looking at valid gaze cues. This is translated into the robot predicting the human understanding (Figure 5.21, bottom image).

In the second experiment the alignment of the robot is tested when the human misunderstands the action.

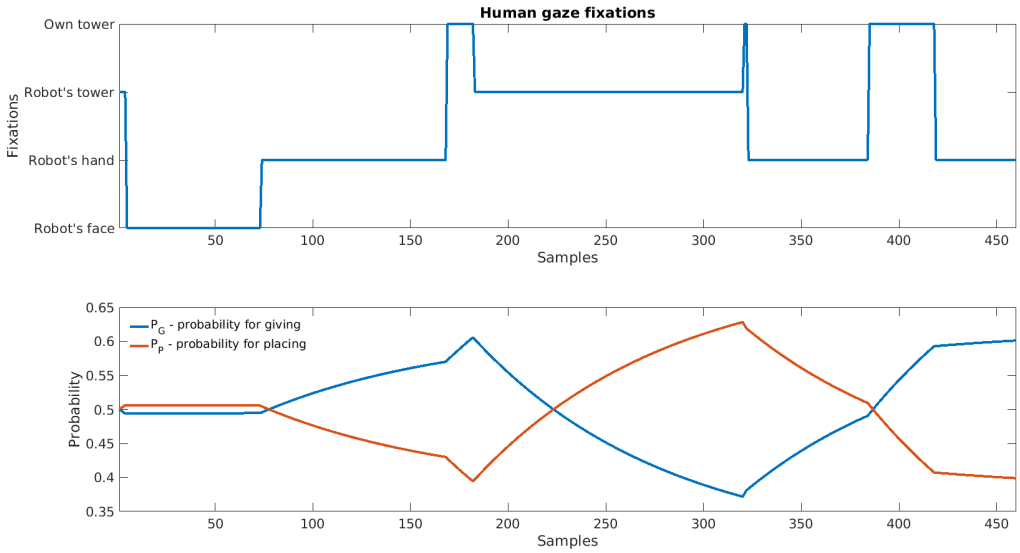


Figure 5.22: On the top is human gaze fixations for the HRI experiment. On the bottom is the robot predictions of the human action.

Figure 5.23 shows the human initially looking at the robot's face and hand. This implies that the human understands the on-going action, as it is seen from the action prediction outcome in Figure 5.22.

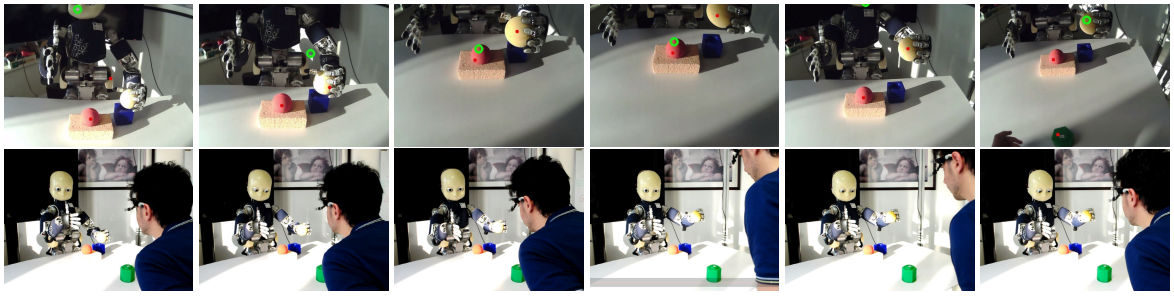


Figure 5.23: A robot interacting with a human that misunderstands the robot's action. The interaction starts with an engaged human on the correct action, then the human misunderstands the robot's action, and hence, mutual alignment is broken. A video showing the interactions is available in video.eccv.2018

The human then switches to fixate the robot's tower. This changes the prediction of the robot, concerning what the human understands, to a *placing* action. This results in the robot retracting the arm, signaling that there is no action alignment, and that the interaction needs to

adapt. The human then looks again at the robot's face and hand, giving the robot the correct prediction of the action. The robot resumes the interaction and finally hands over the object.

5.7 Remarks

This work describes a model of the stochastic gaze behavior of a leader, in a leader-follower social interaction. The gaze fixations are used as an instrument for non-verbal communication, to achieve transparency of the intended actions of an artificial agent. Simultaneously, the agent also reads the human partner's gaze cues to understand the action (s)he performs. Based on this feedback, an agent can plan its motion to align its behavior to the current conditions of the social interaction. The proposed models for gaze behavior and action understanding were integrated in the iCub's robot controller and validated in a HRI scenario with a human in the loop. The iCub's gaze behavior was modeled with two discrete-time Markov chains, to drive the gaze before and after handover. The outcome of the models correlates to the analysis obtained from the HHI experiment data.

Inferring the level of understanding of the action by a human is also based on the HHI experiment data. From these data, an instantaneous probability of the two types of action (*giving* and *placing*) is built. These instantaneous probabilities integrated over time, are used to decide if the human understands the robot's action. Our experiments illustrate how the understanding of the action changes from the correct to the wrong action, and back again to the correct one. When the inferred action is misunderstood, it signals the robot to stop moving the arm toward the handover location, and to go back to the resting position. During that period, the gaze behavior continued to emit cues to communicate the intention of the interaction.

6

Motor Contagion in Human-to-Robot Handovers

> *Reading furnishes the mind only with materials of knowledge; it is thinking that makes what we read ours.*
>
> John Locke,

Contents

Human interaction involves very sophisticated non-verbal communication skills like understanding the goals and actions of others and coordinating our own actions accordingly. Neuroscience refers to this mechanism as motor resonance, in the sense that the perception of another person's actions and sensory experiences activates the observer's brain as if (s)he would be performing the same actions and having the same experiences Natale et al. [2014].

6.1 Introduction

When humans perform actions which involve sharing objects between one another, coordination and understanding are pivotal factors in a successful interaction. A core element in interaction situations is the need and the means of expressing intent. Intent can be communicated directly by comprehensive vocalized sentence or encoded as non-verbal cues through the body, head, or eye movements. Research in corticomuscular and intermuscular coherence in humans report the advantages of non-verbal communication in such interactions. For instance, the mirror neuron system found in humans (and other primates) may have the fundamental function of enabling the preparation of an appropriate complementary response to an observed action. It may explain how two individuals can become so attuned to cooperating in joint actions [Rozzi and Coudé, 2015]. Synchronisation in motor coordination [Pesce Ibarra, 2017] is seen as a biological condition in order to improve efficiency and reliability in HHI. Moreover, synchronisation between two agents is preferred to two individuals systems, to achieve optimal motor control. Further research on psychology [Nowak et al., 2017], cognition, and neuroscience [Hu et al., 2017], reinforces on the idea that social interaction adheres from synchronisation of lower-level elements [Bassetti, 2017].

The contributions of this chapter are threefold: (i) two computational models that describe the behavior of "giver" and "receiver" for handover actions; (ii) a computational model that represents the human-human coordination; (iii) an integrated model implemented into a robotic controller for handover action recognition, and execution, allowing for human-robot coordination.

This chapter begins with the same HHI scenario of previous chapter. A turn-taking game of dyads in action-in-interaction (Appendix B) where the wrist data of both participants is extracted. The DS are formulated representing reaching motions such as the "giver" and "receiver" during a handover. From the data, two computational frameworks employ a state-dependent, time-independent, DS: one corresponding to the person performing the handover (the "giver"), and another to the recipient of the object (the "receiver"). A process of coupling the two DS originates a Coupled Dynamical System (CDS), which relates both participants of the handover. The developed architecture was incorporated in the iCub humanoid robot for a HRI scenario.

6.2 Methodology

Coupled Dynamical Systems

Using CDS enables the integration of two independent DS, learning the coupling function between them ([Shukla and Billard, 2012]). This coupling behavior takes inspiration from the biological studies on motor synchronisation of reach-grasp coupling [Mitz et al., 1991], as well as the coupling between humans [Mörtl et al., 2014].

Let $\xi_x \in \mathbb{R}^{d_x}$ and $\xi_f \in \mathbb{R}^{d_f}$ denote the states of two independent DS. The CDS consist of a master-slave system:

$$\mathcal{P}(\xi^t{}_n, \dot{\xi}^t_n | \theta^g{}_{master}) \tag{6.1}$$

where the master sub-system is the DS of the first motor dynamics. After encoding the master, the next step is to infer the state of the slave conditioned on the master:

$$\mathcal{P}(\Psi(\xi^t{}_n), \xi^t{}_n | \theta^g{}_{coupled}) \tag{6.2}$$

$\Psi : \mathbb{R}^{d_m} \rightarrow \mathbb{R}$ is the coupling function which is a function dependent on the master state. Equation 6.2 allows to encode the dynamics of the slave sub-system:

$$\mathcal{P}(\xi^t{}_n, \dot{\xi}^t_n | \theta^g{}_{slave}) \tag{6.3}$$

$\forall g \in \mathcal{G}$ are the function parameters. θ_{master}, θ_{couple}, and θ_{slave} denote the GMM parameters for each respective sub-system. $\Psi : \mathbb{R}^{d_x} \rightarrow \mathbb{R}$ is the coupling function which is dependent and monotonic on the master state which satisfies $\Psi(\boldsymbol{\xi}^*_x) = 0$.

[Lukic et al., 2014] applied the CDS approach to couple the movement of the human eyes with the arm and the hand when performing point-to-point moving of objects with obstacle avoidance. A CDS with: 3 DS corresponding to the dynamics of the eyes, arm, and hand, and 2 coupling functions to infer the position of the arm with respect to the movement of the eyes (eyes-arm), and to infer the finger configuration of the hand with respect to the location of the arm (arm-hand). In order to achieve human-human motor coordination it is necessary to couple the arm movement of two humans in a HHI involving collaborative tasks. Figure 6.1 presents the CDS architecture for the action-in-interaction task. The CDS architecture of [Lukic et al., 2014] is extended by applying it to an interaction scenario: the first CDS is for the agent that is performing the action-in-interaction, i.e. the leader of the action, and the second CDS is for the agent that is observing the action and participating in the action-in-interaction, i.e. the follower of the action.

It is argued that each agent's CDS has an internal model defined by the intrapersonal coordination of [Lukic et al., 2014]. When interacting with other agents, an external system (the **Coupling System**) performs the motor communication and coordination required for a successful interaction. The end-effector of Agent 2, the "receiver", is then conditioned on the

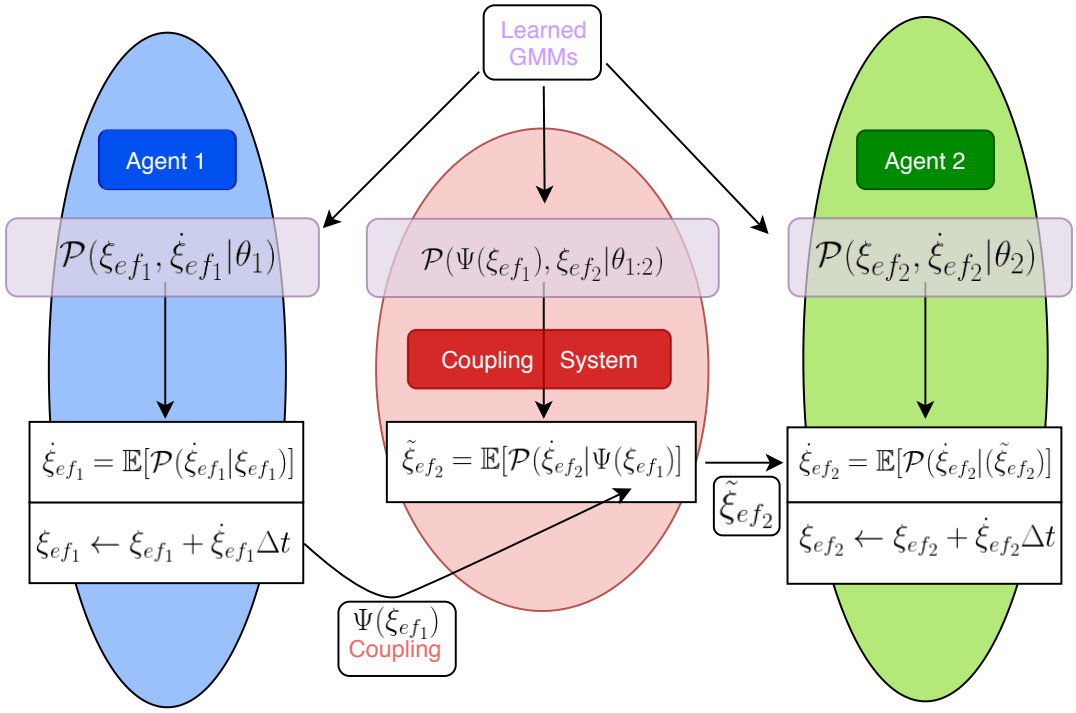


Figure 6.1: Diagram of Human-Human Collaboration System.

end-effector of Agent 1, the "giver":

$$\mathcal{P}(\xi_{ef_1}, \dot{\xi}_{ef_1}|\theta^g{}_1) \tag{6.4}$$

$$\mathcal{P}(\Psi(\xi_{ef_1}), \xi_{ef_2}|\theta^g{}_{1:2}) \tag{6.5}$$

$$\mathcal{P}(\xi_{ef_2}, \dot{\xi}_{ef_2}|\theta^g{}_2) \tag{6.6}$$

where Equation 6.4 represents the end-effector of the intrapersonal coordination of agent 1, which encodes the dynamics of the end-effector and generates the end goal of agent 1. This state is given to the coupling sub-system given by Equation 6.5 and applied to the coupling function $\Psi(\xi_{ef_1})$ in order to infer the state of agent 2. In return agent 2 end-effector position is updated following the dynamics present in Equation 6.6.

6.3 Modelling Human-Human Collaboration System

6.3.1 Dynamics of each Agent

Let ξ_{ef_1} be the "giver"'s right wrist, for simplicity ξ_1, and ξ_{ef_2} the "receiver"'s right wrist as ξ_2. The wrist trajectories for the handover action are shown in Figure 6.2a. $\xi_1 \in \mathbb{R}^d$ and $\xi_2 \in \mathbb{R}^d$, where $d := \{p, h\} = 2$, for proximity and height, respectively. The p-proximity dimension is computed as the euclidean norm of Cartesian coordinates x and y, $\|$x+y$\|_2$. The h-height coordinate is the Cartesian coordinate z.

The *attractor point* of the DS, i.e. the origin point in Figure 6.2b, is the handover meeting

point computed as the average point from all the human trajectories. To compute the GMM parameters the stable estimator of dynamical systems (SEDS) approach [Khansari-Zadeh and Billard, 2011] is used, since it ensures assymptotic stability at the *attractor point*. To generate new trajectories from the learned GMMs, GMR takes a sample t from $\mathcal{P}(\boldsymbol{\xi}^t, \dot{\boldsymbol{\xi}}^t|\theta)$ to provide the desired output.

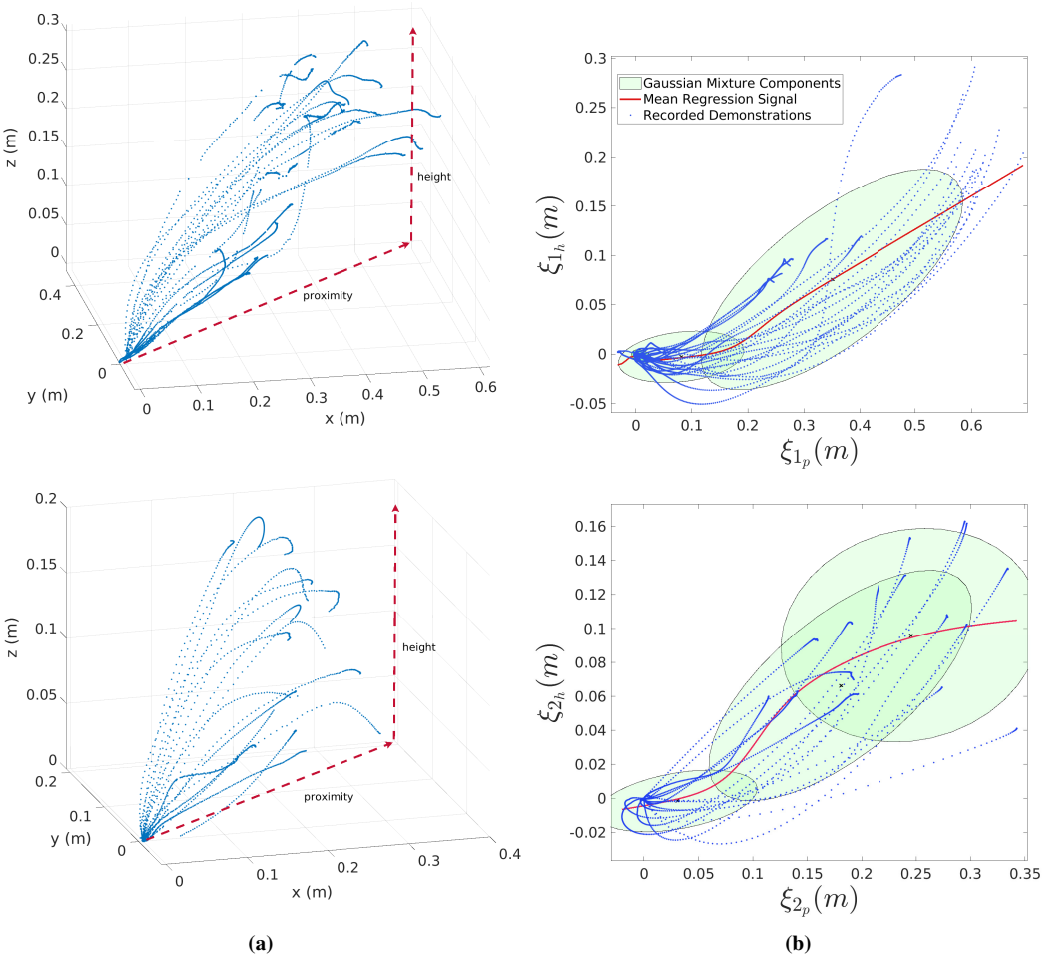


Figure 6.2: The "giver" (top), and "receiver" (bottom) respective learned DS from demonstrations. (a) shows the recorded demonstrations; (b) shows the GMM encoding the desired value of $\xi_{x_h}, \forall x \in \{1, 2\}$, for the height axis (red dashed line), given the current value of $\xi_{y_p}, \exists y = x$, for the proximity axis (red dashed line), as observed in the demonstrations.

6.3.2 Coupling between Agents

The reason for coupling both humans is bolstered by psychologists and neurobiology scientists as an indispensable factor for social interaction [Kelso et al., 2013, Burgoon and Kendon, 1992, Feldman, 2007]. From the analysis of the HHI data it can be concluded that there are only two dimensions of interest that can be extracted from the data: (i) how far away the arms are from each other, (ii) the difference of height. This dimensionality reduction is possible due to the configuration of our HHI experimental setup. In all experiments, seen in Figure 6.3, the dyadic participants were facing each other on opposite sides of a table. As a

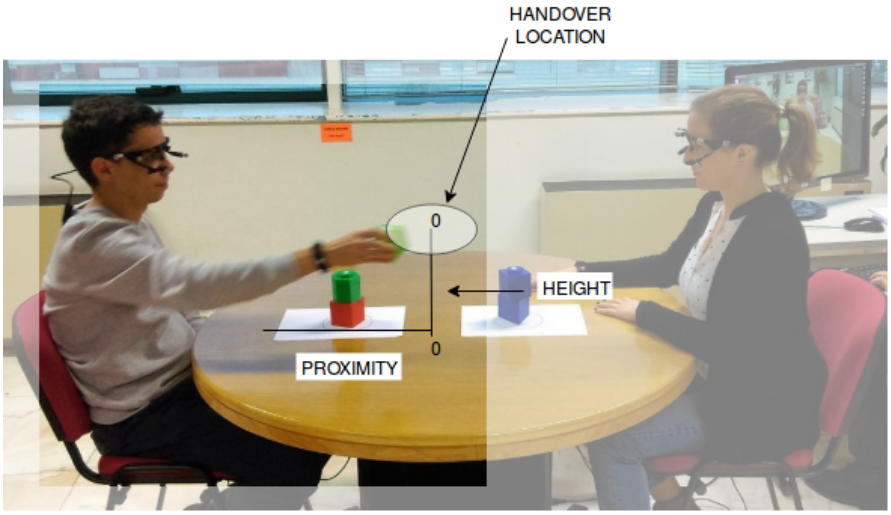


Figure 6.3: Illustration of the coordinate variables y and z as distance and height to the handover location, respectively.

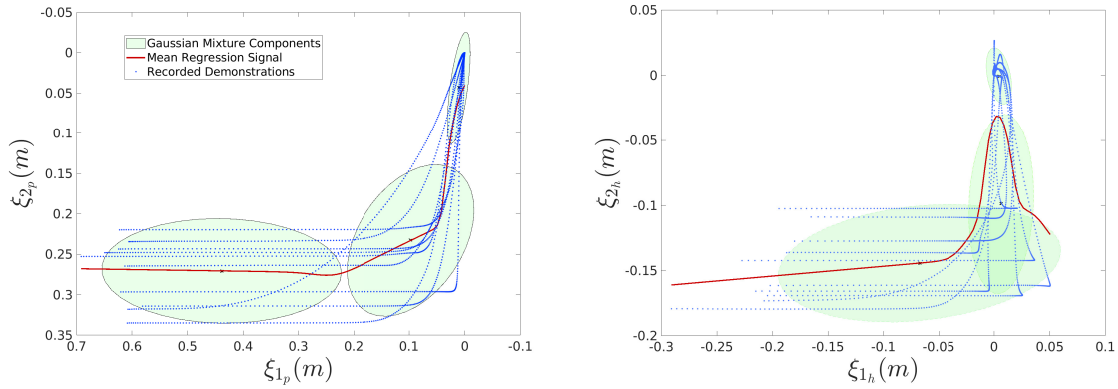
result, the movements of reaching and passing objects were directed forward. The dimension, which can be described as perpendicular to Figure 6.3's image plane, could be removed from the wrist data, reducing the complexity of the problem. Let $\Psi \in \mathbb{R}^2$ be the coupling function that relates ξ_2 conditioned on ξ_1 over $\{p, h\}$ coordinates. The coupling is applied according to

$$\Psi(\xi_{1_p}) = \| \ x + y \ \|_2$$

and

$$\Psi(\xi_{1_h}) = z$$

Figure 6.3 shows an experimental representation of the coordinates used for coupling the motion of the leader and follower's wrist in the handover action.



(a) **(b)**

Figure 6.4: Coupling between "giver" and "receiver" wrists in: (a) the proximity axis $\mathcal{P}(\Psi(\boldsymbol{\xi}_{1_p}), \boldsymbol{\xi}_{2_p}|\theta_p)$, and (b) the height axis $\mathcal{P}(\Psi(\boldsymbol{\xi}_{1_h}), \boldsymbol{\xi}_{2_h}|\theta_h)$, towards the handover meeting point.

Figure 6.4 reveals the correlation between agent's wrist motion for p and h coordinates, respectively. For the s, SEDS is the approach chosen for ensuring asymptotic stability in the

Coupling System. From analyzing the coupling functions some details are relevant. Firstly, the influence of the "giver"'s wrist motion upon the "receiver" is stronger when close to the handover meeting point. Secondly, the h coordinate, i.e. the height of the arm, has a significant impact in closer distances than the p coordinate. Meaning the "receiver" arm's height is only influenced by the "giver"'s motion when the "giver"'s arm is reaching, closely (less than 10 centimeters away), the handover location. Thirdly, for considerable distances (larger than 20 centimetres), the p coordinate can give an indication on the type of action. This might have to do with the setup of the HHI scenario which involves handing over objects to another participant or place it on your own tower near you, i.e. tower was closer to you than the other person.

6.3.3 Alternative Approach to Coupling between Agents

An alternative approach alters the way "giver", and "receiver" are viewed. The dynamics of master-slave does not suit a context of human-human coordination. What is learned from literature is that humans synchronize their movements [Sisbot et al., 2010], and it happens as well between humans and robots [Mörtl et al., 2014]. As such, a different approach must be considered in order to achieve synchronization between the two sides:

$$\mathcal{P}(\xi_g, \dot{\xi}_g|\theta_g) \tag{6.7}$$

$$\mathcal{P}(\Psi(\xi_{couple}), \xi_{couple}|\theta_{couple}) \tag{6.8}$$

$$\mathcal{P}(\xi_r, \dot{\xi}_r|\theta_r) \tag{6.9}$$

where $\mathcal{P}(\xi_g, \dot{\xi}_g|\theta_g)$ is the dynamics of the "giver", and $\mathcal{P}(\xi_r, \dot{\xi}_r|\theta_r)$ is the dynamics of the "receiver". However, the coupling system is defined by $\mathcal{P}(\Psi(\xi_{couple}), \xi_{couple}|\theta_{couple})$, where the variable $couple$ is the relation between "giver" and "receiver"'s wrist position. $\Psi(\xi_{couple})$ denotes the coupling function defined as:

$$\Psi(\xi_{couple}) = |\xi_g - \xi_r| \tag{6.10}$$

where ξ_g and ξ_r are the positions of the wrist of the "giver" and "receiver", respectively. The absolute distance between wrist positions was chosen since when the distance reaches zero, it means the wrists have reached the point of shortest distance, which is assumed as the handover location. Figure 6.5 illustrates the new dimensions for coupling both agents. For DS and CDS, the equilibrium point is the convergence of the system, in a sense, it makes sense that the convergence point of a human-human coordination system of arm movements for a handover action would be the handover location. Moreover, the DS and CDS are robust to perturbations on the input variable, which is convenient due to the oscillatory behavior of the human arm movement.

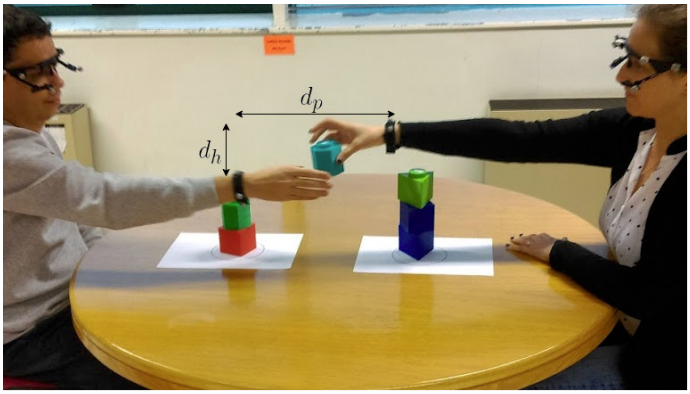


Figure 6.5: The dimensions of the new approach. d_p and d_h are respectively, the distance between wrists, parallel to the floor, and the difference of height, perpendicular to the floor, between wrists.

Since the DS for the "giver" and "receiver" are identical as the previous version in the last section, the focus here is on modelling the coupling behavior between the two using our approach discussed above. From Equation 6.10 the coupling function defined is

$$\Psi(d_p) = \| \, d_h \, \|$$

$d_p = \| \, \xi_{1_p}$ - $\xi_{2_p} \, \|$, where ξ_{1_p} and ξ_{2_p} are the location, parallel to the table, from the handover location to the wrist of the "giver" and "receiver", respectively. $d_h = \| \, \xi_{1_h}$ - $\xi_{2_h} \, \|$, where ξ_{1_h} and ξ_{2_h} are the location, perpendicular to the table, from the handover location to the wrist of the "giver" and "receiver", respectively. The handover location is considered as the final wrist position for the "giver" and "receiver" for each different dyad in each experiment trial. Figure 6.6 shows the learned coupling model. The handover location is not the same for the "giver" and "receiver" since participants may differ in arm's length. Additionally, for the purposes of simplicity, it is assumed that at $d_p = d_h = 0$ the handover takes place.

From the analysis of the coupling model the following can be concluded: the most notable difference compared with the previous approach from Section 6.3 is the single coupling function. In terms of complexity, this approach requires one less computational step to obtain the relation between "giver" and "receiver". The new approach uses the norm difference of the wrist locations (Equation 6.10). The norm is believed to be biologically inspired [Haggard and Wing, 1991]. The other advantage to the previous approach is the fact that it no longer considers the "receiver" as a slave to the "giver" movements. Our new approach considers each to have an impact. This approach has bi-directional usage, in a sense that this coupling function can be used to couple the "receiver" to the "giver" behavior, or the "giver" to the "receiver", depending on which is the one desired to control.

Algorithm for Updating Human-Human Collaboration System

Algorithm 6.1 explains the process Agent 1 $\rightarrow$ **Coupling System** $\rightarrow$ Agent 2 presented in the diagram of Figure 6.1. The process includes generating the next state ξ_1 and velocity

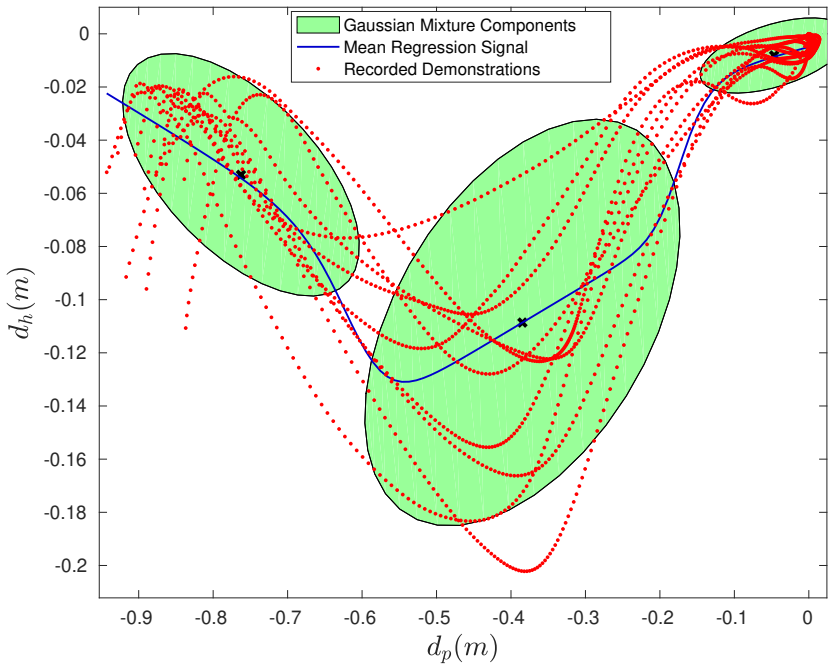


Figure 6.6: Learned CDS between "giver" and "receiver". d_p is the distance between wrists, parallel to the floor, and d_h is the difference of height, perpendicular to the floor, between wrists. The origin is when the wrists are at the nearest distance from the two which it is considered as the handover location.

Algorithm 6.1 CDS for *action-in-interaction*

Input: $\xi_1(0); \dot{\xi}_1^r(0); \xi_2(0); \theta_1; \theta_2; \theta_{1:2} \coloneqq \{\theta^p_{1:2}; \theta^h_{1:2}\}; \Delta t; \epsilon$
Set $t = 0$

while ($\|\dot{\xi}_1(t)\| > \epsilon$ and $\|\dot{\xi}_2(t)\| > \epsilon$) **do**

Update "giver":
$$\dot{\xi}_1(t) = \mathbb{E}[\mathcal{P}(\dot{\xi}_1|\xi_1;\theta_1)]$$
$$\xi_1(t+1) = \xi_1(t) + \dot{\xi}_1(t)\Delta t$$
$$\text{ActRec}(\dot{\xi}_1^r(t), \dot{\xi}_1(t))$$

	First Approach	Alternative Approach
Coupling System:	$\tilde{\xi}_{2_p}(t+1) = \mathbb{E}[\mathcal{P}(\xi_{2_p}\|\Psi(\xi_{1_p});\theta^p_{1:2})]$	$d_p(t) = \parallel \xi_{1_p}(t) - \xi_{2_p}(t) \parallel$
	$\tilde{\xi}_{2_h}(t+1) = \mathbb{E}[\mathcal{P}(\xi_{2_h}\|\Psi(\xi_{1_h});\theta^h_{1:2})]$	$\tilde{\xi}_2(t+1) = \mathbb{E}[\mathcal{P}(\xi_2\|\Psi(d_p(t));\theta^{gr})]$
	$\tilde{\xi}_2(t+1) = (\tilde{\xi}_{2_p}(t+1), \tilde{\xi}_{2_h}(t+1))$	$\tilde{\xi}_2(t+1) = (\tilde{\xi_{2_p}}(t+1), \tilde{\xi_{2_h}}(t+1))$

Update "receiver": $\dot{\xi}_2(t+1) = \mathbb{E}[\mathcal{P}(\dot{\xi}_2|\tilde{\xi}_2;\theta_2)]$
$$\xi_2(t+2) = \xi_2(t+1) + \dot{\xi}_2(t+1)\Delta t$$

$t \leftarrow t+1$

end while

profile $\dot{\xi}_1$ of the "giver"'s motion, this velocity is to be compared with the real "giver"'s velocity $\dot{\xi}_1^r$ for evaluation of the current action (ActRecog explained in Section 8.3.4). To note that the velocity profile $\dot{\xi}_1$ is the generated velocity of the DS model for the learned "giver"'s motion and it is compared with the real (observed) velocity of the human "giver". The state $\xi_1 = (\xi_{1_p}, \xi_{1_h})$ is provided to the **Coupling System** to infer the state of the "receiver"'s wrist $\tilde{\xi}_2$. Following that, it is used to generate the next state ξ_2 and velocity profile $\dot{\xi}_2$ of the "receiver". The cycle is repeated until convergence, i.e. the meeting point has been reached for both agents, to complete the handover. Since the CDS architecture ensures global asymptotic stability, the ϵ convergence parameter can be set as a small number which satisfies that the "receiver" has reached the handover meeting point.

Overall, from Figures 6.2 and 6.4 the conclusions are: (i) the motion of the "giver" and "receiver" are different during human-human coordination of handover actions. It is illustrated in Figures 6.2 (b) that the motion of the "receiver" is more impactful when closer to the handover meeting point, it is hypothesized a reasoning of the "giver"'s intention before interacting; (ii) Figure 6.4 emphasizes this notion and also learns a function that couples the motor movements of the "giver" with the "receiver" for handovers.

6.4 Robot Experiments

In order to evaluate the models of the intrapersonal and the interpersonal coupling motor coordination developed in Section 6.3, experiments of human-to-robot handovers with the iCub humanoid robot are performed. The coupling function selected is the alternative approach as it provides similar results with less complexity.

Experimental Setup

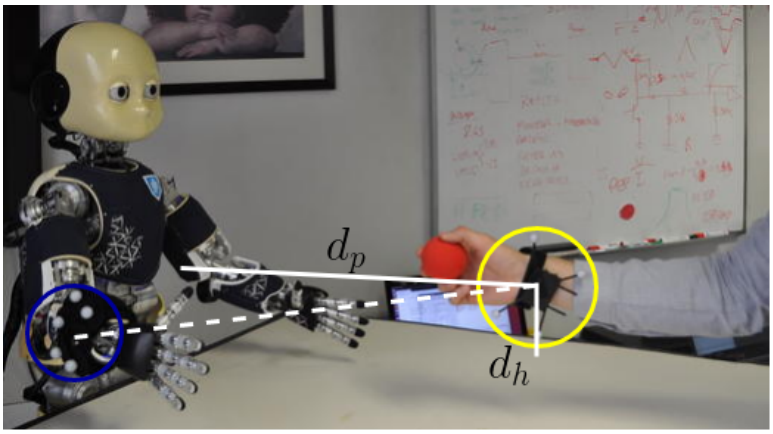


Figure 6.7: Side view of the HRI experiments. The wrists of the robot and human are highlighted in blue and yellow, respectively, to represent the rigid bodies created by the motion capture system.

