TÉCNICO LISBOA

UNIVERSIDADE DE LISBOA
INSTITUTO SUPERIOR TÉCNICO

Robots meet people: socially aware engagement behaviors

João Manuel Caeiro Avelino

Supervisor : Doctor Alexandre José Malheiro Bernardino
Co-Supervisor : Doctor Rodrigo Martins de Matos Ventura
Co-Supervisor : Doctor Leonel Garcia-Marques

Thesis approved in public session to obtain the PhD Degree in Electrical and Computer Engineering

Jury final classification: Pass with Distinction

2023

TÉCNICO LISBOA

UNIVERSIDADE DE LISBOA
INSTITUTO SUPERIOR TÉCNICO

Robots meet people: socially aware engagement behaviors

João Manuel Caeiro Avelino

Supervisor: Doctor Alexandre José Malheiro Bernardino
Co-Supervisor: Doctor Rodrigo Martins de Matos Ventura
Co-Supervisor: Doctor Leonel Garcia-Marques

Thesis approved in public session to obtain the PhD Degree in
Electrical and Computer Engineering

Jury final classification: Pass with Distinction

Jury

Chairperson: Doctor João Manuel Lage de Miranda Lemos, Instituto Superior Técnico, Universidade de Lisboa;
Members of the Committee:

Doctor Estela Guerreiro da Silva Bicho Erlhagen, Escola de Engenharia, Universidade do Minho;
Doctor José Alberto Rosado dos Santos Victor, Instituto Superior Técnico, Universidade de Lisboa;
Doctor Rui Filipe Fernandes Prada, Instituto Superior Técnico, Universidade de Lisboa;
Doctor Iolanda Margarete dos Santos Carvalho Leite, School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Sweden;
Doctor Alexandre José Malheiro Bernardino, Instituto Superior Técnico, Universidade de Lisboa;
Doctor Tomás Alexandre Campaniço Palma, Faculdade de Psicologia, Universidade de Lisboa.

Funding Institutions: Fundação para a Ciência e a Tecnologia

2023

Abstract

Engagement during human-robot first encounters is a challenging task that requires both parties to understand each other's intent to interact without mutual information. Engagement may fail due to people's unawareness of the robot's presence and its desire to interact, or because of people's lack of interest in interacting. In this thesis, we pursue answers to the following research question: "Which theories and methods can enhance the way a mobile social robot engages with humans in first encounters to increase engagement success and improve people's perceptions of the robot?". We concentrate on one-to-one human-robot nonverbal interactions during first encounters with cognitively healthy adults and in public or semi-public places. We explore whether human-human greeting models from social psychology can be an answer to this question. First, we surveyed existing engagement models for mobile robots, analyzed the state-of-the-art of technological solutions for human-robot engagement, and proposed a taxonomy to organize the literature according to well-known greeting models. On the technological side, we developed handshake skills for the Vizzy robot, a pipeline to detect social signals, and methods to detect interaction errors. Our first user study confirmed the social agency of the Vizzy robot and showed how people interact with it. Then, we gathered support for our hypothesis by evaluating a handshake salutation (part of human-human greeting models). People reported increased intentions to help the robot in the future after greetings with a handshake, and increased perceived warmth, likeability, and animacy. The final experiment tested three different greeting models with ablations of a human-human greeting model proposed by Kendon. When the robot acted with the entire model, its interaction intentions were clearer to the human than when only a subset of the model was used, thus empirically demonstrating that a human-human greeting model, namely Kendon's model, is effective in human-robot engagement.

Keywords

social signals, engagement, human–robot interaction, Vizzy robot, Kendon's greeting model

Resumo

Abordar humanos durante primeiros encontros é uma tarefa desafiante para robôs. É necessária a compreensão mútua das intenções de interagir, mesmo sem informação personalizada. Quando os humanos não reparam na presença do robô ou tenham falta de interesse em interagir a abordagem poderá falhar. Nesta tese, procuramos responder à pergunta de investigação: "Que teorias e métodos podem melhorar a forma como um robô social móvel aborda humanos em primeiros encontros, aumentando o sucesso da abordagem e a melhorando a percepção que as pessoas têm do robô?". Focamo-nos em interacções humano-robô não-verbais durante os primeiros encontros com adultos cognitivamente saudáveis em locais públicos ou semipúblicos. Para responder a esta questão, investigámos modelos de cumprimento humano-humano descritos na psicologia social. Inicialmente, revimos modelos de abordagem usados por robôs móveis descritos na bibliografia, o estado da arte das tecnologias usadas na abordagem humano-robô, e propusemos uma taxonomia para organizar a literatura à luz de modelos bem conhecidos. Na parte tecnológica, desenvolvemos um aperto de mão para o robô Vizzy, uma arquitectura para detectar sinais sociais, e métodos para detectar erros de interacção. O nosso primeiro estudo com utilizadores confirmou a agência social do robô Vizzy e mostrou como as pessoas interagem com ele. Depois, reunimos suporte para a nossa hipótese, avaliando um aperto de mão humano-robô (uma fase dos modelos de cumprimento humano-humano). Os participantes que receberam o aperto de mão reportaram maior intenção de ajudar o robô no futuro, e reportaram que era mais caloroso, simpático e animado. Na experiência final testámos três modelos distintos de cumprimento derivados de abluções de um modelo humano-humano proposto por Kendon. O modelo mais completo clarificou as intenções de interacção do robô face aos modelos simplificados. Desta forma, demonstra-se empiricamente que um modelo de cumprimento humano-humano, nomeadamente o modelo de Kendon, é eficaz na abordagem humano-robô.

Palavras Chave

sinais sociais, abordagem (*engagement*), interação humano–robô, robô Vizzy, modelo de cumprimentos de Kendon

Acknowledgments

I thank Fundação para a Ciência e Tecnologia (FCT) for funding this Ph.D. through grants SFRH/BD/133098/2017 and COVID / BD / 152458 / 2022.

I profoundly thank my three advisors for these years. Professor Alexandre, for his scientific knowledge, planning, vision, availability, comments, writing assistance, and the opportunities he has provided before and during this journey. Professor Rodrigo for his scientific and technical knowledge in the experiment preparation and execution, as well as his comments and constructive criticism. And thanks to Professor Leonel for his unique perspectives on psychology as well as the knowledge earned in the Social Cognition classes. I would also like to thank the following professors for their contributions to my development through collaborations, classes, and discussions of ideas: José Santos-Victor, Gaby Odekerken-Schröder, Dominik Mahr, Ana Paiva, João Sequeira, José Gaspar, Andrzej Wichert, Pedro Lima, Jorge Marques, João Pedro Gomes, Ricardo Ribeiro, Dan, and Asim. Thank you for your many constructive criticisms, CAT professors Iolanda Leite and Rui Prada.

I thank the six main pillars of the thesis work's execution. Plinio Moreno, for being a mentor, an inspiration, and a Vizzyzinessess expert. Rui Figueiredo for the technical and scientific training, for introducing me to Vizzy, and for all of his assistance. Filipa Correia and Martina Čaić for their social science teaching and collaboration. Ricardo Nunes and Weverton Macedo for their hardware assistance. Thank you to the entire team for persevering in the face of Vizzy's eccentricities and for your friendship. I'd also like to thank the following colleagues for their contributions to publications, experiments, and software development: Pedro Vicente, Hugo Simão, Heitor Cardoso, Carlos Cardoso, Tiago Paulino, Nuno Duarte, Rui Garcia Figueiredo, Manuel Carvalho, André Gonçalves, Alexandra Gonçalves, Rita Pagaimo, Fernando Loureiro, Laura Santos, Rita Cóias, Rui Livramento, Rafael Sousa, António Ramiro, and Gustavo Brites. And also to the μRoboptics team for their support and understanding when I had to prepare for the thesis defense.

The human contributions of all the people with whom I had the privilege of sharing this stage of life were paramount in this Ph.D. I am grateful to my Vislab colleagues, IST friends, and Alentejo friends for the good times, memes, jokes, and support. I thank my family and Daniela's family for all their emotional support and happy moments. Finally, I want to thank the three people who have always shown me unconditional love, care, and support, illuminating the darkest days: my mother, Maria Catarina, my father, José Avelino, and Daniela.

Agradecimentos

Agradeço o financimento deste doutoramento à Fundação para a Ciência e Tecnologia (FCT) através das bolsas SFRH/BD/133098/2017 e COVID / BD / 152458 / 2022.

Aos meus orientadores o meu muito obrigado. Ao professor Alexandre por todo o conhecimento científico, planeamento de trabalho, visão, disponibilidade, comentários, apoio na escrita e por todas as oportunidades dadas antes e durante este percurso. Ao Professor Rodrigo pelo seu conhecimento científico e técnico na preparação e execução das experiências, pelos comentários e críticas construtivas. E ao professor Leonel pelas suas perspetivas únicas da área da psicologia bem como o conhecimento nas aulas de Cognição Social. Agradeço também aos professores que contribuíram para o meu crescimento através de colaborações, aulas e discussão de ideias: José Santos-Victor, Gaby Odekerken-Schröder, Dominik Mahr, Ana Paiva, João Sequeira, Andrzej Wichert, José Gaspar, Pedro Lima, Jorge Marques, João Pedro Gomes, Ricardo Ribeiro, Dan e Asim. Aos professores membros da CAT, Iolanda Leite e Rui Prada, obrigado pelas muitas críticas construtivas.

Agradeço aos seis pilares fundamentais da execução dos trabalhos da tese. Ao Plinio Moreno por ser um mentor, uma grande inspiração e o especialista em Vizzyzices. Ao Rui Figueiredo pela formação técnica e científica, por me ter apresentado o Vizzy e toda a ajuda prestada. À Filipa Correia e à Martina Čaić pelos seus ensinamentos nas ciências sociais e pela colaboração. Ao Ricardo Nunes e Weverton Macedo pelo suporte ao hardware. A toda a equipa, obrigado por terem lutado de forma resiliente contra todas as idiossincrasias do Vizzy e pela vossa amizade. Quero também agradecer a todos os colegas que colaboraram comigo em publicações, experiências e desenvolvimento de software: Pedro Vicente, Hugo Simão, Heitor Cardoso, Carlos Cardoso, Tiago Paulino, Nuno Duarte, Rui Garcia Figueiredo, Manuel Carvalho, André Gonçalves, Alexandra Gonçalves, Rita Pagaimo, Fernando Loureiro, Laura Santos, Rita Cóias, Rui Livramento, Rafael Sousa, António Ramiro e Gustavo Brites. E também à equipa da µRoboptics pelo apoio à preparação da defesa.

Neste doutoramento foram essenciais os contributos humanos de todas as pessoas com as quais tive o privilégio de partilhar esta fase da vida. Aos meus colegas do Vislab, aos meus amigos do IST e aos meus amigos do Alentejo deixo um grande agradecimento pelos bons momentos, memes, piadas e apoio. À minha família e à família da Daniela, agradeço todo o apoio emocional e os bons momentos. E finalmente, o maior dos agradecimentos às três pessoas que sempre me deram todo o amor, carinho e apoio incondicional, iluminando os dias mais negros: à minha mãe, Maria Catarina, ao meu pai, José Avelino, e à Daniela.

Para ser grande, sê inteiro: nada
Teu exagera ou exclui.

Sê todo em cada coisa. Põe quanto és
No mínimo que fazes.

Assim em cada lago a lua toda
Brilha, porque alta vive.

Ricardo Reis (Fernando Pessoa)

Contents

List of Figures

List of Tables

1 INTRODUCTION

1.1 Motivation

Following the recent technological advances in robotics, mobile social robots are starting to appear in social contexts. We define this class of robots as embodied agents designed to engage in social interaction that can navigate autonomously in their environment, combining the definitions of social robots [1] and mobile robots [2]. Unlike virtual characters on screens, computers, and smartphones, the embodiment of mobile social robots allows them to be proactive members of society and to improve human engagement [3]–[5]. There is a growing interest from both industry and academy in these robots for distinct fields. For instance, industry applications include using them as robotic butlers to approach and guide people in their facilities, greet visitors, and serve food and drinks in restaurants and events. Another important application for these systems is assistance to humans in elderly care centers. Given the unprecedented increasing gap between the supply and demand of care services, researchers have used robots like Vizzy [6], Mbot [7], and GrowMu [8] to help the staff to entertain, persuade, and motivate seniors to participate in activities and physical exercises.

As exemplified, the goals of these robots can be broad, sharing, however, a common task: to meet and engage humans in interaction in a possible first encounter. Engagement during first encounters is a complex task, requiring both parties to understand each others' intent to interact, even though they have not shared personal information before. This mutual understanding is even more crucial when dealing with technology-naive users like elderly care center residents. For instance, during the Augmented Human Assistance project[1], the Vizzy robot visited several elderly care centers in the role of a robot coach that invited and motivated residents for physical exercise with a gaming platform [9], [10]. During this Wizard of Oz (WoZ) experiment, we observed that when the Wizard failed to command the robot with appropriate gaze behaviors, people would get confused, not knowing if it was attempting to interact with them. Some of them also tried to greet the robot with a handshake, highlighting that they expected the social robot to act according to social norms.

[1] http://aha.isr.tecnico.ulisboa.pt/

Underdeveloped engagement attempts, such as systems that navigate towards a point near the target person, are unsatisfactory. Robots may fail either because of people's unawareness of the robot's presence or intentions to interact or because people were just not interested in the interaction. For instance, Figure 1.1 illustrates a scenario where the Vizzy robot attempted to engage with someone during an exhibit. However, it could not appropriately demonstrate its intentions to interact, ultimately being ignored by the target person, who remained completely unaware of the robot's presence and intentions.

Figure 1.1: An example where our Vizzy robot (described in section **??**) attempted to engage with someone unsuccessfully. The robot could not detect the person being completely unaware of its presence nor act to make its intentions clear.

Robotic unsuccessful engagement attempts were not exclusively observed during our experiments but have been formally reported and described in the literature, most notably in the works of Satake and colleagues [11], [12]. We illustrate the four identified types of errors in Figure 1.2. The inability to engage undermines robots' end goals, making them less efficient in their social roles since they cannot even initiate the interaction with the target. In addition, naive attempts to engage with someone may create perceptions of low competence and social skills that indirectly impact people's commitment in subsequent interactions, even when robots manage to initiate it. Thus, improving mobile robots' engagement skills during first encounters is essential for enhanced Human–Robot Interaction (HRI).

1.2 Topic overview

Copresent people can be in unfocused or focused interaction [13], [14]. Unfocused interaction occurs when people are in the presence of each other, like being in a line or a waiting room. For example, people interact by managing their space and movements just by being present. Focused interactions consist of two or more interactants sustaining their attention in a single focus. Examples include chatting, working on a common task, and watching a movie. Following Goffman's [13], [14] and Kendon's [15] theories, Bassetti [16] expands focused interactions into (a) common-focused interactions and (b) jointly-focused interaction. Common-focused interactions have a common but not reciprocal focus of interaction (e.g., supporting a sports team during a game or appreciating art in a museum). In jointly-focused interaction, people perform a mutual activity (e.g., chatting,

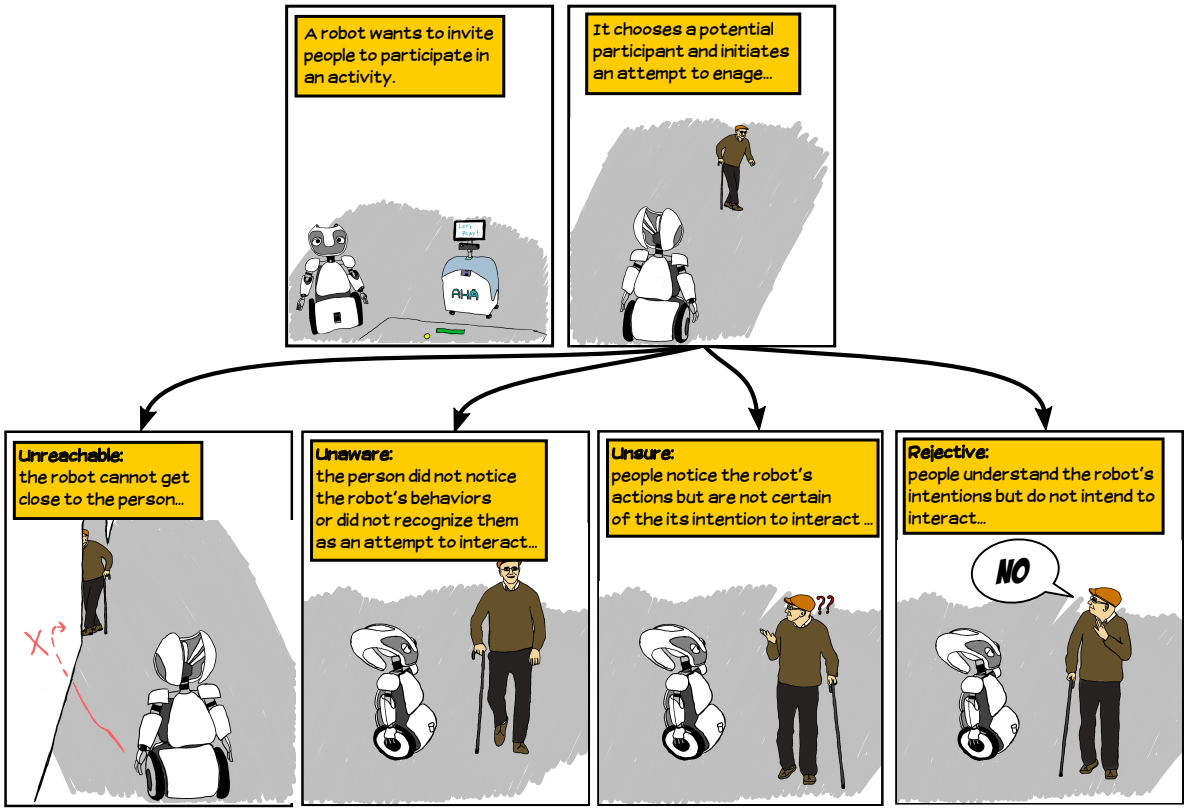


Figure 1.2: Mobile robot approach failure types, as described by Satake et al. [11], [12].

playing a game, or building something). Our work focuses on the transition between unfocused to jointly-focused face-to-face interactions.

When people initiate a face-to-face interaction, they need to acknowledge each other's presence and engage by showing their intentions. Humans share their intents through patterns of verbal and non-verbal behaviors. These have been reported and analyzed in past literature [13], [17], [18] and constitute the greeting protocol, which we discuss in-depth in Chapter 2. In addition, focused interactions need to be maintained, which is easier if people have positive impressions, especially during first encounters. While people generate these impressions in the first moments of observing their peers, they evolve throughout the interaction. Positive evaluations need maintenance based on subtle social cues that indicate whether people's behaviors live up to expectations or not.

The term "engagement" has a broad number of context-dependent definitions. Sidner et al. [19] define it as "(...) the process by which interactors start, maintain, and end their perceived connections to each other during an interaction (...)", which goes in line with the work proposed in this thesis. Additionally, the literature [20] splits engagement into three types: (i) cognitive - psychological devotion to a task; (ii) behavioral - "physical participation and involvement"; and (iii) emotional - the affective responses towards some task. This thesis focuses mainly on the success of a robot in opening the interaction with humans (behavioral engagement) and on the impressions the robot causes on them (emotional engagement). Nonetheless, some of our works also measure the impacts of robot behaviors on people's dedication to the task (cognitive engagement).

1.3 Scope and approach of the thesis

This thesis focuses on one-to-one human-robot non-verbal interactions during first encounters with a human. More specifically, we seek answers for the following research question:

Research question 1 (RQ1)**.** *Which theories and methods can enhance the way a mobile social robot engages with humans in first encounters, increasing engagement success and people's perceptions of the robot?*

This question has a broad range of contexts for which we may obtain distinct answers. Thus, we narrow down our research to first encounters with adults without cognitive impairments in public / semi-public spaces. First encounters, from the robot's perspective, are those where the robot does not have a personalized model of the target's physical and psychological characteristics.

To address this problem we seek inspiration from models from social psychology literature. Humans are relatively competent at engaging with each other, even when dealing with strangers, which may not speak their language. They use non-verbal social signals to communicate their intents to open the interaction and monitor its success or failure. Our hypothesis is that human greeting theories, like Kendon's greeting model and its behaviors, will significantly improve the success of HRI (behavioral engagement) and create a positive impression of the robot in first encounters (emotional engagement). In the context of this thesis, interactions are successful if the robot manages to initiate jointly-focused interactions with target humans and if they do not abandon the robot during the interaction. Measurements of emotional engagement consist of self-reported measures, like questionnaires. We measure cognitive engagement through task-related dimensions, like clicks on interfaces, interaction times, counts of prosocial behaviors (behaviors that people do to help the robot), and self-reports of willingness to repeat tasks in the future.

In our work, we investigate whether we can integrate greeting theories and monitor the interaction in mobile humanoid robots through the integration of both off-the-shelf algorithms and inovative solutions.

1.4 Objectives

The problem of human-robot engagement during first encounters is multidisciplinary, encompassing theories and methods from robotics, artificial intelligence, and social studies. Thus, advances in this field depend on the interconnection of technology methods and HRI studies. The developed methods contribute to this domain by automatically letting robots detect social cues, reason about appropriate social norms, and act in socially acceptable and user-comfortable ways. HRI studies are paramount for HRI fields in two ways. First, they support or reject the applicability of Human–Human Interaction (HHI) theories on HRI. Their insights guide design decisions of methods or establish new interaction research directions. Second, they evaluate whether the proposed methodologies and algorithms are efficient and whether they provide improvements in users' technology acceptance. Due to the nature of our work, we set objectives for both components: (i) methods and (ii) HRI studies.

Our methodological objectives are to develop and integrate methods that implement social theories of greetings and interaction monitoring in a mobile humanoid robot. More specifically, we aim to:

Objective 1 (O1)**.** Develop and integrate methods that can detect people's non-verbal social signals from the robot's onboard sensors and keep track of the interaction.

Objective 2 (O2)**.** Enable mobile social robots to engage with people autonomously while respecting the social scripts reported in the social cognition literature.

Our objectives related to HRI studies intend to support the developed behavioral models and robot capabilities. These objectives consist of:

Objective 3 (O3)**.** Gather evidence that humans respond more positively to mobile social robots that act according to models of engagement observed in human-human interactions (more emotional engagement).

Objective 4 (O4)**.** Evaluate whether mobile social robots following the proposed models are more successful when attempting to open the interaction with people in first encounters (more behavioral engagement).

1.5 Contributions

Our contributions to the problem of human-robot engagement for mobile social robots during first encounters are fourfold. First, this research direction is still recent and unexplored since only in recent years have mobile social robots become robust enough for interaction and deployment in the wild. Since most authors use their own taxonomies and terminologies, it can be challenging to compare engagement models and draw conclusions. Thus, our first contribution is to **propose a taxonomy of engagement behaviors** with which we can compare works in the same light and with standard terminologies and compare the state-of-the-art skills for engagement. For this, we have drawn inspiration from the social cognition literature and theories, and performed a survey of existing works under this taxonomy. In this thesis, this contribution is part of chapters 2 and 3, which we previously published as:

- J. Avelino, L. Garcia-Marques, R. Ventura, and A. Bernardino, "Break the ice: A survey on socially aware engagement for human–robot first encounters", *International Journal of Social Robotics*, vol. 13, no. 8, pp. 1851–1877, Jan. 2021, ISSN: 1875-4805. DOI: 10.1007/s12369-020-00720-2.

The second set of contributions consists of developing **methods for autonomous engagement** for social robots during first encounters. Through these methods, we equipped our social robot Vizzy with the means to initiate interaction with people autonomously following models from the social sciences literature. These methods allow the robot to: (i) **initiate encounters with a handshake**, (ii) **perceive social signals, keep track of the interaction, and act according to Kendon's**

greeting model, and (iii) **detect interaction failures**. This thesis describes these contributions in chapters 4, 5 , and 6. We implemented opensource code for handshakes (Robot Operating System (ROS) tactile drivers for Vizzy, arm/handshake gestures, handgrip tactile-based controllers), robot behavioral control (behavior trees wrapper for ROS), human awareness (3D human pose estimation, tracking, and skeleton completion), and greetings (social signals estimation and greeting control). A significant part of these contributions was published/submitted as the following papers:

- J. Avelino, T. Paulino, C. Cardoso, R. Nunes, P. Moreno, and A. Bernardino, "Towards natural handshakes for social robots: Human-aware hand grasps using tactile sensors", *Paladyn, Journal of Behavioral Robotics*, vol. 9, no. 1, pp. 221–234, Aug. 2018. DOI: 10.1515/pjbr-2018-0017.
- J. Avelino, A. Gonçalves, R. Ventura, L. Garcia-Marques, and A. Bernardino, "Collecting social signals in constructive and destructive events during human-robot collaborative tasks", in *Companion of the ACM/IEEE International Conference on Human–Robot Interaction*, Cambridge, United Kingdom (Online), Mar. 2020.
- M. Carvalho*, J. Avelino*, A. Bernardino, R. Ventura, and P. Moreno, "Human-robot greeting: Tracking human greeting mental states and acting accordingly", in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, Prague, Czech Republic (Online), Sep. 2021.
- F. Loureiro, J. Avelino, P. Moreno, and A. Bernardino, "Detecting human–robot interaction failures through egocentric visual head-face analysis", in *EgoVIP - Egocentric vision for interactive perception, learning, and control, Workshop at IROS 2021*, Prague, Czech Republic (Online), Oct. 2021.

The third set of contributions is a set of **HRI studies**. Through these studies, we observed how people interact with robots during first encounters, sought support for the usage of evaluation metrics, and evaluated robots' behavioral models in people's engagement during interactions. They required us to develop **additional software to support the experiments**: to teleoperate our robot and receive human feedback during interactions, which we open-sourced. We describe these contributions in chapters 7, 8, and 9. Most of these contributions were published as:

- M. Čaić, J. Avelino, D. Mahr, G. Odekerken-Schröder, and A. Bernardino, "Robotic versus human coaches for active aging: An automated social presence perspective", *International Journal of Social Robotics*, vol. 12, no. 4, pp. 867–882, Jul. 2019, ISSN: 1875-4805. DOI: 10.1007/s12369-018-0507-2.
- J. Avelino, H. Simão, R. Ribeiro, P. Moreno, R. Figueiredo, N. Duarte, *et al.*, "Experiments with vizzy as a coach for elderly exercise", in *ACM/IEEE International Conference on Human–Robot Interaction – Workshop on Personal Robots for Exercising and Coaching (PREC)*, Chicago, Illinois, USA, Mar. 2018.
- J. Avelino, F. Correia, J. Catarino, P. Ribeiro, P. Moreno, A. Bernardino, *et al.*, "The power of a hand-shake in human–robot interactions", in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, Madrid, Spain, Oct. 2018.

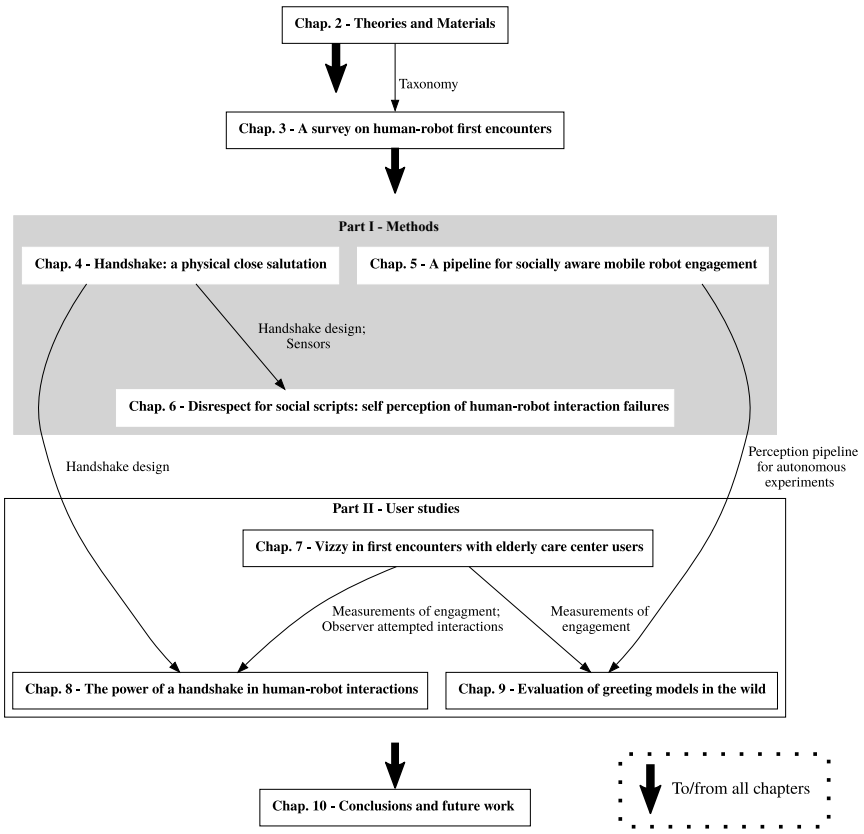


Figure 1.3: Diagram of thesis outline: big rectangles represent the two parts of the thesis, small rectangles represent each Chapter, and arrows represent relationships between chapters.