The setup for the HRI experiments is as follows. A human is giving an object to a humanoid

(Figure 6.7) while the robot receives the object following the coupling behavior mentioned above. The OptiTrack Mocap tracks the arm (i.e. wrists) movements. Two rigid bodies are created from markers: (i) the iCub wrist, (ii) human wrist. These experiments serve to validate the controller's robustness to variability on the human arm movement. The coordinates are sent from the Mocap system and streamed through a Lab Streaming Layer [Kothe] as a YARP port to the humanoid robotic controller.

Robot Controller

The iCub module *Cartesian Interface* [Pattacini et al., 2010] computes the inverse kinematics and joint configuration of the robot to reach the desired Cartesian coordinates given as input. The DS of Agent 2 updates the "receiver"'s wrist ξ_2 (Algorithm 6.1) for p- and h- coordinates which are then sent to the robot controller as the x coordinate, z coordinate, respectively, in the robot reference frame (iCub torso). The robot y coordinate is taken as the y coordinate of the human's wrist, allowing for generalization in the handover location.

Meeting Point Definition

The following assumptions are made during our experiments. Since the dataset contains only information regarding the wrist it is predefined the hand aperture and grasp orientation for the handovers. In order to translate the wrist information to the architecture of our controller, the meeting point needs to be defined. It was set as the furthest region an iCub can reach for a safe and successful grasp without rotating its torso. These assumptions aim to give a one-to-one comparison with the HHI experiments.

Action Recognition

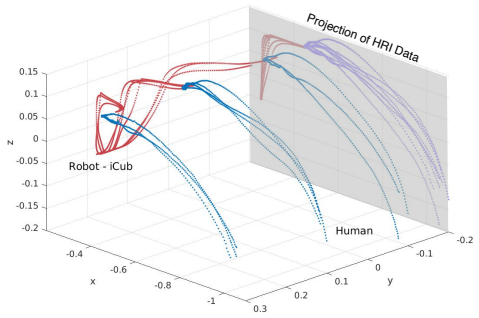
To recognize a handover action from a human the error between the current velocity $\dot{\xi}_1^r$ and the DS generated velocity $\dot{\xi}_1$ is computed, adding the velocity profile direction:

$$\text{ActRec}(\dot{\xi}_1^r(t), \dot{\xi}_1(t)) = - \parallel \dot{\xi}_1^r(t) - \dot{\xi}_1(t) \parallel \text{sign}(\dot{\xi}_1^r(t)) \tag{6.11}$$

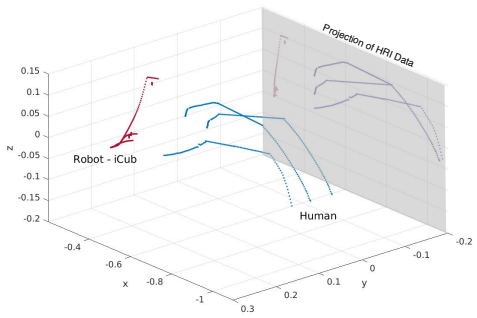
The last term in Equation 6.11 aims at detecting the correct velocity of the wrist. From the velocity direction, it is possible to extract information whether the human is approaching the meeting point or moving away from it.

Results

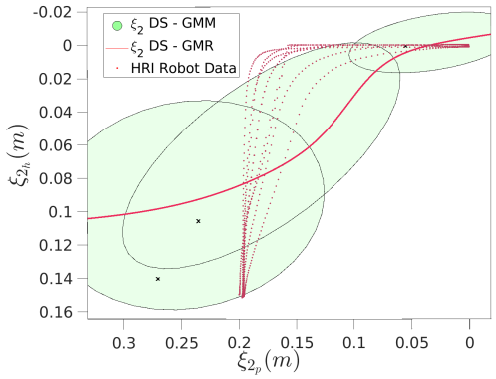
Figure 6.8 shows the results for handover and placing actions where the human is the "giver" and the robot the "receiver". It can be seen from Figure 6.8 (a) that the handover actions were completed successfully. In more detail, (c) and (d) show the projection of the data to the



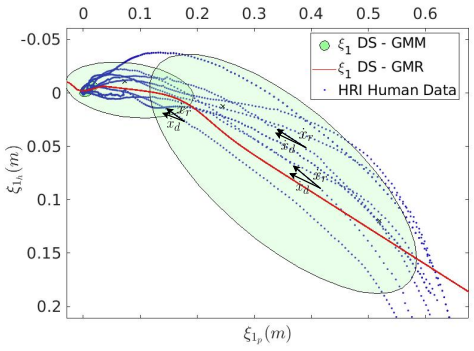
(a) Human-to-robot handovers. Red lines are the iCub wrist trajectories, Blue lines are the human wrist trajectories.



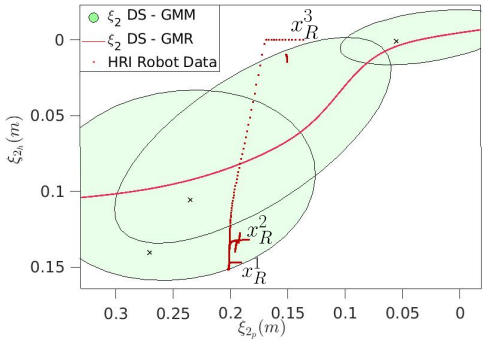
(b) Human performing different types of no Handovers, i.e. placing objects at different places.



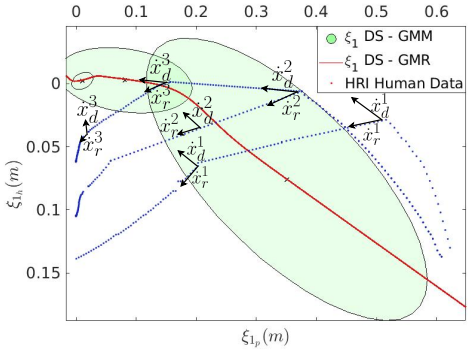
(c)



(d)



(e)



(f)

Figure 6.8: HRI experiments involving a human handing over objects to the iCub or placing them on a table. (c), (d), represents the projection of the HRI data for actions in (a), while (e) and (f) are the projection data for the actions in (b). (c) and (e) compares the robot data with Agent 2's DS, while (d) and (f) compares the human data with Agent 1's DS. (d) and (f) represents the velocity profiles from the real data $\dot{x}_r$ and the generated velocity streamlines from the DS $\dot{x}_d$. (e) and (f) have the trajectories labelled for each different trial, e.g. x_R^1 is the robot's response to the human trajectory $\dot{x}_d{}^1$. The HRI experiments are demonstrated in the complementary video.roman.2019.

respective agent's DS. It shows that the robot follows the coupling function and behaves like a "receiver" as seen in (c) when the human hands over an object. However, for cases where the human behaves differently from what was observed, placing the object instead (Figure 6.8 (b)), the robot can detect and adapt to it. This can be seen from the projections in (e) and (f), which present three different cases of no handovers. The trajectories presented mark the real velocity as x_d^1 and the respective desired velocity from the "giver"'s DS. In the first example of a no-handover, it is quickly detected that the human is not performing a handover, as the real

velocity is to an opposite direction to the desired. The result is a detection of a no-handover from Equation 6.11 which stops the **Coupling System** from generating a new desired location for the robot (x_R^1), moving the robot back to its initial position. In the second example the human performs an action presenting more similarities to a handover. However, from the **Coupling System** the effect is minimized and from Equation 6.11 the incongruities between the real x_r^2 and desired velocity x_d^2 are detected. As an extreme example, the third case was chosen to test the architecture to its limits. It shows the human trying to fool a handover right until the last moment before finishing the movement. This results in the **Coupling System** to update the robot as a "receiver" (x_R^3) until the human alters its trajectory and places the object. These sudden movements cause the action to not be recognized as a handover, i.e. ActRec($\dot{x_d^3}$, $\dot{x_r^3}$)< 0, stopping the motion update and returning the robot back to the initial position. This, of course, it is not desirable, but the goal of this example is to explore the robustness of the architecture to very extreme scenarios of false handovers. To reiterate, the cases that the action is considered a no handover is, when throughout the wrist motion, the "giver"'s velocity profile $\dot{\xi}_1$ is contradictory of the DS "giver" output for a handover.

As for the opposite case, e.g. the human begins by exhibiting a behavior that resembles the placing of an object onto a table and in the middle of the action decides to hand it over to the robot, our method is also capable of adapting the robot's response. This is possible due to the continuous update of Algorithm 6.1 in recognizing the human action from the arm trajectory and adapting the robot's behavior accordingly.

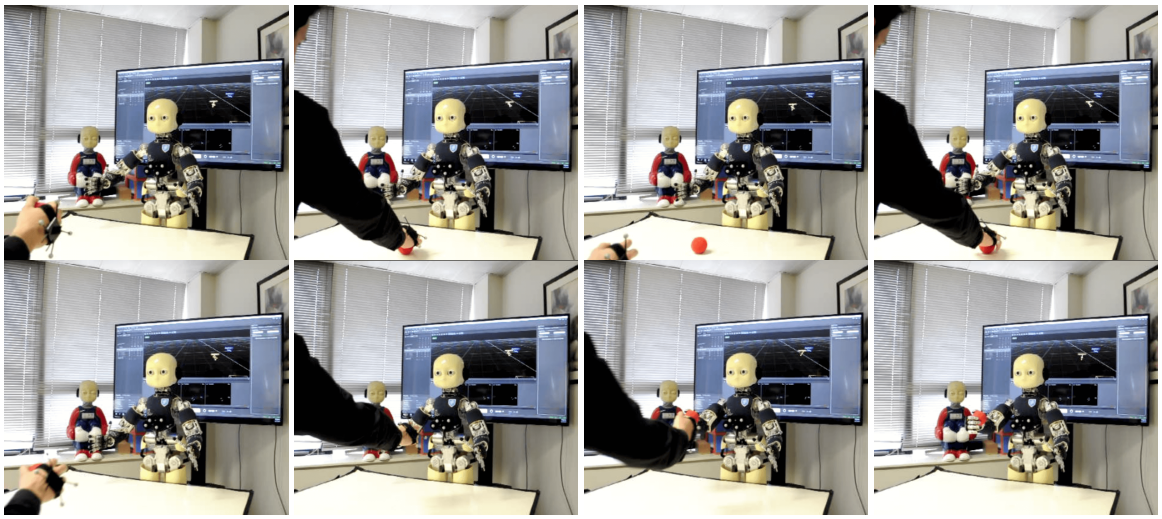


Figure 6.9: HRI experiment involving a human handing over an object to the iCub. This experiment exemplifies the adaptability to human behavior. The human begins by placing the object in front of the robot and, due to the coupling functions, the action is not recognized as handover, so the robot does not interact with the human. Only when the action is recognized as a handover does the robot behave to receive the object. The HRI experiments are demonstrated in the complementary video video.roman.2019.

6.5 Final Remarks

This is an approach for learning an interpersonal coordination model from human demonstrations. The first contribution involves modelling the human kinematic motion during handover of objects. The second contribution aims at modelling the human-human kinematic coordination during a shared action such as a handover. A coupled-DS was adopted to model the intricacies between human wrist motions during a handover movement. Two separate coupling functions were computed from human demonstrations which allows for a detailed understanding of human action intention. The third contribution focused on implementing a controller for a robotic humanoid robot in order to: (i) understand the handover intention from the human wrist motion, and (ii) control a humanoid to behave as observed during the human-human scenarios. Our results support the hypothesis that the robot expresses "human-like" motion during a handover action with a human. Moreover, the developed architecture allows the robot to understand the action intention of humans and decode the handover action. Figure 6.9 shows one of the experiments performed with the robot where the coupling comes into play. During the experiment the human performs three actions: placing an object in front of the robot, picking up the same object, handing over the object to the robot. From the understanding of the behavior of the arm motion in handover actions, it is possible to discriminate between non-handover actions, and handover actions. The internal model of the robot is responsible for understanding the intention of the human, and the coupling model is responsible of adapting the motion of the robot to the observed behavior. The two coupling systems have two distinct ways of coordinating two agents. The first approach defined each agent has an individual model with an additional coupling model that takes information from one (the giver) and sends the desired coupling behavior to the second (the receiver). The second approach is a simplified version of the former where only the coupling model is present. The former only acts when a correct giver's motion is recognized, while the latter approach assumes a correct giver's motion and always adapts the receiver's. The latter approach is less complex with fewer steps to calculate the receiver's motion. Hence it is trade-off between speed and efficiency (the second) vs complexity and accurateness (the first).

Part III

Inferring Object Properties

7

Identify liquid fullness in cups from human *gaze cues*

> *It's not what you look at that matters, it's what you see.*
>
> Henry David Thoreau,

Contents

For collaborative tasks, involving handovers, humans are able to exploit visual, non-verbal cues, to infer physical object properties to modulate their actions. In this chapter, the liquid level of cups is the case used to explore the human non-verbal signals emitted when handling the cups in collaborative scenarios. It is our common experience that transporting a container (such as cups, glasses, or mugs) filled with some liquid is much more delicate than when it is empty. [Mayer and Krechetnikov, 2012] put this common knowledge to the test by examining people walking with a mug filled with coffee. They have found that humans try to avoid spilling the content by either estimating the frequency of sloshing of the liquid (moving the hand so as to counteract the induced slosh), or by slowing down and adopting a more careful manipulation. The choice between these two strategies seems to be related to individual preference. When it comes to programming similar skills in robots we argue that it would be best to choose the latter option, as it will be the most effective to prevent accidental spills.

7.1 Introduction

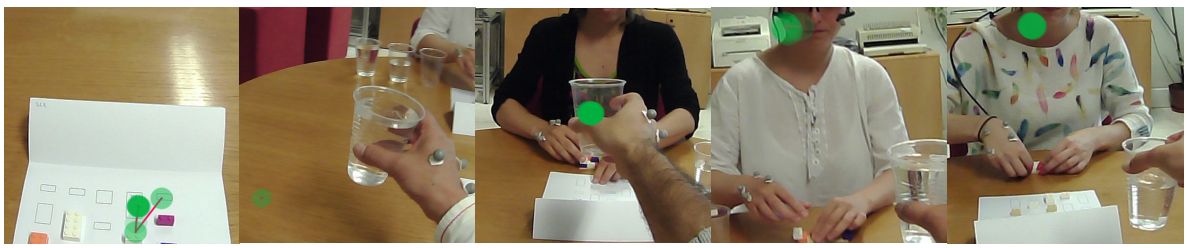
Perceiving the physical characteristics (e.g. mass) of objects manipulated by others is often important to prepare our own actions, as during the handover of heavy objects. Humans can infer such properties, even when they are not visually observable, through the analysis of the motor behavior of another human handling an object [Kjellström et al., 2011]. Understanding the existence of water, or any liquid, in a cup has been a challenging problem in computer vision and robotics. There have been attempts at training large neural networks to classify the level of liquid from a single RGB image [Mottaghi et al., 2017, Modas et al., 2021], or RGB-depth cameras [Do et al., 2016, Do and Burgard, 2019, Schenck and Fox, 2017b]. Alternatively, other approaches required pouring liquid in a cup to detect the liquid level [Schenck and Fox, 2017a,b, Do and Burgard, 2019, Yu et al., 2015]. However, some challenges are extremely difficult to handle, such as occlusions, transparency of the liquid, different types of cups, colors, and the most important case, opaque cups. Most existing approaches struggle with opaque cups as you might not get the chance to view the cup from an advantageous angle that allows to view the liquid inside [Mottaghi et al., 2017, Do et al., 2016], or get the robot to manipulate the cup prior [Do and Burgard, 2019, Schenck and Fox, 2017b,a, Yu et al., 2015]. In most cases the cup or object is handed to the robot without any prior knowledge. The problem tackled here is different from before.

The aim is not through direct visualization of the cup but through observation of human motion. The proposal is a novel perspective that takes into account the human side of the equation. Experiments have shown that human subjects are capable of estimating the weight of an object, through the observation of other people lifting objects with different weights [Alaerts et al., 2010a]. [Wei et al., 2018] used human gaze direction to infer the action being performed. I argue that it is possible to understand the fullness level of a cup, to a certain extent, by observing others manipulating it. This proposal aims at studying the human-to-human

handovers of cups with different water levels [Sanchez-Matilla et al., 2020] and explore the non-verbal gaze cues shared by humans during manipulation. Humans tend to fixate the gaze direction in the regions that are most relevant concerning the executed action [Flanagan et al., 2013]. Therefore, exploring the eye-gaze cues should provide us with the relevant information for classifying water level in cups.

7.2 The Fourth Experimental Setup

The experiment consists of two people sitting on opposite sites of a table and completing a set of instructions hidden in a puzzle set provided to each one of the participants. This puzzle, which can be seen in Figure 7.1 (a) on the bottom region from the point of view perspective, has a set of LEGO® pieces that are to be picked up, one by one, and beneath specific pieces, there are instructions to manipulate the available cups. On each side of the table there are 3 identical cups but with three possible levels of water inside: (i) empty, (ii) 50% full, and (iii) 90% full; which for simplicity is referred as empty, half, and full cup. The action instruction involves manipulating one of the 3 different cups as follows: (i) to grasp it and move it from the initial position (the right of the puzzle) to the left side of the puzzle (final position), or (ii) grasp it and hand it over to the other participant in front of the table. Figure 7.1 (b) shows an example of action (i) and Figures 7.1 (c)-(e) of action (ii). The instruction indicates the type of action and which cup to manipulate. The experiment is finished when both participants pick all the puzzle pieces, building a structure in the process, and all the actions are fulfilled. The pair of participants repeat the experiment a total of 5 consecutive times however the location of the action instructions changes in each repetition. The participants did not know, beforehand, which of the pieces contained an action instruction and the number of actions changes for each trial. This is to prevent any anticipatory behavior by the participants.



(a) (b) (c) (d) (e)

Figure 7.1: HHI experiment: video frames from the head mounted eye tracker field of view camera and corresponding eye-gaze fixation marked in each frame with a green-dot. (a) subject is working on its individual task, (b) moving a cup from right side of the table to the left, (c)-(e) subject handing over a cup to the other participant.

The purpose of the experiments is to record the visuomotor movements of both participants. To collect the sensory information of the human eyes-head-arm movements the Pupil Labs head-mounted glasses, and OptiTrack Mocap systems are used. The Pupil Labs glasses are

worn by each participant providing the eye-gaze movements, and OptiTrack infrared markers are placed on the glasses, as well as the right hand's wrist for capturing and 3D tracking of head-gaze and arm movements. In Appendix C there is more details on the experimental setup, sensors, and data collection.

A total of 6 participants age from 22-30 years old, 5 females and 1 male, all right-handed took part in the experiment. None were members of the lab or the department, and all were naive regarding the purpose of the experiments. A total of 209 cup manipulations are performed: (i) 105 trials are of moving the cups from the right to the left side of the puzzle, (ii) 52 trials are of handing over the cup to the other participant, and (iii) 52 trials of receiving the cup from the other participant (mirror action of (ii)). The dataset is composed of 17, 19, and 16 handovers for empty, half, and full cup, respectively.

7.3 Analysis of Human Eye-Gaze

The eye-gaze fixations is provided by the Pupil Labs Capture system as a 2D pixel location. This location is a representation, in the world camera video reference, of where the participant is looking. Since this is a free-moving reference frame and head-mounted on the participant, the 2D pixel vector points are not useful to understand the non-verbal gaze movements in human-to-human handovers. As a result, it was necessary to process the data acquired from Pupil Labs into meaningful gaze fixations, i.e. eye-gaze cues relevant to the experiments. Henceforth the data was labelled by an independent engineer who followed the sole instruction of identifying the most prevalent gaze fixations in the whole experiment. The engineer did not participate in the makings of this work nor was it aware of the purpose of it.

Fixation	% of Frames
Cup	30 – 40
Own Hand	< 5
Face	10 – 30
Other's Hand	30 – 40
Other's Cup	< 1
Puzzle	NP
Final Position	NP
Outlier	< 1
No Gaze	< 1

Table 7.1: Total percentages on average for all gaze cues during handover actions. NP - Not present.

The most frequent eye-gaze cues are shown in Table 7.1 and correspond to: looking at the cup (Cup), at own hand (Own hand), at the other person's face (Face), at the other person's hand (Other's Hand), at the cup the other person is manipulating (Other's Cup), when picking LEGO® pieces (puzzle), looking ahead to where the cup will be placed (Final Position), none of the above and with no particular meaning (Outlier), and a frame with no gaze fixation (No Gaze).

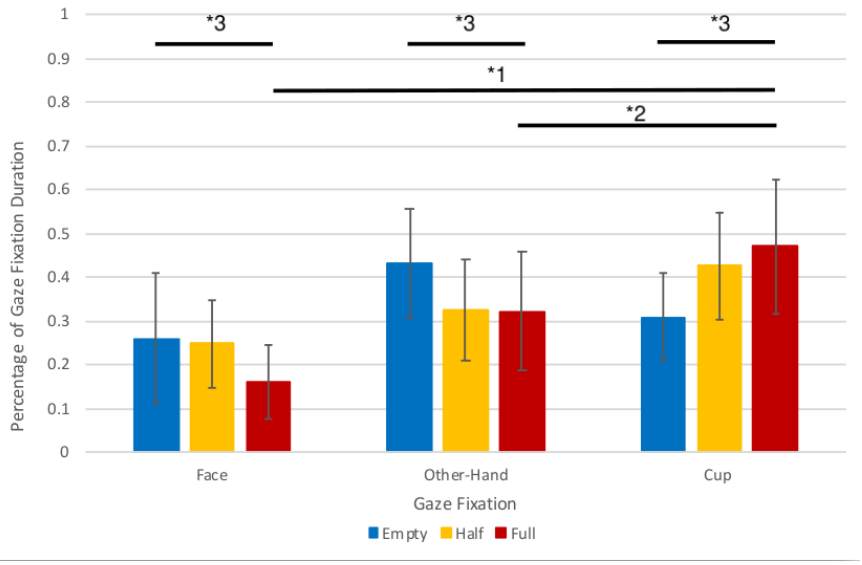


Figure 7.2: The average and standard deviation percentage ([0 - 100%] in a [0 - 1] scale) of eye-gaze cues duration during handover actions

The segmentation of the handover actions is initiated when the participant fixates the cup for grasping and concludes when the handover is completed, which can be identified from the video recordings. These segments were collected for all the participants and the three cups. Table 7.1 shows the percentages of frames present in all the handovers collected for each of the aforementioned eye-gaze cues. It is reasonable to comprehend the reason that some of the cues are not present in the handover situation, e.g. the puzzle refers to moments where the participant is not performing an action, and the Final Position is related to the location where the cup is going to be placed. The Own Hand is uncommon to occur in handovers and when it does happen it is usually during grasping or manipulation of the cup, hence it is as fixating the Cup. As a result, the 3 most relevant eye-gaze cues can be compared for the three possible cups. Figure 7.2 shows the time spent fixating the eye-gaze cues for the three types of the cup during handovers, ignoring the Outliers and No Gaze frames. A Shapiro-Wilk test was performed and demonstrated that the fixations distribution departed significantly from normality (W=0.9121, p=0.0008). As such, non-parametric tests were used in the analysis.

When the cup starts having more and more water, the human has to focus his/her attention on the stability and spilling concern during the handover, hence more time is focused on looking at the cup during the handover. In the Full condition more time is spent fixating the Cup then the Face and Hand and the difference is statistically significant (Wilcoxon test p=0.0059 to Other-Hand *1, and Wilcoxon test p=5.6430e-05 to Face *2). The eye-gaze cues analysis demonstrates that eye-gaze movements have two purposes: visuomotor control and visual-communication control. The former are the visual cues to guide the motor movement, and the latter are the visual cues to communicate intent to others. The former eye-gaze movements happen most often at the beginning of the action to ensure the object is safely stored in the hand and only after the grasp and manipulation are guaranteed to respect the

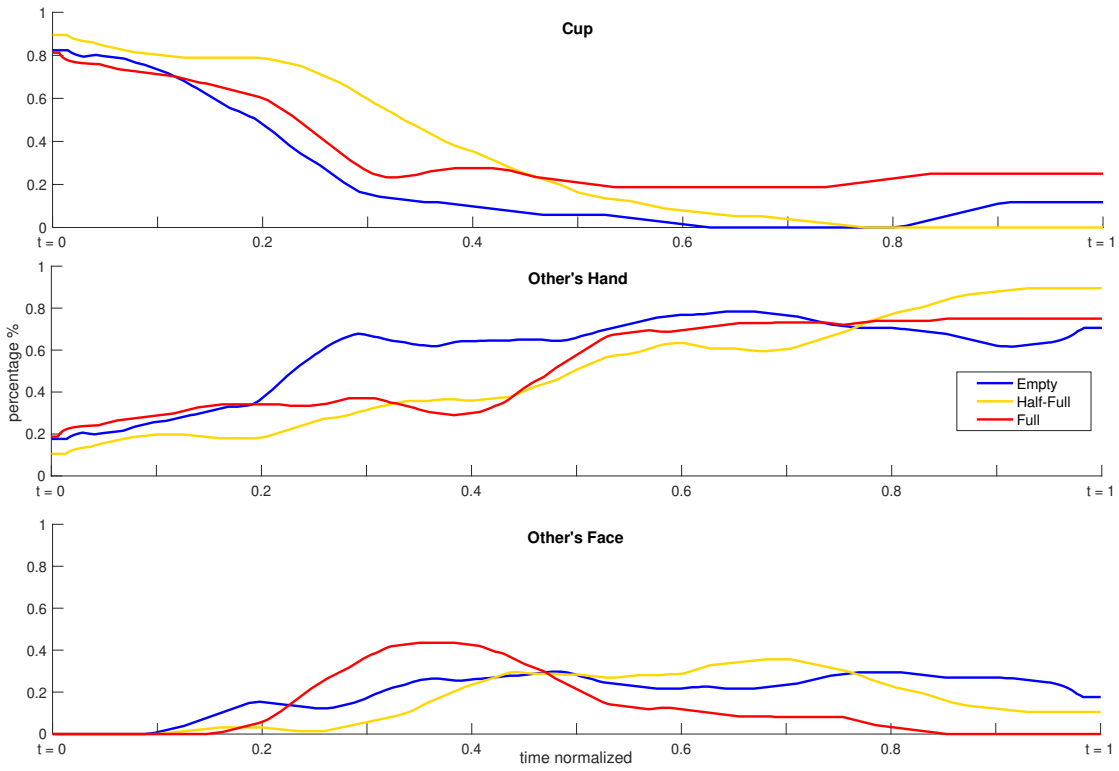


Figure 7.3: The average percentage ([0 - 100%] in a [0 - 1] scale) of eye-gaze cues duration during handover actions along the time sequence.

object conditions then the human moves towards a gaze oriented toward expressing intent. From Figure 7.2 it can be concluded that the emptier the cup, the more time you can spend communicating your intent. The communication intent is expressed by more time spent looking at the subject's face and hand. Friedman test confirms that the cup condition changes significantly the fixation percentage of Face, Hand, and Cup (p=6.4767e-04 *3). For a full cup, the visuomotor control of the action is more important than the visual-communication.

The eye-gaze movements as seen in Figure 7.3 are processed to understand the sequence of events that happen during a handover. From analysing the eye-gaze cues during handovers it can first be concluded that, as previously seen in Chapter 5, the focus, in the beginning, is on fixating the cup. The initial 20% of the duration the Cup is fixated thrice as much than the other two. This is the functional gaze performing the visuomotor control of the arm grasping the cup for a safe transportation. Secondly, the visual-communication only occurs during the transportation and after the visuomotor control check. Additionally, fixating the face does not occur in the visuomotor part, indicating that this is a gaze cue for communicating intent and not for visuomotor guidance. Thirdly, the emptier the cup was the sooner participants started communicating intent and for longer. Figure 7.3 shows that face is more likely to be fixated sooner and continue being present throughout the handover for not-full conditions.

In the full cup condition, there is one evident discrepancy to the other cases. The fixation of someone's face becomes dominant in an small interval of time during the handover. In comparison to the other two conditions, the face continues to be fixated until the end of the

action. From this, the face cue can be imagined as a bell-curve signal in which as the level of water increases the amplitude increases while the width shrinks. This can be translated into an increased difficulty in manipulation so more time has to be spent in performing visuomotor control. This results in less time to communicate intent so the visual-communication is quicker but more pronounced.

7.3.1 Eye-gaze vs Head-gaze cues

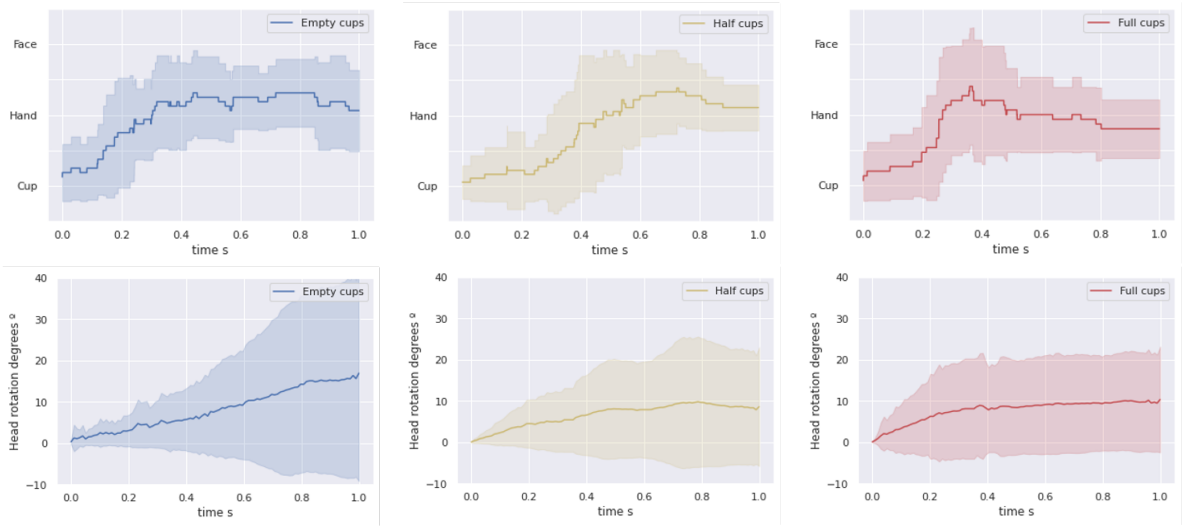


Figure 7.4: Eye movements (top) vs head movements (bottom) for the three cases of water levels.

These data allow us to make a comparative analysis between two non-verbal cues: using eye-gaze cues, against the most common approach in robotics, head-gaze orientation [Claudia and Shah, 2015, Zheng et al., 2015, Kshirsagar et al., 2020]. From the markers placed on the head-mounted eye-tracker, the head orientation during the handover actions can be tracked. The absolute orientation shift of the head is computed for all the participants from the initial point. Figure 7.4 shows the two non-verbal cues over the handover sequence for the three types of cups. The first major difference is the reaction time, where the eyes switch fixation sooner than the head moves. This is in line with the human visuomotor coordination where gaze shows an anticipatory behavior preceding motor movement [Lukic et al., 2014, Johansson et al., 2001]. Secondly, from the eye-gaze data there are three important cues (cup, other's hand, face), however from solely the head orientation that distinction is not available because the spatial configuration difference of all those three cues are too small to be detected. This is simply a limitation on what can be extracted from head movements. On the other hand eye-gaze cues, following the ideas from Chapter 5, have two properties of gaze movements: (i) visual-communication and (ii) visuomotor control. From the analysis of the handovers it can be concluded that: fixating the Cup aims at guiding the grasp and observing the exerted lifting force for potential spilling [Mayer and Krechetnikov, 2012]; fixating the Face aims at expressing handover intent; as for fixating the Other's Hand I hypothesize that this fixation is

an intermediate step between the other two, i.e. endows the two properties, visuomotor control for meeting one's hand with the other's, while at the same time, expressing the intent of the action. From this, it can be concluded that these eye-gaze cues provide valuable information to classify cup manipulations of different levels of water (difficulty) than only head-gaze cues.

7.4 Methodology

This section contains the formalism of a simple Echo State Network (ESN) and the included modifications applied in this work for better performance. For more details, the reader is referred to [Bianchi et al., 2021].

Let's consider a classification problem for a N-dimensional Multivariate Time Series (MTS) with $\mathcal{T}$ time, and for each t there is an observation $\mathbf{u}(t) \in \mathbb{R}^N$. The MTS is represented in compact form as $\mathbf{U}^{\mathcal{T}\times N} = [\mathbf{u}(1), \mathbf{u}(2), ..., \mathbf{u}(\mathcal{T})]^T$. Machine learning techniques for classification are commonly composed of an encoding and a decoding function. The encoder takes the input and generates a representation, whereas the decoder is a predictive model that given the encoder representation, computes the posterior probability of the output. An encoder based on Recurrent Neural Network (RNN) is particularly appropriate to model sequential data [Bianchi et al., 2021]. The encoder can be formulated as:

$$\mathbf{x}(t) = f(\mathbf{u}(t), \mathbf{x}(t-1); \theta_{\mathbf{enc}}) \tag{7.1}$$

where $\mathbf{x}(t)$ is the RNN state at time t that depends on its previous value $\mathbf{x}(t-1)$ and the current input $\mathbf{u}(t)$, $f(.)$ is a nonlinear activation function (e.g. a sigmoid or a hyperbolic tangent), and θ_{enc} are adaptable parameters. Equation 7.1 with an hyperbolic tangent is expressed as:

$$\mathbf{x}(t) = \tanh(\mathbf{W_{in}}\mathbf{u}(t) + \mathbf{W_r}\mathbf{x}(t-1)) \tag{7.2}$$

with $\theta_{\mathbf{enc}} = \{\mathbf{W_{in}}, \mathbf{W_r}\}$. The matrices $\mathbf{W_{in}}$ and $\mathbf{W_r}$ are the weight of the input and recurrent connections, respectively.

From the sequence of the RNN states generated over time, $\mathbf{X} = [\mathbf{x}(1), ...\mathbf{x}(\mathcal{T})]^T$, it is possible to extract a representation $\mathbf{r_U} = r(\mathbf{U})$ of the input $\mathbf{U}$. A common choice is to take $\mathbf{r_U} = x(\mathcal{T})$ since the RNN can embed into its last state all the information required to reconstruct the original input. The decoder maps the MTSq representation $\mathbf{r_u}$ into the output space, which are the class labels $\mathbf{y}$ for a classification task

$$\mathbf{y} = g(\mathbf{r_U}; \theta_{\mathbf{dec}}) \tag{7.3}$$

where $g(.)$ can be a feedforward neural network or a linear model and $\theta_{\mathbf{dec}}$ are the trainable parameters. To avoid the costly operation of backpropagation through time, the reservoir computing (RC) approach takes a radical different direction; it still implements the encoding

function in 7.3, but the encoder parameters are randomly generated and left untrained. To compensate for this lack of adaptability, a large recurrent layer, the reservoir, generates a rich pool of heterogeneous dynamics useful to solve many different tasks. The generalization capabilities of the reservoir mainly depend on three ingredients: 1) a high number of processing units in the recurrent layer, 2) sparsity of the recurrent connections; and 3) a spectral radius of the connection weights matrix $\mathbf{W}_r$. The behavior of the reservoir is controlled by modifying the following hyperparameters: the spectral radius, the percentage of nonzero connections, the number of hidden units, the scaling of the values in $\mathbf{W}_{in}$, which controls the amount of nonlinearity in the processing units and, jointly with spectral radius, can shift the internal dynamics from a chaotic to a contractive regime [Bianchi et al., 2021]. A Gaussian noise with standard deviation ϵ can also be added in the state update function 7.3 for regularization purposes.

ESN is an effective RNN that has attracted substantial interest due to its performance in time-series [Sun et al., 2020]. The core of ESN is a large fixed reservoir, the reservoir is not trained and it contains a large number of randomly and sparsely connected neurons. The determination of the readout weights from the reservoir is the only trainable part, the weights can be obtained simply by linear regression. This basic idea was first clearly spelled out in a neuroscientific model of the corticostriatal processing loop [Dominey and Ramus, 2000]. [Jirak et al., 2020] state that ESNs are viable models for continuous gesture recognition delivering reasonable performance for applications requiring real-time performance in robotic or rehabilitation tasks. ESNs could be an alternative in some tasks that are designed to run on smaller, constrained devices where the size or performance of RNN is limited. In ESNs, the decoder, also known as readout, is usually a linear model

$$\mathbf{y} = g(\mathbf{r_U}) = \mathbf{V}_0\mathbf{r_U} + \mathbf{v}_0. \tag{7.4}$$

The decoder parameters $\theta_{\mathbf{dec}} = \{\mathbf{V}_0, \mathbf{v}_0\}$ can be learned by minimizing a ridge regression loss function

$$\theta^*_{\mathbf{dec}} = \underset{\{\mathbf{V}_0,\mathbf{v}_0\}}{\arg\min} \frac{1}{2} \parallel \mathbf{r_U}\mathbf{V}_0 + \mathbf{v}_0 - \mathbf{y} \parallel^2 + \lambda \parallel \mathbf{V}_0 \parallel^2 \tag{7.5}$$

which admits a closed-form solution. The combination of an untrained reservoir and a linear readout defines the basic ESN model.

A powerful representation of the model space is obtained by first processing each MTS with the same reservoir and then training a ridge regression model to predict the input one step-ahead. The linear model trained to predict the next reservoir state reads

$$\mathbf{u}(t+1) = \mathbf{U}_0\mathbf{x}(t) + \mathbf{u}_0 \tag{7.6}$$

and $\theta_0 = [\mathbf{vec}(\mathbf{U_0}); \mathbf{u}_0] \in \mathbb{R}^{R(R+1)}$ becomes the representation $\mathbf{r_U}$ of the MTS, which then goes to the decoder (classifier) in Equation 7.4.

Due to the high dimensionality of the reservoir, the number of parameters of the prediction model in 7.6 would grow too large, making the proposed representation intractable. Drawbacks in using large representation include overfitting and the high amount of computational resources to evaluate the ridge regression solution for each MTS. In the context of RC, applying Principal Component Analysis (PCA) to reduce the dimensionality of the last reservoir state has shown to improve the performance achieved on the inference task. PCA provides competitive generalization capabilities when combined with RC models and can be computed quickly, due to its linear formulation. PCA projects data on the first D-eigenvectors of a covariance matrix.

RNNs with bidirectional architectures can extract from the input sequence features that account for dependencies very far in time. In RC, a bidirectional reservoir has been used in the context of time series prediction to incorporate future information, only provided during training, to improve the accuracy of the model. In a classification setting, the whole time series is given at once and, thus, a bidirectional reservoir can be exploited in both training and test to generate better MTS representations.

The readout module (decoder) classifies the representations and is either implemented as a linear readout or a support vector machine (SVM), or a multilayers perceptron (MLP). In a standard ESN, the readout is linear and is quickly trained solving a convex optimization problem. The main advantage of ESN over Long short-term Memory (LSTM) is the smaller number of trainable parameters and a simpler training algorithm.

7.5 Modelling Eye-gaze Cues

Let's consider a classification problem for a discrete univariate time-series with $\mathcal{T}$ time, and for each t there is an observation $\mathbf{u}(t) \in \mathbb{D}^{\mathcal{T}\times 1} = \{\text{Cup; Other's Hand; Face}\}$, which in an ESN is the input unit, $x(t) \in \mathbb{R}^{N\times 1}$ denotes the state of the reservoir, and $y(t) \in \mathbb{R}^{M\times 1}$ denotes the output unit. The time-series is represented in compact form as $\mathbf{U}^{\mathcal{T}} = [\mathbf{u}(1), \mathbf{u}(2), ..., \mathbf{u}(\mathcal{T})]^T$. $\mathbf{W_{in}} \in \mathbb{R}^{\mathcal{T}\times N}$ represents the connection weights between the input and hidden layer, $\mathbf{W_{res}} \in \mathbb{R}^{N\times N}$ denotes the connection weights inside the hidden layer. The encoder and decoder functions are formulated as:

$$\begin{aligned} \mathbf{x}(t) &= f(\mathbf{W_{in}}\mathbf{u}(t) + \mathbf{W_{res}}\mathbf{x}(t-1)) \\ \mathbf{y}(t) &= \mathbf{W_{out}}\mathbf{x}(t) \end{aligned} \tag{7.7}$$

where f is a nonlinear function, in this case the $\tanh$ was applied. As mentioned in the previous section, the encoder parameters are randomly generated and left untrained. Only $\mathbf{W_{out}}$, the connection weights between the hidden and output layer, are subject to training

using fast algorithmic closed form solutions like ridge regression

$$\min_{\mathbf{W_{out}}} \| \mathbf{W_{out}}\mathbf{X} - \mathbf{Y} \|_2^2 \tag{7.8}$$

where $\mathbf{W_{out}}$ is commonly referred as the readout weights. To compensate for untrained parameters, a large recurrent layer, the reservoir, generates a rich pool of heterogeneous dynamics. The reservoir has three main hyper-parameters: (i) the spectral radius, i.e. largest eigenvalue, of $\mathbf{W_{res}}$, (ii) the sparsity parameter, i.e. nonzero connections, of $\mathbf{W_{res}}$, and (iii) input scaling of $\mathbf{W_{in}}$. Gaussian noise with standard deviation is also applied in the state update function of Equation 7.7. A PCA projection on the data is performed to extract the first D-eigenvectors of the covariance matrix. Additionally, since RNN with bidirectional architectures can extract features over a long period, ESNs with a bidirectional reservoir has been shown to improve the classification accuracy.

7.5.1 Dataset Results

The full list of hyper-parameters is the following: D-eigenvectors for PCA dimensionality reduction, N neurons, the spectral radius ρ, the sparsity β, input scaling ω, regularization value λ of ridge regression, and Gaussian noise ϵ. The dataset gaze cues sequences are normalized to 100 samples, and the output layer has $M = 3$ for the three types of cups. The hyper-parameter space is explored using grid search and performing 3-fold cross-validation on the whole dataset it reaches $95\% \pm 2\%$ and $72\% \pm 8.5\%$ accuracy in training and testing, respectively. Figure 7.5 shows the prediction results of the ESN, at each time step, for the whole dataset. Since ESN requires a series of observations (comparing to other methods which require a single sample) to regulate the internal state of its reservoir, it is provided around 20% of the sequence at the beginning. The classification result is given by the ESN highest probability output at each time step.

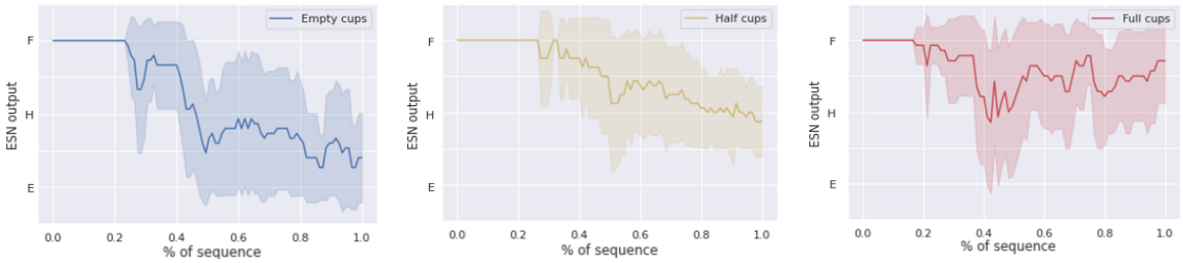


Figure 7.5: Classification results for each cup level (E - empty; H - half-full; F - full) over time by the ESN.

The results in Figure 7.5 show that the ESN classifies all the actions, at the beginning, as full cups. This makes sense since the cup is mostly fixated at the beginning which, without more knowledge, indicates that the handover is challenging (full water level) as only the visuomotor control is present. However, during training it was noticed that the default action would change depending on the random initialization of $\mathbf{W_{res}}$ weights, so it might just be

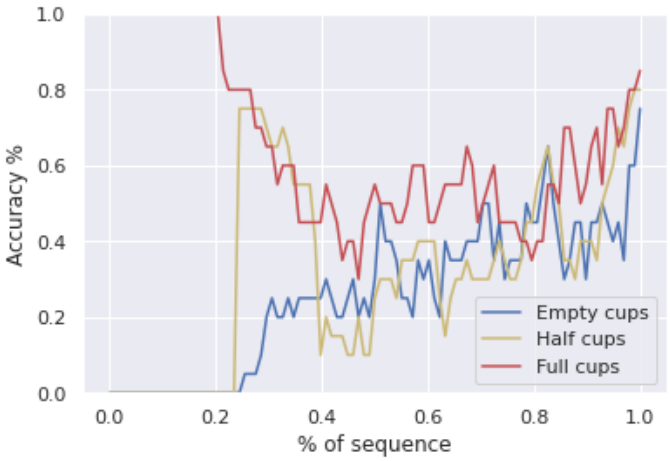


Figure 7.6: ESN output accuracy for the three levels of liquid.

a random coincidence. Although, more often the best accuracy networks would output as initial default action the full cup option. Figure 7.6 illustrates the models accuracy along the handover sequence for the three water cup levels. As stated before, the ESN is provided with around 20% of the sequence at the beginning before prediction and given that most fixate the cup, the accuracy is falsely indicating a full cup classification for all handovers. The model's accuracy achieves good results, of 60% or more for the three cup conditions, at around 90% completion of the handover. This reflects not only the high variance of gaze cues between humans but also the importance of the visual-communication part, which occurs last and goes on until the completion of the handover. Detecting full cups seems to be the easiest, as the accuracy increases sooner and reaches 80%, which could be impacted by the Face cue as it fades the fastest in those conditions (Figure 7.3).

7.6 Robot Experiments

The ESN is capable of classifying unseen handover actions with accuracy more than twice times higher than the chance level (1/3). However, these handover actions are sequences of eye-gaze cues from the human-to-human dataset of Section 7.2. In this section, the model is applied to a human-robot interaction scenario where the system is running online with a humanoid robot. The purpose of this section is to demonstrate the compatibility of the proposed approach to real-robot experiments with online classification of cups with different levels of water.

The HRI controller schematic of the architecture is represented in Figure 7.7. This system is composed of two important blocks which handle the communication between human and robot: (i) the *Gaze Dialogue Model* in Chapter 5, and (i) the Echo State Network. The *Gaze Dialogue Model* is an inter-personal gaze coordination system for human-humanoid interactions. It anticipates the human action by reading human eye-gaze cues and estimating the future human

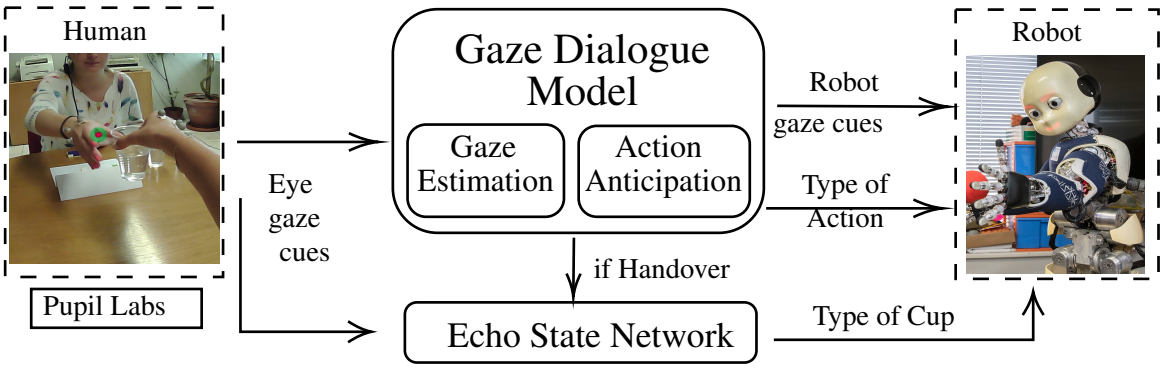


Figure 7.7: Robotic controller for classification of cups with three water levels during handovers.

gaze cues, while at the same time, generating appropriate robot gaze cues and motor control response for the anticipated action (as in Chapter 5). The system distinguishes reliably a human-to-robot handover from a human-object manipulation from human eye-gaze cues. Since the ESN model works exclusively for human-to-robot handovers it seems appropriate to include a lower-level system to discriminate the type of actions and have the ESN model process only the information of handover actions. The gaze cues present in this work are translated to the *Gaze Dialogue Model* gaze cues as follows: Cup is the Brick (i.e. the object), Other-Hand is the Follower's Hand, and Face is the Follower's Face. The gaze cue for the cup is identified as a solid object since the system was trained on a simple block cube, however, for simplicity, the gaze cue for the object is used for detecting the cup. The other gaze cues included in the *Gaze Dialogue Model* are ignored by the ESN model, such as Own Hand, Follower's Tower and Own Tower. The *Gaze Dialogue Model*, at each step, reads the human gaze fixation, updates the action observed (handover or pick-and-place) from the human gaze sequence, it then generates the appropriate robot-as-a-follower gaze fixation, and plans the according robot action (handover or pick-and-place). Figure 7.8 demonstrates a HRI scenario using the robotic controller illustrated in Figure 7.7. In this scenario the human is wearing the Pupil-Labs eye-tracker which computes the human gaze fixations using the VFOA, i.e. Human-in-the-Loop System, as explained in Section 5.4.2. The human hands over a cup to the robot while the *Gaze Dialogue Model* and ESN model are running to predict the action and classify the type of cup. The action begins as in (a) by the human looking at the robot's face, in (b) it can be seen the generated gaze cue from the *Gaze Dialogue Model* where the robot looks at the human. In (c) the human is fixating the robot hand, the robot is looking at the cup and the motion planner is preparing for a handover given the predicted action. In (d) the human continues fixating the robot hand and the robot motion planner approaches the hand of the human even further. In (e) the human fixates the robot's face and the robot starts closing the hand while fixating the cup, and finally in (f) the robot has finished the handover by holding the cup in its hand. When the robot has grasped the cup, the sequence of eye-gaze cues is sent to the Echo State Network block to classify the level of water in a cup.

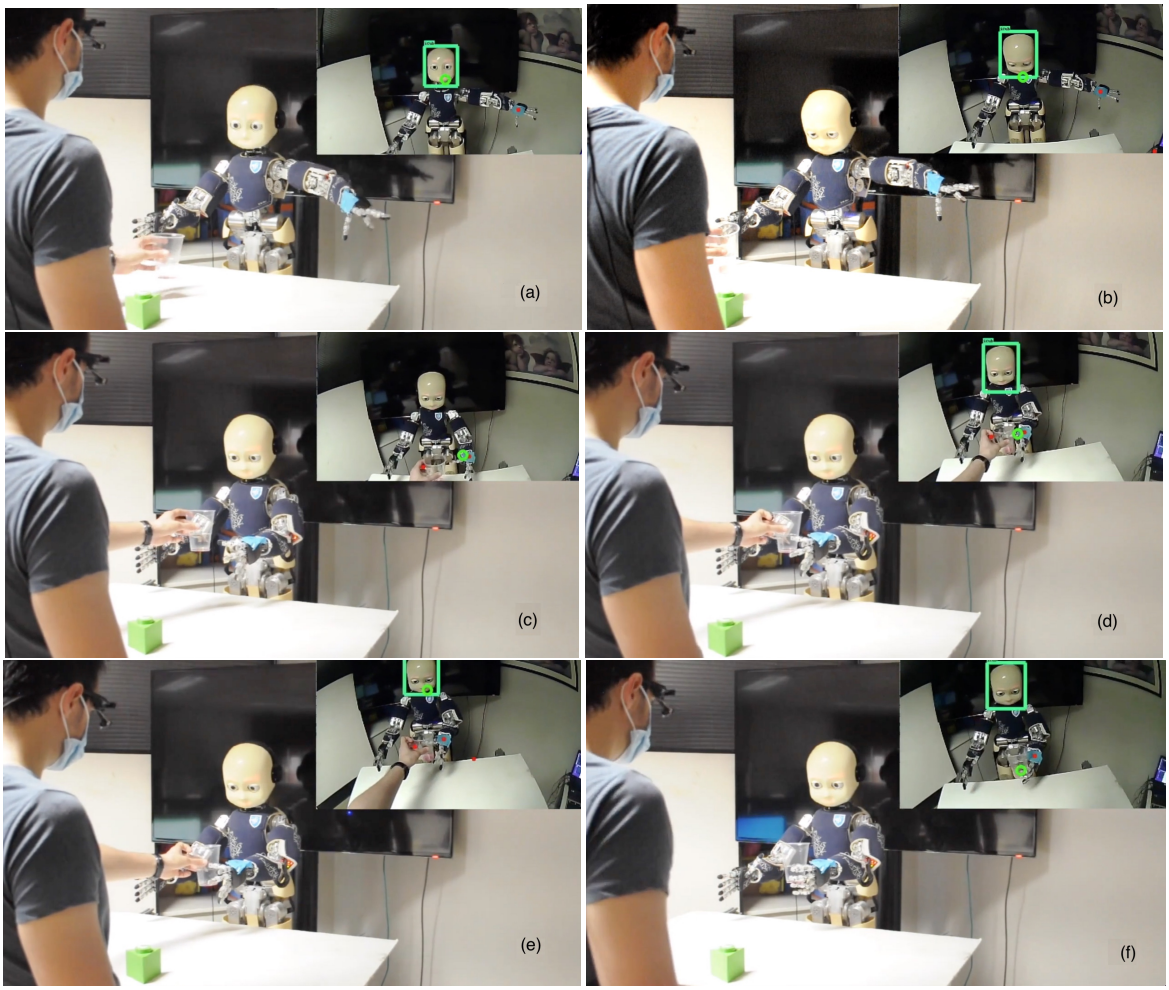


Figure 7.8: The handover of a cup to a robot using the *Gaze Dialogue Model* with integration of the Echo State Network block for classifying whether the cup is empty, half-full or full. Every frame (a)-(f) shows the first person view from the eye-tracker (top right corner). A video of the whole pipeline can be seen in video.icra.ieee-2022.

7.7 Final Remarks

From the human-to-human handover analysis, it can be shown that human behavior adapts to changing properties on cups. Previous works have shown that human motion strategy changes when the object weight is different [Alaerts et al., 2010a] and in the previous chapters it was seen that human eye-gaze movements change according to the desired end-goal, whether it is handover or pick-and-place. In this work the analysis was extended to explore human eye-gaze during handovers of a cup in three conditions: (i) empty, (ii) half-full, and (iii) filled with water. The eye-gaze movement's strategy to perform a handover is altered by an increased level of water inside the cup. As the level of water increases, so thus the risk of spilling, hence the gaze behavior is spent more time on the visuomotor control role, than in the visual-communication role, as seen in Figure 7.2. As the visuomotor control focuses on ensuring a safe grasp and safe transportation of the cup (prevent spilling), and the visual-communication focuses on expressing to others the intent of handing over. The final contribution of this thesis, continues the study of human handling liquid containers, such as cups. Not only the eye-gaze

is affected but the motion behavior also takes into account the risk of spilling, revealing a contrasting strategy when comparing to empty containers.

8

Recognize carefulness by observing human *motion cues*

> *The simple things are also the most extraordinary things, and only the wise can see them.*
>
> Paulo Coelho,

Contents

Endowing robots with the ability to understand human action intentions from non-verbal cues will broaden the robots' use-case scenarios. This ability is especially useful in scenarios where humans and robots need to collaboratively manipulate objects. For example, imagine that a robot and a human are performing handovers of different types of objects. A few questions arise: does the type of object (e.g., fragile vs non-fragile) change the motion of the handover? does the amount of filling (e.g., full vs empty) when handing over a container change the type of behaviour? I hypothesize that non-verbal cues extracted from the human body movement can reveal relevant information regarding the manipulation of the object. In other words, the object's intrinsic physical properties will influence the action execution and, therefore, the non-verbal cues. In HRI, the interpretation of non-verbal cues provides the robot with relevant information concerning the object to grasp, which can be adapted during the interaction, to comply with the object's physical properties.

8.1 Introduction

Studies on human non-verbal cues found that joint kinematics and dynamics of hand manipulation are crucial features for object weight estimation [Rosen et al., 2005], action duration [Hamilton et al., 2007], and absolute velocities [Sciutti et al., 2019]. [Bingham, 1987] mention that the velocity is the key feature to extract when estimating weight from the kinematic motion. Humans have the ability to extract knowledge of objects' weight, fragility, or contents, from motion. [Tomoki OjiSakuragi et al., 2018] estimate object's mass in real-time using data from a single person. They use the kinematic variations of the human body, however it is limited to the same object and just one person. [Senot et al., 2011] state that high-level semantic cues, such as labels on the objects, may influence low-level motor behavior during execution, and if there is a conflict between label and object weight it shows that motor resonance can vanish when a mismatch exists between the expected and observed kinematics. [Alaerts et al., 2010a,b] mention that perceiving kinematic trajectory, associated with lifting heavy and lightweight objects, was sufficient to induce force-related activity modulations in the observer's primary cortex. [Sciutti et al., 2014] proved that subjects can reach a performance in weight recognition from robot observation comparable to that obtained during human observations with no need of training.

[Ortenzi et al., 2021] presented recently a survey on handovers in robotics. The authors looked at human-to-human handovers studies and the current approaches on human-robot handovers either for robot giver (robot-to-human) and robot receiver (human-to-robot). They identify two important phases of the handover: the pre-handover phase, i.e. the approach, and the physical phase, i.e. grasping and releasing. In the pre-handover phase, which is the scope of this chapter, there are several works that proposed strategies for human-robot handovers, inspired in human-to-human handovers. The existing approaches have focused mainly in reproducing human-like motions [Yamane et al., 2013, Sidiropoulos et al., 2018, Maeda et al.,

2017], estimating the handover location [Nemlekar et al., 2019, Widmann and Karayiannidis, 2018], or user satisfaction [Vogt et al., 2018, Pan et al., 2018, Parastegari et al., 2017]. The assumption in the state-of-the-art is that the human handover motion is a purely functional motion. Instead, this thesis argues that it can be modulated (and therefore express) latent features related to the object or the human action intentions. The state-of-the-art does not study human-to-robot handovers where the robot is capable of distinguishing types of handovers: a normal handover, or a challenging handover where the human resorts to perform it with extra care, as presented by [Mayer and Krechetnikov, 2012].

This chapter aims at providing an in depth scope of the manipulation strategies for cups filled with water or completely empty. The focus is not on the effects of the weight, which is a user-dependent variable (the stronger you are the lower the manifestation), but instead on the challenge of transporting liquids. The objective of the analysis is to identify and extract features to recognize human careful and not careful motions during human-human and apply it to human-robot scenarios. Potential applications can be a factory plant, where robots can infer inherent properties of objects from human manipulations, such as fragility or breakableness. Alternatively, a robot caretaker can study the senior residents various levels of musculoskeletal limitations and adapt its motor constraints when assisting them.

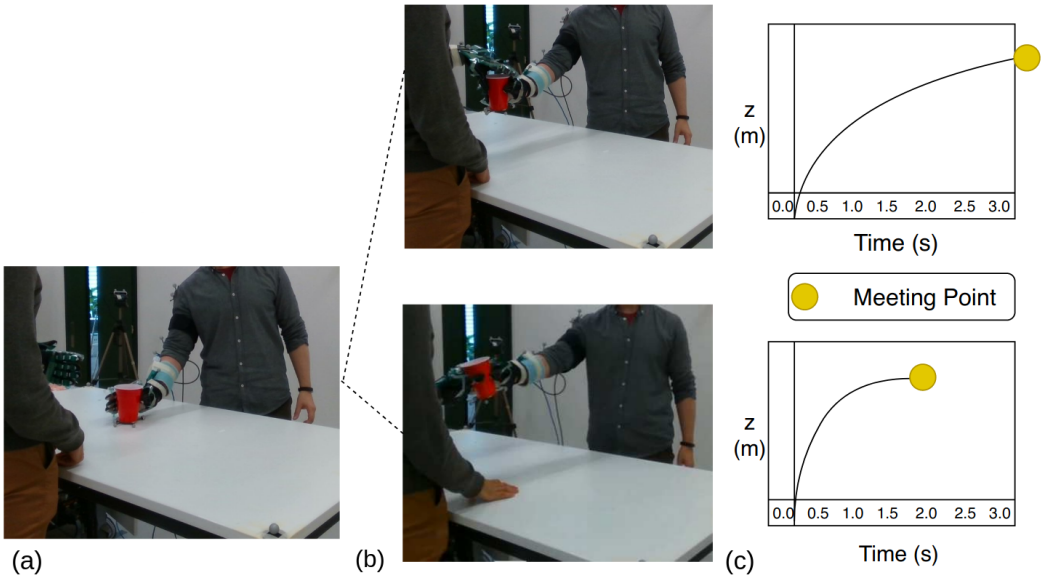


Figure 8.1: Representation of handover actions. (a) t_0 frame of hand-over action; (b) t_f frame is the final frame of a *not careful* motion (bottom) and a *careful* motion (top); (c) the duration of each type of motion.

A recent work has addressed the human manipulation of full and empty cups [Lastrico et al., 2021]. The authors studied the kinematic motion during pick & place and were able to distinguish between careful and not careful motions by inspecting the complete trajectory of the motion. Instead, our work is on human-to-human handovers which adds an interactive variable of "informing" the other partner whether the cup requires extra care to manipulate. Additionally, our approach is not only capable of online classification, without needing to complete the trajectory, but it has been applied to real-time human-robot interactions.

8.2 The Human-to-Human Handover Dataset

The HHI dataset was gathered as a collaboration between École polytechnique fédérale de Lausanne (EPFL) and Karlsruher Institut für Technologie (KIT). It involves two humans interacting with an object whether to grasp and handover to one another, or to manipulate and place it on a table. The focus of this work is solely on the handover motion. Figure 8.2 shows a frame of the handover trajectory of the different cups for each of the participants. A total of 4 participants (male, 25-35 years old, academic employees) took part in the experiments. The experimental task involves grasping a cup from a table and hand it over to a subject on the opposite side that places it back in the table (Figure 8.1).

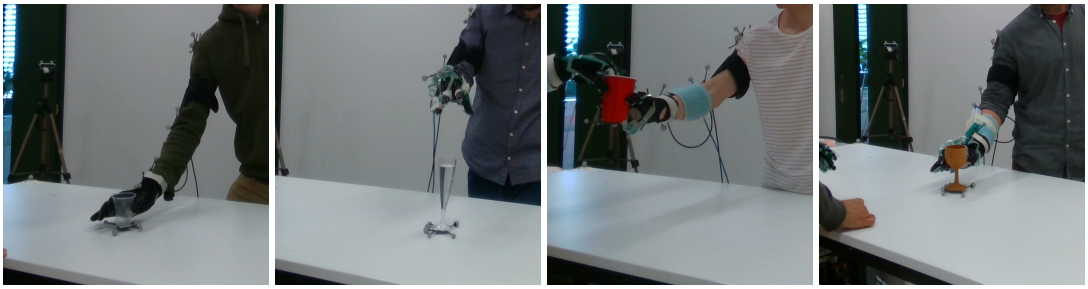


Figure 8.2: The 4 participants with the 4 cups

The handovers of the cups happen under two distinct situations: (i) an empty cup, and (ii) a cup 90% filled with water. This is repeated several times with each pair of participants. There is a total of 157 handovers, 81 of empty cups and 76 of full cups, respectively. Each participant hands-over the different cups to a second participant (also present in the dataset but not analysed during the handover), and each cup is manipulated for both conditions. The cups relevant for this work are the red plastic cup (third picture), the transparent plastic cup (first picture), the champagne plastic cup (second picture), and the opaque wine glass (last picture from Figure 8.2). The handover trajectory is recorded at 120 Hz, taking on average 1-3 seconds, corresponding to 100-300 data points. Each participant had to grasp the cups with their preferred hand (right hand for all) and there were no restriction on the type of grasp. OptiTrack Mocap recorded right-hand wrist's location for each participant as well as the cup's location. The dataset also includes data gloves from the CyberGlove system on the participant's hand not used here. The dataset referred from now as the EPFL-KIT dataset was gathered in collaboration with the High Performance Humanoid Technologies Lab (H2T) of the Karlsruher Institut für Technologie (KIT) [1] [Starke et al., 2019].

To our knowledge, this study reports on the first comprehensive analysis of the effect of an object's physical properties during dyadic human handover. Our approach is also the first that is implemented on a robotic controller that allows for a robot to recognize different manipulation strategies from humans.

[1] http://h2t.anthropomatik.kit.edu/english/index.php

8.2.1 Handover Motion Analysis

The handover motions are 3D Cartesian coordinates over time which begin at the moment of pickup (grasp) and finish when the cup is safely held by the other participant (handover). During the experiments, the participant could grasp the cup irrespective of hand configuration and the handover location could be in a 3D space bounded by the table as seen in Figure 8.2. This gave rise to several different grasp configurations and a disparity in the duration/length of the handover trajectories. Bear in mind that it was not possible to re-grasp the cup or change grasp configuration during the handover, and the cup would have to start upwards on the table for every interaction.

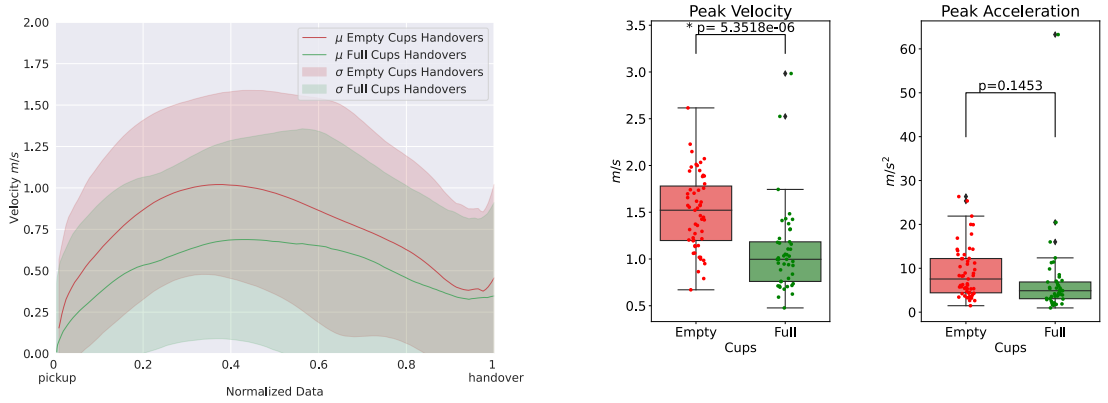


Figure 8.3: The plot shows the velocity mean μ and standard deviation σ of the handover actions of the dataset separated in the two cups conditions (empty and full). The trajectories are normalized. The two box plots represent the peak velocity, on the left, and the peak acceleration, on the right, for each handover action and both conditions. p-values for peak velocities and acceleration of both cup levels are shown on the top each plot. Confirmation of significant difference is highlighted using a star.

Given the degrees of freedom in the cup initial position (pickup), and final position (handover) the handover trajectories have various lengths and durations. In order to analyse the kinematic of the wrist for all handovers for every participant, and every cup in both cup conditions it was decided to normalize all the handovers to a standardized fixed length. Hence it was applied a min-max normalization before reducing the dimensionality to a vector by calculating the euclidean norm of the x, y, and z Cartesian coordinations. This re-scales the data to a [0, 1] dimension where 1 is the final step, referred to as the handover, and 0 is the initial step, representing the moment of pickup. The plot in Figure 8.3 shows the velocities' mean and standard deviation for the human giver's hand throughout the handover in the entire EPFL-KIT dataset distinguishing both empty and full cup conditions.

Throughout our analysis, the common trait of all demonstrations is the typical bell-shape for the velocity profile as humans choose a minimum jerk approach for the hand trajectory. This has been identified in previous works in point-to-point human motion [Fligge et al., 2012] and this behavior manifests, likewise, in object handovers. In contrast, the most notable difference is on the bell-shape peak, i.e. the maximum velocity reached by the human. Analysing the peak velocity box plot in Figure 8.3, the difference is noticeable when distinguishing the cups

by the level of water contents. The one-way ANOVA test revealed a statistically significant difference between the peak velocity of both cup conditions ($F(1,98) = 23.19$, $p < 0.001$). This is fairly straightforward as a cup filled with water presents an additional challenge during manipulation, in other words, transporting the contents inside without spilling or breaking. There is also the added weight of the liquid to the overall mass of the cup, however, it can be argued that the effect of a liquid oscillating inside a cup during human transportation is more impactful in deterring quick and jerky movements than a particularly heavy object. The peak acceleration, as seen in the box plot of Figure 8.3, is not as relevant to differentiate the two cup conditions. The one-way ANOVA test did not reveal a statistically significant difference between the peak acceleration of both cup conditions ($F(1,98) = 2.16$, $p = 0.1453$).

From the kinematics, it is clear that in any handover motion there are two distinct stages, an acceleration stage and a deceleration stage. This is in line with a minimum-jerk motion which starts and ends with the wrist in rest positions. It is also evident the disparity of the two water conditions for those two stages. The empty cup condition is showing a much steeper acceleration and, consequently deceleration, to reach the rest position. This feature, which has been addressed above, can be utilized to differentiate the two types of manipulation strategies: careful and not careful. Given that both stages exhibit the same distinction, it seems appropriate to learn carefulness behavior from the acceleration and deceleration stages, respectively. Another point is related to familiarization with the task and object. As humans repeat an exercise multiple times they tend to gather prior knowledge from past events, and in this particular scenario, estimate the object's mass and the required force to manipulate. As a result, a novel object with an unexpected heavy mass might invoke a slower manipulation in the first trial but after some attempts, there is a pre-activation of muscles and joints to anticipate the requirements which may result in a more natural manipulation. This familiarity procedure is not present when manipulating cups with liquid inside, as the content is visible from the first encounter but the risk of spilling is constantly present. For this reason, the level of water is considered the most important factor.

From the analysis it was detected that manipulating cups full of water would usually originate in a slower, less abrupt motion, compared to empty cups. In the next section, the processed handover segmentations will be used to learn models for detecting different manipulation behaviours. Section 8.4 discusses in depth the different cups and the impact of water contents.

8.3 "Carefulness" Detection Pipeline

This section presents the models for human manipulation of cups in the two conditions: (i) empty cup, or (ii) water level at around 90 %. From the discussion in Section 8.2 it can be argued that there are two possibilities for modelling the human manipulation in the handover context: (i) the acceleration phase, and (ii) the deceleration phase. This has been

shown in pick-and-place actions [Flash et al., 2013] where goal-oriented biological motions are typically a minimum-jerk control problem with an acceleration and deceleration phase [Fligge et al., 2012]. The acceleration phase begins with the object at rest position, the human grasps the object, and then the object increases in velocity as it is lifted up for transportation. The deceleration phase indicates the approach stage to handover the cup to another person, indicated by a gradual decrease in velocity until a stationary state is reached for completion of the handover.

Figure 8.4 is a diagram of the control system for both the acceleration and deceleration phase models. The overall structure is identical for both models, the main differences are the information used from the handover (Training Data) and the Modelling technique itself. The Classifier and the Human-in-the-loop control is identical. The following subsections describe the two modelling techniques, the classifier applied to the control loop and the advantages and disadvantages of both models.

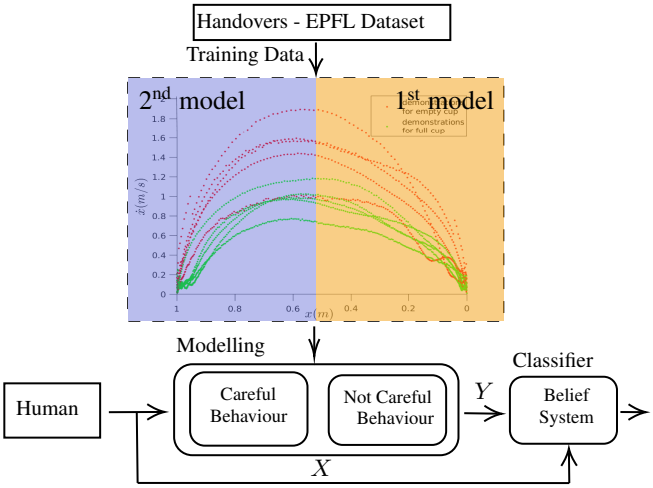


Figure 8.4: Carefulness detection controller loop for both models. The 1st model learns from the deceleration phase of human handovers (right-side of the trajectory - yellow region). The 2nd model learns from the acceleration phase (left-side of the trajectory - blue region).