Although as side projects and not described in this thesis, we also contributed with additional social robotics publications during the Ph.D. activities. These were:

- H. Simão, J. Avelino, N. Duarte, and R. Figueiredo, "Geebot: A robotic platform for refugee integration", in *Companion of the 2018 ACM/IEEE International Conference on Human–Robot Interaction*, Chicago, Illinois, USA, Mar. 2018.
- R. Livramento, J. Avelino, and P. Moreno, "Natural data-driven approaching behaviors of humanoid mobile robots for f-formations", in *IEEE International Conference on Autonomous Robot Systems and Competitions*, Ponta Delgada, Portugal, Apr. 2020.

1.6 Thesis outline

Now, we summarize how we organized this thesis. We visually explain its structure through the diagram in Figure 1.3, while Table 1.1 lists a summary of each Chapter. In Chapter 2, we lay the foundations of our work by proposing a taxonomy, describing Kendon's greeting model, and making a thorough analysis of robot skills for engagement. Since failures are prone to happen even in human-human interactions, we also study concepts related to social feedback to close the loop by detecting them. Further, we enumerate measurements of engagement-related constructs that will be useful for our thesis and present the robotic platform used during our research work. Chapter 3 depicts an exhaustive review of models for mobile robot engagement in the HRI literature, how they

Table 1.1: Summary of each Chapter of the thesis.

Chapter	Summary
2 - Theories and Materials	- Taxonomy - Kendon greeting model - Social awareness skills - Social feedback - Measurements of engagement success - The Vizzy robot
3 - A survey on human–robot first encounters	- Mobile robots engaging with people - State-of-the-art of relevant skills
4 - Handshake: a physical close salutation	- Development and implementation of a handshake for the Vizzy robot
5 - A pipeline for socially aware mobile robot engagement	- Detection of people and information for engagement - HMM greeting state tracking
6 - Disrespect for social scripts: self–perception of human–robot interaction failures	- Perception of failure through tactile sensing - Perception of greeting failure through facial feedback
7 - Vizzy in first encounters with elderly care center users	- Development of WoZ functionality - Qualitative analysis of interactions - People's expectations of robot's actions - Test of robot acceptance and fitness of application - Test of engagement metrics - Identified limitations
8 - The power of a handshake in human–robot interactions	- Study impacts on willingness to help a robot (prosocial behavior) - Impacts on perception of robot characteristics
9 - Evaluation of greeting models in the wild	- Compare the engagement impacts of three greeting models - Evaluate the human–robot interactions

relate to Kendon's greeting model, and their gaps. We also investigate the state-of-the-art of needed robot social skills and how we can integrate them to fulfill our technological objectives. Contents from Chapter 2 and Chapter 3 were previously published in our survey paper [21] (with some novel additions) and support the remainder of our thesis work. Afterward, we split our thesis into two parts.

Part I - Methods: comprises our technological implementation achievements and is composed of chapters 4, 5, and 6. Chapter 4 describes our user-centered design of handshakes for the Vizzy robot, whose contents we published in [22]. We propose a pipeline to estimate the necessary information to initiate an encounter in Chapter 5: human detection, tracking, social signal processing, and greeting phase estimation. We published the greeting phase estimation part in our IROS paper [24]. Then, in Chapter 6, we use features and theories from previous chapters to develop models of self-perception of social expectation violations. Part of this Chapter's results was published in our handshake paper [22], and another part was submitted to the ICMI conference [29].

Part II - HRI studies addresses user expectations and attitudes toward the Vizzy robot during first encounters. Chapter 7 describes a user study where Vizzy and elderly care residents interact for the first time. This Chapter's contributions to this thesis are threefold. The first one is the

development of tools to control the robot in WoZ scenarios. Second, by being our robot's first study with end-users where the robot invites people to perform a task during a first encounter, we could assess the robot's acceptance and social presence, perceptions and opinions, and observations of attempted interactions and issues. Finally, we study the relationships between robot-related self-reported measures and user engagement with the task, which we can use in later chapters to evaluate the developed behavioral models. Parts of this Chapter were previously presented in a workshop at the HRI conference [10] and published in the International Journal of Social Robotics [9]. Chapter 8 formally studies the power of handshakes when opening Human–Robot Interactions. Handshakes are one of the close salutation gestures in Kendon's greeting model (Chapter 2) and were one of the observed engagement attempts during the user study in Chapter 7. We published and presented these results in the IROS2018 conference [26]. We finalize this part with Chapter 9, where we evaluate greeting models through an in-the-wild experiment running a pipeline composed of all the developed models along with this thesis. This Chapter's contents are novel and have not been published before.

Chapter 10 wraps up our thesis, deriving conclusions from all chapters, identifying open questions, and suggesting future work directions.

2

THEORIES AND MATERIALS

This Chapter compiles the theories and definitions related to methods of engagement for mobile robots during first encounters. Our goal is to provide a solid background that supports the remainder of this thesis work. Since this is a recent trend in robotics and artificial intelligence, some constructs and methods may have distinct definitions or be named differently across research works. Thus, we propose a taxonomy of socially aware behaviors based on the social cognition literature to organize and standardize relevant concepts. While building the taxonomy, we also identify the necessary social skills for human-mobile robot engagement in first encounters.

Additionally, we also describe the materials used during our research work, recalling the methods used to measure human engagement.

Parts of this Chapter are an extension of our recent work published in:

J. Avelino, L. Garcia-Marques, R. Ventura, and A. Bernardino, "Break the ice: A survey on socially aware engagement for human–robot first encounters", *International Journal of Social Robotics*, vol. 13, no. 8, pp. 1851–1877, Jan. 2021, ISSN: 1875-4805. DOI: 10.1007/s12369-020-00720-2.

2.1 Taxonomy of socially aware behaviors

2.1.1 Definitions

The start of a pleasant meeting between two agents requires their mutual recognition as social entities and willingness to interact. This interaction is ruled by social norms, which are particularly important in first encounters since they provide a prior model of people's expectations and behaviors without personal data. Thus, they facilitate coordination between interacting parties. Social norms are so important that people are willing to incur self-costs to punish deviant behaviors [30]. In this work, we use the following definition, proposed by Malle et al. [31].

Definition 1.1 (D1.1)**. Social norm:** "... an instruction to (not) perform action A in context C, provided that a sufficient number of individuals in the community (i) indeed follow this instruction and (ii) demand of each other to follow the instruction."

Remark 1.1 (R1.1). In this work, we refer to social norms that emerge from the natural interaction of people, not enforced by a legal system.

Social robots can use social norms to engage with people without violating their expectations. However, this construct does not encode a plan for action sequences by itself. Even though, before any interaction, each party creates visually based impressions supported by norms and culture, this is not sufficient to plan the exact sequence of appropriate behaviors. Social scripts are a construct proposed by Schank that may fill this gap [32]. After people identify the interaction context, they activate a script that embeds social norms and sequences of actions that they should perform along with the interaction. People learn these scripts through the observation of peers in their community. We use the following definition of social scripts, which we adapted from [33], [34].

Definition 1.2 (D1.2)**. Social script:** a mental construct that contains information about the plans and sequences of actions appropriate and expected from the participants of a social situation.

Given the scope of our thesis, we focus on social scripts that may help social robots detect/signal the willingness to initiate interaction. The literature on HHI proposes a script that describes how people behave when meeting someone, described in section 2.2: Kendon's greeting model [35]. During greetings, people interchange social cues that ground their interaction intentions and establish appropriate social norms for present or future interactions [36]. We adapted the following definition from [35], [37]:

Definition 1.3 (D1.3)**. Greeting:** a ritual consisting of a sequence of interaction behaviors observed when people come into another's presence.

Remark 1.2 (R1.2). We make a clear distinction between a greeting and a salutation. The first is a social script composed of multiple behaviors intended to open the interaction. Salutations are the individual gestures or utterances that explicitly signal one's intent to interact (e.g., performing a handshake and saying "Hi").

A greeting social script is far from a one-size-fits-all solution and cannot be followed blindly. Verbal and non-verbal signals vary according to culture and meeting context [38]. As Hall reports, these differences may occur in the management of space, gestures, or salutations. The differences may be so extreme that normal behaviors in one culture may be considered deviant in another. For instance, while the English keep their gaze fixed on another person to show their full attention, Americans find the behavior uncomfortable and advert gaze frequently. Since it is not feasible to encode and list all social norms and scripts for a robot to follow due to the number of possible contexts, it will eventually break modeled or unmodeled social norms. If this happens, the robot must perceive people's verbal [39] or non-verbal [40] social feedback to recover from interaction failures.

Definition 1.4 (D1.4)**. Social feedback:** an evaluative response to a social actor's actions in a specific social context, displayed through social cues.

Remark 1.3 (R1.3). We use the term *social actor* to refer to both humans and social robots.

We believe that closing the loop through the perception of social feedback is crucial to improving how people perceive the social robot. For instance, since the public perceives robots as competent beings, people may interpret failures as incongruent behaviors and recall these more than the positive part of the interaction (incongruency effect [41]–[44]). However, this effect vanishes if people learn a personality trait that explains the incongruent behavior [45]. Thus, it may be possible for a robot to recover from a failure by detecting it and using a recovery behavior that justifies it. We note that, to our knowledge, these effects were only observed in HHI. However, past works state that people tend to anthropomorphize nonhuman agents [46], [47] and may evaluate them as social entities [48].

Our final definition encompasses the complete set of skills necessary to interact with social agents. We make use of Greenspan's [49] definition of *social awareness*.

Definition 1.5 (D1.5)**. Social awareness:** "... the individual's ability to understand people, social events, and the processes involved in regulating social events."

We believe that using this definition of *social awareness* clarifies in which domains we set our research path. Moreover, it establishes a parallel between the developed robot skills and the social cognition skills of humans, allowing researchers to keep track of developments and pinpoint gaps by comparing robots and human cognition.

2.2 A script to open the interaction: the greeting protocol

As introduced in sections 1.2 and 2.1, the greeting protocol is a social script used by social agents to transition from unfocused toward jointly-focused face-to-face interaction. People's certainty of their peers' willingness to interact with them evolves when observing greeting behaviors, and the longer they take to initiate the greeting, the more copresent people will perceive them as not wanting to interact [13], [50]. Even though greetings have culture-specific characteristics - like distances, gaze patterns, or salutation gestures -their basic structure is universal among cultures [51].

Several works in the literature analyze the properties of greetings, either focusing on verbal or non-verbal characteristics. Although the initial literature [13], [18], [52] describes greeting elements with detail, they relied predominantly on common-sense rules and event recalls, lacking a systematic examination. Kendon and Ferber [17] were the first to use video recordings to analyze human behaviors when opening an encounter systematically and build a model. Up to our knowledge, this remains the most widely accepted greeting model, having its observations been replicated in subsequent works [37].

Kendon's greeting model consists of six phases involving distinct sets of social cues exchanged between the participants, depicted in Figure 2.1 and described in the following subsections. While doing so, we highlight the social skills required to implement, which we summarize in Table 2.1 and Table 2.2.

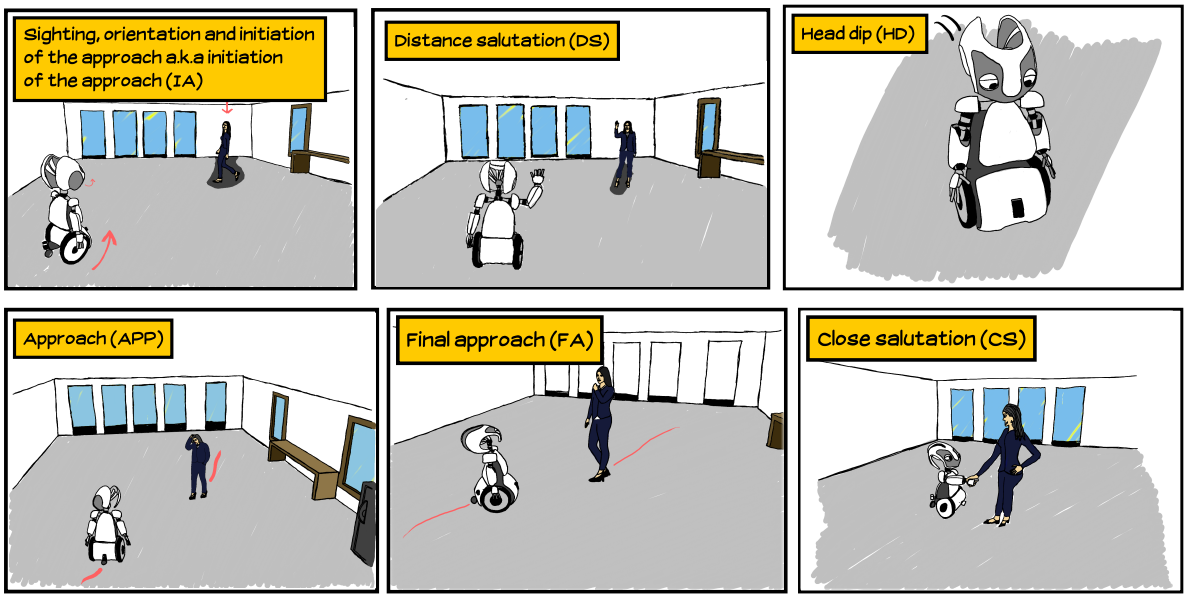


Figure 2.1: The six phases of Kendon's greeting model adapted to human and robot social actors. During the greeting protocol, both actors interchange a set of social cues while approaching each other, as illustrated.

2.2.1 Sighting, orientation, and initiation of the approach (IA)

The first step to initiate the greeting protocol with other social actors is to identify them as someone we wish to interact with and if the conditions are appropriate. This necessity requires an agent to possess a skill set that we call social context detection, which consists of detection, tracking, identification of people, and scene and activity recognition. Additionally, Kedon argues that humans avoid approaching targets before getting noticed. Gaze is usually the primary social signal used to acknowledge the other's presence, making gaze and visual field of attention estimation a needed ability for social agents. Nonetheless, ways to make one's presence recognized depend on the urgency, roles, the greeting's goal, and current activity. People can also call others out and knock on doors, and even the properties of these behaviors may vary. For example, Yoshioka and colleagues [53] observed significant differences in speech distances and approach trajectories depending on how concentrated people perceived the target to be. Thus it is fundamental to be aware of the social context: ongoing human activities, groups, and estimates of people's interruptibility. In addition, humans attempt to minimize the risk of explicit rejection and may send subtle signals of intending to interact. During open-space interactions, Kendon reports the following strategies people use to get their target's attention, which characterize the IA phase:

- Gaze at the targets and wait for gaze signals without reorienting their body toward them.
- To decrease the risk of explicit rejection, synchronize movements with the target's and avert gaze.
- Call out, make gestures, cough, or knock on doors.
- In urgent cases, interrupt the other's activity directly.

2.2.2 Distance salutation (DS)

After acknowledging each others' presence through gaze in the IA phase, both parties signal the official beginning of the greeting through a distance salutation. Two outcomes can follow this stage: the greeting evolves to another phase depending on distance, or the greeting ends in "greetings in passing" scenarios. Thus, it is necessary to keep track of how a greeting evolves. The most common distance salutation forms include a combination of the following actions:

- Wave
- Smile
- Call
- Head movements:
- Nod
- Head toss
- Head lower

Social robots need to detect and execute these actions since both parties perform them. This need translates to requiring skills of gesture recognition and facial expression detection.

2.2.3 Head Dip (HD)

The head dip (HD) is a very short phase where greeters bend their necks forward, lowering their heads. In Kendon's observations, this phase has a higher chance of happening if people need to adjust their body orientation to approach their targets and does not occur after a distance salutation that does not lead to further interaction. Kendon hypothesizes that this behavior represents a shift of attention but cannot conclude whether it functions as a regulatory signal that strengthens the greeting intent or just a way to get the awareness of one's surroundings when starting to approach the other.

2.2.4 Approach (APP)

During the approach phase, people move toward their interaction target. Either both parties move, or one waits while the other approaches. While approaching someone, people need to respect people's personal spaces and may display the following behaviors:

- Gaze aversion (more pronounced in the social actor that walks more)
- Grooming behaviors
- A gesture where the person brings one or both arms forward briefly, called body cross (more likely to happen in the one who walks more)

2.2.5 Final Approach (FA)

The final approach occurs at distances closer than 3.5m and precedes the close salutation. People start deaccelerating in this phase. Unlike the approach phase, its main characteristics are:

- Mutual gazing
- Mutual smiling
- Palm showing gestures
- Verbal salutation

Like the approach phase, social agents need to respect proxemic distances - i.e., follow a socially aware trajectory. In addition, they may need to know how to enter a group of people, as such is a common scenario.

2.2.6 Close Salutation (CS)

The greeting script culminates into the close salutation. The behaviors of this phase signal that both parties understand and agree with each others' intent of opening the encounter. The participants stop walking, orient their hands toward each other, and perform a verbal and non-verbal salutation. Both types are culture-dependent. Some examples of non-verbal salutation gestures include:

- Handshakes
- Waves
- Fist bumps
- Hugs
- Elbow bumps (common during the COVID-19 pandemic)
- Head nods
- Kisses on cheeks
- Bows

Some salutations, like the handshake or kisses, require people to enter each other's personal space. Afterward, people readjust their relative positions, following Hall's proxemics theory [54].

2.2.7 Final considerations

We reiterate that even though the general structure of greetings is stable, the characteristics of individual behaviors may differ. Even in the same culture, there are differences in greeting features related to sex and age, while others are stable. For instance, Astrom [55] did such an investigation for the approach and close salutation phases in a laboratory setup. The results showed that women smiled more, stopped significantly nearer, and performed weaker handshakes than men, while there were no differences in walking speed and handshake duration between genders. Younger participants walked faster than older participants. We also emphasize that Kendon does not argue that all phases occur in every greeting, depending on extra factors like the scene or the acquaintance between social actors. Goffman's [56] states that greeting behaviors attenuate with an increasing acquaintance evolving into a minimal greeting script for mutually well-known peers. This insight highlights the importance of having a complete and detailed greeting model for robots during first encounters, where the greeting tends to be longer. In addition, Schiffrin [50] states

that greetings may not always be linear due to communication failures, like an agent failing to see a social signal from the other, which may lead to repeating some gesture. Thus, we argue that it is also relevant for robots to keep an uncertain estimation of greeting phases. Finally, a social agent may break social norms due to failures in its internal models or sensor/actuator issues, thus, needing to be aware of its peers' social feedback to correct unexpected behaviors.

Table 2.1: Perception skills and greeting phases where they are needed.

	IA	DS	HD	APP	FA	CS
3D detection, tracking & recognition of people	✓	✓	✓[1]	✓[1]	✓	✓
Scene recognition	✓					
Action recognition	✓					
Gaze & VFOA estimation	✓	✓			✓	✓
Facial expression detection	✓	✓			✓	✓
Gesture recognition	✓	✓				✓
Personal & affordance space estimation	✓			✓	✓	
Speech recognition	✓				✓	✓
Greeting tracking with uncertainty	✓	✓	✓	✓	✓	✓

2.3 Social feedback

Even with social norms and scripts knowledge, it is impossible to always act according to people's expectations. Thus, social actors need to detect their actions' impacts on others, a crucial skill to maintain order and protection in society [57]. Besides verbal signals, facial expressions and body movements give feedback to other people about someone's internal states, regulating their behaviors. This skill can be observed in humans as young as 12 months of age, which can perceive facial expressions and use these perceptions to act [58]. In addition, past studies claim that facial expressions are processed and used to reinforce emotional learning by the human amygdala. Thus, given the need for social robots to be aware of their mistakes in Human–Robot Interactions and the importance of facial expressions for social feedback, this thesis also studies ways to implement algorithms that automatically perceive social feedback using facial features.

[1] During these phases the robot looses visual contact with the target. Thus, it needs a belief of people's motion and current position without observations.

[2] Adjust position after a salutation that requires a temporary invation of personal space.

Table 2.2: Actuation skills and greeting phases where they are needed.

	IA	DS	HD	APP	FA	CS
Gaze	✓	✓	✓	✓	✓	✓
Body rotation	✓					
Socially aware navigation				✓	✓	✓[2]
Speech	✓	✓				✓
Head gestures (head dip, nod, among others...)		✓	✓		✓	✓
Non-physical human-robot salutatiations		✓			✓	✓
Physical human-robot salutations (handshake, hug, fist bump, ...)						✓
Facial expressions	✓	✓			✓	✓

2.4 Categorizing atomic skills for engagement

Since engaging with someone is a complex process grounded in a multidisciplinary set of skills, we need to categorize them to have a structured evaluation over their state-of-the-art. We recall that a social robot needs knowledge of social context and social norms, detect social cues and feedback and communicate through non-verbal and verbal actions. We thus found inspiration on Greenspan's [49] theoretical/conceptual model of Social Competence to create a taxonomy of human-robot first encounter skills. Even though this model is part of studies on children with mental disabilities and the literature comprises several theoretical models for Social Competence [59]–[62], we argue that Greenspan's model is a simple yet efficient tool to categorize robots' social skills in first encounters. We focus on Greenspan's Social Awareness competence group, which is composed of three categories of skills: (i) Social sensitivity; (ii) Social insight; (iii) Communication.

2.4.1 Model description

Social sensitivity is composed of skills to perceive and understand social agents, objects, and events. Two sub-components are part of it: social inference and role-taking. Social inference abilities classify situations, gathering, and context, while role-taking skills are those necessary to understand others' points of view and feelings.

The social insight category comprises skills used to interpret, understand, and evaluate the processes that govern social events, and it is composed of three sub-components. Social comprehension is the sub-component related to knowledge of social models and processes, such as norms, social classes, relationships, and reciprocity. Psychological insight is the capability to understand people's motivations and personalities. The third sub-component is Moral judgment, consisting of skills

related to intentionality, morality, and ethics.

Social communication is the set of skills that communicate information and influence other social actors' behaviors and is composed of referential communication and social problem-solving sub-components. While referential communication deals with the verbal and non-verbal skills used to communicate one's thoughts and feelings, social problem solving is the ability to resolve conflicts and influence others' actions.

2.4.2 Assignment of robotic first encounter-related skills to Greenspan's model: automated social awareness

Given Greenspan's model, we now assign the engagement skills to one of the social awareness categories, either using their classic subcomponents or proposing new ones, as needed. Our goal is to keep a simple structure and avoid unnecessarily complex nested sub-categories. We call our taxonomy - automated social awareness for engagement in first encounters.

Here, we propose and address three subcomponents for social sensitivity: social context detection, gaze & VFOA estimation, and role-taking. Social context detection consists of atomic skills that detect/track/recognize people, objects, activities, and facial expressions. Since we are only interested in integrating these skills into perception pipelines, we do not discuss them individually. The role-taking component describes the robot's ability to understand people's reactions and thus social feedback. The remaining sub-components are self-explanatory.

Social insight is split into implicitly and explicitly defined social comprehension. While the first deals with models that encode social norms implicitly (such as socially aware navigation costmaps), the second consists of methods and models where social norms are explicitly defined.

We re-use Greenspan's referential communication and social problem-solving sub-components for the communication category. The first deals with the gestures and dynamics of non-verbal communication through salutations and gaze. Social problem-solving deals with robot behavior adaptation to social feedback.

2.5 Measurements of engagement

Finding support for our thesis requires adequate measures of engagement with the robot and the task it proposes. Past works in the literature used objective and subjective metrics, whoever we could not find a consensual and validated set of metrics for engagement in HRI. Thus, we now overview some examples that report measurements of dimensions of engagement introduced in section 1.2. Then, we propose some materials that will help us measure engagement during the works described in this thesis.

2.5.1 Measuring behavioral engagement

Literature works usually measure behavioral engagement from observations of human actions toward the robot or task. The most straightforward measure is whether the robot successfully made humans act as planned. Notable examples are the works of Satake and colleagues [11], [12] where

the robot's goal was to open the interaction with transients. Thus, their measure was the success rate of robot approaches and the failure rate per engagement failure type (as depicted in Figure. 1.2, Chapter 1). Shi et al.'s [63] goal was to make their robot distribute as many flyers as possible, thus measuring the flyer distribution efficiency. However, although having the person perform the task is essential for the robot's objectives, a simple counter may not be a rich representation of the quality of the encounter. Other works take into account more human behaviors to address this issue. For example, during a robot acceptability study in an elderly care center, Khosla et al. [64] assessed behavioral engagement by measuring the frequency and duration of gestures, people holding the robot, petting it, gaze at the robot, attempted speech, among others. Likewise, other works also consider some of these behaviors either for post-experiment assessment [65], [66] or real-time evaluation of engagement [67]. Finally, other works also used self-reports through questionnaires and interviews with tailored questions for their experiments. In an education HRI experiment, researchers asked students whether they worked on the robotics activities outside of class and if they went beyond class requirements [68]. Past works have also used probes about intentions to perform a task in the future [69].

2.5.2 Measuring emotional engagement

Video analysis was also a tool to assess positive and negative emotional engagement, extracting facial expressions as measures (emotions) as well as audio features (laugher, yelling, voice shaking) [64], [70]. Although these are the most common measures of emotional engagement in our literature search and contain rich information, it is not always possible to perform video analysis in all experimental scenarios. While in some cases, a moving robot's onboard cameras may have too much noise and not always capture the participant, it is difficult to capture a moving participant with fixed cameras or have appropriate angles in others. An alternative may be the usage of self-reported measures related to emotional responses, affection, and anxiety of robots, similar to what researchers do in education research (attitudes toward school and teachers) [71], [72]. Since the HRI literature is scarce on scales that explicitly measure emotional feedback, we propose to use scales previously used in HRI research that measure constructs that fit the definition of emotional engagement stated in section 1.2. For instance, the Godspeed series [73] is a widely used questionnaire in HRI experiments that measures likeability, perceived intelligence, and perceived safety, among other constructs. More recently, the Robotic Social Attributes Scale (RoSAS) [74] was proposed as a tool to measure perceived warmth, competence, and discomfort. The Inclusion of Other in the Self (IOS) [75] is a Likert-like scale that measures closeness was also widely used in past interaction studies [76].

2.5.3 Measuring cognitive engagement

Cognitive engagement is task-related and more challenging to measure. While we can assess behavioral and emotional engagement with a social actor or a task through signals and self-reports (although with limitations), evaluating whether people's cognition efforts go beyond minimal work

is less tangible. Likewise, there are self-report-based methods described in education literature that probe cognitive engagement through interview questions about a task, mainly on education studies [68]. Others use either task analysis or automatic methods. For instance, Miller [77] discusses how reading times and eye gaze features could be used as indicators of cognitive engagement. Atapattu and colleagues [78] developed an automatic detector that uses Doc2vec embeddings of class documents and student discussions. Their method assumes that the greater the cosine distance between both embeddings, the more a student contributed to the topic discussion with additional content instead of strictly following the given texts.

3

A SURVEY ON HUMAN-ROBOT FIRST ENCOUNTERS

We now make a thorough revision over methods related to engagement for mobile robots during first encounters. While performing our literature review, we followed the guidelines of Webster and Watson [79] and vom Broke and colleagues [80], [81].

This Chapter follows a top-down organization. First, we address works that focus on mobile social robots that engage with people in possible first encounters. We describe each engagement model, compare their social script to Kendon's greeting model, and summarize their reported engagement success according to each work metric. Then, the following section focuses on atomic social skills, splitting them into three subsections related to the components of Greenspan's model of social awareness, as defined in section 3.1. Section 3.2 analyses methodologies concerning the social sensitivity component, i.e., those that perceive social cues and context. Then, section 3.3 addresses the social insight component, focusing on works related to modeling social interaction and norms. Finally, section 3.4 evaluates papers under the communication component, focusing on those related to non-verbal communication skills and strategies.

This Chapter extends parts of our previous published work:

J. Avelino, L. Garcia-Marques, R. Ventura, and A. Bernardino, "Break the ice: A survey on socially aware engagement for human–robot first encounters", *International Journal of Social Robotics*, vol. 13, no. 8, pp. 1851–1877, Jan. 2021, ISSN: 1875-4805. DOI: 10.1007/s12369-020-00720-2.

3.1 How robots engage with people

Mobile human-robot engagement is a hot topic in robotic's research. However, even though many research works address parts of this problem, a significant amount solely focuses on socially-aware robot trajectories [82]–[85]. Albeit an important factor for comfortable HRI, the robot needs to express and understand additional social signals to improve its chances of interacting with someone, otherwise risking the failures described by Satake and colleagues [11], [12] (illustrated in Figure 1.2

in Chapter 1). Literature works study several approach models and behaviors to improve the way robots engage with people. The following subsections review and relate literature works where mobile robots engage with people, sparating them by their main goals.

A summary of how past works' distinct taxonomies relate to Kendon's model is depicted in Table 3.1. The results related to the measures of engagement (2.5) are shown in Table 3.2.

3.1.1 Empiric models to reduce engagement failures

Upon facing the engagement failures, Satake et al. [11], [12] designed a system to moderate them. They split the engagement attempt into a three-step model: (i) Finding an interaction target; (ii) Interaction at a public distance; and (iii) Initiating a conversation at a social distance. During step (i), the Robovie robot estimates whether people are busy and if it can reach them, thus attempting to reduce "Unreachable" and "Unwilling" error types. In step (ii), the robot performs a frontal approach to people by computing an interception pose and gazing at the target. The last step (iii) consists of attempting the initiate the interaction with the targets by reacting to their trajectories. If the person is passing through the robot, it reorients its body to reinforce its intention to interact, giving up if the person leaves. The robot continually keeps its body oriented to people if they stop, indicating its desire to interact, and starts the conversation. This method had a success rate of 55.9%, while a naive approach only succeeded 35.1% of the time. This result encompasses a significant decrease of unreachable, unaware, and unsure failures. From this description, we can relate step (i) to Kendon's Initiation of the Approach, step (ii) to Kendon's Approach (but without gaze aversion), and step (iii) to both Final Approach and Close Salutation phases. Shi et al. [63] used a similar three-step model to make Robovie efficiently distribute flyers among passersby. Instead of saluting people, however, the robot extended its arm with a pamphlet and made a verbal offer. This model had a flyer-acceptance rate of 18%, while a human only had managed 10%.

3.1.2 Openness for interaction

Detecting if someone is open for interaction can reduce the chance of rejection when approaching people. This insight guided how Brščić et al. [86] and Kato and colleagues [87] selected people for interaction. Brščić et al.'s goal was to help people that seemed lost. Their model has the (i) Wait & observe, (ii) Approach, and (iii) "Guidance service" phases. During phase (i), the robot uses a classifier to detect if people's trajectories are atypical, a signal that they may be lost and need help. Thus, people would be more receptive to the robot. After selecting the target, the robot enters phase (ii), approaching the target while gazing. In phase (iii), the robot verbally greets the person and offers guidance. Considering Kendon's model, phase (i) corresponds to the Initiation of Approach, phase (ii) to both Approach (without gaze aversion) and Final Approach, and phase (iii) to the Close Salutation. This model achieved a success rate of 87.5%. Kato and colleagues' objective was to have Robovie help people in a store. To design the robot's approach mode, the researchers observed how the store's staff behaved, designing a three-states state machine. State transitions respond to the results of a customer intention estimation algorithm, an SVM that classifies people's

trajectories/poses as "intention to interact", "other distinctive intention", or "uncertain". While the customer's label is "other distinctive intention" the robot remains in the "Free" state, where it remains idle. If the customer's intention is classified as "uncertain", Robovie enters the "Proactively-waiting" state. In this state, the robot turns its body and gazes towards the person but does not approach. When Robovie is sure that the customer intends to interact, it enters the "Collaboratively-initiating" state, approaching the visitor until achieving a distance of less than 1.5m, verbally offering help in the meanwhile. They compared the performance of this model against a model where the robot approaches everyone (Simply-proactive) and another where the robot passively waits for someone to approach it (Passive). The performance is the division of successful approaches by the sum of all attempted approaches and situations where people approached the robot without a response from it. The proposed model achieved the best results (87.2%), followed by the Passive approach (62.9%), and the Simply-proactive model (42.7%).

Yousuf et al. [88] studied how their Robovie RV3 guide robot should approach a group of people to explain an exhibit at a museum. Their system uses proxemics theory and F-Formations to reason how the robot should navigate while respecting social norms. Their approach has three phases. In the first phase, the robot assessed whether people were looking at it or not. If they were, the robot looked back at them and oriented its body toward them, going to phase two. Otherwise, if people were looking at the exhibit, the robot would directly go to phase two. In phase two, the robot moved directly toward the visitors if they were looking at them or toward their visual field of attention if they were looking at the exhibit. After being between 1.0 m to 1.3 m to people, the robot greets them while gazing at them and offers its services. Afterward, the robot moves to the exhibit site and starts the presentation. Reported human questionnaire responses reveal that participants prefer the proposed system over one that approaches the exhibit to explain it, regardless of people's attention.

3.1.3 Getting people's attention

Getting the person's attention and expressing the robot's intention to interact are fundamental skills for engaging robots. An in-the-wild study from Saad et al. [89] used the Pepper robot at a building's entrance performing three levels of enthusiasm: mild (wave), moderate (wave & speech), and high (wave & speech & slight approach movement of 0.3m). The robot selects a previously unselected human that has the robot in the visual field of attention and then behaves according to the level of enthusiasm of the condition. We can draw a parallel between the wave & speech behaviors and Kendon's Distance Salutation. The slight approach movement can be seen as the Approach phase without the gaze aversion since the robot is not close to the person. Their primary metric was the attentiveness score, where a researcher labeled whether people looked at the robot (1) or not (0). There were significant differences, with average scores of 0.84, 0.77, 0.95 for mild, moderate, and high enthusiasms, respectively. Although not used in [89], getting the target's attention through gaze behaviors is the most common strategy in the analyzed works [63], [86], [87], [90]–[92]. In addition, gaze also signals to people that they have robots' attention. For instance, the Pepper robot used in the MuMMER project [91], [92] had the role of direction giver, and although it did not approach

people, it gazed back at nearby people that looked at it and greeted them. This behavior is a Close Salutation without a salutation gesture.

3.1.4 Effects of progressive engagement

Zhao et al. [93] were more focused on the impacts of interaction progression. Their study consisted of a Wizard-Of-Oz experiment where their Xiaodu robot followed a three-stage model: (i) far-field, (ii) mid-field, and (iii) near-field. The robot transits between stages according to people's distances to it, with each one consisting of a distinct set of expressions and utterances. The robot's default behavior is an expressionless LED face. The robot uses a facial expression with raised eyebrows during the Far-Field stage (4.2m to 2.7m). When it enters the Mid-Field stage (2.7m to 1.2m), the robot has the "smiling eyes" expression and issues a verbal salutation ("Hi, how are you?). The robot enters the Near-Field stage when the person is closer than 1.2m, changes the facial expression to "smiling eyes with heart-shaped blush" and initiates dialogue, introducing itself and its goals. We find similarities between the Far-Field stage and Kendon's Initiation of Approach, Mid-Field and Final Approach, and Near-Field and Close Salutation. They automatically detected emotions and used self-reports on emotions and attitudes in a within-participant design study with five approach designs (four variations of the progressive approach and one reactive approach where the robot waits for attempted interaction) to measure engagement. Although they did not find significant differences for objective measures, "progressive approaches" had significantly better scores than "reactive approaches" in self-reported measures.

3.1.5 Explicit applications of Kendon's greeting

Up to our knowledge, Heenan et al. [90] were the only ones to implement a social script explicitly based on Kendon's greeting model before our works. They did so through a state-machine model implemented on the NAO robot. This model relied on a subset of social signals: (i) presence, (ii) head and body orientation, and (iii) location. These signals were extracted with fixed motion-tracking cameras and wearable markers. As soon as the robot detects the presence of someone, it attempts to make eye contact with the person. When the human looks at the robot, it enters the Distance Salutation phase, waving at the person. Afterward, the robot starts the Approach (with gaze aversion) and changes to the Final Approach (gazing at the person) when closer. If, at this point, the human keeps oriented and is at Close Salutation distance, the robot performs a handshake and verbally greets the target. However, the authors only performed a qualitative evaluation of this model.

3.1.6 Strengths and research Gaps

The literature has clearly identified the problem of robots engaging with people. Satake and colleagues' works [11], [12] provide an important categorization of failure types based on real-world experiments, whose contributions are twofold. First, it splits a complex problem into easier challenges. Second, it allows researchers to pinpoint which parts of the interaction are more susceptible

to failures and perform more detailed evaluations. Works related to Kanda and Ishiguro [11], [12], [63], [87] modeled their engagement algorithms based on human-human and human-robot in-the-wild interactions. Thus, their methods' results are highly ecologically plausible. We also highlight that literature models share substantial similarities with Kendon's greeting model. Moreover, they significantly improved people's engagement with robots, even though most systems relied on a limited number of detected social signals, which supports our thesis and research direction.

The first gap we identify is the lack of consensus in the taxonomies among the revised works. These works use distinct terminologies to describe their greeting models, making a direct comparison non-trivial. Nonetheless, a thorough analysis reveals that they are compliant with Kendon's greeting model, and thus, the taxonomy for greetings proposed in Chapter 2. As seen in Table 3.1, these models implement subsets of behaviors described in Kendon's foundational work, Heenan et al.'s being the one closer to the complete model and explicitly addressing it.

The literature also lacks a direct comparison between proposed models. Even though they present significant gains over more naive models, we lack a direct comparison between models in these works, and thus whether performance gains are worth the additional model complexity. Each model used a distinct context and distinct sets of metrics.

Although some models consider cases where people refuse to interact or give up on the interaction, they do not keep track of the greeting process measuring the uncertainty. In truth, the robots react to low-level signals and assume sequential models instead of being aware of the human greeting mental state. This knowledge is relevant for humans since failures to detect social cues may make them return to previous stages of the model or skip others.

Finally, most works highlight that it is challenging to detect complex social signals from a mobile robot's perspective, using people's body positions and orientations captured through sensors in the environment [11], [12], [63], [86], [87], [90]. Nonetheless, recent works [89], [92] fuse distinct signals relying solely upon onboard sensors, although the robots' navigation movements are constrained.

Table 3.1: Relationshionship between literature mobile robot engagement models and Kendon's greeting model.

Reference	Stage of Kendon's model					
	1) Sighting, orientation, and initiation of the approach (IA)	2) The distance salutation (DS)	3) The head dip (HD)	4) Approach (APP)	5) Final Approach (FA)	6) Close Salutation (CS)
Satake et al. [11], [12]	1) **Finding an interaction target:** Select reachable & anticipate willingness to interact. No gesture.	-	-	2) **Interaction at a public distance:** Frontal approach	**3) Initiating a converstation at a social distance:** Nonverbal intention to interact. Recognize acknowledgement	Greet people verbally
Shi et al. [63]	Compute approach utility to select target. Gaze at target.	-	-	Frontal approach target. Continuous gaze.	Reduce velocity with distance. Extend arm. Gaze. Verbally offer flyer.	-
Brščić et al. [86]	**Wait and observe**: Gaze around. Select target. Gaze at person.	-	-	**Approach:** gaze and move toward person.		**Guidance service:** Verbal greeting. Offer guidance.
Kato et al. [87]	**Proactively waiting:** body and gaze oriented at target.	-	-	**Collaboratively-initiating:** move toward person and offer help just before stopping.		-
Saad et al. [89]	**1) Select a target:** select a target that is not engaged.	**2) Draw attention (part 1):** Wave & verbal greeting.	-	**2) Draw attention (part 2):** Small approach movement (0.3 m).	-	-
Foster et al. [91], [92]	Select user paying attention. Gaze.	-	-	N.A. (Human approaches robot)		Gaze & verbal greeting.
Zhao et al. [93] (WoZ. Robot reacts to human approaching)	**Far field:** Raised eyes (facial expresion).	-	-	N.A. (Human approaches robot)	**Mid field:** Smiling eyes. Voice greeting.	**Near field:** Smiling eyes & blush. Voice intro.
Yousuf et al. [88]	1) If people look at the robot: turn robot's face and body toward them, and go to phase 2). If people are looking at exhibit, go directly to phase 2.	-	-	2.a) If people are looking at the robot: move directly toward visitors (stop 1.0 m to 1.3 m distance). If people are looking at exhibit: approach visitors, positioning the robot on their field of view		2.b) Gaze & verbal greeting. Offer explanation about exhibit.
Heenan et al. [90]	**IA:** Idle behaviors. Detect person. Attempt eye contact.	**DS:** Sand. Gaze at person. Wave.	-	**APP:** Avoid eye contact. Move to personal space and then gaze at person.		**CS:** Handshake & gaze & vocal greeting.

3.2 Social sensitivity

We now make a literature analysis of methods that make the robot aware of the social environment: the social sensitivity component of Greenspan's model. First, we address full architectures that detect low-level social signals, like people, objects, poses, and facial expressions (subsection 3.2.1). Then, we focus on the visual field of attention detection (subsection 3.2.2). Finally, subsection 3.2.3 deals with the ability to understand others' feelings and viewpoints (role-taking), in which we focus on the detection of social feedback.

3.2.1 Social context detection

Zaraki et al. [94] developed a system to detect and track a significant set of signals for dialogue-based HRI with a robotic head. Their system uses RGB-D, RGB, illuminance, sound level, and temperature sensors to keep track of the social scene. Through those low-level features, the system performs (i) facial analysis, (ii) identity assignment, (iii) body analysis, and (iv) saliency detection. Facial analysis detects face positions and eye/nose/mouth landmarks, which are used to classify people's gender and estimate their age and facial expressions. QR-codes identify people, and skeleton tracking for states/gestures uses Kinect's skeleton tracking library. In addition, their system also detects image regions that attract the human gaze (saliency). The system produces a meta-scene file that compiles all this information, which can be used for HRI algorithms. This perception pipeline becomes part of a cognitive architecture that controls a robot's head and face in their follow-up work [95].

Triebel et al. [96] describe the SPENCER project's architecture of an airport guide robot. It uses laser and RGB-D data to detect and track people, objects, and groups and map and localize itself in dynamic environments. In addition, it identifies the spokesperson to guide a group of people to their desired destination within the airport, formulating the problem as a Mixed Observability Markov Decision Process.

On [82], Truong and colleagues propose a social navigation algorithm based on proxemics theory, using an RGB-D camera. The extracted features consist of: (i) state (standing, sitting, or moving), (ii) walking velocities, (iii) field of view, (iv) interactions with marked objects, and (v) social interactions. With this data, the algorithm creates a costmap for robot navigation.

Foster et al. [91] developed a system to infer the social context around the Pepper robot through audio-visual sensing for the MuMMER project. From RGB data, they extract 2D skeleton poses using convolutional pose machines (Openpose) [97]. OpenHeadPose [98] uses low-level features extracted from Openpose to estimate people's head pose. An algorithm proposed by Sheiki and Odevez [99] detects the visual field of attention. In parallel, they also use OpenFace's [100] features, along with colors and head poses for face tracking. They feed data captured through a microphone array to a neural network proposed by He et al. [101] that performs speech/non-speech detection and voice localization. Finally, the fuse visual and audio location estimates to detect who is speaking with the robot.

Table 3.2: Engagement results (if available), context, and experiment types for each covered paper.

Reference	Experiment context	Reported results		
		Behavioral Engagement	Emotional Engagement	Cognitive Engagement
Satake et al. [11], [12]	A robot approaches people at a shopping mall to advertise a coffee shop.	**1) Full engagement success:** Naive - 35.1% vs 55.9% - Proposed **2) Unreachable failres:** Naive - 25% vs 3% - Proposed **3) Unaware failures:** Naive - 14% vs 4% - Proposed **4) Unsure failures:** Naive - 24% vs 18% - Proposed **5) Rejections:** Naive - 29% vs 27% - Proposed	-	-
Shi et al. [63]	A robot distributes flyers at a shopping mall.	**Accepted flyer ratio:** Robot - 18% vs 10% Human	-	-
Brščić et al. [86]	A robot approaches people that are seamingly lost to guide them.	**Unreachable:** 29.2% **People reactions when engaged:** Verbally interacted - 54.8%, Listened but did not talk - 32.7%, Left or ignored: 12.5%	-	-
Kato et al. [87]	A robot helps customers looking for help in a store.	**Performance** ($\frac{success}{success+miss+failure}$)**:** Proposed - 87% vs Passive - 62.9% vs Proactive - 42.7%	-	-
Saad et al. [89]	A robot at a building's entrance attempts to attract people's attention. Limited navigation (0.3 m)	**Attentiveness score:** Mild condition - 0.84 vs Moderate condition - 0.77 vs High condition: 0.95.	-	-
Foster et al. [91], [92]	A robot interacting at a shopping center (limited navigation, inside a box). Services: chat, quiz, route description, and route guidance (pointing).	-	-	-
Zhao et al. [93]	Wizard of Oz experiment. A human approaches the robot and it reacts according to the human distance.	-	**Objective measures:** N.S. differences in detected neutral, happy, surprised, and negative emotions. **Questionnaires:** Sig. differences in reported happiness, surprise, and confusin. N.S. in disgust and neutral emotions.	-
Yousuf et al. [88]	A robot approaches museum visitors to explain an exhibit. Controlled study where students pretended to be museum visitors.	-	**Likert-like self-reports (1-7 scale):** 1) Robot greet was proper: Proposed - 5.7 vs 4.4 - Conventional 2) Robot attended adequately: Proposed - 5.2 vs 4.5 - Conventional 3) Overall evaluation: Proposed - 5.3 vs 4.5 - Conventional	-
Heenan et al. [90]	Laboratory experiment where a robot attempts to greet a human.	-	-	-

3.2.2 Gaze and Visual Field of Attention Detection

As highlighted in Chapter 2, gaze patterns are major features both when opening and keeping the interaction with someone. Albeit humans are exceptionally accurate in gaze estimation, robots still struggle with this task. It is, thus, a hot topic in research. OpenFace [102] is an open-source face analysis framework that estimates head poses, facial features, and gaze. Their gaze estimation method is called eye-CLNF (Constrained Local Neural Field) and was trained on a synthetic dataset of photo-realistic renders of human eyes. This algorithm achieves accurate results ($\mu_{error} = 11.12°$) if the eye's image has sufficient resolution (working for crops of at least 20px) but fails when people wear glasses, their eyelids occlude the eye. The algorithm will not work if people are too far away from the camera due to low resolution. Additionally, OpenFace fails to estimate people's heads if they wear masks, undermining gaze estimation.

Head orientation alone can be a very informative proxy for gazer direction when we assume that people's eyes are centered. OpenHeadPose [98] uses OpenFace's head keypoints and feature maps to estimate the full head pose. The network consists of a Convolutional Neural Network with three layers and a fully connected output layer. This method works even when people's faces are partially occluded, as in the case of people wearing masks, and can estimate people's head poses even when they are far away from the image.

Kellnhofer, Recasens et al. [103] proposed an end-to-end for unconstrained gaze estimation from 2D images called *Gaze360*. First, the authors created a rich dataset of people looking at a moving target with known positions. Afterward, they use head crops of the past seven frames fed to a backbone network (ImageNet pre-trained ResNet-18) that extracts features of dimensionality 256. These are fed to a bidirectional LSTM with two layers and a fully connected layer that outputs the spherical coordinates relative to the camera and error quantile. The algorithm needs a pre-processing algorithm to crop centered heads for in-the-wild applications. In the author's default implementation, they use DensePose [104], but other authors [104] have also used OpenPose [97] successfully. *Gaze360* can estimate gaze directions even with partially occluded faces (for instance, with masks), although its in-the-wild implementation depends on the capabilities of the face-cropping algorithm.

Other methods do not focus on gaze direction directly. In truth, the gaze direction is a proxy for the Visual Field of Attention, which is probably a more important cue to reason about interactions and activities. Its estimation was the goal of Masse and colleagues [105]. They formulated the problem with a probabilistic approach, defining target locations (objects or heads) and head orientations as observed random variables and Visual Field of Attention and gaze directions as latent random variables. Later, they extended their method to predict the VFOA when targets are outside of the image [106].

3.2.3 Role-taking

According to Greenspan's model, the role-taking component consists of skills that enable an individual to understand the feelings of others. Here, we relate these habilities to the capacity of perceiving social feedback. Feedback about one's actions closes the interaction loop and regulates

people's actions. These allow people to develop relationships that are fundamental to maintaining social order [57]. Some authors consider that understanding signals of praise and disapproval is a crucial skill to master social intelligence [107]. Although humans naturally exchange social feedback, the robotics literature is still scarce on complex methods to detect and learn from it. Here, we analyze a set of works where robots had to be aware of human implicit and explicit feedback. In the following paragraphs, we organize and present these works according to their application of social feedback. First, we start with research intended to estimate when it is appropriate to attempt interaction. Then, we follow with social feedback-mediated robot approach behaviors. Next, we focus on research on using social feedback to keep people engaged. And finally, we survey works where the goal is to make robots self-aware of interaction failures.

3.2.3.A Interaction appropriateness

Understanding the right time and environment to initiate interactions is the first step to avoiding uncomfortable or failure to start interactions. However, the literature on this topic is scarce at this time of writing. The study by Nigam and Riek [108] acknowledges receiving feedback related to the appropriateness of a robot interruption through a button. Qureshi et al.'s study [109] also falls under this category as the goal was to have a robot learn when to use appropriate interaction skills given grayscale and depth image channels.

3.2.3.B Mediation of robot approach behaviors

Even if it is the right time and place to engage with someone, it is fundamental to keep analyzing people's responses to mediate how robots approach them. The work by Mitsunaga et al. [110] mediates the robot's behaviors, including proxemics, gaze meeting ratio, motion speed, and waiting time, based on natural signals using a Policy Gradient Reinforcement Learning (PGRL) method. Similarly, McQuillin et al. [111] present an algorithm that involves socially acceptable positioning and speed using Deep Reinforcement Learning methods. The robot uses explicit and implicit feedback as a reward, with explicit feedback significantly improving perceived appropriateness and sociability. Explicit feedback relied on a keyword spotting scheme that associates a reward value with positive, neutral, and negative feelings. As for implicit feedback, they use valence values associated with detected facial expressions.

3.2.3.C Keeping people engaged

After starting the interaction, it is essential to keep people engaged to allow the robot to achieve its social objectives. Ritschel and colleagues [112] propose a method of keeping the user engaged during the interaction by adapting the robot's personality through distinct language behaviors. The robot gauges users' engagement with a Dynamic Bayesian Network and adjusts its level of extroversion in response. It makes use of uses found gaze features and head movements. Similarly, Tsiakas et al.'s research [113] involves a robot adapting its speech behaviors based on users' scores and an Inverse Reinforcement Learning method based on EEG signals to keep users interested in a game.

Ahmad's Ph.D. thesis [114] also involves a reinforcement learning-based algorithm to set a robot's personality during a game interaction with children, thereby keeping them engaged. He used a reward signal that matches our definition of social feedback, resulting from eye-gaze toward the robot, facial expressions, verbal responses, and simple gestures.

3.2.3.D Classification of robot failures

During every interaction phase, the robot is error-prone due to sensor, actuator, and software limitations. And since these limitations can potentially harm people, robots need to be aware of their failures from people's feedback. To our knowledge, Trung and colleagues[115] pioneered works in this category, testing distinct combinations of classifiers to categorize reactions to robot failures. They tested distinct combinations of classifiers fed with 3D coordinates of head, shoulders, and neck (data collected in [116]) to classify reactions to robot failures as glstfs or Social Norm Violation (SNV)s. Even though they achieved results over 90 %, the test set contained people used to train their algorithms Kontogiorgos et al. designed a task where a robot generated failures when instructing users to cook non-trivial recipes, using a random forest classifier to classify segments of videos as failures or no failures based on head movements, gaze, and speech features.

3.2.4 Strenghts and gaps

The literature describes impressive perception pipelines that can gather a significant amount of social signals, like the work of Zaraki et al. [94], Lazzeri et al. [95], and Triebel et al. [96]. We can see most of the signals on Table 2.1 present in previous pipelines, although not simultaneously and usually tailored for specific applications. These pipelines are very relevant for our work since, besides providing some of their implementations as open-source code, they also show us how we can integrate these skills. However, these works are not suitable for all robots since they require some sensors that can either be expensive or have excessive dimensions, like RGB-D, LIDAR, temperature sensors, or need users to use extra items like QR codes. Vision-based (RGB) are usually cheaper and appliable even to short robots. However, robots using only RGB images are either in a fixed position or have their movements constrained since spatial awareness is more difficult to achieve, and cameras can be susceptible to motion blur. They need to be adapted to moving robots in order to apply them to our problem. Additionally, individual social sensitivity skills still suffer from high computational requirements and accuracy issues. Thus, efficient perception pipelines should leverage shared low-level perception features to avoid unnecessary recomputations when detecting higher-level social cues. We used this knowledge to propose our perception pipeline of chapter 5.

Regarding Gaze and Visual Field of Attention Detection, there are also accurate methods described in the literature, from which we highlight *OpenFace* [102] and *Gaze360* [103]. Additionally, even though *OpenHeadPose* [98] only considers the head pose, it can be a simple (yet reliable) proxy for gaze most of the time. The open-source implementation of these methods makes them highly appealing for our studies in both Chapter 5 and Chapter 6. On the downside, *OpenFace* will not work for distances farther away than 2m nor with masked people, making it inviable for our problem.

Gaze360 and *OpenHeadPose* work under these conditions but can be exceedingly computationally expensive for a mobile social robot. However, we note that this fact is due to the base algorithms needed either to extract low-level features (*OpenPose* heatmaps or keypoints for *OpenHeadPose*) or crop people's faces (for *gaze360*) since their models are actually quite small (specially OpenHeadPose). Due to motion blur and occlusions, these models are prone to estimation errors for mobile social robots, needing methods to mitigate it. Possible approaches comprise the use of body structure knowledge to infer missing data (proposed in Chapter 5) or the visual suppression method to discard unreliable data [117].

Although the role-taking dimension of social sensitivity is still an underexplored topic, the existing body of knowledge already has very relevant information. First, there is an already established taxonomy related to robot failures [118] and evidence that people show some of these signals through facial expressions, head movements, and some gestures [119]–[121]. Additionally, some researchers [121] found relationships between Facial Action Units (FAU)s and social feedback related to robot errors, which is valuable information to develop an automatic detector. However, this is still an open problem that needs additional context since attitudes to norm violations can be ambiguous, with SNVs being more challenging to detect than Technical Failure (TF)s. For instance, they may express laughter as a response to error situations and norm compliant robot behavior. We also notice that the data used to train these models needs to be ecologically plausible for a robot to be able to receive feedback in the wild. Moreover, there is no relationship between human reactions and measurable quantities (either self-reported scales or physiological data) [122]. Finally, we claim that using raw hand sensor signals to detect social feedback as a successful handshake is insufficient. Outside the laboratory, robots are surrounded by objects that they can mistakenly grab during false detections, which would give the robot wrong feedback. To better disambiguate these scenarios, we believe that social robots should be able to identify these situations through an analysis of haptic feedback. In Chapter 6, we contribute to this topic with two modalities: an automatic method that estimates robot failures using human facial feedback and the robot's actions as context, and a method that uses the robot's tactile sensors to detect whether handshakes succeeded or not.

3.3 Social insight

Robots need to reason about the social signals they perceive to act according to people's expectations. The collection of skills for this task corresponds to Greenspan's social insight component. It encompasses knowledge of social norms, scripts, and models. Our literature survey [21] splits works into two categories: (i) those that implicitly encode social knowledge and those that explicitly do it through social norms. Our research path focuses on methods associated with the second category, for which we are now compiling a list of related works.

3.3.1 Explicitly defined social comprehension

According to our survey, the literature, however, is still scarce in methods that explicitly define social comprehension. Carlucci et al. [123] developed a framework that explicitly encodes social

rules execution into Petri Nets, generating a Petri Net Plan that respects a set of previously defined social norms. They propose a formal definition of social norms for robots to make it possible.

Porfirio and colleagues [124] developed an interface to design interactions with a verification algorithm that tests whether human-designed interaction scripts respect a set of previously encoded social norms. Interactions follow a state-machine-like formulation with a (transition-system) and use Linear Temporal Logic (LTL). Detected human actions fire state transitions while social norms are manually encoded in LTL.

3.3.2 Strengths & gaps

A social robot that follows a human-centered design should be able to perceive and incorporate social norms explicitly, making its behaviors easier more explainable to humans. Thus, interpretable models like Carlucci et al.'s [123] and Porfirio et al.'s [124] may provide stronger safety guarantees. However, these do not learn from the data or demonstrations, thus requiring a human expert to design the interaction.

We believe that because humans can make sense of social norms by observing them or having them explained, this would be a more natural approach for robots. However, this is still an open issue. In our work, we propose a method for modeling how humans begin their interactions inspired by social sciences literature (Kendon's greeting model). Although the proposed model explicitly implements Kendon's greeting model stages, it is also tuned using a data-driven method. This topic is covered in Chapter 5.

3.4 Communication

With the detected social context (section 3.2) together with social insight (section 3.3) agents can have an understanding of the social situation. However, they need the appropriate set of communication skills to act accordingly. This section focuses on literature works related to implementations of non-verbal skills to communicate one's intentions and feelings (subsection 3.4.1 - referential communication).

3.4.1 Referential communication

Non-verbal communication skills are crucial to initiate a successful interaction since robots need to make their intentions clear to people, otherwise risking interaction failures. Social robots need to execute these skills on time to avoid losing people's interest or missing their attention. Shi et al.'s [63] work on flyer distributing robots supports this idea since their robot's success was heavily reliant on the timing of its flyer handing gesture. Their best strategy was to extend the robot's arm while gazing at the target person.

Proper execution of salutations is fundamental to clearly state the robot's intentions to initiate focused interactions with people in a socially acceptable way. As such, they should be able to perform the most common salutations, like the handshake. The handshake was the most common salutation gesture in western civilization. Most literature studies focus on the shaking movement of the

robot's arm, which is a complex problem. For example, Mitsuri et al. [125] profiled the velocity of human wrists during handshake request and response conditions and modeled a transfer function to generate those movements in a robotic arm. They further adapted their model for small-sized robot arms [126]. The joint control of gaze and arm movements for robots requesting a handshake is also a relevant aspect addressed in another of their works [127]. Later, Ota and colleagues [128], [129] focused on the timings and time intervals between the start of a handshake request and the corresponding response. More recently, Mura et al. [130] designed a human-robot handshake controller for their FRANKA robot arm. Their principal focus on this work was handshake stiffness and synchronization. For handshake stiffness and hand closure control, they used pressure sensors from a custom silicon glove, and they used an Extended Kalman Filter to fit human handshake sinusoidal motion parameters. People positively evaluated the handshake. Moreover, they also reported distinct perceived personality qualities when interacting with different motion controllers.