8.3.1 Deceleration Phase

The deceleration phase models the velocity as a function of the distance towards the handover for both situations. This was possible as the extraction of the deceleration phase during the cup manipulation revealed the maximum velocities and its evolution towards the rest stage, i.e. handover completion. Let $x \in D \subset \mathbb{R}^+$ denote the distance of the human wrist towards the handover meeting point. Consider a behavior encoded as a state-dependent DS

$$\dot{x} = \mathbf{f}(x) \tag{8.1}$$

where $\mathbf{f} : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ is a continuous and continuously differentiable function, with a single equilibrium point $\dot{x}_d^* = \mathbf{f}(x^*)$. x^* is set at the origin and it is globally asymptotic stable such

that $\dot{x}^* = \mathrm{f}(x^*) = 0$ which is guaranteed under a Lyapunov function $V(x) : \mathbb{R}^+ \rightarrow \mathbb{R}^+$.

The approach defines each "carefulness" condition, *careful* and *not careful*, as two distinct DS. Each DS is encoded using GMM which defines a joint distribution function $\mathcal{P}({x^t}_n, {\dot{x}^t}_n|\Theta) = \sum_{k=1}^{K} \pi^k \mathcal{N}({x^t}_n, {\dot{x}^t}_n, \mu^k, \Sigma^k)$ over the data as mixture of K Gaussian distributions [Khansari-Zadeh and Billard, 2011], where π^k, μ^k, and Σ^k are, respectively, the prior component, mean, and covariance matrix of the kth Gaussian. x_n^t is nth trajectory of x at time t, and $\dot{x}_n^t$ is its derivative. Figure 8.5 illustrates the position (x) and velocity ($\dot{x}$) relations for *careful* and *not careful* motions. To compute the DS from Equation (8.1) the posterior mean of $\mathcal{P}(\dot{x}_n^t|{x^t}_n)$ is estimated which approximates it to:

$$\hat{\dot{x}} = \sum_{n=1}^{K} h^k(x)(\Sigma_{\dot{x}x}^k(\Sigma_{xx}^k)^{-1}(x - \mu_x^k) + \mu_{\dot{x}}^k) \tag{8.2}$$

where $h^k(x) = \frac{\pi^k \mathcal{N}(x^t, \dot{x}^t, \mu^k, \Sigma^k)}{\sum_{i=1}^{K} \pi^k \mathcal{N}({x^t}_n, \dot{x}_n^t, \mu^i, \Sigma^i)}$, $h^k(x) > 0$, and $\sum_{n=1}^{K} h^k(x) = 1$. The GMMs are computed using the stable estimator of DS (SEDS) approach [Khansari-Zadeh and Billard, 2011].

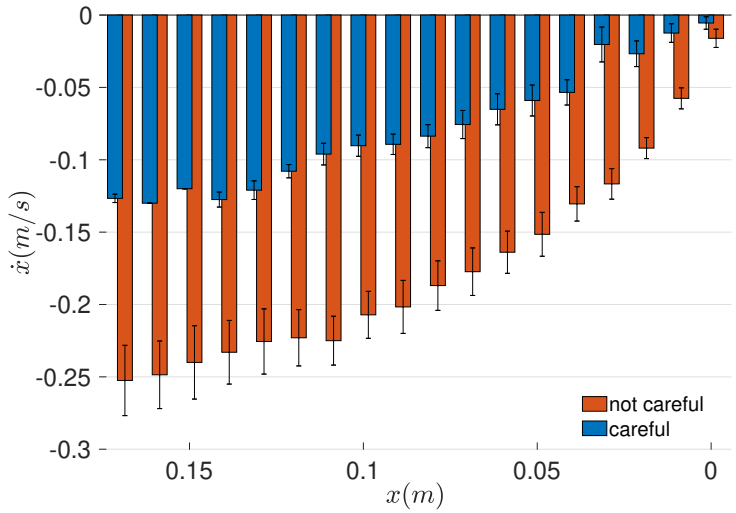


Figure 8.5: Human handover velocities for *careful* and *not careful* behavior in the deceleration phase.

Figure 8.5 shows that the distinction between *careful* and *not careful* is more noticeable in the beginning of the handover than at the end. Based on this observation, the type of behavior is classified *during* the movement instead of recognizing the action only after its completion. By doing so, it is intrinsically embedded an anticipatory capability to the overall pipeline.

8.3.2 Acceleration Phase

In order to learn the latent features in the acceleration phase, a new approach is selected. The reason for opting for a new approach instead of applying the previous model technique is due to not being capable of finding distinct features for the two DS in the acceleration phase. This happens because all handovers, regardless of the empty or full condition, start with zero

velocity and in a stationary position. On account of this, the DS output would render the generated desired velocities of both Careful and Not Careful behavior indistinguishable. As an alternative to GMR to model the acceleration phase of the handovers the covariance matrix of the GMMs is used. The covariance matrix Σ of the GMMs that encode $\dot{x} = \mathbf{f}(x)$ expresses the correlation between the velocity and position in the handover space. The 1st Gaussian represents the steepest phase in the acceleration and from $\Sigma\vec{v} = \lambda\vec{v}$ the 1st eigenvector $\vec{v}$ for both cup conditions is indicative of the direction of largest data variance. The eigenvector components $\vec{v} = [\vec{v}_{\dot{x}}, \vec{v}_x]^T$ are the velocity and distance component respectively, and the

$$\frac{\vec{v}_{\dot{x}}}{\vec{v}_x}$$

gives the inverse of time (velocity divided by position), i.e. the frequency of change of the wrist. As discussed in Section 8.2, the velocity profiles are usually distinct when manipulating empty and filled with water cups, therefore the acceleration model learned the "frequency" of the wrist ($\frac{\vec{v}_{\dot{x}}}{\vec{v}_x}$) for either condition which in the Modelling block represents the Careful and Not Careful behavior.

8.3.3 Classification

For the purpose of classifying human motions, when interacting with either humans or robots, a belief system was implemented. The objective is to compare the human wrist motion of the handover against the learned *careful* vs *not careful* motions. The classification method uses as input the human wrist data (position, velocity) and, at each time step, it computes the desired velocities for the two DS models, the error metric to compare the velocities, and then outputs a belief system of the "carefulness" level. The expression for the classification follows [Khoramshahi and Billard, 2019] and is:

$$\begin{aligned} B &= [b_1, b_2] & \textstyle\sum_{i=1}^{2} b_i = 1 \\ b_i^{t+1} &\leftarrow b_i^t + \dot{b}_i^t \Delta t \\ \dot{b}_i^t &= \epsilon(e^T \mathbf{f}_i + (b_i^t - 0.5)|\mathbf{f}_i|^2) & \epsilon \in \mathbb{R}^+ \\ e &= X - \textstyle\sum_{i=1}^{2} b_i^t Y_i \end{aligned} \tag{8.3}$$

where B provides the belief that the new handover resembles one of the trained models. This is calculated as the error function comparing the real input data and the output of the trained models. The ϵ is the adaptation rate hyperparameter common in both approaches. It weighs the effect of past information on the current step, i.e. memory from the beginning. $\mathbf{f}_i$ is the model output for each of the motion behaviour, b_i^t is the classification output (belief), at time t, for each model $i = 1, 2 :=$ (*not*, *careful*). For real-time classification, the B vector is read at each time step, and when one of the beliefs (b_1^t or b_2^t) reaches 1 (100%) the information is sent to the robot to update its state depending on the HRI scenario in Section 8.5.

For the acceleration model, the real-time "frequency" of the wrist is computed as

$$X = \frac{\dot{x}^t - \dot{x}^{t-1}}{x^t - x^{t-1}}$$

and the generated $Y = \frac{\vec{\dot{x}}}{\vec{x}}$, as the "frequency" of the wrist for Careful and Not Careful. For the deceleration model, the real-time velocity is the current velocity

$$X = \dot{x}^t$$

and the generated $Y = \mathbf{f}$ are the DS velocities. The belief $\dot{b}$ in Equation 8.3 changes to:

$$\dot{b}_i^t = \epsilon(\dot{e}^T \mathbf{f}_i(x) + (b_i^t - 0.5)|\mathbf{f}_i(x)|^2) \tag{8.4}$$

where $\dot{e} = \dot{x} - \dot{x}_d$, $\epsilon \in \mathbb{R}^+$ is the adaptation rate, $\dot{x}_d = \sum_{i=1}^2 b_i^t \mathbf{f}_i(x)$, $\mathbf{f}_i(x)$ is the desired velocity for each DS given the current x. b_i^t is the belief, at step s, for each DS in $i = \{1, 2\}$, and $\sum_{i=1}^2 b_i^t = 1$. Equation 8.4 provides a vector of belief-updates $\dot{B}^t = [\dot{b}_1^t, \dot{b}_2^t]$ which are updated following a winner-take-all process. The winner-take-all aims at favoring the DS model which is considered most similar to the real human motion. The final step is reserved to update the belief $B = [b_1, b_2]$, where $b_i^{s+1} \leftarrow b_i^s + \dot{b}_i^s \Delta t$, for i = 1, 2. The belief system B converges to $b_1 = 1$ or $b_2 = 1$ depending on whether the human motion resembles a *not careful* behavior, or a *careful* behavior, respectively. Figure 8.7 shows the output of B at each time step s for various human trajectories.

Comparing the two phase models

The previous modelling approach (the deceleration phase model) has some limitations. Foremost, it is focused on the latter stage (the deceleration phase), resulting in a later classification. The Belief System classifies the motion at the beginning of the deceleration trajectory where the two DS diverge. However, for handover data outside the trained region, i.e. regions where the velocities are far greater than in the dataset, the data can not be accurately compared with the two DS. This new handover trajectory can occur outside the joint distribution of both DS which would generate unpredictable GMR outputs. One drawback of the second modelling approach (the acceleration phase model) relates to segmentation. It is more challenging to extract the acceleration phase since it involves identifying the precise moment of the pickup which, due to sensor noise and occlusions during grasping, is prone to errors. This problem does not occur in the deceleration phase, making it simpler to extract from the dataset. As a workaround, it was decided to add a low pass filter during training and testing. This low-pass filter ignores the small velocities, which mainly occur right after pickup and in the final stage of the handover (which is not part of the second model). This solution improves the classification accuracy without influencing the real-world performance since the first samples

that are ignored by the low-pass filter are not informative enough to distinguish the motion.

The models discussed in this section, the acceleration and deceleration models, provide two possibilities of understanding human cup manipulations in the presence of varying liquid levels. The next section is reserved for analysing these two models in great detail. It starts by comparing both models on the dataset of Section 8.2, it then evaluates the effect of the ϵ parameter in the classification step, proceeding with an in depth exploration of the novel model, the acceleration phase. This involves studying the impact of different cup materials and properties while testing for other two datasets (QMUL and IST datasets). These datasets will present unseen challenges such as new cups, participants, and new data acquisition techniques.

8.3.4 Robot Control

The robot is represented as a rigid-body with n degrees of freedom described in the m-dimensional Cartesian space. The dynamics of a rigid-body robot are formulated as

$$M(q)\ddot{x}_R + C(q,\dot{q})\dot{x}_R + G(q) = F_c + F_{ext} \tag{8.5}$$

where $q \in \mathbb{R}^n$ denotes the joint configuration and $x_R \in \mathbb{R}^m$ the robot pose. Moreover, $M \in \mathbb{R}^{m\times m}$ is the mass matrix, $C \in \mathbb{R}^{m\times m}$ represents the centrifugal forces, and $G \in \mathbb{R}^m$ denotes the gravitation forces. The terms on the right are respectively the control $F_c \in \mathbb{R}^m$ and external wrenches $F_{ext} \in \mathbb{R}^m$ forces.

For the controller the impedance controller proposed by [Kronander and Billard, 2016] provides the desired stability and passive physical interaction.

$$F_c = -D(\dot{x}_R - \dot{x}_{Rd}) + G(q) \tag{8.6}$$

where $D \in \mathbb{R}^{m\times m}$ is a diagonal with positive entries. $\dot{x}_{Rd}$ is the desired velocity generated by the DS.

Since the desired end-effector's accelerations are known, an inverse dynamics formulation is applied in order to compute the required joint-level torques needed to achieve the target accelerations. The problem is formulated as a quadratic programming problem (QP):

$$\begin{aligned} &\min_{\mathcal{X}} -0.5\mathcal{X}^T\boldsymbol{G}\mathcal{X} + \boldsymbol{g}^T\mathcal{X} \\ &\text{s.t. } \boldsymbol{A}_E\mathcal{X} = \boldsymbol{b}_E \\ &\quad \boldsymbol{A}_I\mathcal{X} \geq \boldsymbol{b}_I \end{aligned} \tag{8.7}$$

The equations of motion and constraint equations for a robot with rigid bodies can be described

as[2]:

$$\begin{aligned} \boldsymbol{M}(\boldsymbol{q})\ddot{\boldsymbol{q}} + \boldsymbol{C}_g(\boldsymbol{q}, \dot{\boldsymbol{q}}) &= \boldsymbol{\tau} \\ \boldsymbol{J}(\boldsymbol{q})\ddot{\boldsymbol{q}} + \dot{\boldsymbol{J}}(\boldsymbol{q}, \dot{\boldsymbol{q}})\dot{\boldsymbol{q}} &= \ddot{\boldsymbol{x}} \end{aligned} \tag{8.8}$$

where $\boldsymbol{q}$ is the full state of the system (including the 6-DOF of the floating base if the robot is not fixed) , $\boldsymbol{M}(\boldsymbol{q})$ is the inertia matrix, $\boldsymbol{C}_g(\boldsymbol{q}, \dot{\boldsymbol{q}})$ is the sum of the gravitational, centrifugal and Coriolis forces, $\boldsymbol{J}$ is the concatenation of the Jacobians of all the contact points, and $\boldsymbol{x}$ is the concatenation of the poses (containing position and orientation) in Cartesian space of all the contacts. x_R is the end effector pose of the robot. The equations of motion are re-written as:

$$\begin{bmatrix} \boldsymbol{M}(\boldsymbol{q}) & -\boldsymbol{S} \end{bmatrix} \begin{bmatrix} \ddot{\boldsymbol{q}} \\ \boldsymbol{\tau} \end{bmatrix} + \boldsymbol{C}_g(\boldsymbol{q}, \dot{\boldsymbol{q}}) = 0 \tag{8.9}$$

where $\boldsymbol{S}$ is a selection matrix where the first 6 rows are all zeros and the rest is the identity matrix. Given this formulation, the state $(\boldsymbol{q}, \dot{\boldsymbol{q}})$ give rise to linear equations for the motion with respect to $\begin{bmatrix} \ddot{\boldsymbol{q}} & \boldsymbol{\tau} \end{bmatrix}^T$. By defining $\mathcal{X} = \begin{bmatrix} \ddot{\boldsymbol{q}} & \boldsymbol{\tau} \end{bmatrix}^T$, it is now possible to formulate the inverse dynamics as a QP problem. In particular, turning the equations of motion to equality constraints ($\boldsymbol{A}_E$ and $\boldsymbol{b}_E$), and turning joint limits and other constraints into inequality constraints ($\boldsymbol{A}_I$ and $\boldsymbol{b}_I$). The desired accelerations of some end-effector are defined by filling $\boldsymbol{G}$ and $\boldsymbol{g}$ appropriately. In this chapter, the desired end-effector accelerations is defined by:

$$\ddot{\boldsymbol{x}}^* = K_p(\boldsymbol{x}_d^* - \boldsymbol{x}) + K_d(\dot{\boldsymbol{x}}_d^* - \dot{\boldsymbol{x}}_d) + \ddot{\boldsymbol{x}}_d^* \tag{8.10}$$

where $\boldsymbol{x}_d^*, \dot{\boldsymbol{x}}_d^*, \ddot{\boldsymbol{x}}_d^*$ are specified by a higher-level controller, can change over time, and define the current task. Depending on the output of the classifier, different gains (K_p, K_d) are chosen to perform the high-level tasks.

8.4 Results for Human Datasets

8.4.1 Evaluation of both models

From the results in Table 8.1 the acceleration model is better than the previous deceleration model and those conclusions are presented below. The evaluation metric was chosen as the classification accuracy of each model when splitting the dataset into types of cups. This means that careful accuracy is how many transportations of full cups are considered careful manipulations, and not careful accuracy is how many empty cups are considered not careful manipulations. Table 8.1 shows that the acceleration model is better at detecting handovers of full cups as careful manipulations. This is desirable given the challenging nature of

[2]The possible contact points are ignored for brevity/clarity as they are not used.

transporting water in cups. Furthermore, it can be concluded that an empty cup does not imply a not careful (careless) manipulation. Although the deceleration model is better at detecting handovers of empty cups as not careful manipulation it comes at a cost of not detecting most full cups as careful. Since the HHI experiment allowed participants to choose a preferable handover strategy, an empty cup restricts the movement less than the same cup filled with water (restriction on the orientation, oscillations, velocity, etc). As a result, the empty cup handover should reflect the user preference, either handover normally (not careful), or restricted (careful), and assuming that the dataset is a fair representation of both types of people, the empty cup should not reflect any preferred carefulness motion.

Type of Cup			**Carefulness Detection**			
			Acceleration Model		Deceleration Model	
Train	**Test**	Real / Predicted	Empty	Full	Empty	Full
Red Cup	Red Cup	Not Careful	0.77	0.17	**1**	0.2
		Careful	0.23	**0.83**	0	0.8
Champagne	Champagne		0.4	0	**0.82**	0.27
			0.6	**1**	0.18	0.73
Transparent Cup	Transparent Cup		0.43	0.33	**0.65**	0.39
			0.57	**0.73**	0.35	0.61
Wine Glass	Wine Glass		**0.57**	0.23	0.5	0.43
			0.43	**0.77**	0.5	0.57

Table 8.1: Train set: One cup; Test set: Same cup. Higher value in the prediction is marked in **bold**

8.4.2 Results for adaptation rate (ϵ) values

The ϵ is the hyperparameter present in the classifier. It is a weighted parameter on the knowledge of past iterations. Figure 8.6 shows that as ϵ increases the careful accuracy drops, while the not careful accuracy increases. When increasing the adaptation rate the system is sensitive to initial noise and spurious data, it reaches a classification quicker (quicker response time) which results in more incorrect decisions. The not careful accuracy increasing as the ϵ rises are the result of being influenced by initial spurious points in the trajectory.

The region for the best accuracy models is the ϵ interval between [0.05 - 0.2] (indicated in Figure 8.6 by the dotted ellipse). This is a trade-off between a faster prediction time and good accuracy. Inside this region, 75% or more of full cups are classified as careful manipulations, and around 50% of empty cups are not careful manipulation (which is ideal given the personal preference choice when handling empty cups).

To evaluate the picked region for the ϵ value, the accuracy is computed for the whole handover trajectory for one of the proposed models. Figure 8.7 plots the accuracy of a sample of handover trajectories in the EPFL-KIT dataset using the deceleration phase model. The results from the plot indicate that our method can generalize and distinguish most situations of

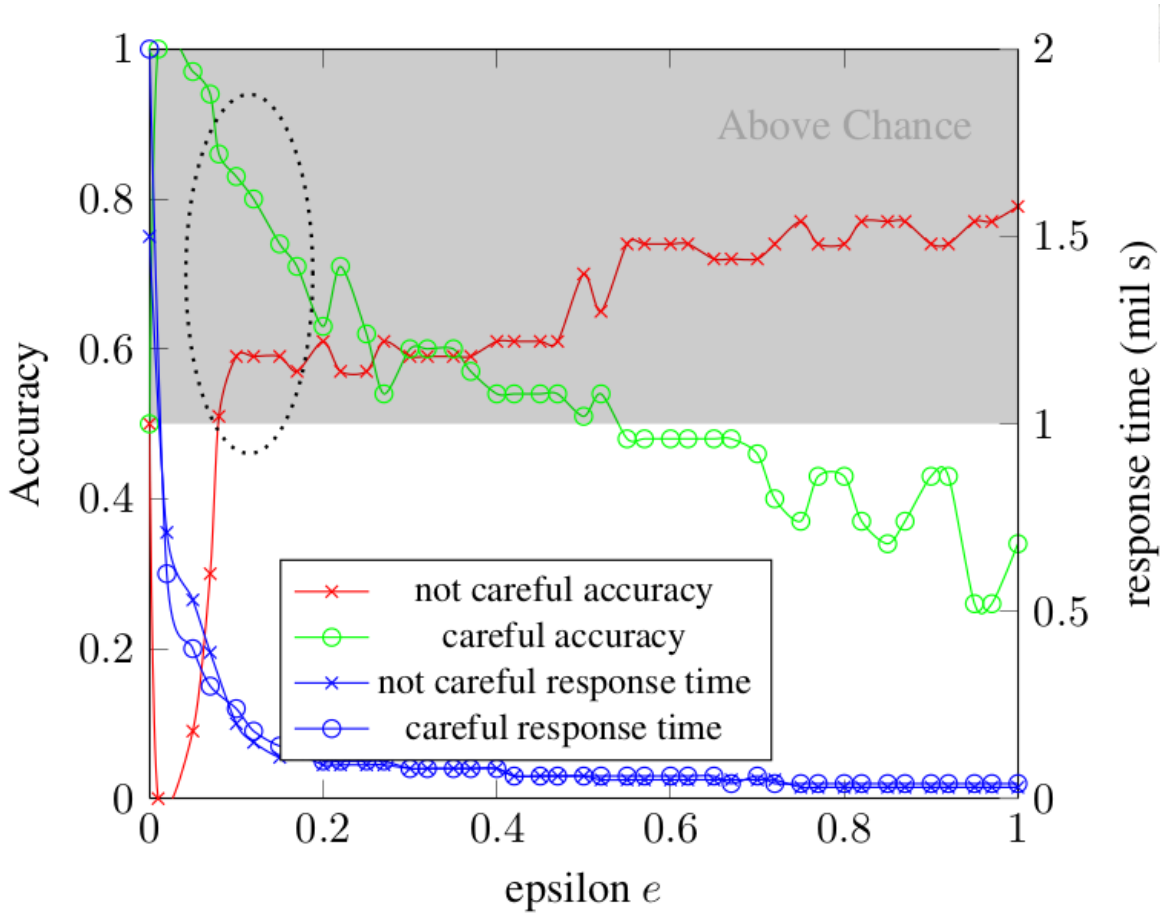


Figure 8.6: The evolution of models accuracy and respective response time of the prediction for each value of epsilon.

careful and *not careful* manipulations of cups, where the *not careful* situation is considered a normal human handover motion. As Figure 8.7 demonstrates, it is only required around 20-40 time steps in the human handover motion to accurately predict the "carefulness" behaviour. Good generalization of the classification is expected to new subjects and cups (not seen during training), as humans follow similar dynamics of reach motion in normal circumstances [Lemme et al., 2015].

8.4.3 Results on type of cups

The following study's objective is to analyse the different properties of cups and the impact on detecting careful manipulations. The analysis comprised of training the models with different samples of the dataset, specifically, splitting the dataset according to different types of cups, and testing the accuracy of the models against other types of cups. Similar to what was achieved in the first experiments of this Section, however, the test set is of handovers from other cup types. From Table 8.2 it can be concluded that unknown full cups are predominantly classified as careful manipulations irrespective of the type of cup the model is trained on. While empty cups, regardless of the type of cups trained on, give rise to a non-preferential manipulation.

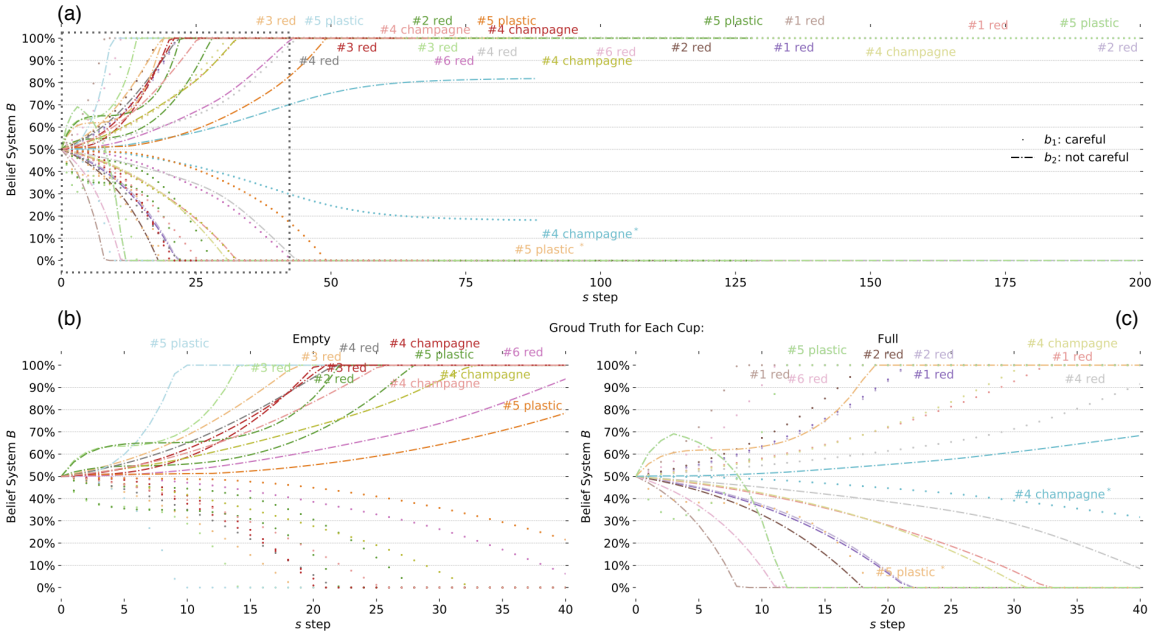


Figure 8.7: The belief system B output. (a) full length of HHI trajectories. #n ${type of cup} is the label of the n participant and the type of cup grasped. Additionally, the position of the label marks the classification result of the final s step of the handover trajectory. (b) and (c) are the highlighted region in (a) for the two conditions: empty, and full, respectively. The * represents the trajectories with a wrong classification.

Training on one plastic type and testing solely on the other plastics give rise to the conclusion that training and testing on the same cup material (such as plastics) achieves the best results since it induces similar characteristics, e.g. risk of breaking, friction, weight, etc. Training the model on glass and testing on plastics it can be concluded that training a model on glass and testing on plastics worsens the likelihood of detecting full cups as careful manipulations. Although the dataset only has one glass cup it can be hypothesized that this could be induced by the risk of breaking, and the fact that glass is heavier than plastic.

Further discussion allows to infer differences when comparing types of plastic cups. Soft plastics are deformable due to their physical structure and material composition, thereby are prone to deforming. This is exacerbated when are filled to the top with a liquid. Rigid plastics are non-deformable as they do not present the same structural flaws as soft plastic, and are non-breakable, contrarily to glass which can break and shatter easily. The model trained on soft plastics (red and transparent cup) and tested on rigid plastics (champagne cup) produce worse careful accuracy. It can be argued that this is the cause of deformability which makes it difficult to handle soft plastics when filled with water compared to rigid plastics. The model learned (intrinsically) the deformability feature in the full cup case and since the testing set does not have cups with that property, it became difficult to detect the full cup cases in the test split. On the other hand, training a model on rigid plastics and testing on soft plastics provide the highest level of careful accuracy (with a bias to plastics only). The rigid plastic is not affected by any of the latent features (deformability or breakability), hence the model did not learn to be extra sensitive. In the testing set, the deformability feature was present (soft plastics only) and since the effect is mainly present in full conditions, the model was capable

Type of Cup			Acceleration Model			
			Training Set		Testing Set	
Train	**Test**	Predicted \ Real	Empty	Full	Empty	Full
Transparent Cup	Red Cup Champagne Wine Glass	Not Careful Careful	0.688 0.312	0.15 0.85	**0.46** 0.54	0.1 **0.90**
Champagne	Transparent Cup Red Cup Wine Glass		0.5 0.5	0.1 0.9	**0.53** 0.47	0.15 **0.85**
Red Cup	Transparent Cup Champagne Wine Glass		0.47 0.53	0 1	**0.5** 0.5	0.16 **0.84**
Wine Glass	Transparent Cup Red Cup Champagne		0.5 0.5	0.25 0.75	**0.59** 0.41	0.17 **0.83**
Champagne	Transparent Cup Red Cup		0.5 0.5	0.1 0.9	**0.55** 0.45	0.08 **0.92**
Transparent Cup Red Cup	Champagne		0.61 0.39	0.15 0.85	**0.625** 0.375	0.3 **0.7**

Table 8.2: One vs Rest Classification. Training set: One cup type; Testing set: Other cup types. Plastic cups; Glass cups

of easily distinguishing the two carefulness levels.

Our conclusion is that a model trained on soft (deformable) plastics learns that full cup actions are extremely difficult, hence these actions for Rigid plastics have a higher likelihood to be classified as not careful since the rigid plastic is considered to be a cup with no inherent challenging properties. A model trained on rigid plastics learns the opposite, considering full cup actions for soft plastics as mostly classified as careful manipulations.

8.4.4 Results of entire datasets

Figure 8.8: Extracted frames of handover actions from the QMUL dataset (the two left most images) and the IST dataset (the two right most images).

Two other scenarios are included to further evaluate the carefulness detection accuracy of the proposed models. One dataset is part of the CORSMAL Containers Manipulation project where participants perform a series of tasks on a set of containers. The tasks involve pouring water, rice, or pasta into containers such as cups or boxes, and initiating a handover towards a

Train	Test	Real / Predicted	Acceleration Model: Training Set Empty	Training Set Full	Testing Set Empty	Testing Set Full
EPFL 10%	EPFL 90%	Not Careful	0.83	0	**0.53**	0.19
		Careful	0.17	1	0.47	**0.81**
EPFL 20%	EPFL 80%		0.89	0.12	**0.5**	0.16
			0.11	0.88	0.5	**0.84**
EPFL 30%	EPFL 70%		0.69	0.09	**0.51**	0.15
			0.31	0.91	0.15	**0.85**
EPFL 40%	EPFL 60%		0.6	0.13	**0.55**	0.1
			0.4	0.87	0.45	**0.9**
EPFL 50%	EPFL 50%		0.61	0.16	**0.5**	0.1
			0.39	0.84	0.5	**0.9**
EPFL 60%	EPFL 40%		0.55	0.08	**0.5**	0.15
			0.45	0.92	0.5	**0.85**
EPFL 10%	QMUL	Not Careful	1	0	**0.5625**	0.2
		Careful	0	1	0.4375	**0.8**
EPFL 20%	QMUL		0.92	0.14	**0.5**	0.2
			0.08	0.86	0.5	**0.8**
EPFL 40%	QMUL		0.94	0.18	**0.5**	0.13
			0.06	0.82	0.5	**0.87**
EPFL 10%	IST	Not Careful	0.45	0.02	**0.65**	0.18
		Careful	0.55	0.97	0.35	**0.82**
EPFL 20%	IST		0.58	0.11	**0.51**	0.22
			0.42	0.88	0.49	**0.78**
EPFL 40%	IST		0.55	0.10	**0.52**	0.23
			0.45	0.89	0.47	**0.77**

Table 8.3: Top: Train set - sample of EPFL; Test set - rest of EPFL dataset. Middle: Train set - sample of EPFL; Test set - QMUL dataset (new people and new cups). Bottom: Train set - sample of EPFL; Test set - IST dataset (new people and transparent cup).

robot. The dataset, referred to as the QMUL dataset[3], includes four cameras, one attached to the human, one attached to the robot, and two looking from each side (left images in Figure 8.8 shows both sides), and one microphone. The cup location is estimated using a multi-view projective geometry which provides a 2D centroid of the cup from the two side cameras at 30 Hz sampling frequency [Xompero et al., 2020]. The other dataset involves participants in pairs interacting with cups where they both perform pick & place and handover actions. The dataset, referred as the IST dataset[4], includes two head-mounted eye trackers (right images in Figure 8.8), one on each participant, and OptiTrack Mocap markers on the head and wrist of the participants to record the motion (recorded at 120 Hz). Table 8.3 shows the results for the three datasets: (i) the EPFL-KIT, (ii) the QMUL, and (iii) the IST. The experiments were accomplished by training the models using varying percentages of the EPFL-KIT dataset. The accuracy results of each model indicate that it can identify most of the full cups handovers as careful manipulations and corroborate the idea that handovers of empty cups are dependent on human preference, and not conditioned by cup properties. An interesting finding is seeing the not careful accuracy in the training set decrease when trained on large datasets while the testing set accuracy for both classifications remains fairly similar. This can be argued as another proof that for large datasets, the model that best generalizes is the one that assumes that empty cups do not necessarily invoke a not careful (natural) manipulation. The conclusions are three

[3] https://corsmal.eecs.qmul.ac.uk/containers_manip.html
[4] https://vislab.isr.tecnico.ulisboa.pt/datasets_and_resources/#hcups_water

fold: (i) the model generalizes well for unknown people and cups, (ii) all 3 datasets results show that for empty cups there is no underlying preference of manipulation (i.e. empty cup is not necessarily not careful), and (iii) the Carefulness detection controller can achieve good accuracy for either precise data points (Mocap markers) and 3D point estimation (from stereo vision).

Discussion

In all datasets, as the amount of data increases in the training set, the best performance models have the classification of empty cups as not-careful near the 50% mark. This makes sense as more data is included in the training set showing more cases where empty cups are manipulated with more freedom - hence the preference of participants is noticeble. In the IST dataset, the effect is visible in every size of the training set. Overall, the model is capable of generalizing to different human-to-human handover scenarios and different data acquisition techniques.

In terms of cup properties, the analysis can be summed by Figure 8.9. It is known that glass is breakable while plastic usually is not. Soft and hard plastic cups are sharing most of the properties, however when filled with water the soft plastic may deform due to the weight of the liquid inside. As a result, soft plastics are characterized as deformable but not-breakable, while hard plastics are not-deformable and not-breakable.

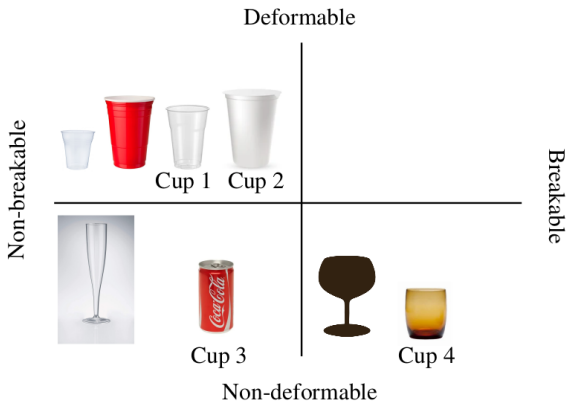


Figure 8.9: An illustration scheme of the important features of cups during manipulation: deformability, and breakability. Deformability is evident solely when filled with water, while breakability is an inherent property of the cup.

From the results of Table 8.3, hard plastic cups are the easiest to manipulate as they do not present a challenge of breakability or deformability. The glass is not-deformable but breakable hence it is more challenging to manipulate than the hard plastic cups. Soft plastic cups are deformable but not-breakable so are also more difficult than hard plastics. It is hard to quantify which of the two challenges has a higher priority when it comes to manipulation strategy. It could be argued that breaking a glass cup is worse and irreversible compared to deforming a cup. However, in this situation, the deformability only manifests when the cup is filled with

water, hence it is fair to conclude that a glass cup is the hardest to manipulate and would influence the manipulation strategy the most.