Handshakes, however, are only one of the possible non-verbal communication gestures to convey one's intention to interact and officially close the greeting protocol. While an option would be to study and manually design behaviors for all cultures, another possibility is to follow a research path where robots learn these behaviors from observation. For instance, Ben Amor et al. [131] proposed two imitation learning algorithms to learn human gestures: (i) Probabilistic Principal Component Analysis-Interaction model, and (ii) Path Map-Interaction Model. They captured data from two human subjects with a motion capture system to train both algorithms. In one of their subsequent works [132], they propose a "learning from demonstrations" method called Interaction Primitives, which learns the dependency between two agents' actions. The algorithm predicts the best motor control signal matching observed gestures through the learned relationship. Motion capture systems, however, can be quite expensive and time-consuming. A more affordable and more natural solution would be to learn these behaviors directly from robots' cheap sensors. A method proposed by Shu et al. [133] uses RGB-D data of HHI to learn socially-compliant action possibilities, which they define as *social affordances*. Their Baxter robot managed to learn handshakes, hand waves, high fives, how to pull a cup up, and how to hand it over.

3.4.2 Strengths & gaps

We found several papers on human-robot handshakes with significant improvements in this multimodal salutation gesture. According to their reports, these methods give robots compliant and comfortable handshakes. We miss, however, studies of the impact of this non-verbal behavior on people's attitudes towards robots. Additionally, since distinct robots have distinct arm and hand designs, they need customized handshake controllers.

Interaction primitives are an interesting concept to synchronize single HRI gestures, having been applied even in human-robot hugs, as shown in the work of Campbell and Yamane [134]. However, during the first encounter between people and a robot, robots may not know which salutation is appropriate since, in a multicultural society, location priors lose their strength. Thus, a socially aware robot should update its belief given real-time observations of people's movements.

Our contributions to social robot's referential communication consist of studies related to the handshake salutation (Chapters 4 and 8).

Part I

Methods

4

HANDSHAKE: DESIGN OF A PHYSICAL CLOSE SALUTATION

As described in Chapter 2, during the close salutation phase, people use a salutation gesture to acknowledge each other's intent to open the interaction. The handshake is a physical salutation widely used in western civilizations. It is a challenging task of Physical Human–Robot Interaction (pHRI) that requires a firm yet comfortable and safe grasp of the human's hand. As reviewed in Chapter 3, section 3.4, robotic handshakes needed to be customized for each hand design since different hands result in distinct pressure points that may harm users. Thus, this Chapter describes our human-centered design approach of a handshake for the Vizzy robot, focusing on comfortable hand-grasps. We start this Chapter with a description of Vizzy and its hand design. Then, we investigate human's preferred hand grips with two experiments with users in the loop, from which we collect grasp finger positions and sensor forces. We use these data to create two handshake approaches: (i) an encoder position-based controller and (ii) a force-based controller, which we compare in a final user study. Both the encoder and force controllers are implemented as a PID.

This Chapter is an extended version of the following paper:

J. Avelino, T. Paulino, C. Cardoso, R. Nunes, P. Moreno, and A. Bernardino, "Towards natural handshakes for social robots: Human-aware hand grasps using tactile sensors", *Paladyn, Journal of Behavioral Robotics*, vol. 9, no. 1, pp. 221–234, Aug. 2018. DOI: 10.1515/pjbr-2018-0017.

4.1 The Vizzy robot

In this thesis, we test our methods on the Vizzy robot. It is a mobile humanoid robot developed by the Institute for Systems and Robotics (ISR-Lisboa/IST). It heights 1.3m and has a differential drive mobile base and humanoid upper torso. The robot has 30 mechanical degrees of freedom (DOF), being able to perform arm/hand gestures, biologically inspired head/gaze movements [135], and navigate in planar surfaces. Vizzy has a friendly marsupial-like design with two four-finger hands (Figure 4.1). Vizzy is unique in the world.

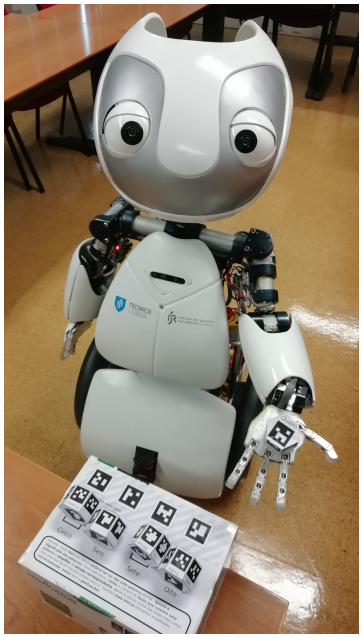


Figure 4.1: Vizzy: the mobile social robot used in this work. This robot was developed at the Institute for Systems and Robotics (ISR-Lisboa/IST).

The robot has a handful of actuators to interact with the world around it. Its head can perform pan and tilt movements, while its eyes can do tilt, vergence, and version movements (5 DOFs). Each one of Vizzy's arms has the following degrees of freedom:

- four shoulder DOF (scapula, flexion, abduction, and rotation),
- one elbow DOF (flexion), one forearm DOF (pronation),
- two wrist DOFs (abduction and flexion).

Vizzy runs the ROS and Yet Another Robot Platform (YARP) middlewares. These communicate with each other using the bridges developed by Aragão et al. [136]. YARP modules are responsible for low-level control of upper torso motors and leverage a large codebase previously developed for the iCub robot, with algorithms that produce biologically inspired gaze patterns and arm motors control. ROS is currently one of the most widely used middleware and is the main robotic software framework to integrate all remaining robot functionalities.

The robot has several characteristics that make it a good fit for research under our thesis topic. First, the robot's humanoid appearance was designed to elicit feelings of social presence during interaction with humans. Moreover, its size is a good compromise between making people feel safe while allowing comfortable standing interactions. Second, the robot's mobility allows it to approach and follow people, a crucial requirement for engagement with moving people. Finally, the robot's degrees of freedom make it possible to mimic the non-verbal human behaviors identified in Kendon's greeting model (section 2.2).

During this thesis, the robot received several skills. Among them are:

- Speech synthesis APIs and web interfaces [1]

[1] `https://github.com/vislab-tecnico-lisboa/vizzy_speech`

- Wizard-of-Oz plugins for motor control through RViz [2]
- Tactile sensors on the robot's fingers [3]
- Gestures for human-robot interaction
- A behavior tree package for planning and execution [4]
- The ability to detect and estimate people's 3D body/head poses [5] [6]
- The ability to approach people according to Kendon's greeting model [7] [8]

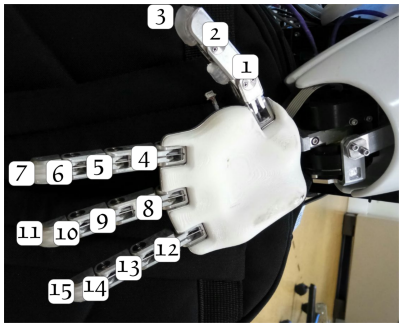
4.2 Robotic hand design

As described in section 4.1, the Vizzy robot has four-fingered hands with human-like palm and finger sizes, capable of grasping objects. Each finger has three joints actuated through a fishing line string attached to the finger's last joint and a pulley in the other end. Thus, fingers are underactuated since one pulley moves three joints. The thumb and index fingers share a single motor, while the middle and ring fingers have individual motors.

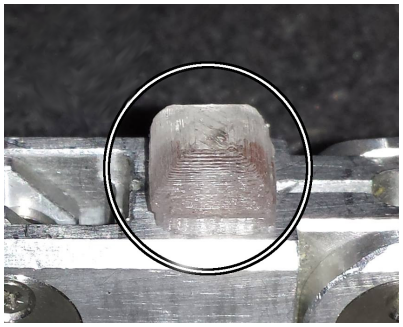
Hand fingers contain force sensors distributed as shown in Figure 4.2a. The thumb has three sensors, while each remaining finger has four sensors. These tactile sensors are composed of a soft elastomer body (Figure 4.2b with a small permanent magnet inside (Figure 4.2c). Below the magnet, there is a magnetic field sensing element (i.e., a Hall-effect sensor). Using a 3-axis Hall-effect sensor allows the measurement of the magnetic field variations in the three axes. When an external force applied to the elastomer changes the magnet's position, making the Hall-effect sensor detect a magnetic field variation relative to the default position, the force sensing system infers the magnitude and direction of the applied force in 3D. An air gap is left between the elastomer and the magnetic sensor to increase the force sensor's sensitivity. These sensors can detect minimum forces in the order of 10 mN. The work of Paulino et al. [137] describes these sensors in more detail. We note that since the sensors on the fourth finger are still in the test phase, we did not use them in our experiments.

The hand design criteria included: (i) Similarity to a human's hand size and (ii) the execution of two types of object grasping: cylindrical and power grasp. Since the hand design did not consider handshake actions, we performed a user human-robot handgrip study to evaluate the plausibility of that kind of interaction. Initially, the robot's hand palm was entirely metallic, but following user feedback on the pilot study (section 4.3), we improved it with a 3D printed plastic white palm cover (as seen in Figure 4.2a).

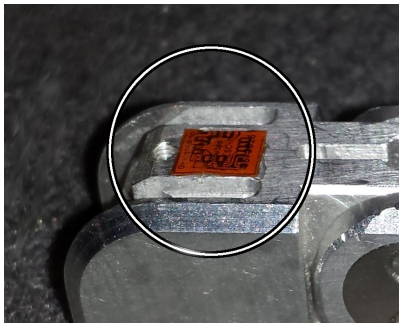
[2] https://github.com/vislab-tecnico-lisboa/vizzy/tree/master/vizzy_rviz_plugins
[3] https://github.com/vislab-tecnico-lisboa/vizzy_tactile_drivers
[4] https://github.com/vislab-tecnico-lisboa/vizzy_behavior_trees
[5] https://github.com/joao-avelino/3d_human_detection
[6] https://github.com/joao-avelino/skeleton-completer-train,https://github.com/joao-avelino/skeleton-completer
[7] https://github.com/vislab-tecnico-lisboa/kendon_greetings/tree/vizzy_real
[8] https://github.com/vislab-tecnico-lisboa/vizzy_approach



(a) The robot's hand with the indexes of the force. This hand version has a 3D printed palm cover for additional user comfort. Sensors 12 to 15 are still on testing phase and will only be used in future work.



(b) Force sensor covered by an elastomer that contains an embedded magnet. When a force is applied, the magnetic field changes.



(c) The Hall effect sensor embedded in the robot's finger measures the surrounding magnetic field, detecting changes resulting from the displacement of the magnet away from its default position.

Figure 4.2: Visual description of the Vizzy's hand and its force sensors.

4.3 Pilot study

The first step to designing a robotic handshake for the Vizzy robot was to assess whether its hands could fit this purpose. Thus, we performed a simple pilot study with users with two central goals. First, we intended to gather an initial coarse guess of preferred human-robot handshake forces. Second, we wanted to get user feedback on comfort, haptic, and mechanical properties. We also highlight that experimenting with users gives us important insights to enhance the experimental design of further experiments.

The population for this experiment consisted of 20 subjects, of which 13 were male and 7 were female. Their ages ranged from 20 to 51 years old ($\mu = 32.95$, $\sigma = 9.26$). These subjects were researchers and staff from the authors' research group with different nationalities and cultures.

4.3.1 Methodology

To prepare the experiment, we set the encoder values of three handgrips, manually controlling the position of each finger while a researcher grasped the robot's hand. We labeled each of these handgrips according to the force perceived force reported by the researcher: strong, medium, and weak.

We asked people to shake hands three times with the robot. The handshake started with an initial position of finger joints (controlled by three motors), followed by a timely closing of the fingers to the final configuration. Each final finger configuration corresponded to one of the handgrips set when preparing the experiment. After executing the three handshakes, we asked the participants to rank the handshake by their preference using three labels (bad, average, and good), mainly considering the strength that conveys an adequate handshake interaction. More specifically, the handgrip strength should be high enough to be engaged in the handshake, and at the same time low enough that did not make the person feel uncomfortable nor causes an injury. We counterbalanced the order of execution to avoid biases due to meetings between participants or expectations related to the sequence. Finally, we asked people's opinions about the handshake experience. We collect the tem-

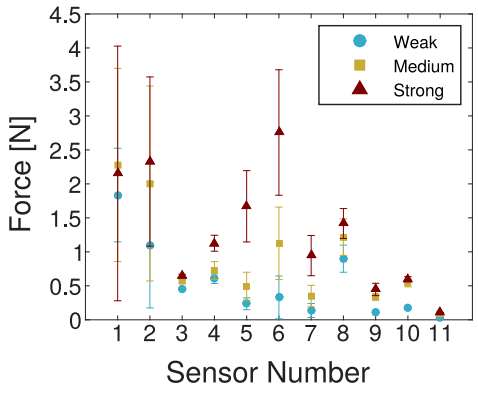
poral sequence of the magnetic flux, the force data from each tactile sensor during the handshake interactions, and encoder values. We note that only data from sensors 1 to 11 was available during these experiments since the remaining were still under calibration.

Analysis of handshake preferences based just on the final position of the fingers of the robot may not be meaningful because different people have different hand sizes. Since the same finger position produces a large variability of forces across different hand sizes, we analyzed the mean and variance of the measures by each force sensor using two different ways of grouping, using the original labels and using participant-provided preference level.

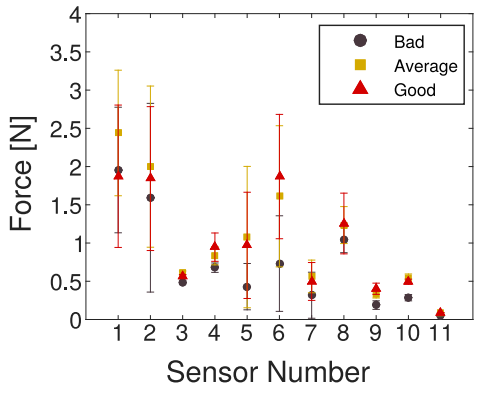
4.3.2 Quantitative measurements

Figure 4.3a depicts the statistics for the force magnitudes applied on each sensor. We can see that sensors 3 and 11 (fingertips), as well as sensors 9 and 10 (middle finger), make almost no contact with participants' hands, which is expectable for two reasons. First, the robot's hand is slightly larger than the average human's. Second, the under-actuated finger limbs move to different final configurations depending on the initial contact points. The most active sensors were 1, 2 and 6. Different human hand sizes may explain the variance of forces measured by these sensors.

We also present the total force magnitudes, corresponding to the sum of forces at the contact points of humans' hands and tactile sensors, in Table 4.1a. Our handgrip force values were lower than those reported in previous works, which we expected since we cannot measure forces in a large portion of the robot's hand - the hand palm. However, our goal was never to compute the total handgrip force but to study the most comfortable force distribution on the available sensors. Statistical information regarding this distribution is presented in Figure 4.3b. We note that people preferred similar forces on sensors 1, 2 and 6 that represent the main contact points of the thumb and pointer fingers. These preliminary results give us an idea of force distributions that can be used as the feedback signal for a handshake grip strength controller.



(a) Average and variance of the force measured in each sensor for each hand grip action (best seen in color).



(b) Average and variance of the force measured in each sensor according to the user preference (best seen in color).

Figure 4.3: Average force distributions.

The collected data suggest a non-significant tendency of female subjects to prefer a slightly larger

grip force than male subjects (see table 4.1a and table 4.1b). A possible explanation for this tendency could be that forces exerted by the three predefined handgrips are a very coarse measurement of force ranges. In addition, the exerted force depends on the combination of the characteristics of the human hand (shape and size) and the elasticity and compliance of the artificial tendons and sensors in the robot hand. Thus, two handgrips had low contact forces leading these participants to prefer the strongest one since there were no better options.

In this experiment, the strong handgrip was selected by 11 persons out of 20 (55 %) where the chance level was 33 %. A one-sample binomial test provides a p-value of 0.0552, showing the results are statically significant. Nonetheless, we would like to point out that the population in this pilot study might not represent the general population.

Table 4.1: Quantitative results related to forces measured with Vizzy's force sensors.

(a) Average forces of handshakes during the pilot experiment. Users graded three pre-defined finger configurations with the "Bad", "Average", and "Good" labels.

	Average force (N)		
	Bad	Average	Good
Female	6.79	7.96	11.41
Male	8.63	13.22	10.50
Total	7.76	11.33	10.81

(b) Preferred handshake over a predefined set of three hangrips with predefined finger positions. We referred to these grips as "Weak", "Medium", and "Strong".

	Preferred handshake (%)		
	Weak	Medium	Strong
Female	0.0	14.3	85.7
Male	15.4	46.1	38.5
Total	10.0	35.0	55.0

4.3.3 Participants' qualitative feedback

Participants reported that, despite the metallic hand, the silicon sensors provided a touch and grip more comfortable than they had expected. Their prior discomfort expectations were due to the robotic looks of the hand. However, they identified some issues with their experience. First, some people felt that the robot's thumb needed adjustments since it exerted significantly more pressure than the remaining fingers. People also reported that all fingers should close simultaneously and felt the absence of the shaking motion of a handshake. In addition, they suggested that the robot's hand palm should have the same tactile feeling as the fingers' sensors. Indeed, the inclusion of tactile sensors in the palm would increase the comfort of the handshake and simultaneously provide added perceptual information to exploit. Participants also noticed the absence of one finger and suggested that the robot should use it regardless of its ability to measure forces.

4.4 User-guided grasp design

The pilot study in section 4.3 highlighted several problems in both experimental and hand design. This experiment addressed some of those issues, attempting to gather more reliable force preferences

and improve the robot's hand for more comfortable handshakes. This experiment follows a co-design approach, where each participant customizes the handgrip.

The main objective of this study is to find the handgrip reference force distribution for control, using the feedback from the participant to set the position of the motors that provide better comfort and handgrip strength. Through customized handgrips, we should be able to collect a more accurate distribution of the forces over the sensors than through the method described in section 4.3. Additionally, we also looked forward to comments on the mechanical properties of the robot's hand from a wider population.

We performed these experiments during an exhibition on bachelor's admission day, an event that allowed us to get more participants. The experiment had a total of 35 participants, 28 male and 7 female, with ages between 12 to 49 ($\mu = 25.38$, $\sigma = 9.69$).

4.4.1 Methodology

Our first step was to improve the robot's hand by adding the missing finger and a palm cover to prevent contact between the human hand and the robot's metallic hand palm. In the absence of the possibility to create a silicon palm cover in due time, we 3D printed a white plastic palm cover instead. A silicon palm with additional new force sensors is currently under development in our laboratory.

Since having predefined hand grasps resulted in a coarse measurement of people's preferences and unbalanced force between the thumb finger and the remaining fingers, we opted for a co-design methodology in which people iteratively command the robot's motors. Following this perspective should increase the overall comfort of people and provide a better fit and hand contact than the previous method. We expected a greater average total force because the finger's positions would fit better with each participant's hand.

The experiment began with the human subject receiving a standard handgrip where the fingers' positions were the average position of handgrips with the "good" label in the pilot experiment. Then this handgrip was customized by subjects by tightening or loosening the robot's fingers through adjustments of the respective motors. This process occurred through the use of a Graphical User Interface (GUI) that controls the position of the fingers. Since the subjects could not use the GUI themselves (they were grasping the robot's hand), they verbally instructed a researcher that controlled the GUI on how to move the robot's fingers. During the finger position adjustment, we recorded the motor encoder position of each finger for the selected handshake grip. Afterward, this customized handshake was performed again on the subject while recording the forces on each tactile sensor, and we asked the participant for confirmation if the handshake was comfortable.

4.4.2 Quantitative measurements

Like in the first experiment, we present the average total force magnitude chosen by the subjects for each sensor. These results are only relative to one handgrip per person, the one with the final customized finger positions. The results are presented in Figure 4.4.

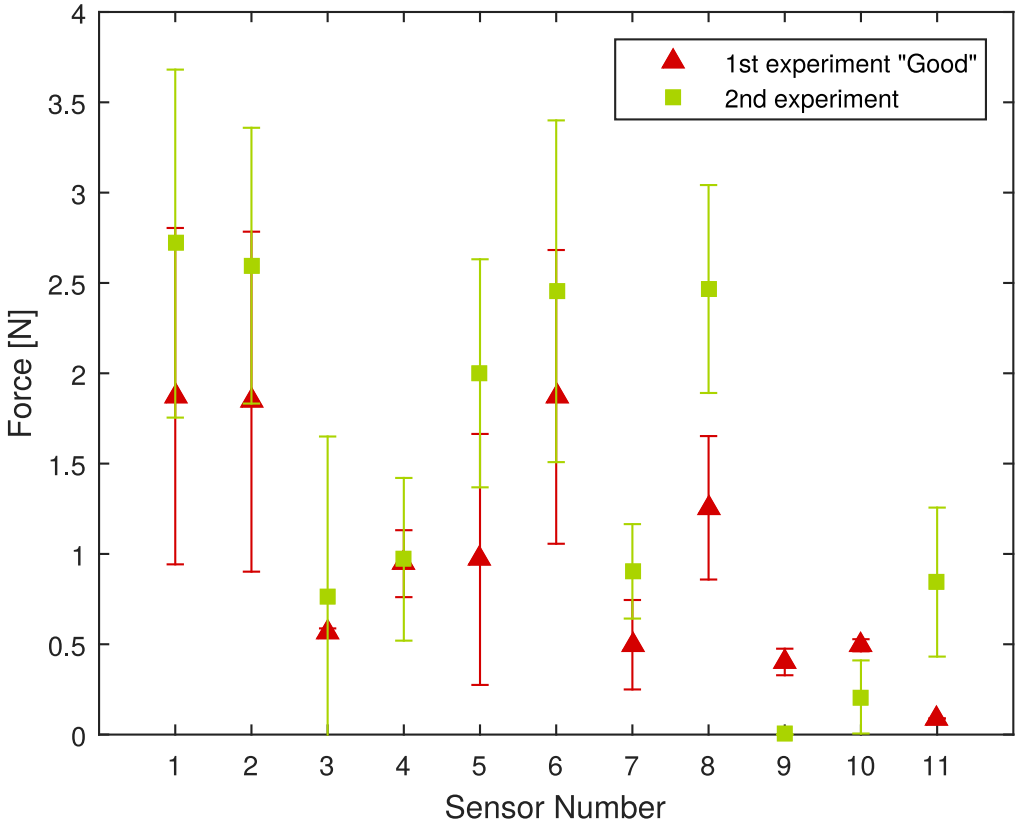


Figure 4.4: Average and variance of the force measured on each sensor for the ideal hand grip chosen by the subjects (blue), and the average of the hand grip evaluated as "good" (green) in the experiment of Section 4.3 (Figure 4.3b). The average sum of sensor forces for customized handgrips was 15.92 N. For males this value was of 16.07 N while for females it was of 15.37 N.

We notice that sensors 1, 2, 5, 6 and 8 are the ones that perform more contact with the hand of the subject, while some of them almost make none. Compared to the first experiment, we see statically significantly larger forces applied practically in all sensors that result in larger overall applied forces, as presented in Figure 4.4 and Table 4.2. Comparison of force sample distributions on each sensor for both handgrasps shows statistically significant differences for most sensors (Sensors 1, 2, 5, 7, 8, 9, 10 and 11 - 8 out of 11 - 72.7 %) on both conditions. Additionally, during the second execution of the handshake, all the users confirmed that the selected handshake is comfortable.

4.4.3 Participants' qualitative feedback

Overall, participants enjoyed the handshake with the robot. Similar to the pilot study, participants were surprised with the comfort of the robot's tactile sensors. The 3D-printed hand palm was also well-rated. Nonetheless, the participants suggested that having the palm with the same sensors would provide an even more comfortable handshake. Additionally, they claimed that knowing the palm also had force sensors would increase their perceived safety.

Table 4.2: Hypothesis tests' statistics and p-values for the distribution of measured forces on Vizzy's hand sensors.

(a) Independent samples t-test statistics and p-values between forces captured by each force sensor during the preferred pre-defined handgrips of the pilot experiment and the user-guided handgrips of the second experiment.

Independent samples t-test		
Sensor	t	p
1	3.13	0.003
2	2.03	0.047
4	0.452	0.653
5	3.680	0.001
6	1.755	0.085
8	6.303	< 0.001
9	-8.626	< 0.001

(b) Mann-Whitney U test statistics and p-values between forces captured by each force sensor during the preferred pre-defined handgrips of the pilot experiment and the user-guided handgrips of the second experiment.

Mann-Whitney U test		
Sensor	t	p
3	315	0.654
7	212	0.022
10	134	< 0.001
11	11	< 0.001

4.4.4 Discussion

This experiment showed that using a co-design approach to collect handgrip force distributions where users have control of the robot's hands gives us significantly finely-tuned data. This observation is supported by the significant differences observed between readings of the pilot experiment and this one. Thus, we argue that even though the methodology of subsection 4.4.1 is slightly more time-consuming than the one from subsection 4.3.1 due to the fine-adjustment process of each finger, it gives us more reliable force distribution preferences.

The addition of 3D-printed hand palm reduced people's complaints about the metallic palm even though people mentioned that the hand grasp would be more comfortable and safer with a silicon palm with tactile sensors.

4.5 Handgrip controllers

The previous experiments allowed us to gather a reference signal for the handshake's handgrip and enhance the robot's hand. We now describe a force-based controller that uses the distribution of average forces per sensor as the reference signal. Afterward, we performed a user study to evaluate the force-based handshake (which we will refer to as "PID HD") against a finger position-based one (which we will refer to as "Fixed HS"). Unlike our previous experiments with people, this one had a full handshake with a predefined arm shake motion.

Each finger of the robot poses a multiple input single output system, with force sensor values as

inputs and motor encoder positions as outputs. However, the physical model of an underactuated finger that opens and closes is not very accurate since the motion of the limbs and the forces applied on the sensors depend on several variables. For instance, the initial point of contact with the object, the object's shape, and other physical properties such as roughness and elasticity play a role in finger motions. Thus, our approach for control is a PID controller [138]. Our controller has the ℓ^2 norm of the forces measured on the sensors for each finger as the process variable:

$$F_{s_i} = \sqrt{F_{x,s_i}^2 + F_{y,s_i}^2 + F_{z,s_i}^2}, \tag{4.1}$$

$$F_{f_j} = \sum_{s_i} F_{s_i}, \forall s_i \in f_j, \tag{4.2}$$

where s_i stands for sensor i and f_j stands for finger j. In Eq. (4.1), F_{s_i} is the magnitude of the force on sensor s_i, and in Eq. (4.2) F_{f_j} is the sum of the sensors mounted on the finger f_j. F_{f_j} is the process variable of the controller, and the set value is obtained from the mean value of the forces F_{x,s_i}, F_{y,s_i} and F_{z,s_i} for every sensor across all the users of the customized handshake study in Section 4.4.

The finger position-based handshake used the average positions of each finger collected during the experiment of section 4.4. It also uses a PID controller to actuate the robot's fingers towards the desired setpoints. We refer to this method as the "Fixed HS" controller during this work.

4.6 Robot's arm's shake motion

A realistic human-robot handshake should autonomously control the arm and hand motor joints using force-based and torque-based feedback. On the one hand, we can define a closed-loop force control using the finger sensors to control the motors of the fingers. On the other hand, our robot's sensing capabilities of the arm joints are limited to closed-loop position and velocity data. Having this limitation in mind, we implemented an arm motion inspired by the human-human handshake model proposed by [139]. This arm motion is thus not compliant with human forces exerted on the robot's arm. In Figure 4.5, we show the evolution of the position of the robot's wrist during the handshake. Note that we control the elbow and wrist joints together to keep the same robot's hand orientation while the robot is performing the arm motion. The handshakes executed by the robot are composed of the arm's shake trajectory of this section and the handgrips under study ("Fixed HS" or "PID HS").

4.7 User evaluation of handshake controllers

We designed a user study to evaluate people's perceptions of the developed handshake methods, comparing a handshake with the "PID HS" handgrip controller to one with the "Fixed HS" controller.

For this experiment, we recruited 38 participants (20 female and 18 male) with ages ranging from 19 to 52 years old ($\mu = 25.39$, $\sigma = 6.34$). The recruited participants belonged to the academic community (students and staff). However, they did not have robotics, electrical, or computer engineering

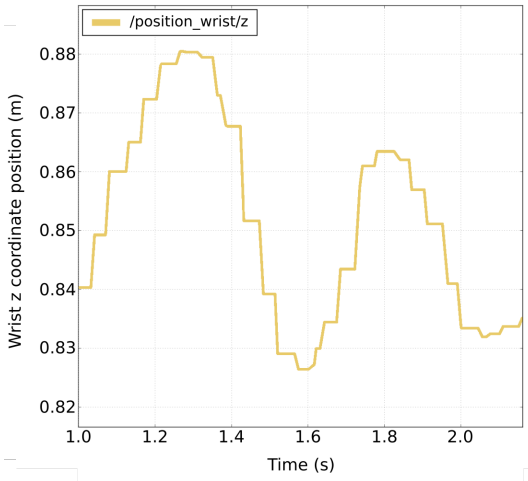


Figure 4.5: Predefined arm motion to perform the handshake. We pre-programmed the robot's arm to mimic the human arm movements reported in [139].

backgrounds, nor did they have any prior knowledge about the experiment, thus avoiding possible biases.

We used the following procedure. First, a researcher would present the robot and assign a code to the participant: "A" or "B". The purpose of these codes was to let the researcher controlling to robot know which handshake to perform first: "A" - "Fixed HS" followed by "PID HS" and "B" - "PID HS" first followed by "Fixed HS". Then the participant was instructed to shake hands with the robot (using the first handgrip controller of the experiment) and answer the questionnaire shown in Table 4.3. Afterward, we told the participant to shake hands with the robot again (the second handgrip controller) and answer the questionnaire about the second handshake. With this questionnaire, we intended to evaluate how people perceive the interaction with the handshakes (with INT items [140]), handshake firmness (FRM), strength (STR), and handshake safety (SFT) [73], [141]. Finally, PE items reported on people's Perceived Enjoyment [142].

4.8 Results

We now analyze the results and compare both handgrips. Since participants went through both conditions, we performed Dependent T-tests to compare results represented by normal data. For non-normal data, we used Wilcoxon tests.

4.8.1 Interaction items

According to the Shapiro-Wilk test, answers regarding the interaction items do not follow a normal distribution. Thus we compare both conditions using the Wilcoxon test. No statistically significant differences were found for any item: INT1 - $Z = -0.176$, $p = 0.860$, INT2 - $Z = -1.117$, $p = 0.264$, INT3 - $Z = -0.730$, $p = 0.466$, INT4 - $Z = -0.192$, $p = 0.848$ (Figure 4.6a). During both conditions the interaction with the robot was not considered scary ($p < 0.001$), and was considered interesting ($p < 0.001$), meaningful ($p < 0.001$) and exciting ($p < 0.001$) when comparing the median

Table 4.3: Post-handshake questionnaire items and respective scales.

(a) Interaction, handshake firmness, and handshake strength questionnaire items.

Code	Item
	How did you feel about your interaction with the robot?
INT1	Scary (1) - Not scary (7)
INT2	Boring (1) - Interesting (7)
INT3	Meaningless (1) - Meaningfull (7)
INT4	Unexciting (1) - Exciting (7)
FRM	The handshake was firm
	Totally disagree (1) - Totally agree (7)
STR	I think the handshake was:
	Too weak (1) - Too strong (7)

(b) Questionnaire items of perceived safety and enjoyment.

Code	Item
	During the handshake I was feeling:
SF1	Anxious (1) - Relaxed (7)
SF2	Agitated (1) - Calm (7)
SF3	Surprised (1) - Quiescent (7)
PE1	I enjoy the robot's handshake
PE2	I find it fun to handshake the robot
PE3	I find the handshake pleasurable
	Totally disgree (1) - Totally agree (7)

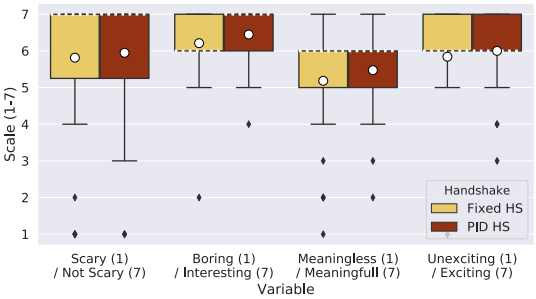
values of these items with the neutral value using a Wilcoxon Signed Rank test.

4.8.2 Firmness and strength items

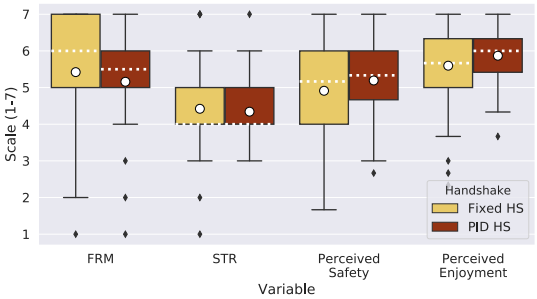
A Shapiro-Wilk test on the firmness (FRM) and strength (STR) items rejected the hypothesis for normal data, which led us to compare the results with the Wilcoxon test. This test yielded no statistically significant differences between conditions. We found a small tendency to find the "Fixed HS" hand grip firmer ($Mdn_{fixed} = 6.0$ and $Mdn_{PID} = 5.5$). Regarding the strength, both conditions had a median value of 4, which is the ideal value (Fig 4.6b). However, a Wilcoxon signed-rank test rejects the hypothesis that the true median of both distributions is equal to 4 ($p < 0.05$ for both conditions). Looking at the mean values, we can see that they are $\mu_{Fixed} = 4.42$ and $\mu_{Fixed} = 4.34$, meaning that people might have found the grips a bit tighter than the ideal.

4.8.3 Perceived safety and perceived enjoyment

The Shapiro-Wilk test could not reject the hypothesis of normal data for the Perceived Safety dimension ($\alpha = 0.804$). Thus, we compared both conditions using a dependent *t-test*. No statistically significant differences were found between conditions ($t = -0.9590, p = 0.3440$). We found a very small tendency to perceive the "PID HS" hand grip safer than the "Fixed HS" hand grip: $\mu_{fixed} = 4.9123$, $\sigma_{fixed} = 1.298$, and $\mu_{PID} = 5.1930$, $\sigma_{fixed} = 1.038$. A One-Sample *t-test* showed that people found both handshakes significantly safe ($t = 4.331$, $p < 0.001$ for the "Fixed HS", and $t = 7.080$, $p < 0.001$ for the "PID HS").



(a) Boxplot with self-reports of interaction items.



(b) Boxplot with self-reports of handshake preceived Firmness, Strength, Safety, and Enjoyment.

Figure 4.6: Boxplots with questionnaire results of people's perceptions of both handshakes. The ∘ symbol represents the mean and the dashed line represents the median of each variable.

Normality analysis of the gathered data regarding the Perceived Enjoyment dimension ($\alpha = 0.847$) using the Shapiro-Wilk test rejected the hypothesis of normal data, leading us to compare both handshakes with the Wilcoxon test. No statistically significant differences were found between conditions ($Z = -0.816, p = 0.414$). Nonetheless, data shows a small tendency towards the "PID" version of the hand grip: $Mdn_{fixed} = 5.667$, and $Mdn_{PID} = 6.0$ (Figure 4.6b). A Wilcoxon Signed Rank test against the neutral value shows that people enjoyed both handshakes ($Z = 672.0$, $p < 0.001$ for "Fixed HS" and $Z = 739.0$, $p < 0.001$ for the "PID HS".

4.9 Discussion and conclusions

In this Chapter, we studied and developed a human-robot handshake for our robot, Vizzy. Due to the absence of torque sensors on the robot's arm's motors, we focused our developments on the handgrip, using a pre-defined shaking motion. We measured the forces in the contact points of robot fingers in real-time.

The pilot study of section 4.3 allowed us to identify initial flaws in the design of the handshake study and initial improvements of the robot's hand, like the hand palm. Relying only on three handgrip finger positions limited the analysis of the user's preferred grip force since the "optimal" grip strength for that user might not be part of this set of grips. Additionally, the forces felt by participants would depend highly on the size and shape of the hand of their hands. Moreover, participants reported that we could improve the hand palm touch comfort by covering it with the same material used for the force sensors. The feeling of the metal surface could be slightly uncomfortable.

In the second experiment (section 4.4), we intended to fix these issues. Using a co-design-like experiment with users' fine control over the robot's fingers allowed us to gather a distribution of forces that significantly differed from the one collected during the pilot experiment. Thus, we believe this method is a more accurate way of collecting people's preferences related to handshakes. The significantly higher overall preferred forces could also result from increased perceived safety and improved comfort the 3D-printed hand palm cover provided. Nonetheless, this is just a hypothesis since the groups in both experiments are different. Although we could not cover the hand palm with silicon, a 3D-printed palm cover solved complaints about the unpleasant metal haptic feeling

but left room for further improvements.

We then compared two handgrip controllers based on the data collected in the experiment of secion 4.7, which had setpoints based on force distributions or encoder position, respectively.

We did not find statistically significant differences between both approaches in any measured items. We could only find small tendencies regarding interaction items towards the "PID HS". Similarly, there was a small tendency to consider the "PID HS" safer and more enjoyable than the "Fixed HS". Participants positively evaluated both handgrips, which leads to the hypothesis that low-end robotic systems can actually perform comfortable handshakes for populations similar to the ones in the present study without the need for touch sensors by following a methodology similar to the one described in section 4.4. Generalization requires further investigation with other robotic platforms and hands. Nonetheless, improving pHRI safety is a paramount advantage of force-based controllers over encoder-based controllers, even if they do not directly impact people's perceptions of handshakes.

5

A PIPELINE FOR SOCIALLY AWARE MOBILE ENGAGEMENT

In this Chapter, we describe a pipeline that detects some relevant social signals identified in Chapter 2. It is composed of off-the-shelf perception algorithms and our extensions to achieve this goal. Additionally, we describe a method to estimate and track the human greeting according to Kendon's greeting model, using these automatically detected social signals.

This Chapter contains an extended version of a paper published in the IROS 2021 conference:

M. Carvalho*, J. Avelino*, A. Bernardino, R. Ventura, and P. Moreno, "Human-robot greeting: Tracking human greeting mental states and acting accordingly", in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, Prague, Czech Republic (Online), Sep. 2021.

5.1 Proposed pipeline

We start this Chapter with an overview of the proposed pipeline for human-robot engagement, which uses Vizzy's left eye camera to capture the scene (Figure 5.1). The system extracts people's keypoints with Openpose and completes the missing joints with our proposed skeleton completer (section 5.3) for each frame. Using these detections, it estimates two human body features: head and body pose. OpenHeadPose estimates the head orientation, but since the model was trained with image-centered heads, we needed to apply the orientation correction developed in subsection 5.2.7 to generalize it to all people on the image.

The perception system uses a homography-based scheme to estimate feet and head 3D positions and the estimated 3D positions of keypoints as a proxy for body orientation. The tracker system uses the Hungarian algorithm based on people's positions for data association and a Kalman Filter with a constant velocity process model to track and compute a smooth estimate of people's poses and velocities.

A simple gesture detector proposed in section 5.6 detects distance/close salutations and handshakes in real-time. We note that the social signal processing relies on accurate images, heavily

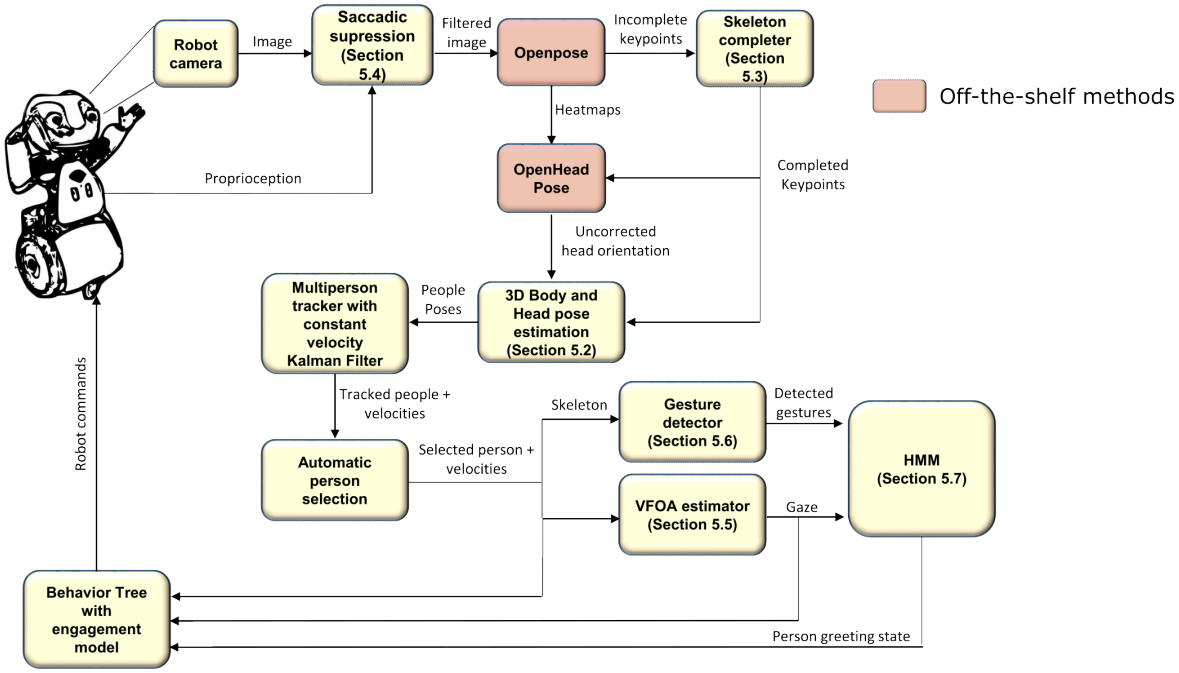


Figure 5.1: The complete pipeline for human-robot engagement.

degrading all estimation if the pipeline receives blurred or saturated images. These are often the results of moving cameras and excessive sunlight. Thus, the beginning of the pipeline consists of a saccadic suppression mechanism that discards images taken under certain motor velocities and with excessive detected blur. We described this mechanism in detail in section 5.4.

We implemented an algorithm consisting of an Hidden Markov Model (HMM) that receives social signals and outputs people's greeting state according to the six greeting phases of Kendon. We describe this method and evaluate it in detail in section 5.7.

Finally, we use Behavior Tree (BT)s to plan and execute the behaviors described in Kendon's greeting model. We developed modular building blocks that can be assembled to compose distinct models. These are described in section 5.8.

5.2 3D pose estimation from monocular RGB images

In this section, we used classic computer vision models to obtain the 3D body and head poses (position + orientation) from 2D pose information on monocular images. We used 2D keypoints extracted from an out-of-the-box 2D person estimation approach, OpenPose, focusing on the 3D estimation of body and head poses instead of individual skeletons' keypoints. Additionally, we use head orientations extracted with OpenHeadPose, another out-of-the-box algorithm that uses OpenPose's keypoints and heatmaps, although it considers that people's heads are image-centered. Our methods use the pinhole camera model, the camera's intrinsic and extrinsic parameters to estimate the 3D body pose from the 2D skeleton keypoints and generalize OpenHeadPose's head orientation estimates for heads in the whole image.

Since this section describes several algorithms and reports their results, we opted to show the

results right after describing each method. Additionally, since all algorithms share the same experimental setup, we describe it next. Although this structure is unusual in research documents, we believe it allows the reader to keep up with the flow of information without disruptions.

5.2.1 Experimental setup

We used the Vizzy robot to collect data in a laboratory setting to evaluate the methods proposed in the following sections. We asked a human subject to orient her body and head toward a red ball we placed in the environment (Figure 5.2). Then, using the front laser readings and the RGB-D depth map, we manually labeled the subject's body pose as well as the position of the red ball. We later calculated the subject's head pose using her known height and the ball's location, assuming that the y-axis of the head was parallel to the ground floor. We recorded image data from the robot's left eye RGB camera, the camera's intrinsic parameters, and the robot's joint states from which we could extract the camera's extrinsic parameters. We asked the subject to stand still in six distinct poses. For each pose, we recorded data with different robot head configurations. This way, we could collect images containing the person in diverse image locations (center/upper/bottom right/left) per person pose (Figure 5.2).

Figure 5.2: Samples from the robot's left camera, illustrating our experimental setup.

During our experiments, we used rectified images and associated intrinsic parameters. Thus, we did not deal with the camera's distortion parameters. Rectification was performed by ROS' *image_proc* package [1]. Our images had a resolution of 640 $\times$ 480 pixels. We only collected samples when the robot's head stopped moving. With this decision, we aimed to avoid blurred images and mismatches between the image and extrinsic parameters (due to image time stamping delays). The robot had its camera at around 1 m above the ground floor.

Finally, when reporting results, we kept the same color scheme for each person's pose among the plots, except in violin plots comparing two methods side-by-side.

5.2.2 Proposed pose estimation pipeline

The proposed pipeline for the 3D body and head pose estimation (Figure 5.3) uses two out-of-the-box state-of-the-art algorithms to extract 2D keypoints and head orientations from RGB images: Openpose and OpenHeadPose. First, Openpose extracts 2D skeleton keypoints from RGB images

[1] `http://wiki.ros.org/image_proc`

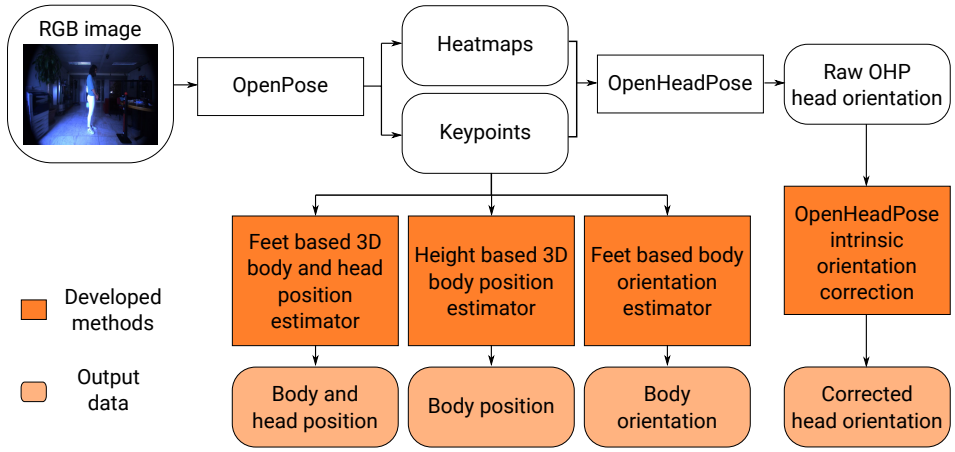


Figure 5.3: Proposed 3D body and head pose estimation pipeline. The pipeline's inputs are individual RGB image frames. Then we use two off-the-shelf algorithms (OpenPose and OpenHeadPose) to extract people's keypoints and biased head orientations. We used keypoints to estimate body poses and head positions using three methods that use different geometric assumptions. Our fourth proposed method corrects OpenHeadPose's estimated head orientations for people whose heads are not image-centered.

and a set of heatmaps encoding the likelihood of each type of keypoint along with the image. We propose two methods to extract people's 3D body and head position. The first one (subsection 5.2.4) uses the assumption that people's feet are on the ground floor to estimate both body and head positions. Instead, the second method (subsection 5.2.5) estimates the body position based on an assumed height. These methods complement each other in situations where feet positions can not be accurately estimated. We also use feet keypoints to estimate human body orientations (subsection 5.2.6) because they are a stable proxy for it, according to Setti et al. [16]. OpenHeadPose uses Openpose's keypoints and heatmaps to estimate the head orientation. However, it assumes that the person's head is image-centered. That assumption leads to wrong 3D head pose estimations for the remaining cases, more pronounced when faces are near image corners. Thus, we also propose a method to correct head orientations using the camera intrinsics to be sure that accurate head orientations do not depend upon people's face location in images (subsection 5.2.7).

5.2.3 Background: pinhole camera model

The pinhole camera model [143] describes the geometric relationships between a point in a 3D scene (p_x, p_y, p_z) and its projection in the camera image plane (x, y). It assumes that the camera aperture is a single point with no lenses, and as such, it does not consider other phenomena like distortion and focus. Thus, its reliability usually degrades for image coordinates farther away from the image center since they suffer from increasing distortion. An approach to counter distortion-related inaccuracies is to estimate a set of coefficients that describe the tangential and radial distortions through camera calibration and then create a new undistorted image [144].

$$\lambda \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \mathbf{K} \left[\begin{array}{c|c} \mathbf{R} & \mathbf{t} \end{array} \right] \begin{bmatrix} p_x \\ p_y \\ p_z \\ 1 \end{bmatrix} \tag{5.1}$$

The model (Equation 5.1) is composed of a matrix of intrinsic parameters $\mathbf{K}$, a matrix of extrinsic parameters that results from the concatenation of a rotation matrix $\mathbf{R}$ and translation vector $\mathbf{t}$ (converts to camera coordinate system), and an arbitrary scale factor λ. Points are described in homogeneous coordinates. The matrix of intrinsic parameters $\mathbf{K}$ is an upper triangular 3×3 matrix described as:

$$\mathbf{K} = \begin{bmatrix} f_x & s & x_0 \\ 0 & f_y & y_0 \\ 0 & 0 & 1 \end{bmatrix} \tag{5.2}$$

$f_{x/y} = F/s_{x/y}$ is computed with the focal length F and the pixel sizes in the metric camera coordinate system along the x and y axis, respectively. The skew coefficient s is non-zero if the axes are not perpendicular. x_0 and y_0 represent the optical center in pixels. As we can see, directly recovering a point's 3D position from its projection in the image plane is impossible because the same projection is shared by an infinite number of points. However, we can address this problem by using geometric constraints. We describe approaches in this direction in the following subsections.

5.2.4 3D body and head position estimation with known feet keypoints

A possible approach to estimate people's position from their image projections is to consider them standing on the ground floor. Thus, assuming a world reference frame placed on the ground, we set $p_z = 0$. By setting $p_z = 0$, we constrained point locations to a plane, simplifying the point projection model into a homography, $\mathbf{H}$. Thus, to obtain the 3D feet coordinates p_x, p_y we use the median value of the known feet image keypoints (x_f, y_f) and solve the following system of linear equations for p_x, p_y, dividing the result by the scale factor λ afterward:

$$\begin{bmatrix} x_f \\ y_f \\ 1 \end{bmatrix} = \mathbf{H} \begin{bmatrix} p_x \\ p_y \\ 1 \end{bmatrix} \lambda \tag{5.3}$$

where:

$$\mathbf{H} = \mathbf{K} \left[\begin{array}{cc|c} \mathbf{r_1} & \mathbf{r_2} & \mathbf{t} \end{array} \right], \quad \mathbf{H} \in \mathbb{R}^{3\times3} \tag{5.4}$$

We can now estimate people's 3D head positions based on their positions on the ground floor. To simplify computations, we can use a reference frame translated to the 2D coordinates of the points on the ground floor. In general, we can represent a point in an arbitrary translated frame as follows:

$$\lambda \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \mathbf{K} \left[\begin{array}{c|c} \mathbf{R} & \mathbf{t}_\delta \end{array} \right] \begin{bmatrix} p_x - \delta_x \\ p_y - \delta_y \\ p_z - \delta_z \\ 1 \end{bmatrix} \tag{5.5}$$

$$\mathbf{t}_\delta = \begin{bmatrix} tx + R_{1,1}\delta_x + R_{1,2}\delta_y + R_{1,3}\delta_z \\ ty + R_{2,1}\delta_x + R_{2,2}\delta_y + R2,3\delta_z \\ tz + R_{3,1}\delta_x + R_{3,2}\delta_y + R_{3,3}\delta_z \end{bmatrix}$$

If we use a reference frame translated by $(\delta_x = p_x, \delta_y = p_y, \delta_z = 0)$ we will get the following overdetermined system:

$$\begin{bmatrix} x_h \\ y_h \\ 1 \end{bmatrix} = \boldsymbol{A} \begin{bmatrix} p_z \\ 1 \end{bmatrix} \lambda \tag{5.6}$$

where:

$$\boldsymbol{A} = \mathbf{K} \left[\begin{array}{c|c} \mathbf{r_3} & \mathbf{t}_{x,y} \end{array} \right], \qquad \boldsymbol{A} \in \mathbb{R}^{3\times 2}$$

$$\mathbf{t}_{x,y} = \begin{bmatrix} tx + R_{1,1}p_x + R_{1,2}p_y \\ ty + R_{2,1}p_x + R_{2,2}p_y \\ tz + R_{3,1}p_x + R_{3,2}p_y \end{bmatrix} \tag{5.7}$$

Re-arraging this system of equations we can get:

$$\begin{bmatrix} 1 \\ 1 \end{bmatrix} p_z = \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \tag{5.8}$$

where:

$$\begin{cases} \alpha = \frac{a_{1,2} - x_h a_{3,2}}{x_h a_{3,1} - a_{1,1}} \\ \beta = \frac{a_{2,2} - y_h a_{3,2}}{y_h a_{3,1} - a_{2,1}} \end{cases} \tag{5.9}$$

The least-squares estimate of p_z is then:

$$\hat{p}_z = \frac{\alpha + \beta}{2} \tag{5.10}$$

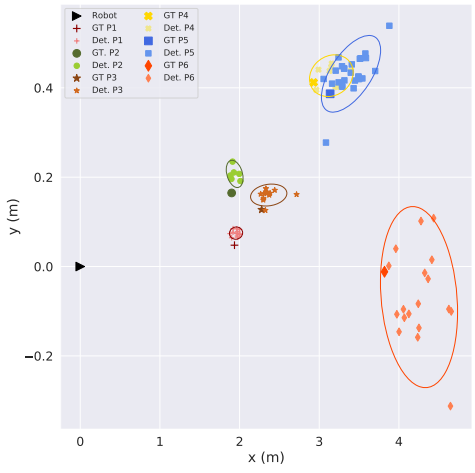
There is, however, a catch to this solution. Because of the projection model, the α estimate of p_z is extremely sensitive to small variations in x_h. We can only approximate the value of x_h up to half a unit because image coordinates are integers 0.5. When implemented with common intrinsic parameters, using the alpha estimate with such error would cause the estimation method to fail. As a result, we estimate p_z as follows:

$$\hat{p}_z = \beta = \frac{a_{2,2} - y_h a_{3,2}}{x_h a_{3,1} - a_{2,1}} \tag{5.11}$$

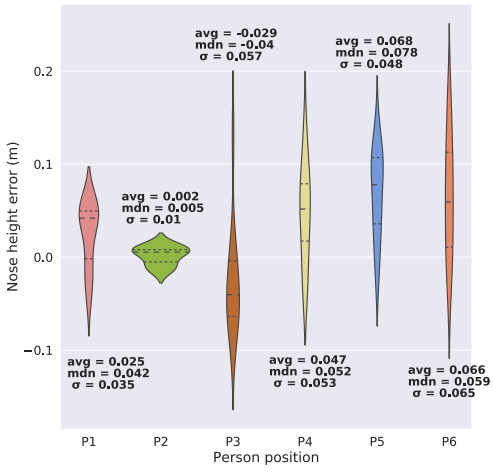
5.2.4.A Results

To evaluate the feet-based body position estimation, we created a 2D plot with the estimated positions on the ground floor (Figure 5.4a). In addition, we plotted the ellipses of two standard deviations. We note that the x and y axis have different scales. As we can see, the error and its standard deviation grow with the distance to the robot, starting from $\mu = 0.05\,\text{m}$, $\sigma = 0.014\,\text{m}$ at a distance of around 1.94 meter to $\mu = 0.448\,\text{m}$ $\sigma = 0.242\,\text{m}$ at a distance of 3.82 m. The growing standard deviation is expected due to the projection model since small pixel distances near the horizon correspond to big distances in the ground plane.

We obtained a similar error behavior for the head position estimate (Figure 5.4b). We expected this result because it is based on the body position estimate. However, because head keypoints are less stable than feet keypoints, we expected some variation in error behavior. For instance, even though P5 is closer to the robot than P6, the standard deviation of its head position estimation error was larger. In general, the average error was less than 0.07 m.



(a) Results for 6 positions estimated with feet information using known feet keypoints. Coordinates are in the robot's base frame.



(b) Results for noise of height estimation after estimating the 3D position of people's feet.

Figure 5.4: Results for body and head position estimation with known feet. Both subfigures use the same color scheme, as stated in subsection 5.2.1.

5.2.5 Estimation of head and body position with unknown feet keypoints

When we do not have feet keypoints, we simplify the projection model into a homography under a different assumption: we assume a known height for the median of a subset of visible 2D keypoints (head keypoints), z_h. The homography matrix $\mathbf{H}$ can be easily obtained if we describe the 3D head point in a reference frame translated by $p_z = z_h$:

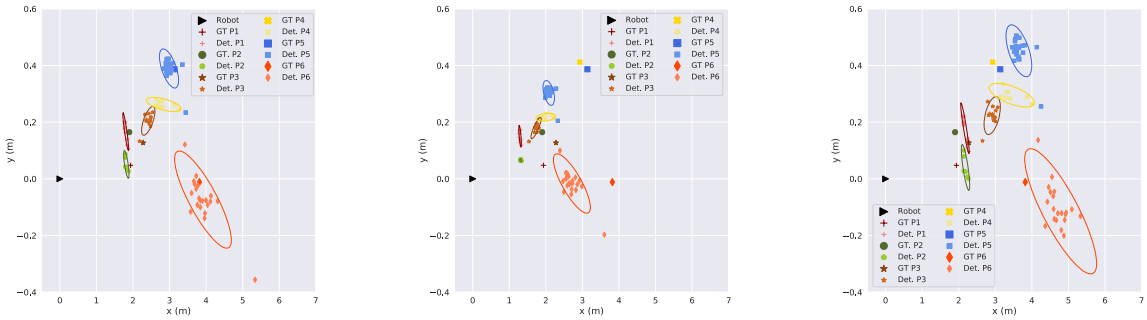
$$\lambda \begin{bmatrix} x_h \\ y_h \\ 1 \end{bmatrix} = \mathbf{K} \left[\begin{array}{c|c} \mathbf{R} & \mathbf{t}_h \end{array}\right] \begin{bmatrix} p_x \\ p_y \\ 0 \\ 1 \end{bmatrix} = \mathbf{H} \begin{bmatrix} p_x \\ p_y \\ 1 \end{bmatrix} \tag{5.12}$$

$$\mathbf{H} = \mathbf{K} \left[\begin{array}{cc|c} \mathbf{r}_1 & \mathbf{r}_2 & \mathbf{t}_h \end{array}\right], \qquad \mathbf{t}_h = \begin{bmatrix} tx + R_{1,3} z_h \\ ty + R_{2,3} z_h \\ tz + R_{3,3} z_h \end{bmatrix}$$

Solving the system of linear equations yields the p_x, p_y coordinates of head and body, with the $p_z = z_h$ and $p_z = 0$ respectively.