8.5 Robot Experiments using the Deceleration Phase Approach

In this section it is detailed the HRI experiment performed to evaluate the deceleration phase approach. The robot platform used is a KUKA LBR iiwa 7-DoF manipulator (14 kg payload)[5] and the Robotiq 2F-85 2-finger gripper[6] is attached to robot's end-effector to perform grasping and manipulation. The HRI scenario is as follows: a human picks a cup from a table and places it on a shelf, as shown in Figure 8.10. The cup's position is provided to the pipeline as the input during human manipulation. The cup's position is set as reference, instead of the human wrist, to simplify the experiments with different participants. The human motion is extracted from the cup's motion and the classifier predicts whether it is in the presence of a *careful* or *not careful* motion. Afterwards, the controller from Section 8.3.4 adapts the robot's motion and gripper to the desired behaviour.

Figure 8.10: Setup outside perspective for the Pick and Place task.

The pick & place motion is generated for the end-effector position using the linear dynamics $\dot{x_d} = K_p(x_R - x_t)$ where $K_p \in \mathbb{R}^3$ is diagonal with positive entries, and x_t is the target location which is specific to the task. The robot end-effector is controlled with Cartesian velocity as input. To complete the robot task three instances of the same dynamics are generated: (i) approaching the picking location, (ii) retreat from shelf, and (iii) approaching the placing location. The switch to the next dynamics happens when x_t (the attractor point) is close enough $\| x_R - x_t \| < \sigma$. Depending on the output of belief system, the robot controller manipulates the cup keeping the gripper's orientation fixed, to prevent spilling (*careful*), or allows for an unrestricted movement (*not careful*).

The HRI experiments involved 4 participants (all male and right-handed, age between

[5] https://www.kuka.com/en-ch/products/robotics-systems/industrial-robots/lbr-iiwa/
[6] https://robotiq.com/products/2f85-140-adaptive-robot-gripper/

Table 8.4: "Carefulness" level predicted on unknown people.

True \ Predicted	Not Careful	Careful
Empty	**0.80**	0.20
Full	0.07	**0.93**

28-34, with some experience with robots) picking and placing the same cup onto the shelf under two conditions: (i) empty, and (ii) full cup. Each participant performed 10 trials for each condition. The cup used was not present in the HHI dataset. The results of the classification in Table 8.4 for new subjects, with a new cup, proves that our pipeline can correctly distinguish *careful* and *not careful* manipulation of cups by only varying one underlying condition: empty vs full of water.

The pipeline invokes the robot to behave according to the desired "carefulness" (i.e. the type of object and filling condition), taking longer in the *careful* case. In the *not-careful* behaviour, the robot grasps the cup and places it a manner that is simpler for the robot's configuration, taking less time to complete the action. The downside is that the robot spills its content. The approach developed allows for real-time HRI where the robot can adapt the behavior of manipulating a cup according to the human motion behaviour.

8.6 Robot Experiments using the Acceleration Phase Approach

In this section it is detailed the HRI experiment performed to evaluate the acceleration phase approach. The model parameters chosen is of the highest accuracy model from Table 8.3 which generalizes best for every cup and dataset. These HRI experiments are three-fold: (i) pick-and-place, (ii) handover task, and (iii) robot assistance. The robot platform used is the Kinova gen3 with a Robotiq gripper attached to the end-effector as seen in Figure 8.11. The Kinova robot was controlled using the kortex_ros[7] package for ROS. For the pick & place and handover tasks the robot follows the linear dynamics $\dot{x_d} = K_p(x_R - x_c)$ where $K_p \in \mathbb{R}^3$ is diagonal with positive entries, and x_c is the cup's location. To complete the robot task two instances of the same dynamics are generated: (i) approaching the cup, (ii) placing the cup in final location. The switch to the next dynamics happens when x_t (the attractor point) is close enough $\| x_R - x_t \| < \sigma$. For each of the three HRI experiments, the objects (cups and box) are tracked by the OptiTrack Mocap system which is streaming the data to ROS at 120 Hz. The model from Section 8.3.2 has the accuracy of Table 8.3. The ϵ value is picked inside the region mentioned in Section 8.4.2.

The control loop system presented in Figure 8.4 is a human-in-the-loop controller running at 120 Hz during the human-robot scenario. For simplification, the human data is processed by the movement of the cup which always starts stationary on a table or the floor in case of the

[7]The official repository to interact with the Kinova robot https://github.com/Kinovarobotics/ros_kortex

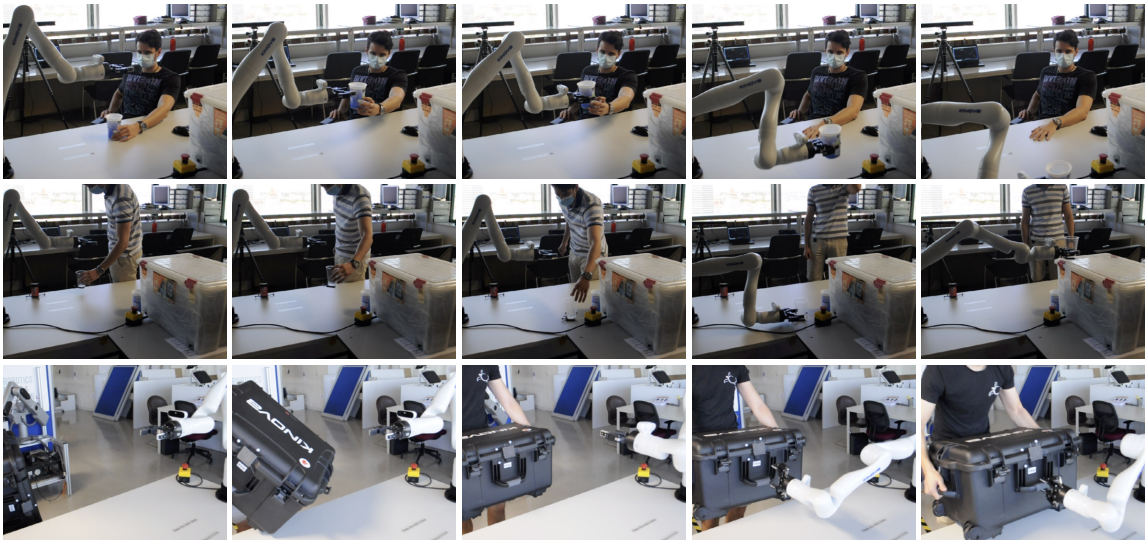


Figure 8.11: Each row of images illustrate the three HRI scenarios where the Carefulness detection controller using the acceleration model is applied. The first row is the human-robot handover, the second row is the human-robot pick & place of cups, and the third row is the robot assistance to a human carrying a heavy box. In the third application setup the human was lifting the box at the right side of the robot to prevent occlusions of the box's markers.

box in scenario (iii). The human-in-the-loop system begins predicting the human manipulation the moment the cup has non-zero velocity. The Belief System classification B then outputs to the robot controller either not-careful (1) manipulation or careful (2) manipulation. This result can then be applied to several HRI scenarios.

8.6.1 Human and Robot Pick and Place

This HRI experiment has been applied to the deceleration phase model (shown in the previous section) and it involves a human picking cups from a table and placing them in another location. In the not-careful manipulation option, the robot transports the cup to a bottom shelf, without worrying about any danger of spilling (tilting the cup). As for the careful option, the robot transports the cup, keeping the orientation fixed while slowing its velocity, and placing it on the top shelf. Table 8.5 shows the successfulness of the human-in-the-loop system of adapting correctly to the present cup conditions. A total of 4 subjects participated in the experiment and manipulated 4 different cups with the two conditions (empty and full of water). Two cups are from the category of non-breakable and deformable properties (soft plastics), and the other two are a rigid plastic cup and a glass cup, respectively. Each participant manipulated the cup 10 times per cup and per condition. As a comparable variable, the cup present in the previous section is also included in the experiments as Cup 1.

8.6.2 Human-Robot Handover

The second scenario involves a human-robot handover of cups where the robot tries to infer, from human motion, whether the cup requires a careful manipulation or not. The main difference to the previous scenario of Section 8.6.1 has to do with the proactiveness

Table 8.5: Results of Pick & Place and Handover experiments. Properties of Cup 1-4 are shown in Figure 8.9

		Not Careful	**Careful**
Cup 1	Empty	**0.65**	0.35
	Full	0.10	**0.9**
Cup 2	Empty	**0.6**	0.4
	Full	0.057	**0.95**
Cup 3	Empty	**0.68**	0.32
	Full	0.22	**0.78**
Cup 4	Empty	**0.55**	0.45
	Full	0.15	**0.85**

of the robot that, instead of waiting for the cup to be placed, meets the human in order to perform the handover. This, as mentioned before, is one of the advantages of analysing the acceleration phase of the motion with regards to the previous method in Section 8.3.1. The results are present in Table 8.5 since no major changes to the controller were implemented. The same subjects participated in both experiments and manipulated the same cups for both conditions. The classification results are calculated during the acceleration phase hence there is no difference in waiting for the object to be picked up or handed over. Nonetheless, the accuracy is the result of all the trials classifications for both pick-and-place and handover actions.

It can be concluded from the pick & place scenario that it reaches good if not better results in detecting cups filled with water as careful manipulations. When comparing the results from the previous deceleration model, the Cup 1 results match the ones observed in the previous work validating both models as good detection mechanisms for human manipulation during pick & place tasks. However, these results also extend to other HRI applications such as handovers and given the architecture of the model the results achieve the same accuracy for both applications. To note, due to the risk of spilling water, the participants' trials ran in real-time but without robot participation. The Carefulness detection controller did output the commands to the robot and Figure 8.11 illustrates how the robot interacts in each scenario in a careful and not careful situation. The supplementary video video.tcds.ieee-2022 shows clearly the different robot responses for each scenario.

8.6.3 Robot Assistance

In the last scenario, the context is changed. On a different note, it was decided to move away from the realm of cups and carefulness manipulation detection and aim for another potential use case. The recognition of different manipulation strategies can also be applied to household activities such as lifting boxes, furniture, appliances, etc. The human is carrying a large box and the robot has to infer if the human requires assistance in lifting the box due to being too heavy. The robot, if it detects that the human is struggling to transport the box would grasp onto the side handle of the box and pull, in order to assist the person in placing the box

on the table. In this case, two instances of the linear dynamics are generated: (i) approach the box, (ii) after holding the box, pull it back to the table. If, on the contrary, the human does not exhibit any challenge in lifting and transporting the box, it would not interact and leave the human unassisted. Table 8.6 shows the results for detecting whether the human was having trouble lifting the box given the human's motion behaviour. This final experiment shows that it is possible to detect, fairly accurately, when a human is having difficulty in moving a large or heavy object by simply observing the motion pattern of the item being moved. Although light objects, similarly to empty cups, do not reflect one particular strategy, which once again is indicative of human preference. We validate with one participant performing 20 trials for each condition as this is solely a proof of concept.

Table 8.6: Results for Robot Assistance experiments.

		Easy	**Hard**
Box	Light	0.78	0.22
	Heavy	0.12	**0.88**

8.7 Remarks

This chapter studied the human hand motion during handovers of cups in two conditions: carrying an empty cup, or a cup filled with water. These experiments explored several datasets with data acquired from different sensors during handovers between two humans, or handovers simulated by a single person. Each dataset had different types of cups with different materials, and several participants manipulated each cup multiple times, in the two conditions mentioned above. The results provide a broad and overall general analysis of the human motion behavior during handovers of cups in two relevant conditions (an empty vs a full water cup). From these two conditions, it is possible to distinguish motion strategies from humans when manipulating cups filled with water. This is a more secure, risk-free, option of moving objects when there is an apparent risk of spilling or danger compared to a normal handover between humans. Based on these findings, two computational models describing careful/careless handovers are developed. These computational models provide the robot with anticipatory knowledge of the type of manipulation, careful or not careful, thus facilitating the robot's motor preparation and the adaptation, prior to the interaction, allowing for a better understanding of the object inherent properties. A link is provided to a video that shows examples of the HRI scenarios working in real-time, the operation of the controller and the online classification of the action's carefulness - video.PropertiesCupsHRI.ieee-2021.

The overall conclusions from the HRI experiments are that the acceleration model clearly shows its advantages over the previous model with its multi-use in different robot applications. While the previous one had been only applied to pick & place due to its limitation of having to set the final meeting point, this new model gives information the moment the object (cup or

box) is picked by the human, making it versatile. As it was mentioned in the introduction, this model can be useful for many robot situations where humans play a vital role. Robots can learn a lot from humans and should take advantage of how humans tackle problems to better understand the world surrounding them. As a result, this robot controller aims to enhance the robot capabilities in understanding object properties from human manipulations.

9

Conclusion

The conclusion of this thesis is organized in four different sections. The first section elaborates on the conclusions reached in the part of the thesis exploring the imitation capabilities of robots to express human-like actions from non-verbal cues. The second section is the part on how humans communicate and how robots can use those cues to communicate their goals and intentions. The third section is a conclusion on the final part of the thesis focused on exploring the intricate details in non-verbal cues that reveal object latent properties during manipulation. Each part has its own conclusions, the current limitations and possible directions of future work. The final section is reserved for answering the research questions.

9.1 Imitating Human Actions

This thesis began by focusing on the importance of the robot's gaze for the overall readability of the coordinated motion. We have shown that merely the gaze information is enough to distinguish the end-goal. In contrast, the arm information alone results in a slower prediction since the subjects have to wait for the arm of the robot to start moving which is slower than the movement of the eyes. This work express the importance of non-verbal cues during a HHI, and the benefit of affording robots with the two-fold capacity: (i) interpreting those cues to read the action intentions of their human counterparts and (ii) to act in a way that is legible and predictable to humans.

Dragan et al. [2013] proposed two types of arm movements (predictable and legible), and demonstrated that a legible arm movement, which is an overemphasised predictable motion of the human arm, can give more information about the action that the human or the robot is going to do. When humans interact with each other, eye gaze movements have to support motor control as well as communication. On one hand, we need to fixate the task goal to retrieve visual information required for safe and precise action-execution. On the other hand, gaze movements fulfil the purpose of communication, both for reading the intention of our interaction partners, as well as to signal our action intentions to others. Sciutti et al. [2018] state the ability of the robot to anticipate human behavior requires a very deep knowledge of the motor and cognitive bases of HHI. There is a need to design robots to predict human needs and design robots to be predictable for humans. We propose an alternative to embed action legibility with overemphasized arm motions, and extend the motion model to incorporate eye gaze information. Our experiments showed that legibility of the robot's actions improves with the integration of human-like eye gaze behavior into the controller. Hence, our work generalizes Dragan et al. [2013], as legibility is achieved through the combination of both human arm, body, and eye-gaze movements. The resulting robot's behavior showed to be legible even for multiple sets of actions such as pick-and-place and handovers. To improve the work we plan to revisit the modelling of the arm in order to better coordinate the overall eyes/head/arm speed. In our implementation, the robot arm controller is slower than the actual human arm motion. The eye-gaze controller was not modelled at the same level of detail as the arm trajectories using GMMs. While this current model could qualitatively reproduce the human gaze-shift behaviors, its "human likeness" would need improvement. Part II tries to tackle this limitation by representing the gaze shift dynamics present in humans.

The second work continues on the same line of observing human non-verbal cues and reproducing them on the robot side to express the same action intention. We analysed human motion during polishing tasks and developed an approach capable of learning the polishing dynamics from human demonstrations and generate the required motor dynamics for the robot to replicate the human motion. We applied the robotic controller in a online human-in-the-loop system and the controller can recognize in real-time the human polishing motion and generate the appropriate robot movements to replicate the recognized polishing strategy.

One of the limitations is assuming that the polishing motion follows an elliptical limit cycle when there are other types of stable limit cycles. For example, The Van der Pol oscillator system van der Pol Jun. D.Sc [1926] can generate strong non-linearities which deform the limit cycle to a non-elliptic shape. Fantuzzi et al. [2016] presents some interesting limit cycles which can extend this work. The second assumption is the limit cycle dynamics keeping the linear velocity around the limit cycle attractor constant. This is definitely not the case in real human polishing motions and it could be relevant to detect and model different velocity strategies (e.g. polishing harder to remove a stain or smoother to shine a porcelain).

9.2 Understanding Human Intention while Expressing Robot Goals

In Part II we introduced a "*Gaze Dialogue*" between two participants working on a collaborative task involving two types of actions: *individual action* and *action-in-interaction*. We recorded the eye-gaze behavior of both participants during the interaction sessions and used the paired eye-gaze data to build the *Gaze Dialogue* model, encoding the interplay of the eye movements during the dyadic interaction. The model also captures the correlation between the different gaze fixation points and the nature of the action. This knowledge is used to infer the type of action performed by an individual. From the model, we designed a humanoid robot controller that provides inter-personal gaze coordination in HRI scenarios. During the interaction, the robot is able to adequately infer the human action from gaze cues, adjust its gaze fixation according to the human eye-gaze behavior, and signal non-verbal cues that correlate with the robot's own action intentions.

The objective of the *Gaze Dialogue Model* is to be a bi-directional non-verbal gaze communication system between a human and a robot. The system achieves that goal by reading the human gaze fixations and, at the same time, generating the appropriate robot gaze fixations. The second contribution in Part II explored the third level of interaction defined by Gallotti et al. [2017] where both agents, the leader and follower mutually adapt, and extended the alignment system in the *Gaze Dialogue Model* to focus on the adaptation of the leader to the responses of the follower. We use the human follower's gaze behavior data for two purposes: to determine (i) whether the follower is involved in the interaction, and (ii) if the follower's gaze behavior correlates to the type of the action under execution. This information is then used to plan the robot leader's actions in order to sustain the leader/follower alignment in the social interaction. During HRI the robot (i) emits non-verbal cues consistent with the action performed, (ii) predicts the human reaction, and (iii) aligns its motion according to the human behavior. This allowed for the robot, as the leader, to understand the engagement of the human follower and adapt its eye-gaze communication to the human responses.

Both previous systems learn from human eye-gaze cues to understand action intention and generate appropriate robot eye-gaze cues, but the former is focused on the robot reaction as a follower and the latter on the robot reaction as a leader during the HRI. The issue lies in the lack of interconnectivity between the two *Gaze Dialogue Model* pieces from Chapter 5. One of the goals of future work is to integrate both systems into one leader-follower response where the robot can adapt to both situations on the fly. Another direction is applying the *giving* and *placing* to other HRI scenarios. This has the goal of exploring the acceptability of the non-verbal communication cues exhibited by the robots to newer scenarios. Additionally, it is of interest to evaluate the quickness of the action prediction system in comparison to other approaches of human action prediction. Quantitative analysis of the human gaze fixations Dehais et al. [2011], as well as the reaction time are some of the metrics to evaluate the prediction capabilities of the model. Further is a more thorough evaluation of the impact of

the gaze behavior controller and motion planning alignment in the quality of the HRI. Future plan forsees enrolling a group of naive subjects in a HRI with the iCub running the gaze behavior model and compare it to an alternative controller. It will allow to analyze on how the human gaze reaction time correlates with the understanding of the robot's action, and the initiation of the arm movement towards the handover location to take the object from the robot. This "mutual" understanding of the action is a field worth exploring in the context of social interaction and collaborative tasks.

Eye trackers in the past were expensive and the outcome gaze-in-world data were difficult to analyze in any automated or semi-automated way. Nowadays with open-source projects it became more readily accessible and easy to use. In the future cameras will be capable of detecting human pupils at human-to-human distance so we can apply this controllers without having to add additional sensors to human subjects.

Human interaction involves very sophisticated non-verbal communication skills like understanding the goals and actions of others and coordinating our own actions accordingly. Regarding the motor resonance between humans and between humans and robots, our last contribution on Part II addresses the motor coordination that occurs in a dyadic interaction. We analyzed and modelled the arm motion cues exchanged between two humans in handover actions. We were able to show that the robot is able to interpret the human wrist motion and infer whether or not the observed action is a "handover", and use the motor resonance model to coordinate its actions with the human partner, during handover actions.

9.3 Inferring Object Properties

Observing how humans interact with each other and how they manipulate objects, offers insight on interaction mechanisms and how these are influenced by the physical properties of the used objects. These insights can support a more informed design of robot controllers meant to manipulate objects in collaboration with humans. In the computer vision and the robotics field there is the desire to discover the unknown characteristics of objects present in the scene. When it comes to cups and glasses one of toughest problems is to estimate the contents inside. The previous attempts deal with specific characteristics of cups and liquids Mottaghi et al. [2017] or particular scenarios. However, not all cups are transparent and the liquid may not be clearly visible (e.g. water). Moreover, you might not get the chance to view the cup from an angle that allows visualization of the liquid inside Do et al. [2016], or get to manipulate the cup to notice that it contains liquid inside Do and Burgard [2019], Schenck and Fox [2017b]. In most cases the cup or object is handed to a robot without any prior knowledge and the robot will need to use its sensors to discover some information from the object. Observing how the human is manipulating it, on the other hand, could provide relevant information on the contents of the cup. The second to last contribution in this thesis proposes a novel approach to infer the level of liquid inside a cup. The analysis of the eye-gaze

cues revealed changes in the handover depending on the level of water inside the cup. A model was proposed capable of learning how to classify three levels of water in a cup from human eye-gaze cues. The model was integrated in a robot controller for online classification of real-time human handovers. This approach has the advantage over previous ones in not being dependent on the cup transparency or liquid's color given that the information used to classify cups is based on the human eye-gaze response to cup manipulation and not on detecting the liquid level inside the cup. However the system is not capable of handling moments during the handover where the human is not fixating anything meaningful. In other words, the model is not susceptible to gaze diversions where no particular fixation is relevant at a certain time. This situation can happen in a natural human-robot scenario either because there is noise in the sensor or the human looked away for a brief moment. The next iteration of the model must learn to ignore this noise and irrelevant cues. If the occurrence is only sporadic then it should be ignored, if on the other hand it is prolonged then it could mean that the human is not paying attention. Moreover, it can even hypothesize that the cup is risk-free (no danger of spilling) hence it can be maneuvered without much worry.

The human-to-robot handover scenario has proven that the learned human-to-human handover of cups with different spilling risk levels is capable of classifying similar cups in a human-in-the-loop online interaction system with a humanoid robot. Future work involves extending the robotic architecture to include a robot-cup manipulation controller that, according to the classified cup, adapts the motor control strategy of the robot arm to prevent spilling. One final note can be pointed to the HRI scenario which, at the moment, does not provide any practical benefit. The robot is present and interacting with the human but there is no clear advantage gathered by recognizing the level of water inside the cup. Future work should be concentrated in discovering robot applications for detecting the liquid level in a cup.

The last contribution in Part III involves studying human-to-human handovers of cups filled with various amount of liquid and textures, and investigates to which extent the manipulation strategy depends on: (i) the individual preference, (ii) whether the cup is filled with water or not, and (iii) the cup physical properties. An analysis of the human giver's hand acceleration, velocity, and position during the handover of different cups under two liquid level conditions (full of water or empty), allows to distinguish between careful and not-careful (normal) manipulation. We quantified to which extent the liquid level inside the cups influences the carefulness level of human manipulation. We concluded that the cups' physical properties, such as fragility, breakability, and deformability, play a role in shaping the carefulness of the manipulation. We applied these findings to human-robot scenarios by developing a robot controller capable of detecting, in real-time, if the human is being more careful than normal, and adapting the robot's approach of interaction accordingly. It was shown that the detection of a careful manipulation, depending of the experimental context, provides the robot with information concerning the human partner's intention to act or need for assistance.

Limitations include the choice for a supervised model that tries to distinguish between

empty and full cups instead of discovering the carefulness behaviors from unseparated data. Although it was able to find two contrasting carefulness motion strategies, it most likely invokes some biases in the model. Perhaps applying an unsupervised learning approach like data clustering would have found other, more correct, carefulness strategies. Another point to improve is related to the classifier, where the introduction of a third output when the motion does not fit any of the two carefulness models (either because there is not enough information or the motion is ambiguous). Introducing the option of *Not sure* could provide an additional evaluation of the model, not only to test the sensitivity of the parameters (ϵ for example) but also to measure the certainty of the model when applied to new data.

9.4 Reply to Research Questions

This thesis contributions and discussions provide the answers to the proposed research questions in Chapter 2.

- RQ1 - *Can robots execute actions and be successfully understood just by imitating human non-verbal cues?*

  Yes. Throughout this thesis human non-verbal cues are used to recognize actions (an individual or *action-in-interaction*), intentions (to handover or not an object), and motion configurations (the polishing or carefulness strategy).

- RQ2 - *Can humans and robots mutually understand each other during interaction simply through non-verbal cues?*

  Yes. The main distinction between individual actions and *action-in-interaction* is the sociological factor that the former does not have and the latter requires for a successful action. From the analysis of the human non-verbal cues it was possible to detect that human eye-gaze cues patterns were significantly different between the two types of actions. The *Gaze Dialogue Model* learned the two eye-gaze behaviors and, given that gaze precedes arm motion, it is capable of recognizing the two actions even before the action is completed.

- RQ3 - *Do human non-verbal cues reveal object properties and can it be detected by robots?*

  Yes. The models developed in this thesis are bio-inspired in HHI and allow for recognition of human actions as well as generation of human-like non-verbal cues for expressing robot actions in HRI. Humanoid robots have the ability of mimicking human eye-gaze and arm-hand cues generated from the bio-inspired models. Non-humanoid robots, although limited in its human-like expressiveness, are still capable of generating arm motion cues which resemble the human's when performing a specific action (e.g. polishing or handover).

- RQ4 - *Can robots use human-like eye-gaze and arm-hand cues to express actions, intentions, and motion profiles?*

  Yes and it is my conviction that it will be a necessity for general-purpose robots. This thesis has shown that it is possible to extract non-verbal eye-gaze and arm-hand cues from humans in HHI that communicate valuable information not only for humans but for robots as well. Current robots used in HRI settings have noisy and limited sensors and actuators. Despite these problems, current robots are already capable of reading human non-verbal cues and expressing human-like non-verbal cues executing actions. As robot hardware evolves, there will be the possibility of reading more minute details in human motion and expressing more complex actions with non-verbal cues.

The take-home-message of this thesis is that it is possible for robots to learn and communicate non-verbally their intentions just like humans do in HHI scenarios. There is still much to do when it comes to robot cognitive abilities to reason about human non-verbal cues in more complex scenarios. Nonetheless, there is no doubt that if humans use non-verbal communication "protocols" in their daily life to navigate in human-centered environments, then robots should possess those same capabilities.

A

Ball Placing and Giving Dataset

Humans have fascinating skills for grasping and manipulation of objects, even in complex, dynamic environments, and execute coordinated movements of the head, eyes, arms, and hands, in order to accomplish everyday tasks. When working on a shared space, during dyadic interaction tasks, humans engage in non-verbal communication, by understanding and anticipating the actions of working partners, and coupling their actions in a meaningful way. The key to this performance is two-fold: (i) a capacity to adapt and plan the motion according to unexpected events in the environment, (ii) and the use of a common motor repertoire and action model, to understand and anticipate the actions and intentions of others as if they were our own. Despite decades of progress, robots are still far from the level of performance that would enable them to work with humans in routine activities.

A.1 Human pick & place and giving action dataset

The dataset was designed with the following goals: (i) study the human eyes and body behavior during manipulation of an object and when interacting with other humans in giving the object; (ii) provide a complete sensor data of a human performing individual and action-in-interaction actions; (iii) provide a dataset with raw first-person view of the human gaze movements in action-in-interaction scenario.

The experiments start with an actor and three participants. For each trial, one actor executes a set of *placing* (pick & place) or *giving* actions directed towards one of the three

(left/middle/right) subjects. The actor was instructed to act as normal as possible when performing those actions. The actor picks the object from the initial position and executes one of these 6 preselected action-configurations (2 actions and 3 spatial directions).

Figure A.1: Illustration of the pick & place and giving action dataset. The figure shows the 3 different video perspectives (left, top-right and self-view) as well as the two video perspectives used for the two questionnaires (top-right and bottom-right).

- ***placing*** on the table to the actor's **left** (P_L), **middle** (P_M), or **right** (P_R),
- ***giving*** the ball to the person on actor's **left** (G_L), **middle** (G_M), or **right** (G_R).

The actions to execute were instructed over an earpiece to the actor so that none of the other participants could know which would be performed next. The order of the actions is randomly selected to prevent the actor from adapting its posture prior to initiation. Every action begins with picking up the ball and ends with the actor placing the ball back to the initial position on the table.

A.2 Questionnaries

The two questionnaires that are an integral part of the research behind the paper "Action Anticipation: Reading the Intentions of Humans and Robots" (Chapter 3) are prepared as a set of randomly distributed questions. For each question the participant is asked to play a video shown to the left on the screen. After watching the video, one out of six possible answer should be provided for the question: "What will the actor do with the red ball after the video stops?" (See Figure A.2) After the answer is provided, the participant moves to the next question. It will repeat the same procedure until the end of the questionnaire.

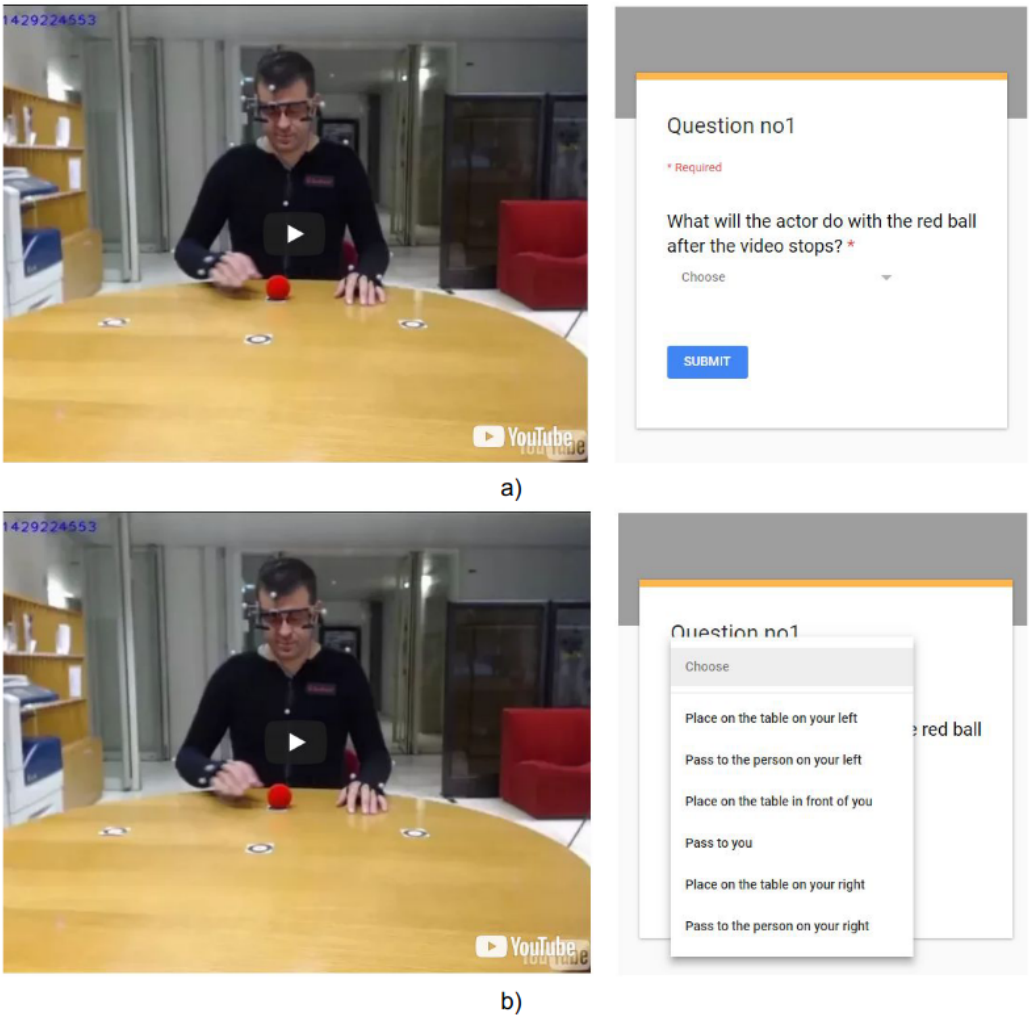


Figure A.2: The illustration of one question a) The snapshot of the screen with the video on the left and the question on the right; b) list of possible answers

A.2.1 Questionnaire 1

The first questionnaire has 24 questions. In each question there are six possible answers:

1. Place on the table on your left
2. Pass to the person on your left
3. Place on the table in front of you
4. Pass to you
5. Place on the table on your right
6. Pass to the person on your right

All the answers are referring to the perspective of the subject watching the video. So “Place on the table on your left” is referring to the white marker present on the table to the left of the video, while ”Pass to you” is when the actor is giving the ball to the center of the camera (representative location of the participant viewing the video). The difference between the videos of the same correct answer is in the length of the action, i.e. video fractions. There were four different video fractions which are:

1. only eye movement (G)
2. eye + head movement (G + H)
3. eye + head + start of the arm movement (G + H + A)
4. almost complete action (G + H + A +)

A.2.2 Questionnaire 2

The second questionnaire has 36 questions. Just like in the questionnaire 1, each question has six possible answers 1-6. The difference between the videos of the same answer was in the length of the videos and the amount of blurring present in the videos. There were three different video fractions, which are:

1. only eye movement (G)
2. eye + head movement (G + H)
3. eye + head + start of the arm movement (G + H + A)

and three different blurred regions illustrated in Figures A.3 - A.5 by a frame of the robot for each of the types of videos:

- n - no blurring (G)
- e - blurred eyes (Blrd G + H)
- eh - blurred eyes and head (Blrd G + Blrd H + A)

Figure A.3: Robot with no blur

Figure A.4: Robot with blurred eyes

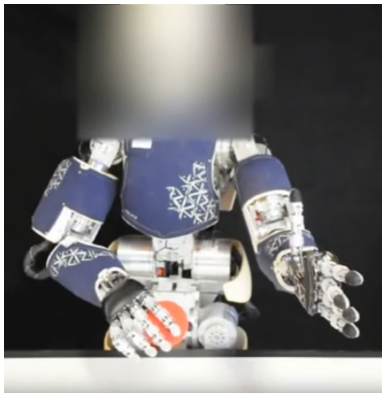


Figure A.5: Robot with blurred Face

A.3 Data from the pick & place and giving action dataset

The actor movements were recorded with an OptiTrack Mocap system, consisting of 12 cameras all around the environment and a suit with 25 markers, placed on the upper torso, arms, and head, that is worn by the actor. The provides position and orientation data of all relevant body parts (head, torso, right-arm, left-arm).

The eye gaze was recorded with the mobile, binocular Pupil-Labs eye tracker Kassner et al. [2014], that allowed us to track the actor's fixation point. To track the head movements with the Mocap system, head markers were placed on the Pupil-Lab system. To record the scene, three video cameras are used to provide different viewing angles that will complement during the evaluation phase. The first camera provides the world-view perspective of the actor from the Pupil Labs eye tracking headset (top-right image in Figure A.1, the small window on top). The second camera records the table top where the actions will take place. This one provides a continuous look at the table and all the actor's movements (Figure A.1 - top right). The third camera was located further from the scene, looking inwards, giving a proper reading of the subject's actions and an outlook of the experiment (Figure A.1 - left).

To collect all the sensory information, the OptiTrack's Motive and Pupil Lab's Pupil Capture software were used. Prior to recording, both sensors were calibrated. Custom software was developed to acquire the video of the actor's action. All the sensory data are captured on distributed machines and data are streamed through the Lab Streaming Layer Kothe for centralised storage and data synchronisation. Acquisitions were performed at the Institute for Systems and Robotics, IST, University of Lisbon, during August 2017.

A total of 120 trials are performed with action-configurations: P_L, P_M, P_R, G_L, G_M and G_R performed 20, 23, 17, 17, 19 and 24 times respectively. The binocular eye gaze tracking system recorded world camera video and eye gaze data at 60Hz, the motion capture system recorded the movements of the body at 120Hz, and video camera facing the actor, recorded video at 30Hz. The data from all sensing systems are streamed and collected at one place, with the timestamps of each sensing system as well as the internal clock information, that is used as a reference to synchronise all sensory flows.

Inside the dataset there are @folders with files and isolated files. In the @Folder:

- @LabeledVideo one can find labeled videos of actor performing one out of six action in randoms order
- @RawGaze one can find raw data of gaze recordings obtained from PupilLabs gaze tracker
- @RawVideo one can find the rawVideo of the @LabeledVideo
- @RawMotionCapture one can find raw joint and body data skeleton of the actor recorded using OptiTrack Mocap system
- @gaze_visualisation_sample one can find video sample of visualisation of gaze tracking system
- @merged_labeled_actions.mat one can find 4 matrices: body gaze joints and merged_cuts body gaze joints are 3D matrices of the size nx200x120, where n is number of coordinates (features), 200 is number of samples, and 120 is number of actions

Inside the dataset, the other files such as:

- merged_cuts is the matrix containing the row with assigned: action, gaze quality, timestamp on labeled video, start and end sample for gaze recording, start and end sample for motion capture recording
- Reference_To_Sync_Data is the synchronization of the timestamps of both sensors (PupilLabs and OptiTrack) with the corresponding timestamp of the LSL network that collected the data.
- outsideView.mp4 is the outside perspective video recordings of the whole experiment. It examines the full experimental setup from an outscope where it is possible to view the actor, the participants, and the sensory hardware setup present in the setup and on the actor.

A.4 Access to the data

The link for the Ball Placing and Giving dataset is publicly available on the instituion website https://vislab.isr.tecnico.ulisboa.pt/datasets_and_resources/#acticipate2

B

Gaze Behavior in Dyadic Interaction Dataset

The dataset contains recordings of adult subjects in dyadic interaction task. During the experiment, the subjects are asked to pick up an object and, based on the randomly defined instructions, to place it on the table in front of her/him or to give the object to a person sitting across the table. If the object is handed over, the second person takes the object and places it on the table it in front of her/him. The goal of this dataset is (i) (ii) (iii) intended to be used to model the behavior of the human's gaze while interacting with another human and implement the model in a controller of a robot for dyadic interaction with a humans.

With this experiment, we want to create a basis for research on how to integrate this coupling in robot's motor control system, in scenarios where both human and robot, share the same space and objects during task execution. For that purpose, the participants are asked to assemble a pair of towers inside a circle on the paper in front of them. Both towers are assembled from 3D printed objects of different shape or color as shown in Figure B.1. The objects are marked with numbers 1-3.

Figure B.1: Objects for assembling the tower.

B.1 Human-human pick & place and handover action dataset

In the beginning, two stacks of three objects are placed next to each participant. A stack of objects is positioned below the table top in order to occlude them from the other person. Next to the stack of objects is given a paper with the desired order of the objects to build the tower. (Figure B.2). When the assembly of towers starts, the participants are asked, one at a time, to pick the first object from the stack. If the number of the object matches the number in their next level of the tower, they should use the object for their tower. Otherwise, they are instructed to give, i.e. handover the object to a teammate. Thus, there are two types of actions the participant can execute: (i) intrapersonal action (pick and place an object on its tower, i.e. placing action) or (ii) inter-personal action (pick and handover an object, i.e. giving action).

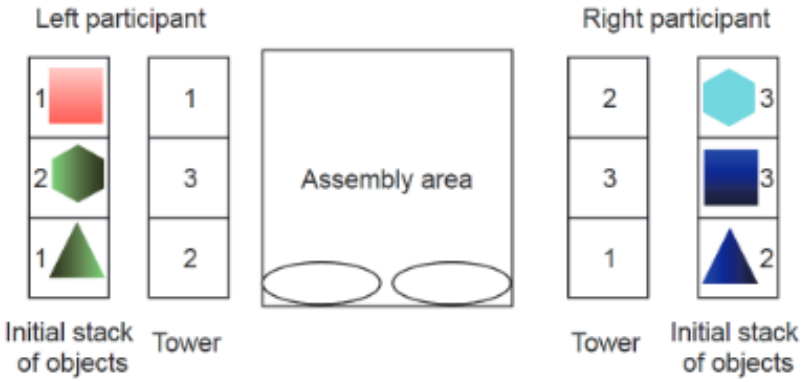


Figure B.2: Illustration of the initial stack of objects and the task given to the participants.

The towers are defined such that in the case of a handover, the object given to another participant is always the matching object for her/his next level in the tower. After an object is positioned in one of the towers, the turn is taken by a second participant. The actions are repeated until all the objects are used and both towers are assembled. Illustration of the progress of the task with the order and the type of action is given in Figure B.3. Once the assembly is finished, the new task and the new initial stack of objects is prepared and given to the participants. Each pair of participants had to repeat the task four times, i.e. to assemble

four different pairs of towers.

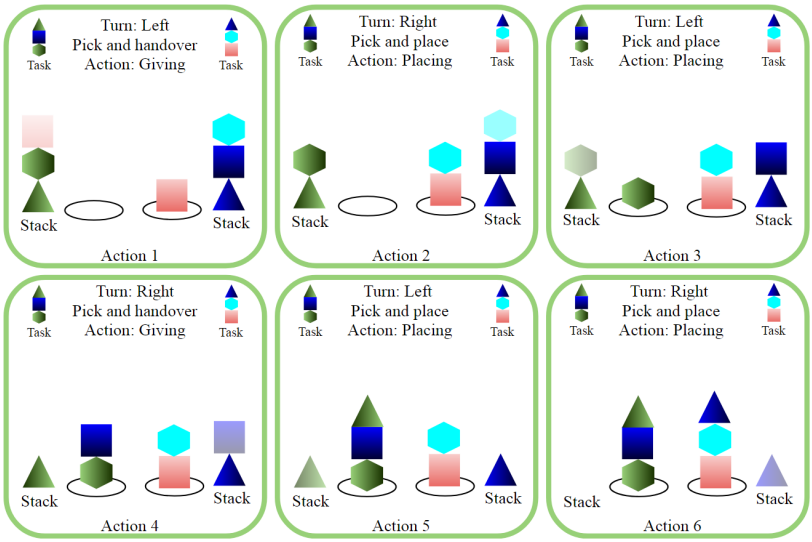


Figure B.3: Example of turn-taking order (left and right participant) and type of actions (pick and place or pick and handover) for assembling two towers.

The four tasks are defined in a way that there is always a different number of giving actions. This is to prevent the subjects to predict the action ahead of time. The goal here is to record a natural, unbiased human gaze behavior. The first task has two giving and four placing actions, the second task has six giving and no placing actions, the third task has no giving and six placing actions and the fourth task has four giving and two placing actions. Thus, during the experiment two participants performed together twelve giving and twelve placing actions.

B.2 Labeling for the pick & place and handover action dataset

When observing or scanning immediate surroundings, human eyes make jerky saccadic movements and stop several times, moving very quickly between each stop. The speed of movement during each saccade cannot be controlled, and the eyes move as fast as they are able Carlson et al. [2009]. To capture such eye movements, in this experiment both participants were wearing Pupil-Labs binocular gaze trackers Kassner et al. [2014]. During the performed actions, participants' head gaze was recorded using Optitrack Mocap system Point [2011]. Hardware and software setup used to acquire dataset is illustrated in Figure B.4.

The Pupil Labs binocular gaze tracker is in the form of glasses equipped with three cameras. Two cameras are recording eyes at 120Hz. A video stream of the egocentric view is recorded at 60Hz. The pupil detection algorithm does not depend on corneal reflection technique and as reported in Kassner et al. [2014] the gaze tracker should work with users who wear contact lenses and eyeglasses. However, we experienced difficulties in calibrating the glasses with such participants, and we had to choose the participants not wearing glasses and contact lenses. Before the recording starts, each participant first calibrates his/her gaze tracker using the screen calibration method. Optitrack Mocap system captures passive opto-reflective spherical

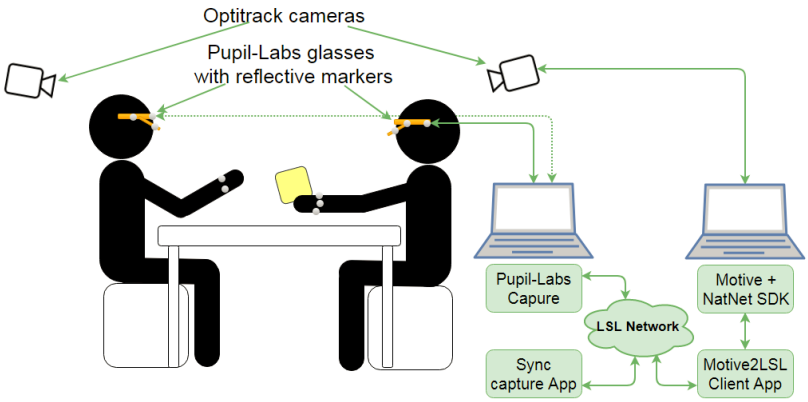


Figure B.4: Experiment hardware and software setup.

markers at 120Hz. To record head gaze we fixate five opto-reflective markers on each glasses. Each group of five markers represented one rigid body whose position and orientation in the reference frame is being recorded. Software setup is composed of following applications. For gaze data recording we used Pupil Labs Capture. For recording the body movements the Motive software platform is used. Since we want to capture synchronous data of head and eye gaze it was necessary to merge the input from two sensory systems. For that purpose is used Lab streaming layer (LSL) library Kothe. LSL is designed to be a system for unified collection of measurement time series of various sensing equipment that handles both the networking, time-synchronization, (near-) real-time access and optionally the centralized collection. In order to use LSL, we developed a Motive2LSL application that captures the broadcasted position of the markers and rigid bodies tracked within Motive software platform. Another application we developed is Sync capture application that receives the data measurements from two Pupil-Labs glasses and Optitrack cameras and records those data together with timestamps of the measurements into a file with synchronization timestamps.

We have acquired the data of three pairs, i.e. six participants. Participants were adults between 25 and 40 years of age. Acquisitions were performed at the Institute for Systems and Robotics, IST, University of Lisbon, during January 2018. The dataset contains:

- a video stream of the egocentric view with associated timestamps for both glasses,
- pupil data with gaze positions, pupil positions and its timestamps,
- position and orientation of rigid bodies representing head gaze with its timestamps and
- synchronization file with timestamps of gaze tracking and motion tracking data.

In each recording, subjects had to perform 6 actions in one task, and each pair of subjects had 4 tasks. Thus, we collected the 24 actions in dyad scenario, i.e. 48 gaze and head motion for each pair (24 observer's movements and 24 performer's movements). The recordings are repeated for three different pair of humans, and thus we collected 72 actions and 144 gaze and head movements. Figure B.5 illustrates the recorded gaze movement of the eyes during the

experiment for two different actions (placing and giving) and for two different roles (performer and observer) of the subject.

B.3 Access to the data

The link for the Gaze Behavior in Dyadic Interaction dataset is publicly available on the instituion website https://vislab.isr.tecnico.ulisboa.pt/datasets_and_resources/#acticipate1

The structure of the dataset is as follows:

- @RawData folder contains the Pupil-Labs and OptiTrack raw data:
- @Dyad_1, @Dyad_2, and @Dyad_3 are the folders for each of the pair of participants:
  - @RawPupilLabs one can find raw data of gaze recordings obtained from PupilLabs gaze tracker
  - @RawOptiTrack one can find raw joint and body data sceleton of the actor recorded using OptiTrack Mocap system
- @SegmentedData folder contains the PupilLabs segmented into different actions
- @Dyad_1, @Dyad_2, and @Dyad_3 are the folders for each of the pair of participants:
  - @L folder refers to the participant on the left
  - @R folder refers to the participant on the right
    - @A##_$$$$-$$$$?ŁŁ folder contains one action
      - A## is the id of the action
      - $$$$-$$$$ is the initial frame and end frame
      - ? is the type of action: G - for giving action P - for placing action
      - ŁŁ mentions the direction of the action: RL - giving action from the right participant to the left participant LR - giving action from the left participant to the right participant LL - left participant placing object RR - right participant placing object
  - 0 folder contains all the data in Raw

Figure B.5: Illustration of data set with an example given by an image sequences showing the gaze of a performer/observer during placing and giving actions (green circle represent the recorded gaze points, yellow line represent interpolation between recorded gaze points).

C

Human Manipulation of Cups with Water Dataset

The dataset contains recordings of adult subjects in dyadic interaction task. During the experiment, the subjects are asked to pick up a cup which can have different water levels (empty, half-full, and full) and, based on the randomly defined instructions, to place it on the table or to give the cup to a person sitting across the table. If the cup is handed over, the second person takes the cup and places it on the table. The goal of this dataset is (i) (ii) (iii) intended to be used to model the behavior of the human's gaze while inqbilhas0 teracting with another human and implement the model in a controller of a robot for dyadic interaction with a humans. Human-Human Interaction scenario where a cup with are manipulated to perform two types of actions: handover and pick-and-place.

C.1 Human-human pick & place and handover action dataset

The two participants start with a Lego board game indicated by the orange box in Figure C.1. Each participant needs to pick each Lego piece individually and build a puzzle at their choosing. The top-left frame in Figure C.2 shows the board game and a subject picking the first Lego piece. A base Lego puzzle board is provided in the beginning to both participants. The top-right frame in Figure C.2 illustrates the subject building their Lego puzzle on the

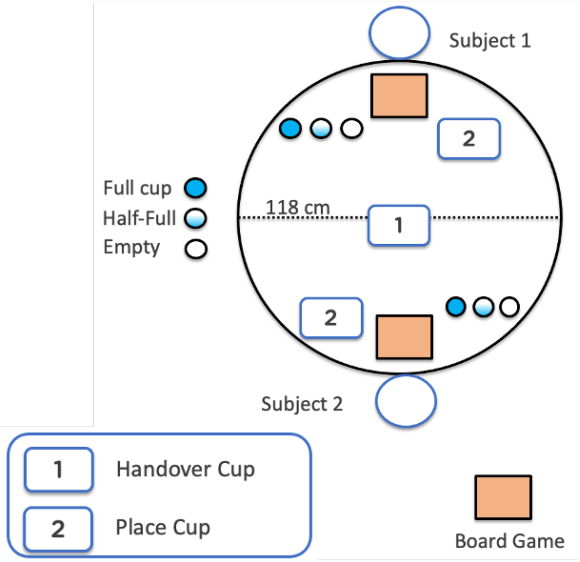


Figure C.1: Human-human Experimental Setup.

base Lego puzzle board. The bottom-left frame in Figure C.2 shows that the second subject is also building the LEGO puzzle. Predefined Lego pieces will have secondary tasks for the participants to complete. These pieces locations are selected by the experimenters unknown to the subjects and each new experiment the locations of these particular Lego pieces are changed to prevent anticipatory strategies by the participants. The tasks involve either to pass one of the cups to the other participant (handover), or to move the cup to a specific location (pick & place). The bottom-right frame in Figure C.2 shows on handover occurring between two participants. The cups, at the start of each experiment, begin at the predefined locations illustrated in Figure C.1. One experiment is deemed complete when all Lego pieces are picked by both participants and the board is empty of pieces.

This dataset is composed of data from two sensors mounted on the two humans participating in the experiment: head-mounted eye-tracker and infra-red markers for motion capture. The dataset provides video, eye, and gaze information from the PupilLabs system, and 3-D position and 4-D quaternion information of the wrist and head of the two humans in the whole experiment. There were a total of 6 participants in 3 pairs performing the experiments. The cup present is a transparent plastic and was identical for all the 3 water level conditions and thoroughout the experiments. There is a total of 6 experiments accomplished per pair of participants and in each experiment there is 6 actions present. This a total of 36 actions per pair of subjects and 108 actions included in the whole dataset. The PupilLabs data is for both participants giving two perspectives of the same actions, hence there is 216 eye-gaze data in total for the dataset. The OptiTrack is recorded at 120 Hz. PupilLabs is recorded at 30 Hz.

Figure C.2: Frames from the PupilLabs world camera of the HHI experiments.

C.2 Labeling for the pick & place and handover of cups dataset

The dataset is segmented for the purpose of analyzing the human eye-gaze behavior during the handover of cups with different water levels. The data is thoroughly examined by an expert which is instructed to mark and label the following instants:

- Events
  - Start
  - Object picked
  - Object handover
  - Object placed
  - End
- Gaze Cues
  - Team-mate's Face
  - Team-mate's Hand (with or without cup)
  - Team-mate's Cup
  - My cup
  - My hand (with or without cup)

  - Lego task
  - Cup final position
  - Outlier
  - No gaze

The Start and End events are marked at specific moments in the experiment. The Start event occurs when the subject receives the command to perform one of the actions. The command is considered received when one of Lego pieces unveils an action instruction. The End event occurs when the action is completed and the subject returns to the Board game task. This event takes place when the subject pick a new Lego piece. The events related to the object refer to specific actions performed: picking the cup, handing over, and placing. The Gaze cues refers to the moments where the subject is fixating one of the aforementioned regions of interest. These regions were selected as the most frequent in the HHI scenario. Outlier refers to the fixations which do not match to any of the previous gaze cues and the fixations did not resemble anything meaningful to the experiment. No gaze refers to the frames where there is no registered fixation.

For each segment there are two types of actions, pick & place (1) or handover (2), two types of subjects, leader (1) or follower (2), each subject can only be on one side of the table, top (1) or bottom (2), and action involves one of the three cups, empty (1), half-full (2), or full cup (3). The side of the table which the subject is located refers to the configuration in Figure C.1 where top and bottom is Subject 1 and 2, respectively. The top-left Figure C.2 shows the Figure C.1 configuration where the frame is from Subject 2's perspective viewing the table and Subject 1.

C.3 Access to the data

The link for the Human Manipulation of Cups with Water dataset is publicly available on the instituion website https://vislab.isr.tecnico.ulisboa.pt/datasets_and_resources/#hcups_water

The structure of the dataset is as follows:

- @RawData folder contains the PupilLabs and OptiTrack raw data:
  - @RawPupilLabs one can find raw data of gaze recordings obtained from PupilLabs gaze tracker 00# # - refers to the id of the participant. The pairs are 1-2; 3-4; 5-6 The PupilLabs folder includes all the raw data provided by the PupilLabs Capture System (raw videos, gaze information, eyes information, etc) For more information on the specifications, please consult the github repo: https://docs.pupil-labs.com/core/software/pupil-player/#raw-data-exporter

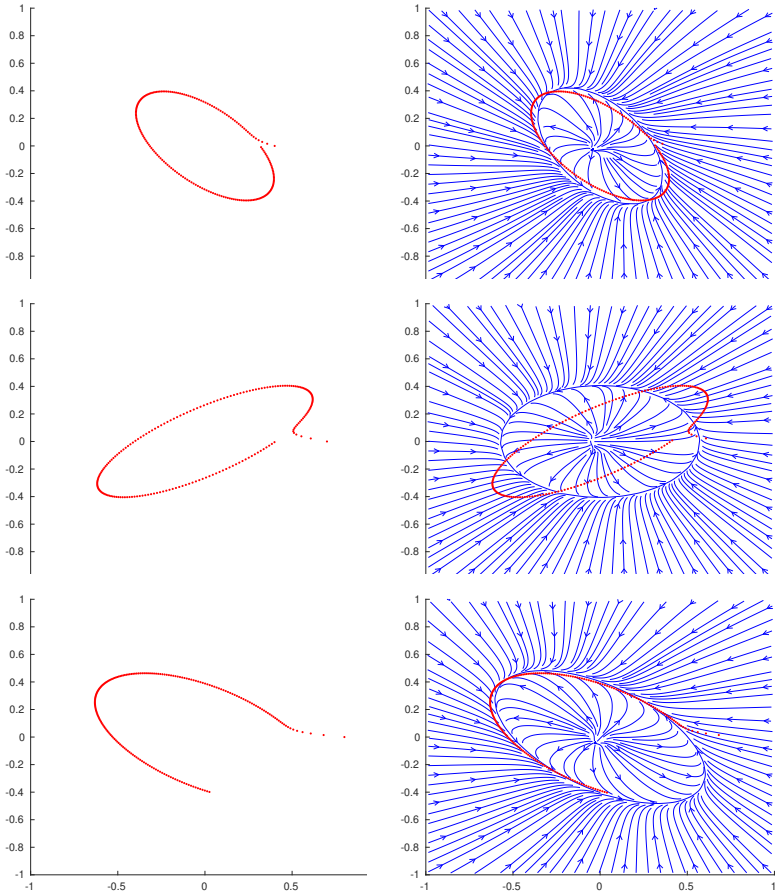
- @RawOptiTrack one can find raw joint and body data skeleton of the actor recorded using OptiTrack Mocap system
  - Take 2018-07-25 ##.##.## PM
  - ##.##.## refers to .tak and .csv files of each pair of participants
    - 02.29.14 - pair 1-2
    - 02.55.40 - pair 3-4
    - 03.58.38 - pair 5-6
  - In each .csv file there are 4 Rigid bodies and associated infra-red markers:
    - RB1, RB2 are the rigid bodies of the head of each participant (fixed in the PupilLabs head-mounted eye-tracker)
    - RB3, RB4 are the rigid bodies of the right-wrist arm of each participant.

  Columns in .csv file:
  - RB1 - 7-9 (X,Y,Z) 3-6 (QX,QY,QZ,QW)
  - RB2 - 31-33 (X,Y,Z) 27-30 (QX,QY,QZ,QW)
  - RB3 - 19-21 (X,Y,Z) 15-18 (QX,QY,QZ,QW)
  - RB4 - 51-53 (X,Y,Z) 47-50 (QX,QY,QZ,QW)
- @SegmentationLabels is a folder with excel files for each of the PupilLabs folders. Each excel file has information on the specific label that was fixated by the participant wearing the eye-tracker and the frame that it happens.
  - 00#
  - # - refers to the id of the participant. The pairs are 1-2; 3-4; 5-6
- @readme_labels.xlsx explains in detail the information provided by the manual segmentation

D

Additional Results from Polishing Motions

D.1 More Examples of Simulated Polishing Trajectories

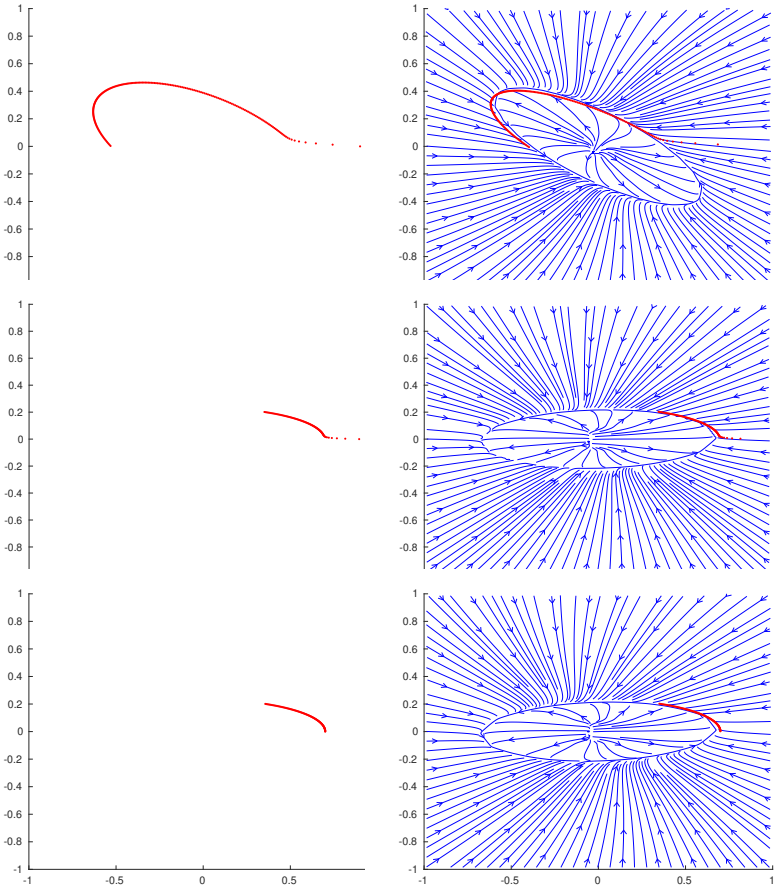


Final Objective	517.15
α	10.225
radius	0.501113
ω	1.33541
a	1.97952
b	1.0014
θ	-5.61548

Final Objective	1037.1177
α	10
radius	0.572102
ω	0.487011
a	1
b	1.4075
θ	6.28319

Final Objective	1173.6798
α	10.4534
radius	0.692466
ω	0.987008
a	1.95595
b	1.0048
θ	0.978913

Figure D.1: The top example is a circle with the $\Theta = [20, 0.5, \pi/2, 1, 2, -pi/4]$, middle has $\Theta = [20, 0.7, \pi/2, 3, 1, -pi/3]$, and bottom $\Theta = [20, 0.7, \pi/3, 2, 1, \pi/3]$.



Final Objective	244.3786
α	10
radius	0.697733
ω	0.563692
a	2.95326
b	1
θ	0.992359

Final Objective	14.0678
α	15.0331
radius	0.681471
ω	0.429468
a	1
b	3.15208
θ	0.0306536

Final Objective	13.6495
α	10
radius	0.677477
ω	0.463362
a	1
b	3.17935
θ	0.0362522

Figure D.2: The top example is a circle with the $\Theta = [20, 0.7, \pi/4, 3, 1, pi/3]$, middle has $\Theta = [20, 0.7, \pi/6, 1, 3, 0]$ with just half the data points as the previous one, and bottom $\Theta = [20, 0.7, \pi/6, 1, 3, 0]$ without the initial points outside the circle.

D.2 KUKA polishing while being perturbed

Figure D.3: Compliant controller running during the limit cycle DS on the KUKA robot. Human perturbs the robot in several directions.

Bibliography

Henny Admoni. *Nonverbal Communication in Socially Assistive Human-Robot Interaction*. PhD thesis, Yale University, 2016.

Henny Admoni, Thomas Weng, Bradley Hayes, and Brian Scassellati. Robot nonverbal behavior improves task performance in difficult collaborations. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 51–58, Christchurch, New Zealand, March 2016. IEEE. ISBN 978-1-4673-8370-7. doi: 10.1109/HRI.2016.7451733. URL http://ieeexplore.ieee.org/document/7451733/.

Salvatore M Aglioti, Paola Cesari, Michela Romani, and Cosimo Urgesi. Action anticipation and motor resonance in elite basketball players. *Nature neuroscience*, 11(9):1109–1116, 2008. ISSN 1097-6256. doi: 10.1038/nn.2182. ISBN: 1097-6256, 1097-6256.

Kaat Alaerts, Patrice Senot, Stephan P. Swinnen, Laila Craighero, Nicole Wenderoth, and Luciano Fadiga. Force requirements of observed object lifting are encoded by the observer's motor system: A TMS study. *European Journal of Neuroscience*, 31(6):1144–1153, 2010a. ISSN 0953816X. doi: 10.1111/j.1460-9568.2010.07124.x.

Kaat Alaerts, Stephan P. Swinnen, and Nicole Wenderoth. Observing how others lift light or heavy objects: Which visual cues mediate the encoding of muscular force in the primary motor cortex? *Neuropsychologia*, 48(7):2082–2090, 2010b. ISSN 00283932. doi: 10.1016/j.neuropsychologia.2010.03.029. URL http://dx.doi.org/10.1016/j.neuropsychologia.2010.03.029. Publisher: Elsevier Ltd.

Franco Amati and Susan E. Brennan. Chapteramp;160;2. eye gaze as a cue for recognizing intention and coordinating joint action. In *Eye-tracking in Interaction*, pages 21–46. John Benjamins, 2018. URL https://www.jbe-platform.com/content/books/9789027263469-ais.10.02ama.

Pierre Andry, Arnaud Blanchard, and Philippe Gaussier. Using the Rhythm of Nonverbal Human–Robot Interaction as a Signal for Learning. *IEEE Transactions on Autonomous Mental Development*, 3(1):30–42, March 2011. ISSN 1943-0604, 1943-0612. doi: 10.1109/TAMD.2010.2097260. URL http://ieeexplore.ieee.org/document/5664771/.

Salvatore M. Anzalone, Sofiane Boucenna, Serena Ivaldi, and Mohamed Chetouani. Evaluating the Engagement with Social Robots. *International Journal of Social Robotics*, 7(4):465–478, August 2015. ISSN 1875-4791, 1875-4805. doi: 10.1007/s12369-015-0298-7. URL http://link.springer.com/10.1007/s12369-015-0298-7.

João Avelino, Leonel Garcia-Marques, Rodrigo Ventura, and Alexandre Bernardino. Break the ice: a survey on socially aware engagement for human–robot first encounters. *International Journal of Social Robotics*, 13(8):1851–1877, 2021.

Sven Bambach. A Survey on the Cognitive Basis of Visual Attention in Real-World Behavior. 2013.

Sven Bambach, David J. Crandall, Linda B. Smith, and Chen Yu. An egocentric perspective on active vision and visual object learning in toddlers. In *2017 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, pages 290–295, Lisbon, September 2017. IEEE. ISBN 978-1-5386-3715-9. doi: 10.1109/DEVLRN.2017.8329820. URL http://ieeexplore.ieee.org/document/8329820/.

Shray Bansal, Mustafa Mukadam, and Charles L Isbell. Interaction-Aware Planning via Nash Equilibria for Manipulation in a Shared Workspace. In *ICRA 2019 Workshop on Human Movement Science for Physical Human-Robot Collaboration*, pages 1–2, 2019.

Jimmy Baraglia, Maya Cakmak, Yukie Nagai, Rajesh Rao, and Minoru Asada. Initiative in robot assistance during collaborative task execution. *ACM/IEEE International Conference on Human-Robot Interaction*, 2016-April:67–74, 2016. ISSN 21672148. doi: 10.1109/HRI.2016.7451735. Publisher: IEEE ISBN: 9781467383707.

Chiara Bassetti. Chapter 2 - social interaction in temporary gatherings: A sociological taxonomy of groups and crowds for computer vision practitioners. In Vittorio Murino, Marco Cristani, Shishir Shah, and Silvio Savarese, editors, *Group and Crowd Behavior for Computer Vision*, pages 15 – 28. Academic Press, 2017. ISBN 978-0-12-809276-7. doi: https://doi.org/10.1016/B978-0-12-809276-7.00003-5. URL http://www.sciencedirect.com/science/article/pii/B9780128092767000035.

Cristina Becchio, Luisa Sartori, Maria Bulgheroni, and Umberto Castiello. Both your intention and mine are reflected in the kinematics of my reach-to-grasp movement. *Cognition*, 106(2):894–912, February 2008a. ISSN 00100277. doi: 10.1016/j.cognition.2007.05.004. URL https://linkinghub.elsevier.com/retrieve/pii/S0010027707001400.

Cristina Becchio, Luisa Sartori, Maria Bulgheroni, and Umberto Castiello. The case of Dr. Jekyll and Mr. Hyde: A kinematic study on social intention. *Consciousness and Cognition*, 17(3):557–564, September 2008b. ISSN 10538100. doi: 10.1016/j.

concog.2007.03.003. URL https://linkinghub.elsevier.com/retrieve/pii/S1053810007000207.

Cristina Becchio, Luisa Sartori, and Umberto Castiello. Toward You: The Social Side of Actions. *Current Directions in Psychological Science*, 19(3):183–188, June 2010. ISSN 0963-7214, 1467-8721. doi: 10.1177/0963721410370131. URL http://journals.sagepub.com/doi/10.1177/0963721410370131.

Philipp Beckerle, Gionata Salvietti, Ramazan Unal, Domenico Prattichizzo, Simone Rossi, Claudio Castellini, Sandra Hirche, Satoshi Endo, Heni Ben Amor, Matei Ciocarlie, Fulvio Mastrogiovanni, Brenna D. Argall, and Matteo Bianchi. A human-robot interaction perspective on assistive and rehabilitation robotics. *Frontiers in Neurorobotics*, 11(MAY):1–6, 2017. ISSN 16625218. doi: 10.3389/fnbot.2017.00024.

James Scott Bell. *Write great fiction-plot & structure*. Penguin, 2004.

Francesca Biagini and Massimo Campanino. *Discrete Time Markov Chains*, pages 81–87. Springer International Publishing, Cham, 2016. ISBN 978-3-319-07254-8. doi: 10.1007/978-3-319-07254-8_6. URL https://doi.org/10.1007/978-3-319-07254-8_6.

Filippo Maria Bianchi, Simone Scardapane, Sigurd Lokse, and Robert Jenssen. Reservoir Computing Approaches for Representation and Classification of Multivariate Time Series. *IEEE Transactions on Neural Networks and Learning Systems*, 32(5):2169–2179, May 2021. ISSN 2162-237X, 2162-2388. doi: 10.1109/TNNLS.2020.3001377. URL https://ieeexplore.ieee.org/document/9127499/.

Geoffrey P. Bingham. 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, 1987. ISSN 00961523. doi: 10.1037/0096-1523.13.2.155.

Christopher Bodden, Daniel Rakita, Bilge Mutlu, and Michael Gleicher. Evaluating intent-expressive robot arm motion. *25th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN 2016*, pages 658–663, 2016. ISSN 10709878. doi: 10.1109/ROMAN.2016.7745188. arXiv: NIHMS150003 ISBN: 9781509039296.

Cynthia Breazeal and Brian Scassellati. Robots that imitate humans. *Trends in Cognitive Sciences*, 6(11):481–487, 2002. ISSN 13646613. doi: 10.1016/S1364-6613(02)02016-8. ISBN: 1364-6613.

Rechele Brooks and Andrew N. Meltzoff. The importance of eyes: How infants interpret adult looking behavior. *Developmental Psychology*, 38(6):958–966, 2002. ISSN 1939-0599, 0012-1649. doi: 10.1037/0012-1649.38.6.958. URL http://doi.apa.org/getdoi.cfm?doi=10.1037/0012-1649.38.6.958.

Judee K. Burgoon and Adam Kendon. Conducting Interaction: Patterns of Behavior in Focused Encounters. *Contemporary Sociology*, 21(2):256, March 1992. ISSN 00943061. doi: 10.2307/2075490. URL http://www.jstor.org/stable/2075490?origin=crossref.

Baptiste Busch, Jonathan Grizou, Manuel Lopes, and Freek Stulp. Learning Legible Motion from Human–Robot Interactions. *International Journal of Social Robotics*, 9(5):765–779, November 2017. ISSN 1875-4805. doi: 10.1007/s12369-017-0400-4. URL https://doi.org/10.1007/s12369-017-0400-4.

Judith Butepage, Hedvig Kjellstrom, and Danica Kragic. Anticipating Many Futures: Online Human Motion Prediction and Generation for Human-Robot Interaction. *Proceedings - IEEE International Conference on Robotics and Automation*, pages 4563–4570, 2018. ISSN 10504729. doi: 10.1109/ICRA.2018.8460651. arXiv: 1702.08212 ISBN: 9781538630815.

Sylvain Calinon, Florent Guenter, and Aude Billard. On Learning, Representing, and Generalizing a Task in a Humanoid Robot. *IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics)*, 37(2):286–298, April 2007. ISSN 1083-4419. doi: 10.1109/TSMCB.2006.886952. URL http://ieeexplore.ieee.org/document/4126276/.

Neil R Carlson, Donald Heth, Harold Miller, John Donahoe, and G Neil Martin. *Psychology: the science of behavior*. Pearson, 2009.

Monica S Castelhano, Mareike Wieth, and John M Henderson. I See What You See: Eye Movements in Real-World Scenes Are Affected by Perceived Direction of Gaze. *Attention in Cognitive Systems. Theories and Systems from an Interdisciplinary Viewpoint*, 4840: 251–262, 2007. ISSN 0302-9743. doi: 10.1007/978-3-540-77343-6. URL http://www.springerlink.com/index/10.1007/978-3-540-77343-6. ISBN: 978-3-540-77342-9.

Thierry Chaminade, Erhan Oztop, Gordon Cheng, and Mitsuo Kawato. From self-observation to imitation: Visuomotor association on a robotic hand. *Brain Research Bulletin*, 75(6):775–784, April 2008. ISSN 03619230. doi: 10.1016/j.brainresbull.2008.01.016. URL https://linkinghub.elsevier.com/retrieve/pii/S0361923008000130.