5.2.5.A Results

We now evaluate the results when assuming a known height for the person. Here, we made three assumptions: (i) a height of 1.70 m - the subject's true height; (ii) a height of 1.825 m - an error of -0.12 m; and (iii) an height of 1.528 m - an error of 0.172 m. Assumptions (ii) and (iii) were based on the tallest average height and the shortest average height of adults reported in a study with 1.86 million subjects in 200 countries [145]. Figure 5.5 shows the distribution of estimated positions based on the known height assumption and the ellipses of two standard deviations. As expected, the method is more reliable when the person's height is known, quickly degrading when the assumed height deviates from the truth. As we can see, if the assumed height is less than the true height, the person's position is estimated to be closer to the true value, whereas the opposite occurs if the assumed height is overestimated.



(a) Assumed height: 1.70m (true height)

(b) Assumed height: 1.528m

(c) Assumed height: 1.825m

Figure 5.5: Results with three height assumptions: the target's true height (5.5a), the mean shortest height for women (5.5b) and the mean tallest height for men (5.5c).

We also directly compared how both body position estimation methods face each other in the violin plot in Figure 5.6. When the subject's height is accurately known, both approaches perform similarly. The feet-based estimation is better for shorter distances, while the height-based estimation is better for longer distances. The feet-based estimation performs better than the height-based one at all distances if the person's height is not precisely known.

5.2.6 Body orientation estimation

We use information from people's feet to estimate their body orientation. As reported in the literature [16], [35], feet layout is more reliable than head and shoulders as a proxy for body orientation. We use OpenPose's feet keypoints for this task (Fig 5.7a). Our method assumes that people are standing perpendicular to the ground floor. This assumption allows us to compute the 3D position of feet keypoints and limits the orientation estimation problem to the angle around only the base's z axis. First, we estimate the 3D position of each keypoint using the homography $\mathbf{H}$ from the ground floor to the camera plane as described in subsection **??**. We then create vectors (2 per foot) connecting the heel to the big and small toes, as shown in Figure 5.7a. Afterward, we extract these $\vec{v}^1 ... \vec{v}^4$ vectors' angles with the base's x axis:

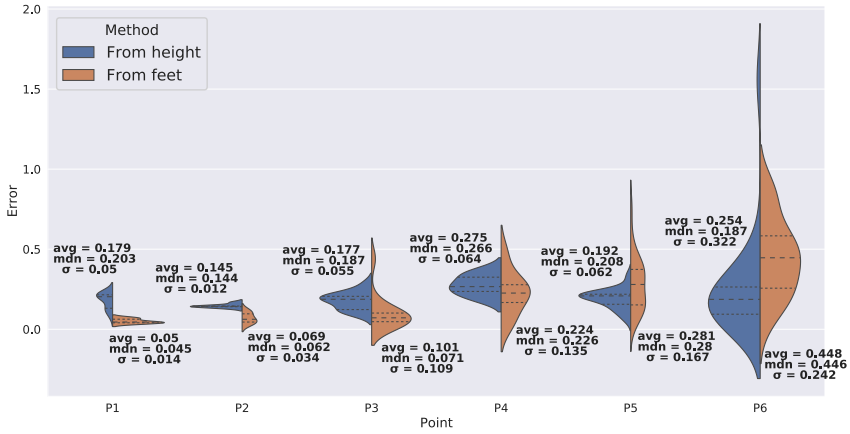
$$\theta_n = \arctan2(v_y^n, v_x^n) \tag{5.13}$$

The final body orientation around the base's z axis is the circular mean of these four angles:

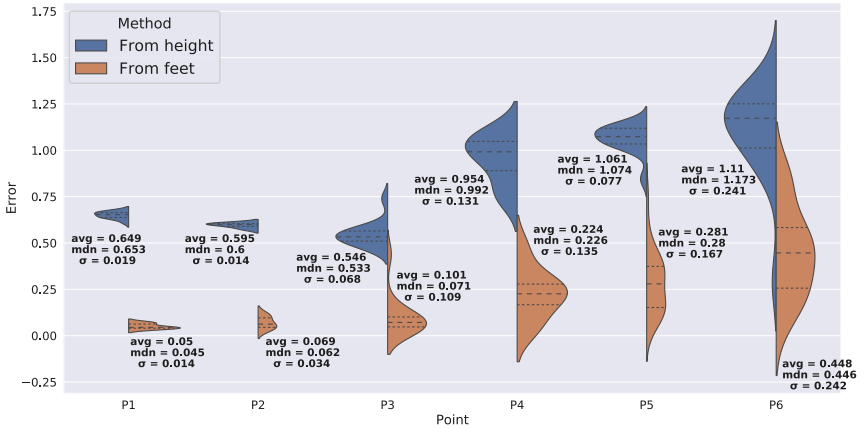
$$\theta^{avg} = \arctan2(\frac{1}{4}\sum_{i=1}^{4}\sin(\theta^i), \sum_{j=1}^{4}\cos(\theta^j)) \tag{5.14}$$

5.2.6.A Results

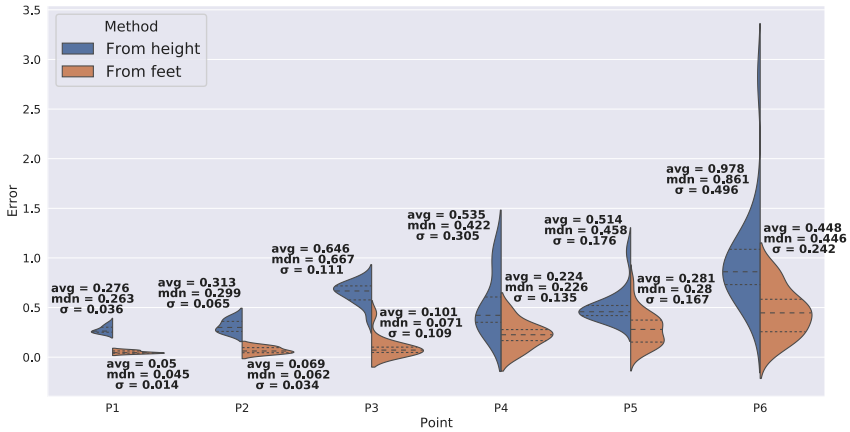
The results for body orientation estimation through feet (subsection **??**) are presented in the violin plot in Fig. 5.7b. The average and median errors are lower than 12° for these configurations. Similar to the previous results, the orientation error also grows with the subject's distance to the



(a) Feet estimation vs estimation with assumed height of 1.70m (true height).

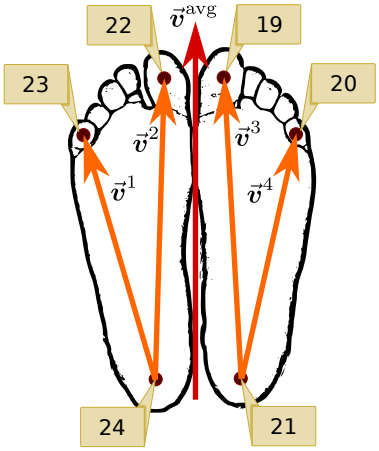


(b) Feet estimation vs estimation with assumed height of 1.528m (assumption error: 0.172m).

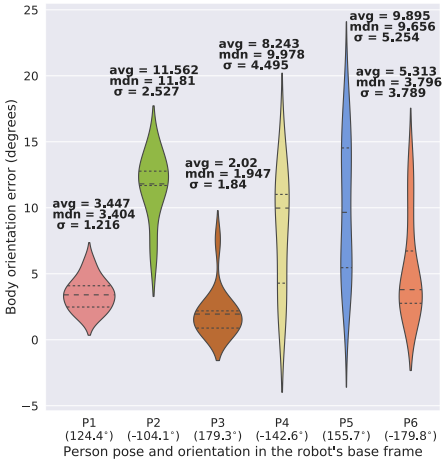


(c) Feet estimation vs estimation with assumed height of 1.825m (assumption error: -0.125m)

Figure 5.6: Comparison of methods to estimate people's 3D position. Feet based estimation vs height based estimation distance error.



(a) Feet keypoints and vectors used to compute body orientation.



(b) Results for the estimation of body pose orientation.

Figure 5.7: Feet keypoints and computed vectors for body orientation estatimation and their results for the 6 distinct person poses.

camera. The results also suggest that the error is lower for poses where the subject faces the camera (P3, P6). Then, it increases when the subject rotates away from this configuration. We can see these phenomena when comparing P3 with P1/P2 and P6 with P4 and P5.

5.2.7 3D head orientation correction

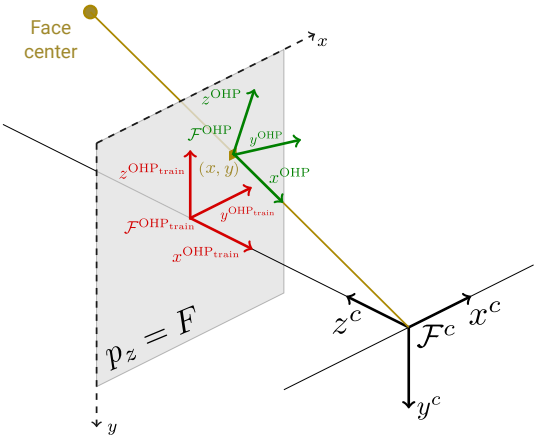
OpenHeadPose leverages OpenPose's estimated keypoints and generated heatmaps to estimate human head orientation. Trained on image-centered faces, it outputs the estimation relatively to a frame with the $x-axis$ aligned with the optical axis ($\mathcal{F}_{OHP_{train}}$), pointing toward the optical center. However, during inference on images in the wild, faces are unlikely to be image-centered, leading to wrong head orientation estimates if we consider the same frame used during training, since their projections on the image plane are distinct even when they have the same orientation in the real world (see Figure 5.8a). The solution is to compute a reference frame for each non-centered face in the image that replicates the training conditions: x-axis connecting the face center to the optical center. We can compute this frame in at least two ways: (i) using the 3D position of the face and camera; or (ii) using the camera intrinsics. Since the first option is more prone to errors, we propose to use the camera intrinsic parameters. Through this method we will obtain a rotation matrix that will convert orientations outputed from OpenHeadPose the the camera frame.

Equation 5.1 can be rewritten as follows:

$$\begin{cases} x = \frac{f_x p_x + s p_y}{p_z} + x_0 \\ y = \frac{f_y p_y}{p_z} + y_0 \end{cases} \tag{5.15}$$

and re-arranged as:

(a) An illustrative example showing multiple objects with the same 3D shape, the same orientation, but clearly distinct projection shapes on the image. If an orientation estimation algorithm assumes these objects to be image-centered, the results will be incorrect. This issue happens with OpenHeadPose, which only correctly handles image-centered faces.



(b) Camera model and OpenHeadPose training frame $\mathcal{F}^{\text{OHP}_{\text{train}}}$ and corrected frame $\mathcal{F}^{\text{OHP}}$. Since the algorithm was trained image-centered face crops, we need to compute a reference frame $\mathcal{F}^{\text{OHP}}$ during inference for non-centered faces.

Figure 5.8: Projection example and coordinate frames.

$$\begin{cases} p_x = \frac{p_z(x - x_0) - \frac{s\,p_z}{f_y}(y - y_0)}{f_x} \\ p_y = \frac{p_z(y - y_0)}{f_y} \end{cases} \tag{5.16}$$

We set x^{OHP} as the x axis of the frame which we use to rotate (thus correcting) the estimated head orientation given by OpenHeadPose. To obtain this axis, we use the 3D coordinates of the projection of the head center in the image plane. Thus $x^{\text{OHP}} \propto (-p_x, -p_y, -p_z)$ up to a positive scale factor. From Eq. 5.16 we get the following unnormalized x_u^{OHP} axis:

$$\begin{aligned} \mathbf{x}_u^{\text{OHP}} &= \begin{bmatrix} \frac{p_z(x_0 - x) - \frac{p_z s}{f_y}(y_0 - y)}{f_x}, & \frac{p_z(y_0 - y)}{f_y}, & -p_z \end{bmatrix}^T \\ &\propto \begin{bmatrix} \frac{(x_0 - x) - \frac{s}{f_y}(y_0 - y)}{f_x}, & \frac{(y_0 - y)}{f_y}, & -1 \end{bmatrix}^T \end{aligned} \tag{5.17}$$

And finally, we obtain $\mathbf{x}^{\text{OHP}} = \mathbf{x}_u^{\text{OHP}} / \|\mathbf{x}_u^{\text{OHP}}\|$.

We obtain the $\mathbf{y}^{\text{OHP}}$ axis by constraining it to be perpendicular to the plane containing $\mathbf{x}^{\text{OHP}}$, and $\mathbf{y}^{\text{CAM}}$, which we can obtain with the cross product of these vectors:

$$\mathbf{y}^{\text{OHP}} = \frac{\mathbf{x}^{\text{OHP}} \times \mathbf{y}^{\text{CAM}}}{\|\mathbf{x}^{\text{OHP}} \times \mathbf{y}^{\text{CAM}}\|} \tag{5.18}$$

The same can be done to obtain the $\mathbf{z}^{\text{OHP}}$ axis, constraining it to be perpendicular to the plane containing $\mathbf{x}^{\text{OHP}}$ and $\mathbf{y}^{\text{OHP}}$, which are orthogonal unit vectors:

$$\mathbf{z}^{\text{OHP}} = \mathbf{x}^{\text{OHP}} \times \mathbf{y}^{\text{OHP}} \tag{5.19}$$

With these frames we can now build the rotation matrix which we use to rotate orientations from OpenHeadPose's frame to the camera frame:

$$_{\text{CAM}}\mathbf{R}^{\text{OHP}} = \left[x^{\text{OHP}}, y^{\text{OHP}}, z^{\text{OHP}}\right] \tag{5.20}$$

5.2.7.A Results

We now evaluate how our OpenHeadPose orientation corrector performs in the experimental setup. First, we removed two poses, P1 and P3, since the subject's head was not stable in these configurations. Additionally, even though we attempted to compute the ground truth orientation using the person's head position and the ball, small errors can produce significant bias. Since each sample consists of a distinct camera angle and OpenHeadPose inaccuracies grow with the face's distance to the image center, we can evaluate the correction by analyzing the variance. As such, we centered the data by removing the average error from the results for each pose, focusing only on the variance of estimations. These results are presented in Fig. 5.9. They show a reduction in head orientation variance when using the intrinsic correction for all poses, more pronounced for closer distances.

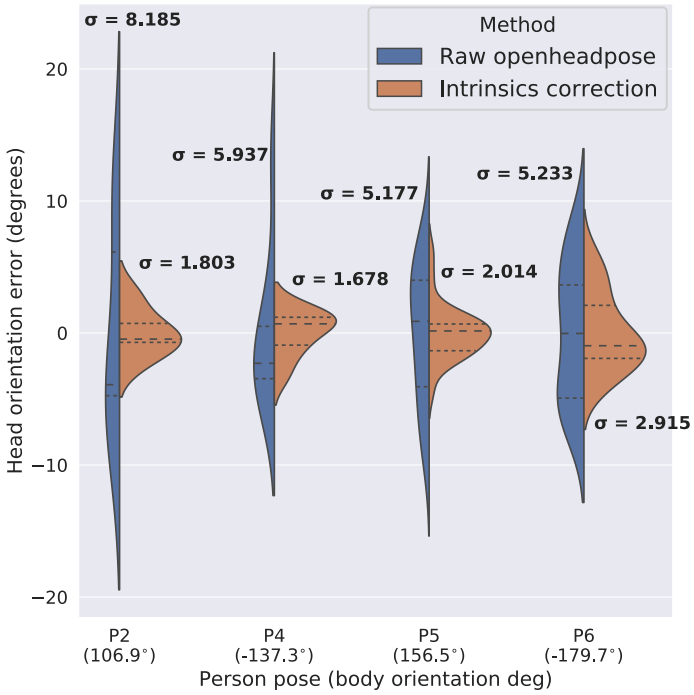


Figure 5.9: Results for head orientation correction.

5.3 Awareness of unseen body parts: completing body keypoints

Even though Openpose can detect human keypoints accurately, some detections may fail due to challenging conditions. These include pedestrian movements, far-away people, partial occlusions caused by furniture and other people, and when people and robots are interacting close to each

other, which makes the camera unable to capture the complete body (frequent during the Close Salutation, for instance). However, the absence of this information can hinder the performance of other algorithms of perceptual pipelines or even make them fail. For example, the methodology of section 5.2 requires either feet or head keypoints. Moreover, when the robot needs to redirect its gaze to perform eye contact, it needs to be aware of the location of people's faces even when it cannot see them.

Thus, we need a method that completes the missing information, allowing the robot to be aware of unseen body parts. Given the human body's structure, we intend to fill the missing data with the most likely values, even if the true ones are impossible to recover due to the high number of possible human poses.

5.3.1 Previous methods

Several methods are used in the literature to impute missing data in datasets. The simpler ones consist of filling absent points with a constant, either preset values (like -1 or 0) or a statistic of those points in the dataset (mean, median, or mode). However, these approaches are unreliable for our pipeline. For instance, people's feet are one of the most occluded parts. If we completed their keypoints with fixed values, we would lose the possibility to estimate people's body orientation using the methods of subsection 5.2.6.

A deep learning method proposed by Carissimi et al. [146] uses the correlations between keypoints to address the problem. Their formulation states this challenge as a denoising problem, proposing Denoising Autoencoders to solve it. The method achieves interesting visual and quantitative results, being able to complete people's skeletons with the missing parts. However, three main limitations made it challenging to fit it into our pipeline. First, we did not find an open-source implementation of their model. Second, their model does not consider feet keypoints, which are essential for our pipeline when it estimates body orientation. We would need to implement and train the whole model with a large-scale dataset that contains feet keypoints. This training was a challenging task at the time since the default versions of COCO [147] and MPII [148] datasets did not have this information. The *Human Foot Keypoint Dataset* [2] contained a small subset of COCO features just feet keypoints used to extend OpenPose in this matter. Even though we could use this information and research to expand the work of Carissimi et al. [146], it could be too time-consuming to do it from scratch. Finally, we intended to have a method that does not consume additional GPU resources since the pipeline already contains two GPU-intensive algorithms (Openpose and OpenHeadPose). We note that this is a real-time application where everything needs to run smoothly.

Thus, given the above limitations, we developed our skeleton completer scheme following a probabilistic-based approach.

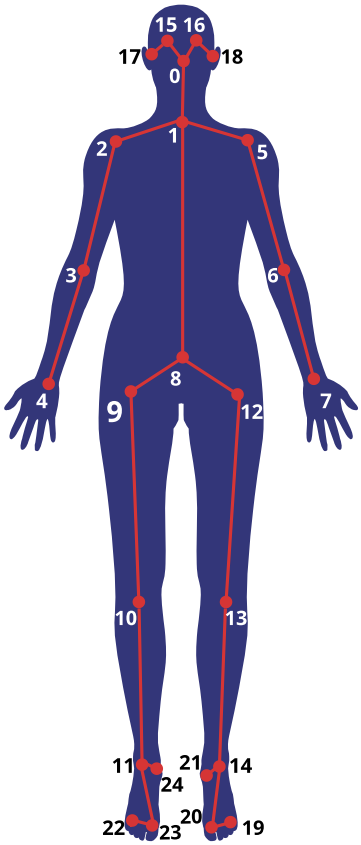


Figure 5.10: Openpose numbered keypoints.

5.3.2 Methodology

We propose to capture the structure of human poses by fitting the joint Probability Density Function (PDF) of their image projections and use the PDF conditioned on observed data to impute missing data, completing the skeleton. For simplicity, we use a Multivariate Normal distribution to approximate the human poses PDF. However, raw keypoint coordinates depend on people's distance, reference frame, cover the whole image, and their relationships are highly non-linear and difficult to interpret. Thus, instead of using keypoint coordinates $\mathbf{k} = (k_x^1, k_y^1, ..., k_x^M, k_y^M)$ to represent a pose, we use the limb attributes $\mathbf{l} = (l_\rho^1, cos(l_\theta^1), sin(l_\theta^1), ..., l_\rho^N, cos(l_\theta^N), sin(l_\theta^N))$, where l_ρ^n is the limb length normalized to the torso and l_θ^n is the limb angle with the image $\mathbf{y}$ axis, for limb n. A limb connects a pair of keypoints, and in this case $N = 24$ limbs (Fig. 5.10). If the torso is unavailable, we use another limb as the basis for normalization (left/right upper arm, shoulders, neck, right/left femur, or tibia) corrected by a scale factor α. This factor is the average proportion of the limb's length to the torso's length on the training set.

Given a Multivariate Normal distribution of which represents the joint PDF of poses $p(\mathbf{l})$ with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, we use the mean vector of $p(\mathbf{l}_m \mid \mathbf{l}_o)$, $\boldsymbol{\mu}_{\mathbf{l}_m|\mathbf{l}_o}$ to impute the missing attributes ($\mathbf{l}_m$) conditioned on the observed attributes ($\mathbf{l}_o$). We compute this conditional mean with:

$$\boldsymbol{\mu}_{\mathbf{l}_m|\mathbf{l}_o} = \boldsymbol{\mu}_{\mathbf{l}_m} + \boldsymbol{\Sigma}_{\mathbf{l}_m,\mathbf{l}_o}\boldsymbol{\Sigma}_{\mathbf{l}_o,\mathbf{l}_o}\left(\mathbf{l}_o - \boldsymbol{\mu}_{\mathbf{l}_o}\right) \tag{5.21}$$

[2] https://cmu-perceptual-computing-lab.github.io/foot_keypoint_dataset/

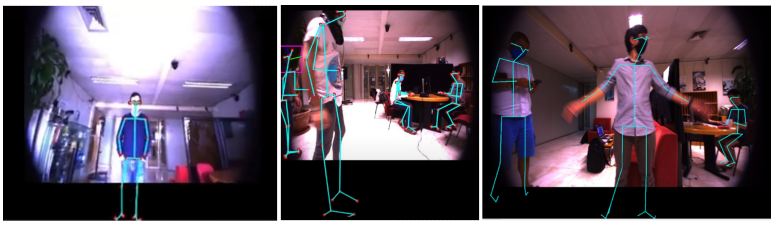
where μ_i is composed of the i elements of μ and $\Sigma_{i,j}$ is the matrix composed of rows i and columns j of Σ.

Afterward, the method unnormalizes the limb lengths and obtains the missing keypoints using the available ones, limb lengths, and their angles with the y axis. We used only the full poses from the *Human Foot Keypoint Dataset* to fit the joint PDF of pose attributes.

5.3.3 Qualitative results

We now illustrate some cases with the resulting keypoints of this method. As we can see in Figure 5.11a, the algorithm can generate plausible keypoints for the critical keypoints of our pipeline (feet and head) where they are missing. The generated feet keypoints are congruent with the body pose, allowing the pipeline to estimate it even when the feet are not visible. Additionally, the estimated positions of head keypoints inform the robot about people's head position when it needs to perform visual contact.

However, the system has some limitations. As seen in Figure 5.11b, head keypoints seem deformed when people turn their backs to the camera. We can also see some errors that may indicate the model's limitations since we are approximating the distribution of human poses to a single Multivariate Gaussian Distribution. While we claim that the results are sufficient for our application, we believe it is possible to improve the model using Mixtures of Gaussians since, in theory, Gaussian Mixture Models are universal approximators of densities [149].



(a) Visual example where the skeleton completer outputed plausible results.

(b) Visual example where the skeleton completer outputed implausible results

Figure 5.11: Qualitative skeleton completer results.

5.4 Biologically inspired suppression of noisy data: saccadic suppression

Social signal processing relies heavily on accurate images. Moreover, our pipeline is heavily reliant on the robot's proprioception when it estimates people's pose information in 3D. However, it is subject to two sources of errors that may completely undermine estimations: image noise and mismatches between image and proprioception timestamps. These issues happen when the robot's sensors move too quickly, for instance, during gaze shifts and when the robot's base rotates too quickly. To reduce the impact of these events, we revisit the concept of saccadic suppression [117].

Unlike the previous work, we extended the image suppression mechanism to use multiple triggers for suppression. First, we use the velocity of head motors and the angular velocity of the robot's base. The system discards the current image if these values are larger than a set threshold. Otherwise, we analyze the image blur. The Laplacian operator highlights fine details in images and is a method to detect image blur. It is given by:

$$\nabla^2 f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} \tag{5.22}$$

We evaluate the variance of the Laplacian, discarding the image if it is smaller than a set threshold. The rationale is that if the variance is very low, there is a low spread of responses, indicating that the image has very few edges. Thus, the fewer edges a real-world image contains, the more blurred it is.

5.5 Visual field of attention estimation

To detect whether people are looking at the robot or not, we relied on the head orientation. Although software like OpenFace [102] can estimate gaze direction using people's eyes, it only works at short distances (around 1 m). Our application scenario deals with longer distances, making this method infeasible during most of the greeting protocol. As such, we used people's head orientation as a proxy of gaze. From the literature [150], we know that the human visual field of attention is associated with an angle of $\theta = 60°$. As such, we use a cone that models the likelihood of gazing at the robot with the following expression: $G = \|g_h - f_r\|_{L2} / r$, where g_h is the center of the cone's base in the plane that contains the robot's face, f_r is the center of the robot's face, and r is the radius of the cone's base.

5.6 Greeting gesture detection

The computer vision literature is rich in image-based gesture detectors. These are common even in commercial applications like the Kinect. Since most available methods require additional equipment (like a depth sensor) or neural networks, we opted to develop a lightweight solution. Using both arms' image projections, we use a sliding temporal window of the limb lengths and angles (computed in section 5.3) to create a time series.

We use these time series as features of three classes: (i) no gesture, (ii) wave, and (iii) handshake. To detect them, we use a K–Nearest Neighbors (KNN) classifier previously trained with gestures toward the robot. The detection is performed in real-time, comparing the time series in the KNN database of training samples with the current window, using the Dynamic Time Warping (DTW) metric. If the result is higher than a threshold, we consider that no gesture was detected. In addition, we also classify a gesture as a close or distance salutation. If the classifier detects a wave or a handshake closer than 2 m we classify it as a close salutation. Otherwise, if it detects a wave at a further distance, we consider it is a distance salutation. We ignore handshake detections of people farther than 2 m.

5.7 Human greeting tracking with Hidden Markov Models

In this section, we propose a model that estimates and keeps track of people's interaction mental states according to Kendon's greeting model. The model considers people's social signals to update its belief of people's current interaction state. This information is noteworthy since people may miss some of the robot's signals during the interaction, leading the robot to reconfirm its intentions by repeating them. Moreover, greeting phases may occur in distinct orders or be absent from the interaction [35], [56], and some social signals may happen in multiple greeting phases. Thus, an estimate of the uncertainty of the current greeting phase is valuable for risk-aware decision-making.

We modeled the estimation problem as an HMM [151] to track greeting interactions. HMMs are probabilistic models defined by a set of states which are not directly observable (hidden) and possible observations whose emission probabilities depend on the current state. For this problem, the hidden states represent the six greeting phases, while the observations are composed of the detected human social signals. We compared two approaches to fit the HMM in this problem: (i) manual computation of the parameters from Kendon's observations and (ii) use of a data-driven approach with observations from video data.

5.7.1 Background: Hidden Markov Models

HMMs [151] are probabilistic temporal models that can be used for state estimation of discrete states. We used a Gaussian Hidden Markov Model (GHMM), in which the observations are real values corrupted by Gaussian noise. The model is composed of a set of unobservable (hidden) states $\mathbf{X}$, a state transition model $P(\mathbf{X}_{k+1} \mid \mathbf{X}_k)$, observations $\mathbf{O}$, and the emission probabilities $P(\mathbf{O}_k \mid \mathbf{X}_k)$.

An HMM with continuous observations is defined as a tuple $\langle \mathbf{X}, \mathbf{A}, \mathbf{O}, \mathbf{B}, \boldsymbol{\pi} \rangle$. The initial state probabilities are represented with vector $\boldsymbol{\pi}$. Matrix $\mathbf{A}$ represents the state transition model, with its entries A_{ij} representing the probability of transition from state i to state j. A GHMM assumes that observation noise follows a Normal distribution. Thus, $\mathbf{B}$ represents the set of emission pdfs described by: i) A matrix $\mathbf{M}$, with the mean values for each observation in each state; and ii) a covariance matrix $\mathbf{C}$ with the covariance of observations in each state.

The parameters of an HMM can be learned using the Expectaction–Maximization (EM) algorithm from a set of observations [152], [153].