Wesley P Chan, Matthew K X J Pan, Elizabeth A Croft, and Masayuki Inaba. Characterization of Handover Orientations used by Humans for Efficient Robot to Human Handovers. *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1–6, 2015. doi: 10.1109/IROS.2015.7353106. Publisher: IEEE ISBN: 9781479999941.

Balasubramaniyan Chandrasekaran and James M. Conrad. Human-robot collaboration: A survey. *Conference Proceedings - IEEE SOUTHEASTCON*, 2015-June(June):1–8, 2015. ISSN 07347502. doi: 10.1109/SECON.2015.7132964. Publisher: IEEE ISBN: 9781467373005.

Raja Chatila, Erwan Renaudo, Mihai Andries, Ricardo-Omar Chavez-Garcia, Pierre Luce-Vayrac, Raphael Gottstein, Rachid Alami, Aurélie Clodic, Sandra Devin, Benoît Girard, and Mehdi Khamassi. Toward self-aware robots. *Frontiers in Robotics and AI*, 5, 2018. ISSN 2296-9144. doi: 10.3389/frobt.2018.00088. URL https://www.frontiersin.org/articles/10.3389/frobt.2018.00088.

P Claudia and Julie A Shah. Fast target prediction of human reaching motion for cooperative human-robot manipulation tasks using time series classification. In *International Conference on Robotics and Automation*, 2015. ISBN 1050-4729. doi: 10.1109/ICRA.2015.7140066. ISBN: 9781479969234.

Wesley Collier, Michael Gleicher, Sean Andrist, David Shaffer, and Bilge Mutlu. Look together: analyzing gaze coordination with epistemic network analysis. *Frontiers in Psychology*, 6(July):1–15, 2015. doi: 10.3389/fpsyg.2015.01016.

E. S. Cross, D. J.M. Kraemer, A. F. d. C. Hamilton, W. M. Kelley, and S. T. Grafton. Sensitivity of the Action Observation Network to Physical and Observational Learning. *Cerebral Cortex*, 19(2):315–326, February 2009. ISSN 1047-3211, 1460-2199. doi: 10.1093/cercor/bhn083. URL https://academic.oup.com/cercor/article-lookup/doi/10.1093/cercor/bhn083.

Kelly Cummings. *Nonverbal communication and first impressions*. PhD thesis, Kent State University, 2011.

Antonio R. Damasio, Daniel Tranel, and Hanna C. Damasio. Somatic markers and the guidance of behavior: Theory and preliminary testing. In *Frontal lobe function and dysfunction.*, pages 217–229. Oxford University Press, New York, NY, US, 1991. ISBN 0-19-506284-1 (Hardcover).

Frédéric Dehais, Emrah Akin Sisbot, Rachid Alami, and Mickaël Causse. Physiological and subjective evaluation of a human-robot object hand-over task. *Applied Ergonomics*, 42(6): 785–791, 2011. ISSN 18729126. doi: 10.1016/j.apergo.2010.12.005. URL http://www.sciencedirect.com/science/article/pii/S0003687011000044. Publisher: Elsevier Ltd ISBN: 0003-6870.

Chau Do and Wolfram Burgard. Accurate pouring with an autonomous robot using an RGB-D camera. *Advances in Intelligent Systems and Computing*, 867:210–221, 2019. ISSN 21945357. doi: 10.1007/978-3-030-01370-7_17. arXiv: 1810.03303 ISBN: 9783030013691.

Chau Do, Tobias Schubert, and Wolfram Burgard. A probabilistic approach to liquid level detection in cups using an RGB-D camera. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2075–2080, Daejeon, South Korea, October

2016. IEEE. ISBN 978-1-5090-3762-9. doi: 10.1109/IROS.2016.7759326. URL http://ieeexplore.ieee.org/document/7759326/.

Peter Ford Dominey and Franck Ramus. Neural network processing of natural language: I. Sensitivity to serial, temporal and abstract structure of language in the infant. *Language and Cognitive Processes*, 15(1):87–127, February 2000. ISSN 0169-0965, 1464-0732. doi: 10.1080/016909600386129. URL http://www.tandfonline.com/doi/abs/10.1080/016909600386129.

Marek Doniec, Ganghua Sun, and Brian Scassellati. Active Learning of Joint Attention. *IEEE-RAS International Conference on Humanoid Robots*, pages 34–39, 2006. doi: 10.1109/ICHR.2006.321360. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4115577. ISBN: 1-4244-0199-2.

Anca D. Dragan, Kenton C T Lee, and Siddhartha S. Srinivasa. Legibility and predictability of robot motion. *ACM/IEEE International Conference on Human-Robot Interaction*, 1: 301–308, 2013. ISSN 21672148. doi: 10.1109/HRI.2013.6483603. ISBN: 9781467330558.

Guillaume Dumas, Jacqueline Nadel, Robert Soussignan, Jacques Martinerie, and Line Garnero. Inter-Brain Synchronization during Social Interaction. *PLoS ONE*, 5(8): e12166, August 2010. ISSN 1932-6203. doi: 10.1371/journal.pone.0012166. URL https://dx.plos.org/10.1371/journal.pone.0012166.

Claudia Elsner, Marta Bakker, Katharina Rohlfing, and Gustaf Gredebäck. Infants' online perception of give-and-take interactions. *Journal of Experimental Child Psychology*, 126: 280–294, 2014. ISSN 00220965. doi: 10.1016/j.jecp.2014.05.007. URL http://dx.doi.org/10.1016/j.jecp.2014.05.007. Publisher: Elsevier Inc.

W. Erlhagen, A. Mukovskiy, E. Bicho, G. Panin, C. Kiss, A. Knoll, H. van Schie, and H. Bekkering. Goal-directed imitation for robots: A bio-inspired approach to action understanding and skill learning. *Robotics and Autonomous Systems*, 54(5):353–360, 2006. ISSN 09218890. doi: 10.1016/j.robot.2006.01.004. ISBN: 09218890.

Luciano Fadiga, Laila Craighero, and Etienne Olivier. Human motor cortex excitability during the perception of others' action. *Current Opinion in Neurobiology*, 15(2):213–218, 2005. ISSN 09594388. doi: 10.1016/j.conb.2005.03.013. ISBN: 0959-4388.

Merle T. Fairhurst, Petr Janata, and Peter E. Keller. Leading the follower: An fMRI investigation of dynamic cooperativity and leader-follower strategies in synchronization with an adaptive virtual partner. *NeuroImage*, 84:688–697, 2014. ISSN 10959572. doi: 10.1016/j.neuroimage.2013.09.027. URL http://dx.doi.org/10.1016/j.neuroimage.2013.09.027. Publisher: Elsevier Inc. ISBN: 1053-8119.

Jing Fan, Dayi Bian, Zhi Zheng, Linda Beuscher, Paul A. Newhouse, Lorraine C. Mion, and Nilanjan Sarkar. A Robotic Coach Architecture for Elder Care (ROCARE) Based on

Multi-User Engagement Models. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(8):1153–1163, August 2017. ISSN 1534-4320, 1558-0210. doi: 10.1109/TNSRE.2016.2608791. URL http://ieeexplore.ieee.org/document/7565740/.

G. Fantuzzi, D. Goluskin, D. Huang, and S. I. Chernyshenko. Bounds for Deterministic and Stochastic Dynamical Systems using Sum-of-Squares Optimization. *SIAM Journal on Applied Dynamical Systems*, 15(4):1962–1988, January 2016. ISSN 1536-0040. doi: 10.1137/15M1053347. URL http://epubs.siam.org/doi/10.1137/15M1053347.

Ruth Feldman. Parent–Infant Synchrony: Biological Foundations and Developmental Outcomes. *Current Directions in Psychological Science*, 16(6):340–345, December 2007. ISSN 0963-7214, 1467-8721. doi: 10.1111/j.1467-8721.2007.00532.x. URL http://journals.sagepub.com/doi/10.1111/j.1467-8721.2007.00532.x.

Nadia Figueroa. *From High-Level to Low-Level Robot Learning of Complex Tasks: Leveraging Priors, Metrics and Dynamical Systems*. PhD thesis, EPFL, 2019.

J. R. Flanagan, G. Rotman, A. F. Reichelt, and R. S. Johansson. The role of observers' gaze behaviour when watching object manipulation tasks: predicting and evaluating the consequences of action. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1628):20130063–20130063, 2013. ISSN 0962-8436. doi: 10.1098/rstb.2013.0063. URL http://rstb.royalsocietypublishing.org/cgi/doi/10.1098/rstb.2013.0063. ISBN: 1471-2970 (Electronic)\r0962-8436 (Linking).

Tamar Flash, Yaron Meirovitch, and Avi Barliya. Models of human movement: Trajectory planning and inverse kinematics studies. *Robotics and Autonomous Systems*, 61(4):330–339, 2013. ISSN 0921-8890. doi: https://doi.org/10.1016/j.robot.2012.09.020. URL https://www.sciencedirect.com/science/article/pii/S0921889012001741.

Nadine Fligge, Joseph McIntyre, and Patrick van der Smagt. Minimum jerk for human catching movements in 3D. In *2012 4th IEEE RAS & EMBS International Conference on Biomedical Robotics and Biomechatronics (BioRob)*, pages 581–586, Rome, Italy, June 2012. IEEE. ISBN 978-1-4577-1200-5 978-1-4577-1199-2 978-1-4577-1198-5. doi: 10.1109/BioRob.2012.6290265. URL http://ieeexplore.ieee.org/document/6290265/.

Chris D. Frith. The role of metacognition in human social interactions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1599):2213–2223, 2012. ISSN 14712970. doi: 10.1098/rstb.2012.0123.

Vittorio Gallese. The manifold nature of interpersonal relations: the quest for a common mechanism. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 358(1431):517–528, March 2003. ISSN 0962-8436, 1471-2970. doi:

10.1098/rstb.2002.1234. URL https://royalsocietypublishing.org/doi/10.1098/rstb.2002.1234.

Vittorio Gallese, Luciano Fadiga, Leonardo Fogassi, and Giacomo Rizzolatti. Action recognition in the premotor cortex. *Brain*, 119(2):593–609, 1996. ISSN 0006-8950, 1460-2156. doi: 10.1093/brain/119.2.593. URL https://academic.oup.com/brain/article-lookup/doi/10.1093/brain/119.2.593.

M Gallotti, M T Fairhurst, and C D Frith. Alignment in social interactions. *Consciousness and Cognition*, 48:253–261, 2017. ISSN 1053-8100. doi: https://doi.org/10.1016/j.concog.2016.12.002. URL http://www.sciencedirect.com/science/article/pii/S1053810016303749.

Philippe Gaussier and Alexandre Pitti. Reaching and Grasping : what we can learn from psychology and robotics. *Hal*, pages 1–11, 2017.

Ioanna Georgiou, Cristina Becchio, Scott Glover, and Umberto Castiello. Different action patterns for cooperative and competitive behaviour. *Cognition*, 102(3):415–433, 2007. ISSN 00100277. doi: 10.1016/j.cognition.2006.01.008.

Matthew Gombolay, Anna Bair, Cindy Huang, and Julie Shah. Computational design of mixed-initiative human–robot teaming that considers human factors: situational awareness, workload, and workflow preferences. *International Journal of Robotics Research*, 2017. ISSN 17413176. doi: 10.1177/0278364916688255.

Michael Grant and Stephen Boyd. *MATLAB Optimization Toolbox*. The MathWorks Inc., 2018. The MathWorks, Natick, MA, USA.

Jessica J. Green, Marissa L. Gamble, and Marty G. Woldorff. Resolving conflicting views: Gaze and arrow cues do not trigger rapid reflexive shifts of attention. *Visual Cognition*, 21(1): 61–71, January 2013. ISSN 1350-6285, 1464-0716. doi: 10.1080/13506285.2013.775209. URL http://www.tandfonline.com/doi/abs/10.1080/13506285.2013.775209.

Elena Corina Grigore, Kerstin Eder, Anthony G. Pipe, Chris Melhuish, and Ute Leonards. Joint action understanding improves robot-to-human object handover. *IEEE International Conference on Intelligent Robots and Systems*, pages 4622–4629, 2013. ISSN 21530858. doi: 10.1109/IROS.2013.6697021. Publisher: IEEE ISBN: 9781467363587.

Abhinav Gupta, Aniruddha Kembhavi, and Larry S. Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009. ISSN 01628828. doi: 10.1109/TPAMI.2009.83. ISBN: 0162-8828.

Patrick Haggard and Alan M. Wing. Remote responses to perturbation in human prehension. *Neuroscience Letters*, 122(1):103–108, January 1991. ISSN 03043940. doi: 10.1016/0304-3940(91)90204-7. URL https://linkinghub.elsevier.com/retrieve/pii/0304394091902047.

Antonia F.De C. Hamilton, D. W. Joyce, J. R. Flanagan, C. D. Frith, and D. M. Wolpert. Kinematic cues in perceptual weight judgement and their origins in box lifting. *Psychological Research*, 71(1):13–21, 2007. ISSN 03400727. doi: 10.1007/s00426-005-0032-4.

Clint Hansen, Paula Arambel, Khalil Ben Mansour, Véronique Perdereau, and Frédéric Marin. Human–Human Handover Tasks and How Distance and Object Mass Matter. *Perceptual and Motor Skills*, 124(1):182–199, 2017. ISSN 1558688X. doi: 10.1177/0031512516682668.

Syed Khursheed Hasnain, Ghiles Mostafaoui, and Philippe Gaussier. A Synchrony-Based Perspective for Partner Selection and Attentional Mechanism in Human-Robot Interaction. *Paladyn, Journal of Behavioral Robotics*, 3(3):156–171, 2012. ISSN 2081-4836. doi: 10.2478/s13230-013-0111-y. URL http://www.degruyter.com/view/j/pjbr.2012.3.issue-3/s13230-013-0111-y/s13230-013-0111-y.xml. ISBN: 1323001301.

Syed Khursheed Hasnain, Ghiles Mostafaoui, Robin Salesse, Ludovic Marin, and Philippe Gaussier. Intuitive human robot interaction based on unintentional synchrony: A psychoexperimental study. *2013 IEEE 3rd Joint International Conference on Development and Learning and Epigenetic Robotics, ICDL 2013 - Electronic Conference Proceedings*, 2013. ISSN 978-1-4799-1036-6. doi: 10.1109/DevLrn.2013.6652569. ISBN: 9781479910366.

Mary Hayhoe and Dana Ballard. Eye movements in natural behavior. *Trends in Cognitive Sciences*, 9(4):188–194, 2005. ISSN 13646613. doi: 10.1016/j.tics.2005.02.009. ISBN: 1364-6613.

Rafi Hayne, Ruikun Luo, and Dmitry Berenson. Considering avoidance and consistency in motion planning for human-robot manipulation in a shared workspace. *Proceedings - IEEE International Conference on Robotics and Automation*, 2016-June:3948–3954, 2016. ISSN 10504729. doi: 10.1109/ICRA.2016.7487584. ISBN: 9781467380263.

Erin E Hecht, Lauren E Murphy, David A Gutman, John R Votaw, David M Schuster, Todd M Preuss, Guy A Orban, Dietrich Stout, and Lisa A Parr. Differences in Neural Activation for Object-Directed Grasping in Chimpanzees and Humans. *Journal of Neuroscience*, 33 (35):14117–14134, 2013. ISSN 1529-2401. doi: 10.1523/jneurosci.2172-13.2013. ISBN: 0270-6474.

B.J. Hedge, B.S. Everitt, and C.D. Frith. The role of gaze in dialogue. *Acta Psychologica*, 42(6):453–475, November 1978. ISSN 00016918. doi: 10.

1016/0001-6918(78)90033-1. URL https://linkinghub.elsevier.com/retrieve/pii/0001691878900331.

P. M. Hilt, P. Cardellicchio, E. Dolfini, T. Pozzo, L. Fadiga, and A. D'Ausilio. Motor Recruitment during Action Observation: Effect of Interindividual Differences in Action Strategy. *Cerebral cortex (New York, N.Y. : 1991)*, 30(7):3910–3920, 2020. ISSN 14602199. doi: 10.1093/cercor/bhaa006.

Guy Hoffman. Evaluating Fluency in Human-Robot Collaboration. *IEEE Transactions on Human-Machine Systems*, 49(3):209–218, 2019. ISSN 21682305. doi: 10.1109/THMS.2019.2904558. Publisher: IEEE.

Yi Hu, Yinying Hu, Xianchun Li, Yafeng Pan, and Xiaojun Cheng. Brain-to-brain synchronization across two persons predicts mutual prosociality. *Social Cognitive and Affective Neuroscience*, 12(12):1835–1844, 2017. ISSN 17495024. doi: 10.1093/scan/nsx118.

Chien-Ming Huang, Sean Andrist, Allison Sauppé, and Bilge Mutlu. Using gaze patterns to predict task intent in collaboration. *Frontiers in psychology*, 6(July):1049, 2015a. ISSN 1664-1078. doi: 10.3389/fpsyg.2015.01049. URL http://journal.frontiersin.org/article/10.3389/fpsyg.2015.01049/abstract.

Chien-Ming Huang, Maya Cakmak, and Bilge Mutlu. Adaptive Coordination Strategies for Human-Robot Handovers. *Robotics: Science and Systems XI*, 2015b. ISSN 2330765X. doi: 10.15607/RSS.2015.XI.031. ISBN: 9780992374716.

Nicholas Humphrey. *Consciousness Regained*. Oxford University Press, 1984.

Marco Iacoboni and Mirella Dapretto. The mirror neuron system and the consequences of its dysfunction. *Nature Reviews Neuroscience*, 7(12):942–951, December 2006. ISSN 1471-003X, 1471-0048. doi: 10.1038/nrn2024. URL http://www.nature.com/articles/nrn2024.

Marco Iacoboni, Istvan Molnar-Szakacs, Vittorio Gallese, Giovanni Buccino, John C Mazziotta, and Giacomo Rizzolatti. Grasping the Intentions of Others with One's Own Mirror Neuron System. *PLoS Biology*, 3(3):e79, February 2005. ISSN 1545-7885. doi: 10.1371/journal.pbio.0030079. URL https://dx.plos.org/10.1371/journal.pbio.0030079.

Takashi Ikegami and Hiroyuki Iizuka. Turn-taking Interaction as a Cooperative and Co-creative Process. *Infant Behavior and Development*, 30(2):278–288, May 2007. ISSN 01636383. doi: 10.1016/j.infbeh.2007.02.002. URL https://linkinghub.elsevier.com/retrieve/pii/S0163638307000239.

Mert Imre, Erhan Oztop, Yukie Nagai, and Emre Ugur. Affordance-based altruistic robotic architecture for human–robot collaboration. *Adaptive Behavior*, 2019. ISSN 17412633. doi: 10.1177/1059712318824697. ISBN: 1059712318.

Félix Ingrand and Malik Ghallab. Deliberation for autonomous robots: A survey. *Artificial Intelligence*, 247:10–44, 2017. ISSN 00043702. doi: 10.1016/j.artint.2014.11.003. URL http://dx.doi.org/10.1016/j.artint.2014.11.003. Publisher: Elsevier B.V.

Johann Issartel. Interpersonal motor coordination: From human–human to human–robot interactions. *Interaction Studies*, 10(3):479–504, 2009. ISSN 1572-0373. doi: 10.1075/is.10.3.09mar. URL https://benjamins.com/catalog/is.10.3.09mar.

Johann Issartel, Ludovic Marin, and Marielle Cadopi. Unintended interpersonal co-ordination: "can we march to the beat of our own drum?". *Neuroscience Letters*, 411(3):174–179, January 2007. ISSN 03043940. doi: 10.1016/j.neulet.2006.09.086. URL https://linkinghub.elsevier.com/retrieve/pii/S0304394006009505.

Serena Ivaldi, Sebastien Lefort, Jan Peters, Mohamed Chetouani, Joelle Provasi, and Elisabetta Zibetti. Towards Engagement Models that Consider Individual Factors in HRI: On the Relation of Extroversion and Negative Attitude Towards Robots to Gaze and Speech During a Human–Robot Assembly Task. *International Journal of Social Robotics*, 9(1):63–86, January 2017. ISSN 1875-4805. doi: 10.1007/s12369-016-0357-8. URL https://doi.org/10.1007/s12369-016-0357-8. arXiv: 1508.04603 Publisher: Springer Netherlands ISBN: 9781457710056.

Lorenzo Jamone, Emre Ugur, Angelo Cangelosi, Luciano Fadiga, Alexandre Bernardino, Justus Piater, and Jose Santos-Victor. Affordances in Psychology, Neuroscience, and Robotics: A Survey. *IEEE Transactions on Cognitive and Developmental Systems*, 10(1):4–25, March 2018. ISSN 2379-8920, 2379-8939. doi: 10.1109/TCDS.2016.2594134. URL http://ieeexplore.ieee.org/document/7523298/.

Nathanaël Jarrassé, Vittorio Sanguineti, and Etienne Burdet. Slaves no longer: Review on role assignment for human-robot joint motor action. *Adaptive Behavior*, 22(1):70–82, 2014. ISSN 10597123. doi: 10.1177/1059712313481044.

T. Jellema, C.I. Baker, B. Wicker, and D.I. Perrett. Neural Representation for the Perception of the Intentionality of Actions. *Brain and Cognition*, 44(2):280–302, November 2000. ISSN 02782626. doi: 10.1006/brcg.2000.1231. URL https://linkinghub.elsevier.com/retrieve/pii/S0278262600912314.

Doreen Jirak, Stephan Tietz, Hassan Ali, and Stefan Wermter. Echo State Networks and Long Short-Term Memory for Continuous Gesture Recognition: a Comparative Study. *Cognitive Computation*, October 2020. ISSN 1866-9956, 1866-9964. doi: 10.1007/s12559-020-09754-0. URL https://link.springer.com/10.1007/s12559-020-09754-0.

Roland S Johansson, G. Westling, A. Bäckström, and J Randall Flanagan. Eye-hand coordination in object manipulation. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 21(17):6917–6932, 2001. ISSN 1529-2401. ISBN: 1529-2401 (Electronic)\n0270-6474 (Linking).

Marcel Adam Just and Patricia A Carpenter. Eye fixations and cognitive processes. *Cognitive Psychology*, 8(4):441–480, October 1976. ISSN 00100285. doi: 10.1016/0010-0285(76)90015-3. URL https://linkinghub.elsevier.com/retrieve/pii/0010028576900153.

Marcel Adam Just and Patricia A Carpenter. A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87(4):27, 1980.

Moritz Kassner, William Patera, and Andreas Bulling. Pupil: An Open Source Platform for Pervasive Eye Tracking and Mobile Gaze-based Interaction. *arXiv:1405.0006 [cs]*, April 2014. URL http://arxiv.org/abs/1405.0006. arXiv: 1405.0006.

Richard Kelley, Alireza Tavakkoli, Christopher King, Monica Nicolescu, and Mircea Nicolescu. Understanding Activities and Intentions for Human-Robot Interaction. In Daisuke Chugo, editor, *Human-Robot Interaction*. IntechOpen, Rijeka, 2010. doi: 10.5772/8127. URL https://doi.org/10.5772/8127.

J.A. Scott Kelso, Guillaume Dumas, and Emmanuelle Tognoli. Outline of a general theory of behavior and brain coordination. *Neural Networks*, 37:120–131, January 2013. ISSN 08936080. doi: 10.1016/j.neunet.2012.09.003. URL https://linkinghub.elsevier.com/retrieve/pii/S0893608012002286.

Christian Keysers, Riccardo Paracampo, and Valeria Gazzola. What neuromodulation and lesion studies tell us about the function of the mirror neuron system and embodied cognition. *Current Opinion in Psychology*, 24:35–40, December 2018. ISSN 2352250X. doi: 10.1016/j.copsyc.2018.04.001. URL https://linkinghub.elsevier.com/retrieve/pii/S2352250X18300332.

Hassan K Khalil. *Nonlinear systems; 3rd ed.* Prentice-Hall, Upper Saddle River, NJ, 2002. URL https://cds.cern.ch/record/1173048. The book can be consulted by contacting: PH-AID: Wallet, Lionel.

S. Mohammad Khansari-Zadeh and Aude Billard. Learning stable nonlinear dynamical systems with Gaussian mixture models. *IEEE Transactions on Robotics*, 27(5):943–957, 2011. ISSN 15523098. doi: 10.1109/TRO.2011.2159412. ISBN: 1552-3098.

Mahdi Khoramshahi and Aude Billard. A dynamical system approach to task-adaptation in physical human–robot interaction. *Autonomous Robots*, 43(4):927–946, April 2019. ISSN 0929-5593, 1573-7527. doi: 10.1007/s10514-018-9764-z. URL http://link.springer.com/10.1007/s10514-018-9764-z.

Mahdi Khoramshahi, Ashwini Shukla, Stï¿½phane Raffard, Benoï¿½t G. Bardy, and Aude Billard. Role of gaze cues in interpersonal motor coordination: Towards higher affiliation in human-robot interaction. *PLoS ONE*, 11(6):1–21, 2016. ISSN 19326203. doi: 10.1371/journal.pone.0156874.

Hedvig Kjellström, Javier Romero, and Danica Kragić. Visual object-action recognition: Inferring object affordances from human demonstration. *Computer Vision and Image Understanding*, 115(1):81–90, January 2011. ISSN 10773142. doi: 10.1016/j.cviu.2010.08.002. URL https://linkinghub.elsevier.com/retrieve/pii/S107731421000175X.

Nathan L. Kluttz, Brandon R. Mayes, Roger W. West, and Dave S. Kerby. The effect of head turn on the perception of gaze. *Vision Research*, 49(15):1979–1993, 2009. ISSN 00426989. doi: 10.1016/j.visres.2009.05.013. URL http://dx.doi.org/10.1016/j.visres.2009.05.013. Publisher: Elsevier Ltd ISBN: 0042-6989.

Mark L Knapp, Judith A Hall, and Terrence G Horgan. *Nonverbal communication in human interaction*. Cengage Learning, 2013.

Günther Knoblich and Natalie Sebanz. Evolving intentions for social interaction: From entrainment to joint action. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1499):2021–2031, 2008. ISSN 09628452. doi: 10.1098/rstb.2008.0006.

Idil Kokal, Valeria Gazzola, and Christian Keysers. Acting together in and beyond the mirror neuron system. *NeuroImage*, 47(4):2046–2056, 2009. ISSN 10538119. doi: https://doi.org/10.1016/j.neuroimage.2009.06.010. URL http://www.sciencedirect.com/science/article/pii/S1053811909006211. Publisher: Elsevier Inc. ISBN: 1095-9572 (Electronic)\n1053-8119 (Linking).

Kyveli Kompatsiari, Francesca Ciardo, Vadim Tikhanoff, Giorgio Metta, and Agnieszka Wykowska. On the role of eye contact in gaze cueing. *Scientific Reports*, 8(1):17842, December 2018. ISSN 2045-2322. doi: 10.1038/s41598-018-36136-2. URL http://www.nature.com/articles/s41598-018-36136-2.

Kyveli Kompatsiari, Francesca Ciardo, Vadim Tikhanoff, Giorgio Metta, and Agnieszka Wykowska. It's in the Eyes: The Engaging Role of Eye Contact in HRI. *International Journal of Social Robotics*, 13(3):525–535, June 2021. ISSN 1875-4791, 1875-4805. doi: 10.1007/s12369-019-00565-4. URL https://link.springer.com/10.1007/s12369-019-00565-4.

Ivana Konvalinka and Andreas Roepstorff. The two-brain approach: how can mutually interacting brains teach us something about social interaction? *Frontiers in Human Neuroscience*, 6(July):1–10, 2012. doi: 10.3389/fnhum.2012.00215.

C Kothe. Lab streaming layer (lsl). *https://github.com/sccn/labstreaminglayer*.

D. Kourtis, N. Sebanz, and G. Knoblich. Predictive representation of other people's actions in joint action planning: An EEG study. *Social Neuroscience*, 8(1):31–42, 2013. ISSN 17470919. doi: 10.1080/17470919.2012.694823.

Klas Kronander and Aude Billard. Passive Interaction Control With Dynamical Systems. *IEEE Robotics and Automation Letters*, 1(1):106–113, January 2016. ISSN 2377-3766, 2377-3774. doi: 10.1109/LRA.2015.2509025. URL http://ieeexplore.ieee.org/document/7358081/.

Christopher Krupenye, Fumihiro Kano, Satoshi Hirata, Josep Call, and Michael Tomasello. Great apes anticipate that other individuals will act according to false beliefs. *Science*, 354(6308):110–114, 2016. ISSN 0036-8075. doi: 10.1126/science.aaf8110. URL http://www.sciencemag.org/cgi/doi/10.1126/science.aaf8110.

Alap Kshirsagar, Melanie Lim, Shemar Christian, and Guy Hoffman. Robot Gaze Behaviors in Human-to-Robot Handovers. *IEEE Robotics and Automation Letters*, 5(4):6552–6558, October 2020. ISSN 2377-3766, 2377-3774. doi: 10.1109/LRA.2020.3015692. URL https://ieeexplore.ieee.org/document/9165096/.

Ryo Kuboshita, Takashi X. Fujisawa, Kai Makita, Ryoko Kasaba, Hidehiko Okazawa, and Akemi Tomoda. Intrinsic brain activity associated with eye gaze during mother–child interaction. *Scientific Reports*, 10(1):18903, December 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-76044-y. URL http://www.nature.com/articles/s41598-020-76044-y.

Alison Kuiper. Intercultural communication: a contextual approach james w. neuliep. thousand oaks, ca: Sage, 2006, 479 pages. *Business Communication Quarterly*, 71(4):516–518, 2008.

Minae Kwon, Sandy H. Huang, and Anca D. Dragan. Expressing Robot Incapability. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pages 87–95, Chicago IL USA, February 2018. ACM. ISBN 978-1-4503-4953-6. doi: 10.1145/3171221.3171276. URL https://dl.acm.org/doi/10.1145/3171221.3171276.

Fanny Lachat, Laurent Hugueville, Jean-Didier Lemaréchal, Laurence Conty, and Nathalie George. Oscillatory Brain Correlates of Live Joint Attention: A Dual-EEG Study. *Frontiers in Human Neuroscience*, 6(June):1–12, 2012. doi: 10.3389/fnhum.2012.00156.

Stephen R H Langton, Roger J Watt, and Vicki Bruce. Do the eyes have it? Cues to the direction of social attention. *Trends in Cognitive Sciences*, 4(2):10, 2000.

Stephen R.H. Langton, Helen Honeyman, and Emma Tessler. The influence of head contour and nose angle on the perception of eye-gaze direction. *Perception and Psychophysics*, 66 (5):752–771, 2004. ISSN 00315117. doi: 10.3758/BF03194970. ISBN: 0031-5117.

Linda Lastrico, Alessandro Carfì, Alessia Vignolo, Alessandra Sciutti, Fulvio Mastrogiovanni, and Francesco Rea. Careful with That! Observation of Human Movements to Estimate Objects Properties. In *Human-Friendly Robotics 2020*, pages 127–141, Cham, 2021. Springer International Publishing. ISBN 978-3-030-71356-0.

Yann Lecun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553): 436–444, 2015. ISSN 14764687. doi: 10.1038/nature14539.

A. Lemme, Y. Meirovitch, M. Khansari-Zadeh, T. Flash, A. Billard, and J. J. Steil. Open-source benchmarking for learned reaching motion generation in robotics. *Paladyn, Journal of Behavioral Robotics*, 6(1):30–41, 2015. ISSN 2081-4836. doi: 10.1515/pjbr-2015-0002.

Daniel Lewkowicz, Yvonne Delevoye-Turrell, David Bailly, Pierre Andry, and Philippe Gaussier. Reading motor intention through mental imagery. *Adaptive Behavior*, 21(5): 315–327, 2013. ISSN 10597123. doi: 10.1177/1059712313501347. ISBN: 1059-7123.

Quan-Lin Li. *Markov Reward Processes*, pages 526–573. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010. ISBN 978-3-642-11492-2. doi: 10.1007/978-3-642-11492-2_10. URL https://doi.org/10.1007/978-3-642-11492-2_10.

Yanan Li, Keng Peng Tee, Rui Yan, Wei Liang Chan, and Yan Wu. A Framework of Human-Robot Coordination Based on Game Theory and Policy Iteration. *IEEE Transactions on Robotics*, 32(6):1408–1418, 2016. ISSN 15523098. doi: 10.1109/TRO.2016.2597322.

Manuel Lopes and José Santos-Victor. Visual learning by imitation with motor representations. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 35(3):438–449, 2005. ISSN 10834419. doi: 10.1109/TSMCB.2005.846654. ISBN: 1083-4419.

Tamara Lorenz, Alexander Mörtl, Björn Vlaskamp, Anna Schubö, and Sandra Hirche. Synchronization in a goal-directed task: Human Movement Coordination with each other and robotic partners. *RO-MAN 2011 - The 20th IEEE International Symposium on Robot and Human Interactive Communication*, 2011. ISBN: 9781457715723.

Luka Lukic, José Santos-Victor, and Aude Billard. Learning robotic eye–arm–hand coordination from human demonstration: a coupled dynamical systems approach. *Biological Cybernetics*, 108(2):223–248, April 2014. ISSN 0340-1200, 1432-0770. doi: 10.1007/s00422-014-0591-9. URL http://link.springer.com/10.1007/s00422-014-0591-9.

Ruikun Luo and Dmitry Berenson. A framework for unsupervised online human reaching motion recognition and early prediction. *IEEE International Conference on Intelligent Robots and Systems*, 2015-Decem:2426–2433, 2015. ISSN 21530866. doi: 10.1109/IROS.2015.7353706. ISBN: 9781479999941.

Guilherme J. Maeda, Gerhard Neumann, Marco Ewerton, Rudolf Lioutikov, Oliver Kroemer, and Jan Peters. Probabilistic movement primitives for coordination of multiple human–robot collaborative tasks. *Autonomous Robots*, 41(3):593–612, March 2017. ISSN 0929-5593, 1573-7527. doi: 10.1007/s10514-016-9556-2. URL http://link.springer.com/10.1007/s10514-016-9556-2.

Roland Mangold. *Informationspsychologie: Wahrnehmen und Gestalten in der Medienwelt*. Springer-Verlag, 2015.

Nikolaos Mavridis. A review of verbal and non-verbal human–robot interactive communication. *Robotics and Autonomous Systems*, 63(1):22–35, 2015. ISSN 0921-8890. doi: https://doi.org/10.1016/j.robot.2014.09.031. URL https://www.sciencedirect.com/science/article/pii/S0921889014002164.

H. C. Mayer and R. Krechetnikov. Walking with coffee: Why does it spill? *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 85(4):1–7, 2012. ISSN 15393755. doi: 10.1103/PhysRevE.85.046117.

José R. Medina, Felix Duvallet, Murali Karnam, and Aude Billard. A human-inspired controller for fluid human-robot handovers. *IEEE-RAS International Conference on Humanoid Robots*, pages 324–331, 2016. ISSN 21640580. doi: 10.1109/HUMANOIDS.2016.7803296. ISBN: 9781509047185.

Andrew N Meltzoff. Born to Learn: What Infants Learn from Watching Us. *Pediatric Institute Publications*, page 10, 1999.

Andrew N. Meltzoff. The 'like me' framework for recognizing and becoming an intentional agent. *Acta Psychologica*, 124(1):26–43, January 2007. ISSN 00016918. doi: 10.1016/j.actpsy.2006.09.005. URL https://linkinghub.elsevier.com/retrieve/pii/S0001691806001211.

Andrew N Meltzoff and M Keith Moore. Imitation of facial and manual gestures by human neonates. *Science*, 198(4312):75–78, 1977.