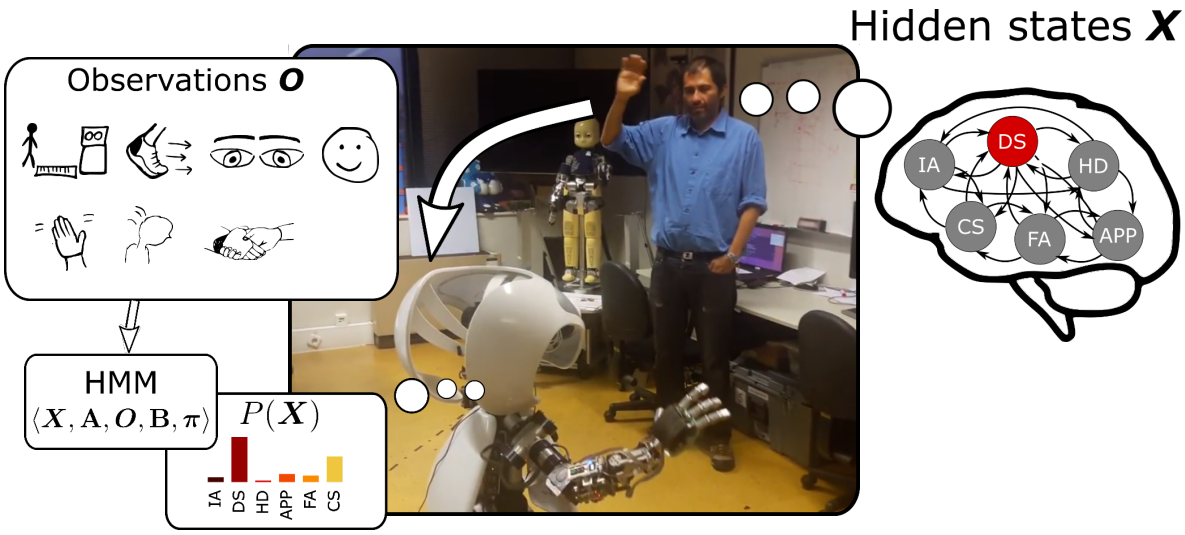


Figure 5.12: Estimation of people's greeting state based on an Hidden Markov Model that encodes Kendon's greeting model and on observed social signals.

5.7.2 Methodology

We modeled the human's mental state of the greeting ritual as a HMM (see Figure 5.12), where each greeting phase corresponds to one hidden state. We assume that the observed features have continuous values and are subject to Gaussian noise to estimate the states accurately, relying on Kendon's reports to choose relevant observations. Thus, the observation vector is composed of the following values:

- *D:* distance from the target to the robot's base
- *V:* norm of the target's velocity in the world frame
- *G:* signals if the target is looking at the robot
- *S:* the probability that the target is smiling
- *G:* three values that represent the occurence of gestures related to the *distance salutation* (DSal), like waving, to the *head dip* (HDip), and to the *close salutation* (CSal), like the handshake.

We used the ROS middleware to implement the proposed solution in the Vizzy robot, making its communication with other pipeline components easier. We restate that we used two approaches to create the HMM: (i) a handcrafted approach and (ii) a data-driven approach. We used the Python package *hmmlearn* [3] for the latter. This node performs the forward algorithm to update the current belief during inference. Then, it publishes the most likely phase to a ROS topic.

5.7.3 Datasets

For the data-driven model, we needed human data illustrating greetings. We used two datasets that contain human-acted greetings for this purpose: the AVDIAR Dataset [154] and the UoL 3D Social Interaction Dataset [155].

[3] https://hmmlearn.readthedocs.io/

The AVDIAR dataset contains stereo sequences of greetings. Additionally, it provides a file with the 2D head position of all the participants in the video and the calibration information of the stereo cameras. To obtain the remaining necessary data, we employed a series of automatic and labeling procedures described in more detail in [24], [156].

The second dataset already provided the 3D position of both greeters at each frame. We used the provided head orientation to compute the gaze direction while using the same procedures to fill the missing data as in the AVDIAR dataset, correcting inconsistent gaze values and labeling the smile and movement features.

After preparing the datasets, we obtained a total of 33 complete greeting sequences with their observations, with a sampling interval of 0.2 seconds.

5.7.4 Model fitting

5.7.4.A The handcrafted model with Kendon's observations

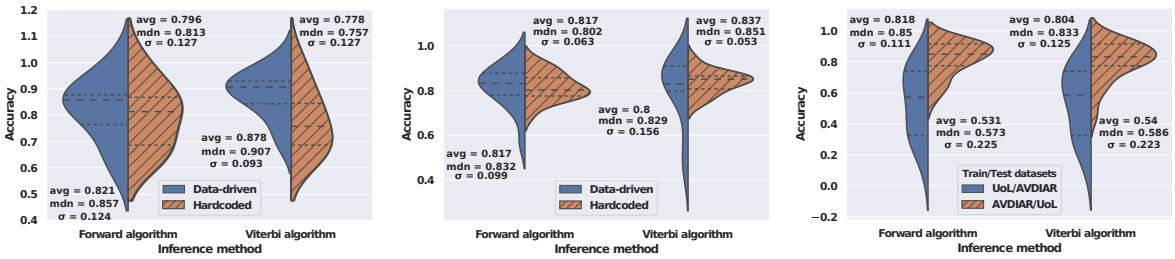
To evaluate the model under different conditions, we used distinct train/test splits. Approach *A* entailed using the longest sequences for training and the remainder for testing for testing (25 %-75 % split). Approach *B* used 15-fold cross-validation. Finally, we also attempted to train and test on separate datasets (approach *C*). These were the input for the EM algorithm, together with the desired number of states (6). We set the algorithm to stop when it reached convergence (log-likelihood lower than 0.01) or reached 100 iterations, returning the parameters for our HMM.

Because the EM algorithm is unsupervised, its output is a model with six states clustered according to the information provided rather than the ordered Kendon's six phases. As a result, we organized the matrices such that each generated state had the label of its most similar phase, in the order of Kendon's model's greeting phases: IA, DS, HD, APP, FA, and CS.

We should note that because the datasets' environments differ from those in Kendon's observations (indoors vs. outdoors), we should expect differences in the distance and speed features.

5.7.4.B Model fitting with a data-driven approach

We used two prediction algorithms to estimate greeting phases with our HMMs, which are fundamentally distinct: the forward algorithm [151] and the Viterbi algorithm [157]. The first is an online algorithm that receives a sequence of observations and iteratively calculates the probability of each state for each index in the sequence. On the other hand, the Viterbi algorithm is an offline method that computes the most likely state path corresponding to the sequence of observations. As a result, we can only use the forward algorithm to predict states in real-time scenarios. Nonetheless, a robot can use the Viterbi algorithm to compare its past actions with the most likely previous state path, detecting its own errors and avoiding repetitions. Observations are processed at a rate of 5 Hz to allow the model to change states without noticeable time gaps.



(a) Experiments with training/test approach *A*: use the longest sequences for training and the remaining for testing.

(b) Experiments with training/test approach *B*: training and testing with 15-fold cross validation.

(c) Experiments with training/test approach *C*: accuracy measurements when training and testing with distinct datasets.

Figure 5.13: Violin plots of experimental results with the UoL and AVDIAR datasets.

5.7.5 Results

We computed greeting sequence estimation accuracies on the test sets of each train/test split approach that we mentioned in subsubsection 5.7.4.B by comparing the estimated state with the labeled one.

Figure 5.13a shows the results for the train/test split of condition *A*. We can see that performing the forward algorithm with the data-driven model tends to outperform the handcrafted one with this train/test split. Nonetheless, we note that a paired samples *t-test* ($\mathrm{t}(7) = 0.584, \mathrm{p} > 0.05$, normality not rejected by a Shapiro-Wilk test on differences, with $\mathrm{p} > 0.05$) did not reveal a statistically significant difference between both methods (likely to happen with few samples). Both models' high accuracy suggests that they could capture the structure of the greeting ritual. The difference in accuracy between the data-driven and handcrafted models may be explained by differences between the environments of the datasets used to fit the model (indoors) and Kendon's observations (outdoors). When the Viterbi algorithm is used, the differences between handcrafted and data-driven models become more apparent, which could be due to the same reason. Even so, a Sign test ($\mathrm{p} > 0.05$, normality of differences rejected by a Shapiro-Wilk test on differences with $\mathrm{p} = 0.015$, distribution of differences not symmetric) did not reveal significant differences between median accuracies.

Figure 5.13b shows the results for train/test condition *B*. As expected, the average and median accuracy of the data-driven model is slightly lower than in the previous condition. As the size of the training set is small, randomly choosing the sequences resulted in some cases where the training data missed important information to identify the six desired states, which resulted in a higher variance in the results.

Finally, Figure 5.13c shows the results of testing and training the data-driven model on distinct datasets, condition *C*. The model fitted in the AVDIAR dataset generalized to the UoL dataset, but not the other way around. This result is consistent with the AVDIAR dataset having longer sequences and thus more information than the UoL dataset.

5.7.6 Conclusions

The results demonstrated that an HMM built using Kendon's notes could keep track of the greeting state even when the test environments differed from those used in Kendon's studies (indoors v.s. outdoors). When the training data contained the longest sequences, fitting the model with a data-driven method improved the results even more. Further tests with training and testing on different datasets indicate that the state estimation model can generalize for indoor scenarios but requires datasets with complete sequences. As a result, while the data-driven model has the potential to adapt to differences in greeting sequences due to different contexts, researchers must ensure that they collect long greeting sequences.

5.8 Planning and behaving

5.8.1 Background: Behavior Trees

BTs [158] are a representation of plans and execution models, whose function is to structure the switching between different tasks (represented as nodes) in an autonomous agent, such as a robot. They are tree like structures composed of control flow, condition, and action nodes. We define a BT following the formulation on [159] that defines it as a directed acyclic graph $G(V, E)$, where V are the tree nodes and E are the directed edges. The root node generates ticks at a given rate that flow through the tree according to the control flow nodes and the result of leaf nodes (which can be either *SUCCESS*, *FAILURE*, *RUNNING*). A BT is reactive, allowing good handling of unexpected changes and errors by being able to check every condition and roll back to a previous task of the sequence, quickly and efficiently. Moreover, they can be composed of modular subtrees which can be developed and tested separately. Example applications include UAV control systems [160] and collaborative robotics with industrial robots [161].

5.8.2 Implemented behaviors

Since we want to have Vizzy following engagement models based on Kendon's greeting model, we developed six modular subtrees. Choosing how to execute each module depends on the model under test, i.e., how decisions are made given a set of social signals and how many of Kendon's greeting phases the model comprises. For that, we developed a ROS package for Vizzy[4] that uses the *BehaviorTree.CPP* library[5]. We describe the behaviors implemented in these six modules as follows:

1. *Initiation of Approach:* the robot orients its body toward the target, followed by a direct gaze.
2. *Distance Salutation:* while gazing at the person, the robot performs a waving gesture.
3. *Head Dip:* the robot performs a movement that consists of setting a head orientation lower than the target's face.
4. *Approach:* the robot starts two parallel branches of the BT. One branch plans and executes a frontal approach toward the target's personal space. On the other branch, it adverts gaze by

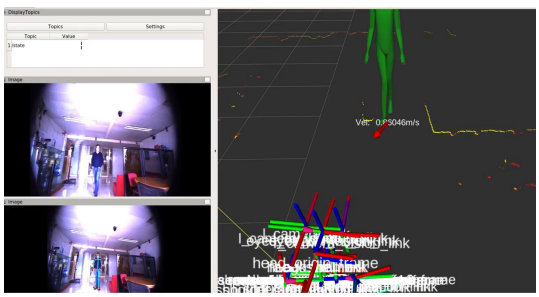
[4] https://github.com/vislab-tecnico-lisboa/vizzy_behavior_trees
[5] https://github.com/BehaviorTree/BehaviorTree.CPP

looking at a position lower than the target's face.

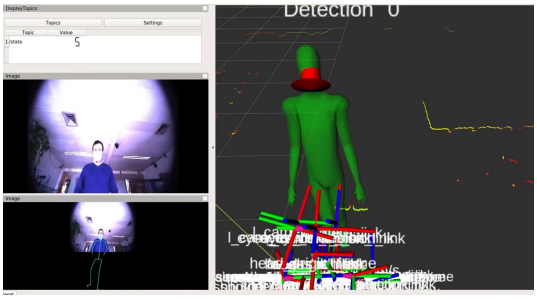
5. *Final Approach:* the robot uses the same approaching logic as in the previous phase, however, it changes its gaze display to look directly at the target.
6. *Close Salutation:* the robot executes a salutation (wave, handshake, other), while combining it with a direct gaze and a verbal greeting.

Regarding navigation, the robot uses two distinct approaches. When the navigation goal is a pose on the map (like returning to the initial condition), the robot uses the *move_base* planner with the *A** algorithm for global planning and the *Elastic Bands* algorithm for local path planning and control. When approaching people, the robot uses a purely reactive controller based on the algorithm proposed by Noubakhsh and Siegwart [162], with the robot stopping if an obstacle was too close. We could not use *move_base* to approach moving people since the planner was too slow to keep up with people's speed toward the robot.

5.9 Discussion and conclusions



(a) Preliminary test subject at the starting the greeting with the robot.



(b) Preliminary test subject during the final approach

Figure 5.14: Visual example with the complete pipeline running. Each figure represents a distinct point in time where a test subject initiated the interaction with Vizzy. In the top left corner of each image, the HMM estimated the current human greeting phase (numbered from 1 - IA to 6 - CS). Each picture shows the person's body pose in green and the head as a red arrow. Another red arrow at the person's feet represents the estimated velocity. Finally, we can see the skeleton completer in the bottom left of Figure 5.14b, which was crucial for the pipeline.

This Chapter presented a pipeline for autonomous human-robot greetings. As seen in Figure 5.14, the architecture can detect and track people, their poses, their velocities, gestures, and the current greeting phase. We could run it in a single machine equipped with an Nvidia GTX 1070 GPU at around 12 frames per second. Regardless, creating this architecture was a challenging task. We observed that each module is subject to noise and amplifies errors from previous modules, requiring careful filtering and tuning. It can be frail in scenarios with crowds and non-uniform illumination. We also noticed that image suppression needs to be carefully tuned. Otherwise, we risk discarding too many images and losing relevant social signals. This issue is especially relevant with time-based reliant information (like the gesture detector) or the computation of people's velocities.

Nonetheless, under normal circunstances, this pipeline allows a mobile social robot to perform Kendon's greeting model.

6

DISRESPECT FOR SOCIAL SCRIPTS: SELF-PERCEPTION OF HUMAN–ROBOT INTERACTION FAILURES

Neither humans nor robots can always comply with social scripts and expectations. Even though they are paramount interaction guidelines for interaction between social agents, they are susceptible to failures related to perception, judgment, behavioral errors, or in the robotics case, TFs. We argue that, during first encounters, interaction errors are even more likely to occur given the lack of information between parties. For instance, even though models like Kendon's greeting model can guide a robot to open the interaction with a target person, due to the considerable complexity of the set of exchanged social signals and culturally dependent gestures, two types of error are bound to happen. First, the group of norms the robot follows may be inappropriate for people interacting with it, e.g., due to cultural differences or personality mismatches. Breaking people's expectations this way constitutes a SNV. Second, due to sensors, hardware, and software issues, the robot's ongoing actions may fail to complete making it behave erratically. These failures are classified as TFs. Giuliani et al. [120] proposed this nomenclature. Since these events are bound to happen and may hinder or even invalidate interactions, it is of paramount importance that social robots can be aware of them and react accordingly. This information complements social norms and scripts, making the interaction more robust to failures.

In this Chapter, we worked towards creating the self-perception of robot interaction failures by developing automatic error detectors and classifiers. We did so in two modalities. First, we focused on detecting whether an interaction error occurred during a specific phase of the greeting protocol: the close salutation with the handshake. We used the robot's tactile sensors to detect whether it grasped a human's hand successfully or not. The detector consisted of a KNN classifier that compared multidimensional time-series information from the robot's tactile sensors using DTW.

In the second part of this Chapter, we focused on non-verbal visual information that a robot

can extract with its own sensors and its actions. However, such systems need rich datasets of ecologically plausible human reactions and self-reported feedback. Current datasets of expressions [163]–[167] do not contain human self-reports and reactions to robot actions. Other systems gathered laughter reactions to a comedian robot's jokes [168]. Although several works regarding human-robot expressions exist[169], we did not find any works or available datasets with reactions and self-reports to robot constructive/destructive behaviors. Thus, we designed a HRI study to collect non-acted human reactions to Vizzy's actions in a controlled environment. It consisted of a cognitive board game task where the robot supported the player with speech and gestures. While in one condition, the robot had a kind personality, and its help was critical to making the participant win, in the other condition, the robot was grumpy and ultimately responsible for the participant's loss by clumsily destroying the whole progress. We used a post-experiment questionnaire mainly for manipulation check and to get people's self-reported feedback on the experiment.

Afterward, we designed an algorithm to detect and classify Vizzy's failures from users' reactions and the robot's actions. Mobile social robots need to interact with people in the wild, without personal information, relying only on their own sensors, and having resource constraints. Furthermore, noise and phenomena like the cocktail party effect [170] render audio signals useless in uncontrolled scenarios. Thus, we focused on non-verbal visual data that a robot can extract with its own RGB sensors and logs of its actions to design an algorithm that detects and classifies failures from users' reactions and the robot's actions. We built our proposed method upon known features as classifiers from HRI and machine learning that, as far as we know, have not been applied together for this problem. We focused on the user's head features (head pose, gaze direction, FAUs, and emotions) and the robot's speech and arm gestures. The algorithm uses a cascade of Random Forest algorithms that first detect when an error occurs and then classify it as SNV or TF on a frame-by-frame basis.

This Chapter is an extended version of a publication in Paladyn, Journal of Behavioral Robotics, the Late-Breaking Report for the 2020 Human–Robot Interaction conference, and the EgoVIP workshop of the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems:

J. Avelino, T. Paulino, C. Cardoso, R. Nunes, P. Moreno, and A. Bernardino, "Towards natural handshakes for social robots: Human-aware hand grasps using tactile sensors", *Paladyn, Journal of Behavioral Robotics*, vol. 9, no. 1, pp. 221–234, Aug. 2018. DOI: 10.1515/pjbr-2018-0017.

J. Avelino, A. Gonçalves, R. Ventura, L. Garcia-Marques, and A. Bernardino, "Collecting social signals in constructive and destructive events during human-robot collaborative tasks", in *Companion of the ACM/IEEE International Conference on Human–Robot Interaction*, Cambridge, United Kingdom (Online), Mar. 2020.

F. Loureiro, J. Avelino, P. Moreno, and A. Bernardino, "Detecting human–robot interaction failures through egocentric visual head-face analysis", in *EgoVIP - Egocentric vision for interactive perception, learning, and control, Workshop at IROS 2021*, Prague, Czech Republic (Online), Oct. 2021.

In addition, we submitted part of this Chapter to the International Conference of Social Robotics as:

F. Loureiro, J. Avelino, P. Moreno, and A. Bernardino, "Self-perception of interaction errors

through human non-verbal feedback and robot context", in *International Conference on Social Robotics*, Florence, Italy, Dec. 2022.

6.1 Self-perception of mistakes: handshake failures

This section describes our proposed algorithm to detect whether the robot grasped a human hand from its tactile sensors. Then, it shows our evaluation methods, where we used distinct non-hand objects. Unlike the second part of this Chapter, in this part, the robot does not account for people's social feedback. We claim that this error detector is a verification mechanism for social scripts, specifically, the handshake salutation.

6.1.1 Methodology

We formulate this task as a problem where the robot needs to discriminate if the multidimensional time series of tactile sensing data corresponds to a grasp on a human hand or another object. The data consists of the temporal sequences of tactile sensing readings from 11 hand sensors.

Our approach consisted of a supervised machine learning method, the KNN algorithm. Since each sample had a different duration, which was uncontrolled, we used the DTW [171] algorithm to compute a comparison metric between them. The DTW is a method that achieves good results in classifying time series in small datasets [172], which made us believe that it is appropriate since the dataset of handshakes was the one we collected in the pilot study of Chapter 4 (section 4.3). Thus, the number of handshake samples was insufficient to estimate the parameters of advanced learning algorithms such as deep neural networks, avoiding over-fitting the model.

During our experiments, we used two different classifiers: one based on forces (N) and another based on fields (Oe) of each of the eleven sensors. The first used the time series X, Y, and Z cartesian components of the force, while the second used the time series of the three components of the magnetic flux.

6.1.2 Dataset

The dataset is composed of temporal sequences of tactile sensing readings from the 20 participants of the pilot study of section 4.3 and from 11 (non-hand) objects. We collected the data during the whole grasp and release, storing the raw value of the magnetic flux (Oersted, Oe) and the force (N) estimated from the magnetic flux. Since each participant experienced three handshake primitives (*weak*, *medium*, and *strong*) we had a total of 60 points with the "hand" label. We also executed the same three primitives on each object and three empty grasps, ending up with 36 grasps with the "no hand" label. The selected objects are shown in Figure 6.1, which covered both rigid 0 and 4 and deformable objects 1, 2, 3, 5, 6, 7, 8 and 9. It is also worth mentioning that these objects did not have magnetic properties, which could influence the results due to the principles behind the functioning of the used sensors.

For training and evaluation, we randomly sampled 49 and 26 object grasps for training (80 % of the initial dataset), and 11 handshakes and 8 "no hand" grasps for testing (remaining 20 % of the

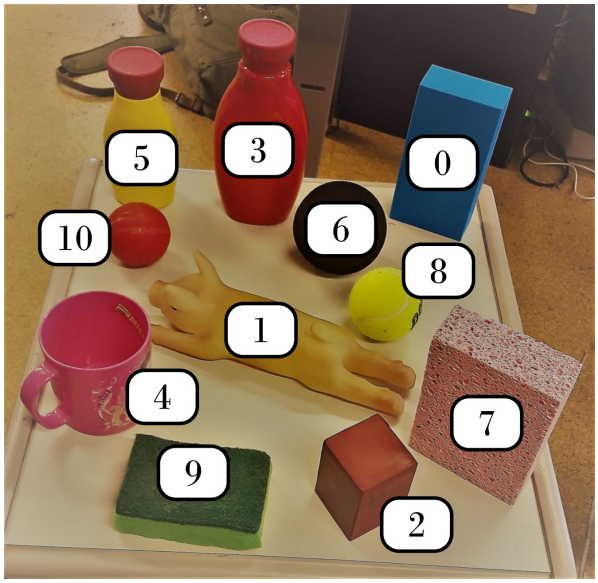


Figure 6.1: Objects used as the "no hand" class.

dataset). During each split, we forced the two empty grasps that resulted from the *weak* and *strong* primitives to be part of training and the *medium* to be part of testing. The goal was to have training data for the special cases where the sensors do not contact anything.

6.1.3 Hyper-parameter tuning and cross-validation

To tune the hyper-parameter K, we perform 7-fold cross-validation and choose the K (on a set from 1 to 27, where K is odd) that yields the lowest average miss-classification error. We use 7 folds because it is the greatest common divisor of the number of handshakes (49) and the number of "no hand" events (28) for training, allowing us to have easy splits.

6.1.4 Results

Given the small size of the dataset, to test our method appropriately, we performed 15 iterations of the following procedure:

- Randomly split the complete dataset into training and test sets as described in Subsection 6.1.2.
- Perform the 7-fold cross-validation on the training set to find the optimal value for the K parameter.
- Test the classifier trained on the complete training set on the test set using the optimal K.

The resulting values of the evaluation are summarized in Table 6.1 and 6.2. We can see that both classifiers yield inspiring results, with the field-based classifier outperforming the force-based classifier. We can infer from these results that the field-based classifier failed 3 of 19 test items in the worst scenario, and the force-based classifier failed 6 out of 19 test items in the worst case scenario.

Table 6.1: Handshake classification results for the field-based classifier on each test run, and the overall miss-classification error mean.

Iteration	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Mean
Optimal K	1	1	3	1	1	3	1	1	1	3	1	1	3	5	3	-
Miss-class. error	0.1579	0	0.1053	0	0	0	0.0526	0	0	0	0.0526	0	0	0.16	0	**0.0351**

Table 6.2: Handshake classification results for the force-based classifier on each test run, and the overall miss-classification error mean.

Iteration	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Mean
Optimal K	11	11	21	13	23	17	13	13	13	13	11	13	9	1	11	-
Miss-class. error	0.21	0.31	0.16	0.16	0.16	0.26	0.32	0.16	0.16	0.21	0.21	0.05	0.1	0.1	0.19	**0.19**

6.1.5 Conclusions and discussion

This study showed promising results in the task of detecting whether the robot grasped a human hand or not during the handshake. Overall the field-based classifier showed better performance than the force-based classifier. However, it does not generalize (like the force-based classifier) to other systems since the magnetic fields depend on each sensor's construction. In future works, we intend to use additional sensors from the robot's fingers and the future sensors of the hand palm. Additionally, we want to create a larger dataset with more handshakes and a wider variety of objects. We note, however, that KNN with DTW becomes computationally expensive when increasing the training set size and feature size. Thus, future works may need to invest significant time in feature selection. Larger and richer datasets may also open the door for more sophisticated methods, which may be an interesting path to follow.

6.2 People's reactions to others' actions: signals of social feedback

As seen in the previous section, a robot can perceive interaction errors by programming it to compare the expected outcome of its actions against the signals measured by its sensors. However, social scripts change over time, and thus do people's expectations and behaviors. Learning and adapting one's behavior in a society is a fundamental mechanism when establishing and evolving the social contract. Mutual knowledge of how one's behaviors impact others maintains the social order and individual's protection [57]. This knowledge is present in the ability to predict others' feelings and grows according to their feedback. Facial expressions and body movements are much more than a representation of an internal state of mind. They actively give feedback to other individuals, regulating their behavior. As soon as 12 months of age, human infants can perceive facial expressions and use their signals to act [58]. Past studies argue that the human amygdala processes facial expressions and uses them to reinforce emotional learning [173], hitting that head and facial features are pivotal for this task. Albeit critical, these features can also be ambiguous since people may react to events unrelated to the robot or can be affected by people's moods. Thus, we believe that the robot should consider the context of the interaction, emphasizing its actions when evaluating the

situation.

6.3 Collection of a dataset of human reactions to constructive and destructive robot actions

Such a system requires a dataset of rich and ecologically plausible human reactions. Still, current datasets are either collected using actors or lack information about robots' actions and context. For instance, the Multi-PIE [163] and CK+ databases [164] have a rich collection of facial expressions collected from several different views but lack spontaneity since people were acting. The CK+ authors also recorded additional spontaneous smiles but without context. Other datasets, like the DISFA+ [165], the MMI datasets [166] and the BP4D [167] contain spontaneous expressions and some self-reports. But none result from interactions with an embodied entity, contain any valence information regarding one, or the actions that originated said reactions.

Since, as far as we know, there was no systematic, reliable, and efficient method to collect the necessary data, our first step was to develop a systematic data collection methodology based on HRI during a collaborative task.

6.3.1 Requirements for data collection methods

To successfully collect quality data, we set the following set of experimental requirements:

R1 Reliability, repeatability, and systematicness.

R2 Recorded data resembled the one "sensed" in other environments, so that developed algorithms can be used in the real world.

R3 Social rewards were expressed naturally: more specifically, we expected approval expressions with smiles and gestures and disapproval gestures and facial expressions that revealed some degree of frustration.

R4 The existence of a tool that can label the social reward associated with the audio-visual data.

R5 A way to record the robot's actions.

Complying with these requirements is challenging since a controlled environment usually implies a laboratory experiment, but it hinders ecologically plausible reactions from people.

6.3.2 Experimental design

To collect ecologically plausible data on human reactions to robot behaviors and associated valence, we developed a dyadic between-subjects design interaction paradigm where a human-robot team plays a cognitive board game for the chance of winning a bar of chocolate. The experiment happened in a closed room where the participant was left alone with the Vizzy robot (Figure 6.2). We used a touch-screen laptop and a board game with 14 hexagonal prism blocks, detected with computer vision. The laptop tracked the game state in real-time with an external webcam. It also

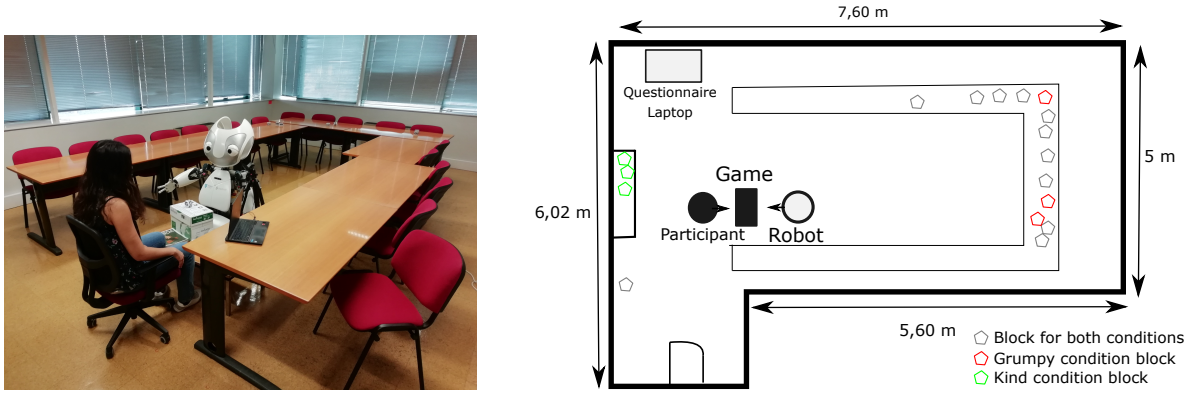


Figure 6.2: Room and block distribution at the beginning of the experiment.