Giorgio Metta, Lorenzo Natale, Francesco Nori, Giulio Sandini, David Vernon, Luciano Fadiga, Claes von Hofsten, Kerstin Rosander, Manuel Lopes, José Santos-Victor, Alexandre Bernardino, and Luis Montesano. The iCub humanoid robot: An open-systems platform for research in cognitive development. *Neural Networks*, 23(8-9):1125–1134, October 2010. ISSN 08936080. doi: 10.1016/j.neunet.2010.08.010. URL https://linkinghub.elsevier.com/retrieve/pii/S0893608010001619.

Ar Mitz, M Godschalk, and Sp Wise. Learning-dependent neuronal activity in the premotor cortex: activity during the acquisition of conditional motor associations. *The Journal of Neuroscience*, 11(6):1855–1872, June 1991. ISSN 0270-6474, 1529-2401. doi: 10.1523/

JNEUROSCI.11-06-01855.1991. URL https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.11-06-01855.1991.

Apostolos Modas, Alessio Xompero, Ricardo Sanchez-Matilla, Pascal Frossard, and Andrea Cavallaro. Improving filling level classification with adversarial training. *arXiv:2102.04057 [cs]*, February 2021. URL http://arxiv.org/abs/2102.04057. arXiv: 2102.04057.

Derek E. Montgomery, Leslie M. Back, and Christy Moran. Children's Use of Looking Behavior as a Cue to Detect Another's Goal. *Child Development*, 69 (3):692–705, June 1998. ISSN 00093920, 14678624. doi: 10.1111/j.1467-8624.1998.tb06237.x. URL https://onlinelibrary.wiley.com/doi/10.1111/j.1467-8624.1998.tb06237.x.

AJung Moon, Daniel M. Troniak, Brian Gleeson, Matthew K.X.J. Pan, Minhua Zheng, Benjamin A. Blumer, Karon MacLean, and Elizabeth A. Croft. Meet me where i'm gazing: how shared attention gaze affects human-robot handover timing. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, pages 334–341, Bielefeld Germany, March 2014. ACM. ISBN 978-1-4503-2658-2. doi: 10.1145/2559636.2559656. URL https://dl.acm.org/doi/10.1145/2559636.2559656.

Quentin Moreau, Lucie Galvan, Tatjana A. Nazir, Yves Paulignan, Sara M. Scharoun, Kelly A. Scanlan, Pamela J. Bryden, Quentin Moreau, Lucie Galvan, Tatjana A. Nazir, Yves Paulignan, Sara M. Scharoun, Kelly A. Scanlan, and Pamela J. Bryden. Hand and grasp selection in a preferential reaching task: The effects of object location, orientation, and task intention. *Frontiers in Psychology*, 7(MAR):1–8, 2016. ISSN 16641078. doi: 10.3389/fpsyg.2016.00360. URL http://link.springer.com/article/10.1007/BF00227183. Publisher: Elsevier Ltd ISBN: 1471-003X (Print)\r1471-003X (Linking).

Roozbeh Mottaghi, Connor Schenck, Dieter Fox, and Ali Farhadi. 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, October 2017. IEEE. ISBN 978-1-5386-1032-9. doi: 10.1109/ICCV.2017.207. URL http://ieeexplore.ieee.org/document/8237469/.

Roy Mukamel, Arne D. Ekstrom, Jonas Kaplan, Marco Iacoboni, and Itzhak Fried. Single-Neuron Responses in Humans during Execution and Observation of Actions. *Current Biology*, 20(8):750–756, April 2010. ISSN 09609822. doi: 10.1016/j.cub.2010.02.045. URL https://linkinghub.elsevier.com/retrieve/pii/S0960982210002332.

Selma Musić and Sandra Hirche. Control sharing in human-robot team interaction. *Annual Reviews in Control*, 44:342–354, 2017. ISSN 13675788. doi: 10.1016/j.arcontrol.2017.09.017. Publisher: Elsevier Ltd.

Alexander Mörtl, Tamara Lorenz, Björn N.S. Vlaskamp, Azwirman Gusrialdi, Anna Schubö, and Sandra Hirche. Modeling inter-human movement coordination: Synchronization governs joint task dynamics. *Biological Cybernetics*, 106(4-5):241–259, 2012. ISSN 03401200. doi: 10.1007/s00422-012-0492-8. ISBN: 0340-1200.

Alexander Mörtl, Tamara Lorenz, and Sandra Hirche. Rhythm patterns interaction - Synchronization behavior for human-robot joint action. *PLoS ONE*, 9(4), 2014. ISSN 19326203. doi: 10.1371/journal.pone.0095195.

Jacqueline Nadel, Arnaud Revel, Pierre Andry, and Philippe Gaussier. Toward communication: First imitations in infants, low-functioning children with autism and robots. *Interaction Studies*, 5(1):45–74, 2004. ISSN 1572-0373. doi: 10.1075/is.5.1.04nad. URL http://www.jbe-platform.com/content/journals/10.1075/is.5.1.04nad.

Yukie Nagai and Katharina J Rohlfing. Can Motionese Tell Infants and Robots "What to Imitate"? In *Proceedings of the 4th International Symposium on Imitation in Animals and Artifacts*, pages 299–306, 2007.

Vipul Nair, Paul Hemeren, Alessia Vignolo, Nicoletta Noceti, Elena Nicora, Alessandra Sciutti, Francesco Rea, Erik Billing, Francesca Odone, and Giulio Sandini. Action similarity judgment based on kinematic primitives. *arXiv*, 2020. arXiv: 2008.13176.

Elena Natale, Irene Senna, Nadia Bolognini, Ermanno Quadrelli, Margaret Addabbo, Viola Macchi Cassia, and Chiara Turati. Predicting others' intention involves motor resonance: EMG evidence from 6- and 9-month-old infants. *Developmental Cognitive Neuroscience*, 7 (November):23–29, 2014. ISSN 18789293. doi: 10.1016/j.dcn.2013.10.004. URL http://dx.doi.org/10.1016/j.dcn.2013.10.004. Publisher: Elsevier Ltd ISBN: 1878-9293.

Heramb Nemlekar, Dharini Dutia, and Zhi Li. Object Transfer Point Estimation for Fluent Human-Robot Handovers. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 2627–2633, Montreal, QC, Canada, May 2019. IEEE. ISBN 978-1-5386-6027-0. doi: 10.1109/ICRA.2019.8794008. URL https://ieeexplore.ieee.org/document/8794008/.

Roger D Newman-Norlund, Matthijs L Noordzij, Ruud G J Meulenbroek, Harold Bekkering, and F C Donders. Exploring the brain basis of joint action: co-ordination of actions, goals and intentions. *Social neuroscience*, 2(768418105):48–65, 2007. ISSN 1747-0919. doi: 10.1080/17470910701224623. URL http://www.tandfonline.com/loi/psns20. ISBN: 1747-0927 (Electronic)\n1747-0919 (Linking).

Andrzej Nowak, Robin R. Vallacher, Michal Zochowski, and Agnieszka Rychwalska. Functional synchronization: The emergence of coordinated activity in human systems. *Frontiers in Psychology*, 8(JUN):1–15, 2017. ISSN 16641078. doi: 10.3389/fpsyg.2017.00945.

Etienne Olivier, Marco Davare, Michael Andres, and Luciano Fadiga. Precision grasping in humans: from motor control to cognition. *Current Opinion in Neurobiology*, 17(6):644–648, 2007. ISSN 09594388. doi: 10.1016/j.conb.2008.01.008.

Valerio Ortenzi, Akansel Cosgun, Tommaso Pardi, Wesley P. Chan, Elizabeth Croft, and Dana Kulic. Object Handovers: A Review for Robotics. *IEEE Transactions on Robotics*, pages 1–19, 2021. ISSN 1552-3098, 1941-0468. doi: 10.1109/TRO.2021.3075365. URL https://ieeexplore.ieee.org/document/9444288/.

Harriet Over and Merideth Gattis. Verbal imitation is based on intention understanding. *Cognitive Development*, 25(1):46–55, January 2010. ISSN 08852014. doi: 10.1016/j.cogdev.2009.06.004. URL https://linkinghub.elsevier.com/retrieve/pii/S0885201409000501.

Oskar Palinko, Francesco Rea, Giulio Sandini, and Alessandra Sciutti. Robot reading human gaze: Why eye tracking is better than head tracking for human-robot collaboration. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5048–5054, October 2016. doi: 10.1109/IROS.2016.7759741. ISSN: 2153-0866.

Matthew K.X.J. Pan, Elizabeth A. Croft, and Günter Niemeyer. Evaluating Social Perception of Human-to-Robot Handovers Using the Robot Social Attributes Scale (RoSAS). *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction - HRI '18*, pages 443–451, 2018. ISSN 21672148. doi: 10.1145/3171221.3171257. URL http://dl.acm.org/citation.cfm?doid=3171221.3171257. ISBN: 9781450349536.

Sina Parastegari, Bahareh Abbasi, Ehsan Noohi, and Milos Zefran. Modeling human reaching phase in human-human object handover with application in robot-human handover. *IEEE International Conference on Intelligent Robots and Systems*, 2017-Septe:3597–3602, 2017. ISSN 21530866. doi: 10.1109/IROS.2017.8206205. ISBN: 9781538626825.

Daniel Parks, Ali Borji, and Laurent Itti. Augmented saliency model using automatic 3D head pose detection and learned gaze following in natural scenes. *Vision Research*, 116:113–126, 2015. ISSN 18785646. doi: 10.1016/j.visres.2014.10.027. URL http://dx.doi.org/10.1016/j.visres.2014.10.027. Publisher: Elsevier Ltd ISBN: 0042-6989.

Aftab Patla and Joan Vickers. How far ahead do we look when required to step on specific locations in the travel path during locomotion? *Experimental Brain Research*, 148(1): 133–138, January 2003. ISSN 0014-4819, 1432-1106. doi: 10.1007/s00221-002-1246-y. URL http://link.springer.com/10.1007/s00221-002-1246-y.

U Pattacini, F Nori, L Natale, G Metta, and G Sandini. 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, Taipei, October 2010.

IEEE. ISBN 978-1-4244-6674-0. doi: 10.1109/IROS.2010.5650851. URL http://ieeexplore.ieee.org/document/5650851/.

Giulia Perugia, Maike Paetzel-Prüsmann, Madelene Alanenpää, and Ginevra Castellano. I Can See It in Your Eyes: Gaze as an Implicit Cue of Uncanniness and Task Performance in Repeated Interactions With Robots. *Frontiers in Robotics and AI*, 8:645956, April 2021. ISSN 2296-9144. doi: 10.3389/frobt.2021.645956. URL https://www.frontiersin.org/articles/10.3389/frobt.2021.645956/full.

Luigi S. Pesce Ibarra. Synchronization matters for motor coordination. *Journal of Neurophysiology*, 119(3):767–770, 2017. ISSN 0022-3077. doi: 10.1152/jn.00182.2017.

Mark Pfeiffer, Ulrich Schwesinger, Hannes Sommer, Enric Galceran, and Roland Siegwart. Predicting Actions to Act Predictably : Cooperative Partial Motion Planning with Maximum Entropy Models. *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2096–2101, 2016. ISSN 21530866. doi: 10.1109/IROS.2016.7759329. ISBN: 9781509037612.

Natural Point. Optitrack. *Natural Point, Inc*, 2011.

C. Press, J. Cook, S.-J. Blakemore, and J. Kilner. Dynamic Modulation of Human Motor Activity When Observing Actions. *Journal of Neuroscience*, 31(8):2792–2800, February 2011. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.1595-10.2011. URL https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.1595-10.2011.

Claudia Pérez-D'Arpino and Julie A. Shah. Fast motion prediction for collaborative robotics. *IJCAI International Joint Conference on Artificial Intelligence*, 2016-Janua:3988–3989, 2016. ISSN 10450823.

Lawrence R. Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. In *Readings in Speech Recognition*, pages 267–296. Elsevier, 1990. ISBN 978-1-55860-124-6. doi: 10.1016/B978-0-08-051584-7.50027-9. URL https://linkinghub.elsevier.com/retrieve/pii/B9780080515847500279.

Robin Rasch, Sven Wachsmuth, and Matthias Konig. A Joint Motion Model for Human-Like Robot-Human Handover. In *2018 IEEE-RAS 18th International Conference on Humanoid Robots (Humanoids)*, pages 180–187, Beijing, China, November 2018. IEEE. ISBN 978-1-5386-7283-9. doi: 10.1109/HUMANOIDS.2018.8624967. URL https://ieeexplore.ieee.org/document/8624967/.

A. Revel and P. Andry. Emergence of structured interactions: From a theoretical model to pragmatic robotics. *Neural Networks*, 22(2):116–125, 2009. ISSN 08936080. doi: 10.1016/j.neunet.2009.01.005. URL http://dx.doi.org/10.1016/j.neunet.2009.01.005. arXiv: 1707.03042 Publisher: Elsevier Ltd.

Paola Ricciardelli, Emanuela Bricolo, Salvatore M. Aglioti, and Leonardo Chelazzi. My eyes want to look where your eyes are looking: Exploring the tendency to imitate another individual??s gaze:. *NeuroReport*, 13(17):2259–2264, December 2002. ISSN 0959-4965. doi: 10.1097/00001756-200212030-00018. URL http://journals.lww.com/00001756-200212030-00018.

Michael J. Richardson, Kerry L. Marsh, and R. C. Schmidt. Effects of Visual and Verbal Interaction on Unintentional Interpersonal Coordination. *Journal of Experimental Psychology: Human Perception and Performance*, 31(1):62–79, 2005. ISSN 1939-1277. doi: 10.1037/0096-1523.31.1.62. URL http://doi.apa.org/getdoi.cfm?doi=10.1037/0096-1523.31.1.62. ISBN: 0191-5886\n1573-3653.

Jelena Ristic, Chris Kelland Friesen, and Alan Kingstone. Are eyes special? It depends on how you look at it. *Psychonomic Bulletin & Review*, 9(3):507–513, September 2002. ISSN 1069-9384, 1531-5320. doi: 10.3758/BF03196306. URL http://link.springer.com/10.3758/BF03196306.

Giacomo Rizzolatti and Laila Craighero. THE MIRROR-NEURON SYSTEM. *Annual Review of Neuroscience*, 27(1):169–192, July 2004. ISSN 0147-006X, 1545-4126. doi: 10.1146/annurev.neuro.27.070203.144230. URL http://www.annualreviews.org/doi/10.1146/annurev.neuro.27.070203.144230.

Giacomo Rizzolatti, Leonardo Fogassi, and Vittorio Gallese. Neurophysiological mechanisms underlying the understanding and imitation of action Giacomo Rizzolatti, Leonardo Fogassi and V. *Nature Neuroscience*, 2(September):1–10, 2001a. ISSN 1471-003X. doi: 10.1038/35090060.

Giacomo Rizzolatti, Leonardo Fogassi, and Vittorio Gallese. Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature Reviews Neuroscience*, 2(9):661–670, September 2001b. ISSN 1471-003X, 1471-0048. doi: 10.1038/35090060. URL http://www.nature.com/articles/35090060.

S. Robla-Gomez, Victor M. Becerra, J. R. Llata, E. Gonzalez-Sarabia, C. Torre-Ferrero, and J. Perez-Oria. Working Together: A Review on Safe Human-Robot Collaboration in Industrial Environments. *IEEE Access*, 5:26754–26773, 2017. ISSN 21693536. doi: 10.1109/ACCESS.2017.2773127.

Alessandro Roncone, Ugo Pattacini, Giorgio Metta, and Lorenzo Natale. A Cartesian 6-DoF Gaze Controller for Humanoid Robots. In *Robotics: Science and Systems XII*. Robotics: Science and Systems Foundation, 2016. ISBN 978-0-9923747-2-3. doi: 10.15607/RSS.2016.XII.022. URL http://www.roboticsproceedings.org/rss12/p22.pdf.

Jacob Rosen, Joel C Perry, Nathan Manning, Stephen Burns, and Blake Hannaford. The Human Arm Kinematics and Dynamics During Daily Activities – Toward a 7 DOF Upper

Limb Powered Exoskeleton. *12th International Conference on Advanced Robotics*, page 8, 2005.

Stefano Rozzi and Gino Coudé. Grasping actions and social interaction: Neural bases and anatomical circuitry in the monkey. *Frontiers in Psychology*, 6(July):1–19, 2015. ISSN 16641078. doi: 10.3389/fpsyg.2015.00973. arXiv: 1011.1669v3 ISBN: 1664-1078.

Lucia Maria Sacheli, Matteo Candidi, Enea Francesco Pavone, Emmanuele Tidoni, and Salvatore Maria Aglioti. And Yet They Act Together: Interpersonal Perception Modulates Visuo-Motor Interference and Mutual Adjustments during a Joint-Grasping Task. *PLoS ONE*, 7(11), 2012. ISSN 19326203. doi: 10.1371/journal.pone.0050223.

Lucia Maria Sacheli, Emmanuele Tidoni, Enea Francesco Pavone, Salvatore Maria Aglioti, and Matteo Candidi. Kinematics fingerprints of leader and follower role-taking during cooperative joint actions, 2013. Issue: 4 ISSN: 00144819.

Ricardo Sanchez-Matilla, Konstantinos Chatzilygeroudis, Apostolos Modas, Nuno Ferreira Duarte, Alessio Xompero, Pascal Frossard, Aude Billard, and Andrea Cavallaro. Benchmark for Human-to-Robot Handovers of Unseen Containers With Unknown Filling. *IEEE Robotics and Automation Letters*, 5(2):1642–1649, April 2020. ISSN 2377-3766, 2377-3774. doi: 10.1109/LRA.2020.2969200. URL https://ieeexplore.ieee.org/document/8968407/.

Luisa Sartori, Cristina Becchio, Bruno G. Bara, and Umberto Castiello. Does the intention to communicate affect action kinematics? *Consciousness and Cognition*, 18(3):766–772, 2009. ISSN 10538100. doi: 10.1016/j.concog.2009.06.004. URL http://dx.doi.org/10.1016/j.concog.2009.06.004. Publisher: Elsevier Inc. ISBN: 1053-8100.

Connor Schenck and Dieter Fox. Reasoning About Liquids via Closed-Loop Simulation. *arXiv:1703.01656 [cs]*, June 2017a. URL http://arxiv.org/abs/1703.01656. arXiv: 1703.01656.

Connor Schenck and Dieter Fox. Visual closed-loop control for pouring liquids. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2629–2636, Singapore, Singapore, May 2017b. IEEE. ISBN 978-1-5090-4633-1. doi: 10.1109/ICRA.2017.7989307. URL http://ieeexplore.ieee.org/document/7989307/.

Stefanie Schuch and Steven P Tipper. On observing another person's actions: influences of observed inhibition and errors. *Perception & psychophysics*, 69(5):828–837, 2007. ISSN 0031-5117. doi: 10.3758/BF03193782. ISBN: 1943-3921.

Alessandra Sciutti and Nicoletta Noceti. Guest Editorial A Sense of Interaction in Humans and Robots: From Visual Perception to Social Cognition. *IEEE Transactions on Cognitive and Developmental Systems*, 10(4):839–842, 2018. ISSN 2379-8920. doi: 10.1109/TCDS.2018.2883166. URL https://ieeexplore.ieee.org/document/8567856/.

Alessandra Sciutti, Ambra Bisio, Francesco Nori, Giorgio Metta, Luciano Fadiga, and Giulio Sandini. Robots can be perceived as goal-oriented agents. *Interaction Studies. Social Behaviour and Communication in Biological and Artificial Systems*, 14(3):329–350, December 2013. ISSN 1572-0373, 1572-0381. doi: 10.1075/is.14.3.02sci. URL http://www.jbe-platform.com/content/journals/10.1075/is.14.3.02sci.

Alessandra Sciutti, Laura Patanè, Francesco Nori, and Giulio Sandini. Understanding object weight from human and humanoid lifting actions. *IEEE Transactions on Autonomous Mental Development*, 6(2):80–92, 2014. ISSN 19430604. doi: 10.1109/TAMD.2014.2312399.

Alessandra Sciutti, Martina Mara, Vincenzo Tagliasco, and Giulio Sandini. Humanizing human-robot interaction: On the importance of mutual understanding. *IEEE Technology and Society Magazine*, 37(1):22–29, 2018. ISSN 02780097. doi: 10.1109/MTS.2018.2795095.

Alessandra Sciutti, Laura Patanè, and Giulio Sandini. Development of visual perception of others' actions: Children's judgment of lifted weight. *PLoS ONE*, 14(11):1–15, 2019. ISSN 19326203. doi: 10.1371/journal.pone.0224979.

Natalie Sebanz and Guenther Knoblich. Prediction in Joint Action: What, When, and Where. *Topics in Cognitive Science*, 1(2):353–367, April 2009. ISSN 17568757, 17568765. doi: 10.1111/j.1756-8765.2009.01024.x. URL https://onlinelibrary.wiley.com/doi/10.1111/j.1756-8765.2009.01024.x.

Natalie Sebanz, Harold Bekkering, and Günther G??nther Knoblich. Joint action: Bodies and minds moving together. *Trends in Cognitive Sciences*, 10(2):70–76, 2006. ISSN 13646613. doi: 10.1016/j.tics.2005.12.009. arXiv: f ISBN: 1364-6613, 1364-6613.

Emmanuel Senft, Paul Baxter, James Kennedy, Séverin Lemaignan, and Tony Belpaeme. Supervised autonomy for online learning in human-robot interaction. *Pattern Recognition Letters*, 99:77–86, 2017. ISSN 01678655. doi: 10.1016/j.patrec.2017.03.015. URL https://doi.org/10.1016/j.patrec.2017.03.015. Publisher: Elsevier B.V.

Patrice Senot, Alessandro D'Ausilio, Michele Franca, Luana Caselli, Laila Craighero, and Luciano Fadiga. Effect of weight-related labels on corticospinal excitability during observation of grasping: A TMS study. *Experimental Brain Research*, 211(1):161–167, 2011. ISSN 00144819. doi: 10.1007/s00221-011-2635-x.

Samira Sheikhi and Jean-Marc Odobez. Recognizing the Visual Focus of Attention for Human Robot Interaction. In David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Doug Tygar, Moshe Y. Vardi, Gerhard Weikum, Albert Ali Salah, Javier Ruiz-del Solar, Çetin Meriçli, and Pierre-Yves Oudeyer, editors, *Human Behavior Understanding*, volume 7559, pages 99–112. Springer

Berlin Heidelberg, Berlin, Heidelberg, 2012. ISBN 978-3-642-34013-0 978-3-642-34014-7. doi: 10.1007/978-3-642-34014-7_9. URL http://link.springer.com/10.1007/978-3-642-34014-7_9. Series Title: Lecture Notes in Computer Science.

Qiming Shen. *Motor Interference and Behaviour Adaptation in Human-Humanoid Interactions*. PhD thesis, School of Computer Science, Faculty of Engineering and Information Sciences, University of Hertfordshire., 2012.

Shervin Javdani, Henny Admoni, Stefania Pellegrinelli, Siddhartha S. Srinivasa, and J. Andrew Bagnell. Shared Autonomy via HindsightOptimization for Teleoperation andTeaming. *Journal of Vibration and Control*, 37(7):717–742, 2019. doi: 10.1177/0278364918776060. arXiv: 1706.00155v1.

A Shukla and A Billard. Coupled dynamical system based hand-arm grasp planning under real-time perturbations. *International Conference on Robotics Science and Systems, RSS 2011*, 7:313–320, 2012. ISSN 2330765X. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-84959289300&partnerID=40&md5=af385f3a276443230928ea1c3a37687a. ISBN: 23307668 (ISSN); 9780262517799 (ISBN).

Antonis Sidiropoulos, Efi Psomopoulou, and Zoe Doulgeri. A human inspired handover policy using Gaussian Mixture Models and haptic cues. *Autonomous Robots*, pages 1–16, 2018. ISSN 15737527. doi: 10.1007/s10514-018-9705-x. URL https://doi.org/10.1007/s10514-018-9705-x. Publisher: Springer US.

F. Simion, L. Regolin, and H. Bulf. A predisposition for biological motion in the newborn baby. *Proceedings of the National Academy of Sciences*, 105(2):809–813, January 2008. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0707021105. URL http://www.pnas.org/cgi/doi/10.1073/pnas.0707021105.

Emrah Akin Sisbot, Luis F. Marin-Urias, Xavier Broquère, Daniel Sidobre, and Rachid Alami. Synthesizing Robot Motions Adapted to Human Presence. *International Journal of Social Robotics*, 2(3):329–343, 2010. ISSN 1875-4805. doi: 10.1007/s12369-010-0059-6. URL https://doi.org/10.1007/s12369-010-0059-6. ISBN: 1875-4791.

Alan Slater and Rachel Kirby. Innate and learned perceptual abilities in the newborn infant. *Experimental Brain Research*, 123:90–94, 1998.

Julia Starke, Konstantinos Chatzilygeroudis, Aude Billard, and Tamim Asfour. On Force Synergies in Human Grasping Behavior. In *2019 IEEE-RAS 19th International Conference on Humanoid Robots (Humanoids)*, pages 72–78, Toronto, ON, Canada, October 2019. IEEE. ISBN 978-1-5386-7630-1. doi: 10.1109/Humanoids43949.2019.9035047. URL https://ieeexplore.ieee.org/document/9035047/.

Kyle Strabala, Min Kyung Lee, Anca Dragan, Jodi Forlizzi, Siddhartha S Srinivasa, Maya Cakmak, Willow Garage, Vincenzo Micelli Universi, Degli Studi, Di Parma, Siddhartha S. Srinavasa, Maya Cakmak, and Vincenzo Micelli. Towards Seamless Human-Robot Handovers. *Journal of Human-Robot Interaction (2013)*, 1(1):112–132, 2012. ISSN 2163-0364. doi: 10.5898/jhri.v2i1.114.

Freek Stulp, Jonathan Grizou, Baptiste Busch, and Manuel Lopes. Facilitating intention prediction for humans by optimizing robot motions. *IEEE International Conference on Intelligent Robots and Systems*, 2015-Decem:1249–1255, 2015. ISSN 21530866. doi: 10.1109/IROS.2015.7353529. ISBN: 9781479999941.

Chenxi Sun, Moxian Song, Shenda Hong, and Hongyan Li. A Review of Designs and Applications of Echo State Networks. *arXiv:2012.02974 [cs]*, December 2020. URL http://arxiv.org/abs/2012.02974. arXiv: 2012.02974.

Leila Takayama, Doug Dooley, and Wendy Ju. Expressing thought: improving robot readability with animation principles. In *Proceedings of the 6th international conference on Human-robot interaction - HRI '11*, page 69, Lausanne, Switzerland, 2011. ACM Press. ISBN 978-1-4503-0561-7. doi: 10.1145/1957656.1957674. URL http://portal.acm.org/citation.cfm?doid=1957656.1957674.

Michael Tomasello and Malinda Carpenter. Shared intentionality. *Developmental Science*, 10(1):121–125, January 2007. ISSN 1363755X, 14677687. doi: 10.1111/j.1467-7687.2007.00573.x. URL https://onlinelibrary.wiley.com/doi/10.1111/j.1467-7687.2007.00573.x.

Rei Tomoki OjiSakuragi, Yasutoshi Makino, and Hiroyuki Shinoda. *Weight Estimation of Lifted Object from Body Motions Using Neural Network*, volume 1. Springer International Publishing, 2018. ISBN 978-3-319-93399-3. doi: 10.1007/978-3-319-93399-3. URL http://dx.doi.org/10.1007/978-3-319-93399-3_1. Publication Title: Proc. EuroHaptics 2018 Issue: June.

Panagiota Tsarouchi, Sotiris Makris, and George Chryssolouris. Human–robot interaction review and challenges on task planning and programming. *International Journal of Computer Integrated Manufacturing*, 29(8):916–931, 2016. ISSN 13623052. doi: 10.1080/0951192X.2015.1130251. URL http://dx.doi.org/10.1080/0951192X.2015.1130251. Publisher: Taylor & Francis.

Kuo-Shih Tseng and Bérénice Mettler. Analysis of Coordination Patterns between Gaze and Control in Human Spatial Search. *IFAC-PapersOnLine*, 51(34):264–271, 2019. ISSN 24058963. doi: 10.1016/j.ifacol.2019.01.041. URL https://linkinghub.elsevier.com/retrieve/pii/S2405896319300436.

M.a. Umiltà, E. Kohler, V. Gallese, L. Fogassi, L. Fadiga, C. Keysers, and G. Rizzolatti. I Know What You Are DoingA Neurophysiological Study. *Neuron*, 31(1):155–165, 2001. ISSN 08966273. doi: 10.1016/S0896-6273(01)00337-3. URL http://linkinghub.elsevier.com/retrieve/pii/S0896627301003373. ISBN: 0896-6273 (Print).

Balth. van der Pol Jun. D.Sc. Lxxxviii. on "relaxation-oscillations". *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):978–992, 1926. doi: 10.1080/14786442608564127. URL https://doi.org/10.1080/14786442608564127.

David Vernon, Michael Beetz, and Giulio Sandini. Prospection in Cognition: The Case for Joint Episodic-Procedural Memory in Cognitive Robotics. *Frontiers in Robotics and AI*, 2(July):1–14, 2015. ISSN 2296-9144. doi: 10.3389/frobt.2015.00019. URL http://journal.frontiersin.org/Article/10.3389/frobt.2015.00019/abstract. ISBN: 2296-9144.

A. Vignolo, N. Noceti, A. Sciutti, F. Rea, F. Odone, and G. Sandini. The complexity of biological motion. In *2016 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, pages 66–71, Cergy-Pontoise, France, September 2016. IEEE. ISBN 978-1-5090-5069-7. doi: 10.1109/DEVLRN.2016.7846792. URL http://ieeexplore.ieee.org/document/7846792/.

Alessia Vignolo, Nicoletta Noceti, Francesco Rea, Alessandra Sciutti, Francesca Odone, and Giulio Sandini. Detecting Biological Motion for Human–Robot Interaction: A Link between Perception and Action. *Frontiers in Robotics and AI*, 4(June), 2017. ISSN 2296-9144. doi: 10.3389/frobt.2017.00014. URL http://journal.frontiersin.org/article/10.3389/frobt.2017.00014/full. arXiv: 1011.1669v3 ISBN: 9788578110796.

David Vogt, Simon Stepputtis, Bernhard Jung, and Heni Ben Amor. One-shot learning of human–robot handovers with triadic interaction meshes. *Autonomous Robots*, 42(5):1053–1065, 2018. ISSN 15737527. doi: 10.1007/s10514-018-9699-4. URL https://doi.org/10.1007/s10514-018-9699-4. Publisher: Springer US.

Lei Wang, Jiehui Zheng, Shenwei Huang, and Haoye Sun. P300 and Decision Making under Risk and Ambiguity. *Computational Intelligence and Neuroscience*, 2015:1–7, 2015. ISSN 1687-5265. doi: 10.1155/2015/108417. Publisher: Hindawi Publishing Corporation.

Weitian Wang, Rui Li, Zachary Max Diekel, Yi Chen, Zhujun Zhang, and Yunyi Jia. Controlling object hand-over in human-robot collaboration via natural wearable sensing. *IEEE Transactions on Human-Machine Systems*, 49(1):59–71, 2019. ISSN 21682291. doi: 10.1109/THMS.2018.2883176. Publisher: IEEE.

Ping Wei, Yang Liu, Tianmin Shu, Nanning Zheng, and Song-Chun Zhu. Where and Why are They Looking? Jointly Inferring Human Attention and Intentions in Complex Tasks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6801–6809, Salt Lake City, UT, June 2018. IEEE. ISBN 978-1-5386-6420-9. doi: 10.1109/CVPR.2018.00711. URL https://ieeexplore.ieee.org/document/8578809/.

Timothy N. Welsh, Digby Elliott, J. Greg Anson, Victoria Dhillon, Daniel J. Weeks, James L. Lyons, and Romeo Chua. Does Joe influence Fred's action? Inhibition of return across different nervous systems. *Neuroscience Letters*, 385(2):99–104, 2005. ISSN 03043940. doi: 10.1016/j.neulet.2005.05.013. ISBN: 0304-3940.

Dominik Widmann and Yiannis Karayiannidis. Human Motion Prediction in Human-Robot Handovers based on Dynamic Movement Primitives. In *2018 European Control Conference (ECC)*, pages 2781–2787, Limassol, June 2018. IEEE. ISBN 978-3-9524269-8-2. doi: 10.23919/ECC.2018.8550170. URL https://ieeexplore.ieee.org/document/8550170/.

Amanda L. Woodward, Jessica A. Sommerville, Sarah Gerson, Annette M.E. Henderson, and Jennifer Buresh. Chapter 6 The Emergence of Intention Attribution in Infancy. In *Psychology of Learning and Motivation*, volume 51, pages 187–222. Elsevier, 2009. ISBN 978-0-12-374489-0. doi: 10.1016/S0079-7421(09)51006-7. URL https://linkinghub.elsevier.com/retrieve/pii/S0079742109510067.

Agnieszka Wykowska, Thierry Chaminade, and Gordon Cheng. Embodied artificial agents for understanding human social cognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1693):20150375, 2016. ISSN 0962-8436. doi: 10.1098/rstb.2015.0375.

Alessio Xompero, Ricardo Sanchez-Matilla, Andrea Cavallaro, and Ricardo Mazzon. Corsmal containers manipulation. *CIS, School of Electronic Engineering and Computer Science, Queen Mary University of London*, 2020. URL http://corsmal.eecs.qmul.ac.uk/containersmanip.html.

Hiroki Yamamoto, Atsushi Sato, and Shoji Itakura. Eye tracking in an everyday environment reveals the interpersonal distance that affords infant-parent gaze communication. *Scientific Reports*, 9(1):10352, December 2019. ISSN 2045-2322. doi: 10.1038/s41598-019-46650-6. URL http://www.nature.com/articles/s41598-019-46650-6.

Hiroki Yamamoto, Atsushi Sato, and Shoji Itakura. Transition From Crawling to Walking Changes Gaze Communication Space in Everyday Infant-Parent Interaction. *Frontiers in Psychology*, 10:2987, January 2020. ISSN 1664-1078. doi: 10.3389/fpsyg.2019.02987. URL https://www.frontiersin.org/article/10.3389/fpsyg.2019.02987/full.

Katsu Yamane, Marcel Revfi, and Tamim Asfour. Synthesizing object receiving motions of humanoid robots with human motion database. In *2013 IEEE International Conference on Robotics and Automation*, pages 1629–1636, Karlsruhe, Germany, May 2013. IEEE. ISBN 978-1-4673-5643-5 978-1-4673-5641-1. doi: 10.1109/ICRA.2013.6630788. URL http://ieeexplore.ieee.org/document/6630788/.

Avrahm Yarmolinsky. *The Unknown Chekhov: Stories and Other Writings Hitherto Untranslated*. Noonday Press, 1954.

Lap-Fai Yu, Noah Duncan, and Sai-Kit Yeung. 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, December 2015. IEEE. ISBN 978-1-4673-8391-2. doi: 10.1109/ICCV.2015.88. URL http://ieeexplore.ieee.org/document/7410445/.

Z. Yucel, A. A. Salah, C. Mericli, T. Mericli, R. Valenti, and T. Gevers. Joint Attention by Gaze Interpolation and Saliency. *IEEE Transactions on Cybernetics*, 43(3):829–842, June 2013. ISSN 2168-2267, 2168-2275. doi: 10.1109/TSMCB.2012.2216979. URL http://ieeexplore.ieee.org/document/6320663/.

Andrea Maria Zanchettin, Nicola Maria Ceriani, Paolo Rocco, Hao Ding, and Björn Matthias. Safety in Human-Robot Collaborative Manufacturing Environments: Metrics and Control. *IEEE Transactions on Automation Science and Engineering*, 13(2):882–893, 2016. ISSN 15455955. doi: 10.1109/TASE.2015.2412256.

Minhua Zheng, AJung J. Moon, Elizabeth A. Croft, and Max Q.H. Meng. Impacts of Robot Head Gaze on Robot-to-Human Handovers. *International Journal of Social Robotics*, 7(5): 783–798, 2015. ISSN 18754805. doi: 10.1007/s12369-015-0305-z. Publisher: Springer Netherlands.