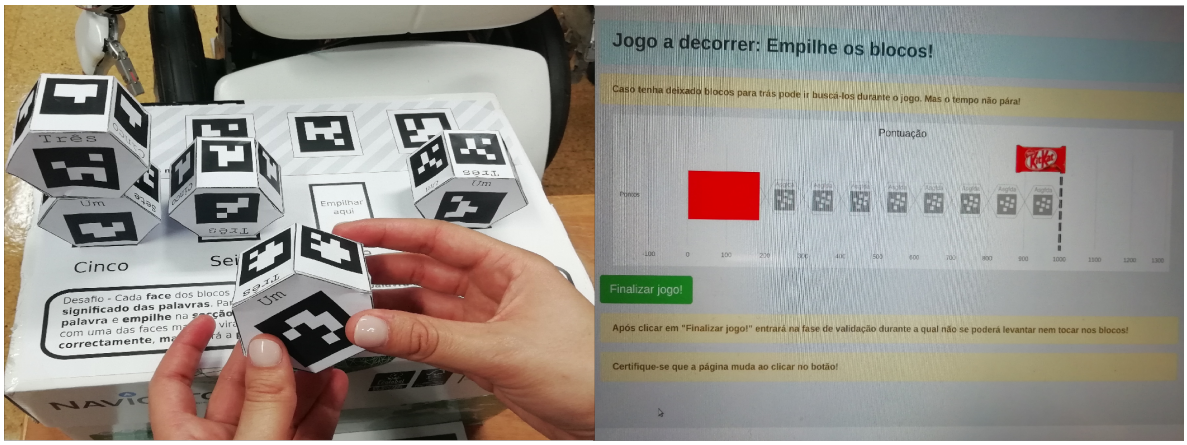


Figure 6.3: The cognitive game used in this experiment: the board, individual blocks and the score interface

provided audio-visual feedback to the team through an on-screen animated bar (current score), the number of needed blocks, and 8bit sounds (Figure 6.3). The robot detected the participant and got all the game's data (through Wi-Fi). It was fully autonomous in this experiment and could speak, move its arms, and gaze at targets. We recorded the scene with (i) the laptop's embedded webcam, (ii) the robot's right-eye camera, and (iii) the robot's chest camera.

6.3.3 Main activity

The proposed activity had three phases.

1. **Preparation:** the participant had 30 s to collect scattered blocks in the room (totaling 14). The robot tracks time and alerts the participant to go back before running out of time.
2. **Game:** the participant had to stack blocks according to their total number of faces with a written word. The board had four areas (5, 6, 7, and 8). Thus, the number of faces with words varied between 5 and 8 per block. The team lost points with time and won/lost 100 points per correct/incorrect block. One could fetch forgotten blocks, but time did not stop. It was impossible to win with less than 11 blocks.
3. **Validation:** the block construction needed to remain stable for 1 min, or the team would lose.

There were two conditions during the data collection experiment, where the robot had two distinct personalities: **(i) Kind Robot** and **(ii) Grumpy Robot**. These conditions differed in the way the robot communicated with participants ("Kind": supportive, team player; "Grumpy": selfish, disdainful) and in the robot's role in the game's outcome. The "Kind Robot" made the person win a seemingly lost game, while the "Grumpy Robot" made the person lose an easy-to-win game (with a critical action). We manipulated the game's difficulty by scattering the blocks differently before starting the experiment, as follows:

- **Kind:** there were 10 blocks on the tables (easy to see) and 3 blocks on a whiteboard (on the left of Figure 6.2). The 3 blocks were visible, but we used attention tricks to make them stay unnoticed: block height [174], cognitive overloading (timed task) [175], and lack of saliency [176]. Nonetheless, if the participant found them, the robot became "Grumpy".
- **Grumpy:** there were 13 blocks on the tables (easy to see).

The 14^{th} block was above the whiteboard for both cases, and it was hard to see. The robot's critical actions were:

- **Kind robot's critical action:** the participant puts their last block and notices it is impossible to win. The robot looks around and points at the missing blocks. With them, they can win.
- **Grumpy robot's critical action:** the participant had enough points to win and proceeded to the validation phase. The robot pointed at the 14^{th} block, mocked the participant, and clumsily knocked down all the blocks. The team lost, and the robot blamed the participant.

After the activity, participants rated the robot's behavior (1-bad to 4-good) and answered if their score was enough to win the game before and after the validation phase (error check). Then, they filled up a questionnaire.

6.3.4 Questionnaire

The questionnaire was composed of two parts: a) manipulation check and b) robot evaluation. The objective of part a) is to check whether the experiment events occurred as planned and if the participants perceived that the robot's actions were critical for the game's outcome. Part a) was composed of the following items:

- "Did you win the prize?"
- "When clicking on the *Submit* button, was the score high enough to win the prize?"
- "At the end (after the validation of 1 min) was the score high enough to win the prize?"
- "I would be able to win the prize if the robot was not present (1 to 6) Likert-like scale: 1 = Totally disagree - 6 = Totally agree."

In the questionnaire's part b), we evaluated people's perceptions of the robot. Our goal here was to gather valence information that could be used, in future works, to infer people's perceptions of the robot from their reactions. All items are shown in Table 6.3. First, we evaluated the perceived

likeability of the robot with items extracted from the Godspeed questionnaire [73]. We evaluated the Perceived Competence dimension with items from the RoSAS questionnaire [74]. For the Social Attraction dimension, we use items proposed by a previous paper on robot personality [177]. We used the individual items also in Table 6.3 to get finer insights into the robot's personalities. Finally, the questionnaire tested for "connection" between the participants and the robot through the "Inclusion of Other in the Self (IOS) Scale" [75]. This scale uses a set of pictures with two circles representing the emotional overlap between robot and participant.

Likeability (1-6 bipolar scale, $\alpha = .91$)			Social attraction (1-6 Likert scale, $\alpha = .94$)
Dislike (1)	...	Like (6)	I think I could spend good moments with this robot
Unfriendly (1)	...	Friendly (6)	I think I could establish a good personal relationship with this robot
Unkind (1)	...	Kind (6)	I would like to spend more time with this robot
Unpleasant (1)	...	Pleasant (6)	
Awful (1)	...	Nice (6)	

Proposed traits (1-6 Likert Scale)	Perceived Competence (1-6 Likert Scale, $\alpha = .95$)
Annoying	Capable
Honest	Responsive
Helpful	Interactive
Preachy	Reliable
Cooperative	Competent
Irritating	Intelligent
Awkard	

Table 6.3: Questionnaire items related to the robot.

6.3.5 Rationale

Most of the robot's behaviors were event-driven and reactive to the game state and movements. The justification for using an autonomous robot instead of a Wizard-Of-Oz approach is threefold. First, once tested, the system's timings and behaviors will be more accurate and predictable than a human operator's. With this decision, we sacrifice flexibility but guarantee repeatability and variable control. The experimental session occurred in a spacious closed room where each participant interacted with the robot alone. With these decisions, we meet requirement R1 (reliability and repeatability). For the collected data to be usable in other environments (R2), we captured it through the robot's onboard cameras. Obtaining natural reactions is a challenge for laboratory experiments. To address it, we took several steps to induce the participant to react. The most obvious decision was

to have a real robot with a physical game instead of a virtual avatar, leveraging the effects embodiment [5], [178]. Then we purposely bias the person into liking or disliking the robot through one of its personalities. The objective of this bias was to increase the reactions' amplitude. Finally, the robot's critical actions occurred during two pivotal conditions: a) when participants think the goal is no longer attainable, and b) when they think winning is guaranteed. We hope to address requirement R3 (natural expressions) with these design choices. We tackle requirement R4 (social reward labeling) using the robot behavior classification prompt and the questionnaire and requirement R5 (record robot's actions) by recording speech, head, and arm movements in real-time.

6.3.6 Participants

We collected data from 24 participants aged 17 to 29 years old ($\mu = 21.54, \sigma = 2.93$), 15 male and 9 female. 14 participants experienced the "Kind robot" condition and 10 experienced the "Grumpy robot" condition. There were 6 participants whose responses to the error check questions were unexpected. 5 participants on the "Kind robot" condition did not follow the rules. In one "Grumpy robot" experiment, the robot failed to knock down the blocks. Hence, we excluded these samples, resulting in 9 participants per condition.

6.3.7 Results

6.3.7.A Quantitative data analysis

We summarize the questionnaire results in Figure 6.4. We tested the data for normality with the Shapiro-Wilk test, using parametric tests for data that does not violate the normality condition. Every "Kind robot" participant rated the robot's behavior with the maximum value (4). The "Grumpy robot's" rating had the following descriptive statistics: $\mu_{\text{grumpy}} = 2.67, \sigma_{\text{grumpy}} = 1.23$. Given the null σ of "Kind robot" data, we used the One-Sample Wilcoxon Signed Rank Test, with significant results. However, the positive rating for the "Grumpy robot" was a surprise. A Mann-Whitney U for "winning without the robot" ($\mu_{\text{kind}} = 2.44$, $\sigma_{\text{kind}} = 1.81$, and $\mu_{\text{grumpy}} = 5.11$, $\sigma_{\text{grumpy}} = 1.17$) showed that the robot's impact on the game was understood. We averaged the items for perceived likeability ($\alpha = 0.91$, $\mu_{\text{kind}} = 5.49$, $\sigma_{\text{kind}} = 0.49$, and $\mu_{\text{grumpy}} = 3.71$, $\sigma_{\text{grumpy}} = 0.99$) and performed an Independent samples t-test that showed significant differences. We also averaged social attraction items ($\alpha = 0.94$, $\mu_{\text{kind}} = 4.59$, $\sigma_{\text{kind}} = 1.51$, and $\mu_{\text{grumpy}} = 4.25$, $\sigma_{\text{grumpy}} = 1.52$), and an Independent samples t-test showed no significant differences. Finally, we did not obtain significant differences in the IOS scale ($\mu_{\text{kind}} = 3.78$, $\sigma_{\text{kind}} = 0.83$, and $\mu_{\text{grumpy}} = 3.44$, $\sigma_{\text{grumpy}} = 1.33$) with a Mann-Whitney U test.

6.3.7.B Qualitative data analysis

In qualitative video analysis, we could identify less expressive reactions to the "Kind robot" with smiles (Figure 6.5a), neutral expressions, and acknowledgment gestures (Figure 6.5b). Reactions to "Grumpy robot" were more extreme, with laughter, shocked faces (Figure 6.5c), and perplexed hand

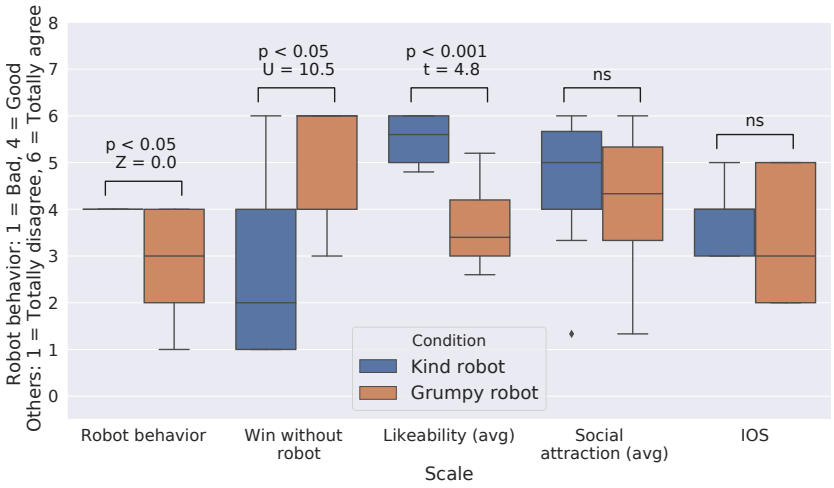


Figure 6.4: Questionnaire results.

gestures (Figure 6.5d). However, a considerable number of participants continued the game without noticeable expressions.

6.3.8 Final considerations on the data collection methodology

Results showed that people recognized the robot's role in winning/losing the game and that the "Grumpy robot" was less likable than the "Kind robot," as expected and desired. Videos also showed some expected behaviors in both conditions. However, people rated the "Grumpy robot's" behavior higher than expected, with positive (> 2) median and mean values. Social attraction, IOS, and videos suggest that, somehow, many people enjoyed interacting with the "Grumpy robot." We guess that "Grumpy robot"'s unexpected ill manners are more characteristic of humans than robots and might charm people. We intend to verify and devise strategies to mitigate this effect (like having a better prize).

This procedure has, however, some limitations. The first limitation is that each session is quite time-consuming, taking around 20 min. The second one is that these reactions were limited to one robot. While people could, in theory, react the same way to the actions of other humanoid robots, it would be interesting to either confirm this hypothesis or gather such data to help algorithms that will use these data to generalize. Third, the scope of the interaction failures is limited to a one-to-one HRI, where the robot destroys/helps people's progress. There are many other situations where the robot can cause SNVs and TFs that require researchers' attention, like invading someone's personal space or colliding with people and obstacles. Would people use similar reactions in those situations? Thus, we must find ways to bootstrap the data collection methodology while fulfilling the requirements.

6.4 Dataset labels and description

During the remainder of this Chapter, we will focus on the subset of negative social feedback (i.e., people's reactions to the robot's negative behaviors). Since the goal was to create an error detection

(a)

(b)

(c)

(d)

Figure 6.5: Example reactions to the critical action of the "Kind robot" (Figure 6.5a) and Figure 6.5b and to the critical action of the "Grumpy robot" (Figure 6.5c and Figure 6.5d), captured through the robot's chest camera.

and classification algorithm from the robot's egocentric perspective (i.e., using the robot's onboard cameras), we did not use video data recorded on the laptop.

Even though the experiment was highly structured (i.e., we knew at which instant the robot misbehaved/helped the user), we still needed to label events as SNVs or TFs for three main reasons. First, we did not know the exact time intervals when people reacted, even though we knew when the robot acted. Second, this work focuses on detecting robot failures from human non-verbal social feedback. Thus, we needed data where people did indeed react to the robot, discarding events where people ignored/did not notice its behaviors. Finally, we intended to find additional SNVs and TFs from unplanned events which we could leverage.

Each annotation contains failure type and the time interval of people's reactions, beginning as soon as we detected that people started reacting and ending when it began to fade. We considered two events as TF: (i) when the robot's arm destroyed assembled blocks and (ii) when there were unanticipated Text–To–Speech (TTS) issues that confused the participants. As for SNVs, they were composed of rude robot speech utterances. The remainder of the dataset was labeled as *No error*. We have acquired approximately 4 h of data. Data collected from the robot's RGB camera consisted of 10 452 frames with SNVs, 20 848 frames with TFs, and 169 107 frames without errors. We note that SNVs and TFs can occur simultaneously.

Since the number of *No error* frames is significantly higher than the number of failures and the dataset is small, we augmented the number of error frames by performing horizontal image flips on those frames.

6.5 Error detection and classification

6.5.1 Related work

Previous studies have addressed how people react to robot failures in the last decade [115], [116], [120], [179]–[182], attempting to identify relevant features. They focused on joint task execution during cooking or building tasks [116], [180], [181], identifying gaze shifting, head movements, facial expressions, and speech as prominent features. In addition, people's emotions evolve with human-human and HRI experiences, which they express through their faces. While some studies noticed how people changed their mood/emotions according to what they were experiencing with the robots [180], [181] along the experiments, as far as we know, people have not developed automatic error detection and classification algorithms using them. However, these social signals can be ambiguous, with the same expression having two possibly opposite meanings (e.g., laughter can be related to both positive and negative feedback [183]). Thus, we believe that context is relevant to disambiguating such cases. However, we do not know past studies that investigated this hypothesis.

Regarding automatic error detectors proposed in past works, Trung and colleagues [115] tested distinct combinations of classifiers fed with 3D coordinates of head, shoulders, and neck. They collected and described their data collection method in their previous work [116]. The goal was to use human reactions to detect robot failures and classify them as TFs or SNVs. They tested

Table 6.4: Summary of error detection and classification features and characteristics related to the scope of this work: automatic detection and classification of interaction errors and SNVs and TFs using non-verbal features through the robot's onboard sensors.

	Gaze	Head movements	Body movements	Speech	Emotions	FAU	Hand gesture	3D body pose	3D head pose	Robot's actions
Tested on automatic systems	Yes [182], [184]	Yes [182]	No	Yes [182], [184]	No	Yes [184]	No	Yes [115]	Yes [182]	No
Used to detect errors	Yes [182], [184]	Yes [182]	No	Yes [182], [184]	No	Yes [184]	No	Yes [115]	Yes [182], [184]	No
Used to classify as SNV and TF	No	No	No	No	No	No	No	Yes [115]	Yes [182]	No
Non-verbal features	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes
Easy to extract with onboard small sensors	Yes [102]	No ([182] used a hat with markers)	Yes [97]	No (subjects in [182], [184] used a microphone)	Yes (section 6.5.2)	Yes [102]	Yes [185]	No ([115] used an external Kinect)	Yes [102]	Yes

several classic classifiers (KNN, Naive Bayes, Random Forest), obtaining accuracy results higher than 90%. However, they noted that the test set contained the same people from the training set, claiming that the classifier could be used in real-life scenarios if the robot has seen those subjects before. Kontogiorgos et al.'s recent work [184] used lexical, visual, and acoustic features to detect conversational failures. They also classify them as 5 types related to their conversational task, but not as SNV or TF. Their visual features were gaze, FAU, and head pose.

Table 6.4 summarizes some characteristics of features used in related works with automatic failure detectors and classifiers, relating them to the scope of this thesis.

Thus, the **baseline** of our study used a Random Forest algorithm with the non-verbal features previously used in automatic error detectors and classifiers [115], [182] that the robot's onboard sensors can extract. Therefore, the baseline features are the 3D head pose (position and roll/pitch/yaw) and gaze features (one normalized direction vector per eye and the average x/y direction in radians). Although Kontogiorgos and colleagues [184] used FAUs in their recent study, we did not consider them part of the baseline since we intended to test them against emotions during the ablation studies.

6.5.2 Proposed pipeline

We propose a system that detects and classifies robot interaction errors according to the taxonomy of Giuliani et al. [120], i.e., as SNVs or TFs. Since our goal is to use this algorithm in the real robot in future real-world scenarios, we followed an egocentric perspective, relying only on the robot's onboard sensors. We propose a two-step model that first detects errors and then classifies them on a frame-by-frame basis. Both detector and classifier are Random Forests that use visual head/face features composed of OpenFace's [102] extracted features and features related to the robot context.

The first set of features contains the baseline features used by previous automatic error detection: 3D head pose (position and roll/pitch/yaw) and gaze features (one normalized direction vector per eye and the average x/y direction in radians). Additionally, we introduced FAUs (a vector of 17 Action Unit intensities and 18 Action Unit presences [102]) and Emotions (the dominant emotion of Ekman's [186] set or neutral). For emotion detection, we propose our "AverageAU" method that uses

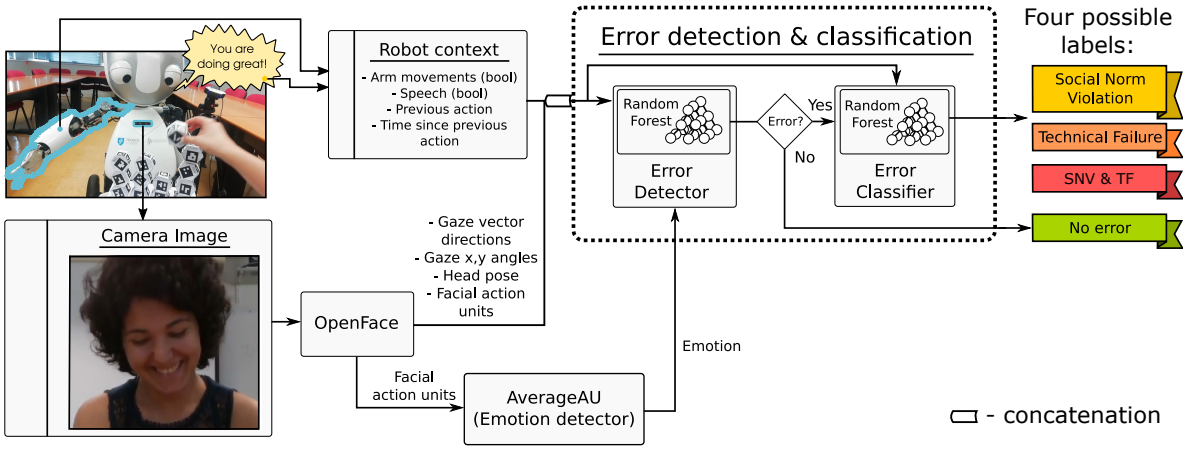


Figure 6.6: Proposed frame-by-frame system for error detection and classification using the robot's onboard sensors. The two sources of information are the robot's camera and proprioception. We extracted visual features with OpenFace and concatenated them with the robot context features of arm movements and speech. In addition, we detect the current emotion from FAUs. The system performs error detection and classification in two steps. First, a Random Forest classifier fed with all features detects whether an error occurred. If it did not detect one, the system outputs the *No error* label. Otherwise, a second Random Forest classifier uses all features except the user's emotion to classify the error as *SNV*, *TF*, or both at the same time.

the association between FAUs proposed in [187]. We can represent this method with Equation 6.1:

$$\hat{m} = \begin{cases} m^* = \arg\max_m f(m) = \frac{\sum_{n=1}^{N_m} \mathbf{I}(\mathrm{FAU}_m(n))}{N_m} & \text{for} \quad f(m^*) > \tau \\ \text{Neutral} & \text{Otherwise} \end{cases} \tag{6.1}$$

where $\hat{m}$ is the estimated emotion, $\mathbf{I}(j)$ is the intensity of FAU j, $\mathbf{FAU}_m$ is a vector of size N_m with the ids of action units associated to emotion m, and τ is a threshold. Table 6.5 lists the FAUs related to each emotion, according to Ekman [187]. For instance $f(m = \text{Hapinness}) = \frac{\mathbf{I}(6)+\mathbf{I}(12)}{2}$. Our previous experiments found an optimal value of 0.8 for τ. This method allows us to leverage data extracted with OpenPose without needing additional computationally intensive emotion detectors. Plus, it achieved 74.92 % accuracy in the CK+ dataset [164]. It also achieved 94.3 % and 75.42 % accuracy in two representative cases of our dataset. Further evaluation of emotion detectors is outside the scope of this thesis. For more information please check the Master's Thesis of Fernando Loureiro [188]. Since we compute the emotions from the FAUs, one could say that we are repeating features unnecessarily. However, by computing emotions, we added structural information that the machine learning algorithms do not need to learn, thus making the learning procedure potentially easier. We note that unlike Kontogiorgos et al. [182], we did not use the accumulated number of head movements over a time window since this information was prone to noise when using OpenFace. Unlike Kontogiorgos et al. [182] our goal was to have a system that could operate without external motion capture markers. Moreover, we also did not use their speech features (word count/response time/questions) since (i) we focus on non-verbal features and (ii) speech recognition is unreliable in non-controlled environments (like public places).

Table 6.5: FAUs used to detect each basic emotion according to Ekman [187].

m	Anger	Disgust	Fear	Happiness	Sadness	Surprise
$\mathbf{FAU}_m$	4, 5, 7 and 23	9, 15 and 16	1, 2, 4, 5, 20 and 26	6 and 12	1, 4 and 15	1, 2, 5 and 26

Our second set of features is related to the robot context. In fact, Kontogiorgos et al. [182] highlight the role of contextual features to which their annotators had access, unlike the automatic system. And a subset of the overall context is composed of the robot's actions. Indeed, if we think about HHI, we usually reason about our actions when we receive negative feedback from someone. Thus, we introduced the following robot context features: (i) current action, (ii) previous action, and (iii) time since the previous action. We encode actions as two boolean vector entries that signal whether there was an arm movement, speech, or none. The current and previous actions can consist of speech and arm movements simultaneously.

As shown in Figure 6.6, the first step is to extract and compute the complete set of visual and context features, which we concatenate. Then, a Random Forest error detector uses all features to classify them as *error* or *no error*. After the Random Forest error detector, we used a median filter on its results, with a window of 30 frames, which equates to 2 s in our system. The algorithm outputs *no error* if that is the result of median filtering. Otherwise, it uses a second Random Forest to classify the error as SNV, TF, or both. Unlike the error detector in the first stage, the error classifier Random Forest does not use emotions as features (but uses all the remaining ones). We do not use emotions for error classification since they did not improve the algorithm's accuracy during our tests. Thus, the final output of our algorithm is either *SNV*, *TF*, *SNV & TF*, or *No error*.

6.5.3 Experimental setup

We split the experimental setup into two parts. First, we evaluated the complete error detection and classification pipeline, focusing on the additional new features and the median filter. Then, we separately assessed the performance of the error detector and error classifier modules and checked out which features make a difference in their performance. In this thesis, we do not focus on classifier selection and assume that Random Forest is the best non-deep learning method for the task. Classifier selection was extensively studied in the past works [182] [115], in our workshop paper [25], and [188]. Random Forest was overall the most accurate, even when compared to outlier detection methods (Isolation Forest). Since this thesis follows a more human-centric approach, we focus on which of the social signals that robots detect are more meaningful to understand that something went wrong.

For training and evaluation, we randomly split the dataset as 75 % for training and 25 % for testing. The splitting process ensures that samples from participants in the training set do not appear on the test set. Since our thesis focuses on first encounters, we want the robot to "see" people in the test set for the first time. During the experiments, we evaluated each condition 30 in random dataset splits (except when we explicitly state otherwise). We call each of these individual

evaluations a *run*.

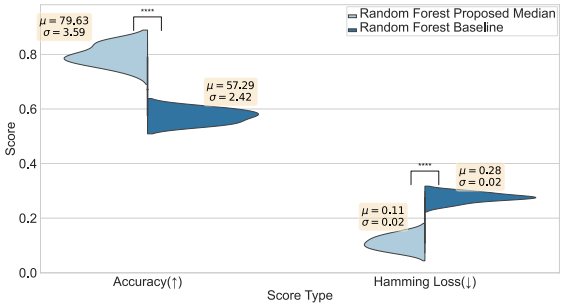
6.5.4 Results

6.5.4.A Evaluation of the proposed pipeline

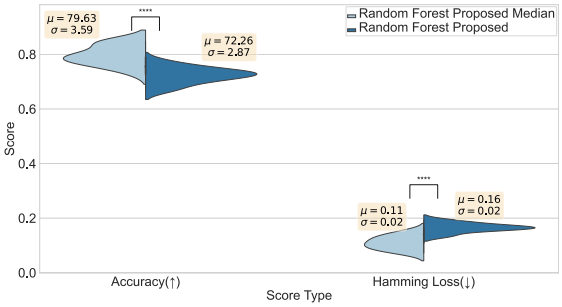
The complete pipeline addresses a multilabel problem, with four possible labels (*SNV*, *TF*, *TF & SNV*, and *No error*). In addition, the dataset is unbalanced, having significantly more samples without errors. Thus, the two metrics used for evaluation are the hamming loss and balanced accuracy.

Our algorithm (Figure 6.6) without the median filter achieved an average balanced accuracy of 79.63 % while the baseline (with only the functional features from the state-of-the-art) only achieved 57.29 %. A t-test showed a statistically significant difference ($t(14) = 22.74$, $p < 0.0001$) between conditions. We obtained a large effect size (Cohen's $d = 7.29$). The proposed method's hamming loss was also statistically significantly lower than the baseline ($t(14) = 26.53$, $p < 0.0001$, Cohen's $d = 7.903$). We illustrate the results on Figure 6.7a.

Adding the median filter as shown in Figure 6.7b produced a significant improvement in the accuracy, as assessed through a paired-samples t-test ($t(14) = 9.64$, $p < 0.0001$). It increased from 72.26 % to 79.63 %, with a large effect size (Cohen's $d = 2.27$). Additionally, there was a significant decrease in the Hamming Loss ($t(14 = 10.19)$, $p < 0.0001$) with a large effect size (Cohen's $d = 2.6$).



(a) Comparing the proposed algorithm with the features used in previous works.



(b) Comparing the proposed algorithm with and without median filter.

Figure 6.7: Comparing the proposed algorithm with the features used in previous works. ↑ - higher scores are better; ↓ - lower scores are better

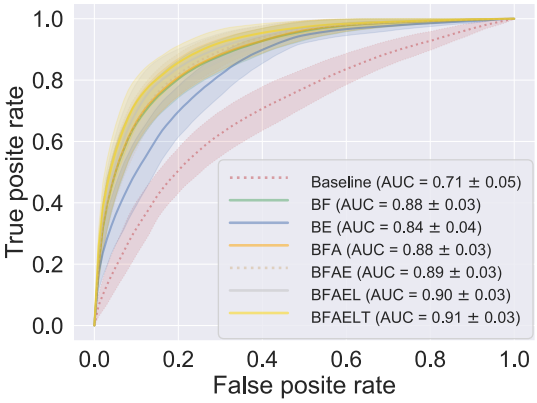
6.5.4.B Subsets of features for ablation studies

In the following subsections, we present ablation studies where we analyze the impact of distinct feature sets on the performance of error detection and error classification. The feature sets consisted of:

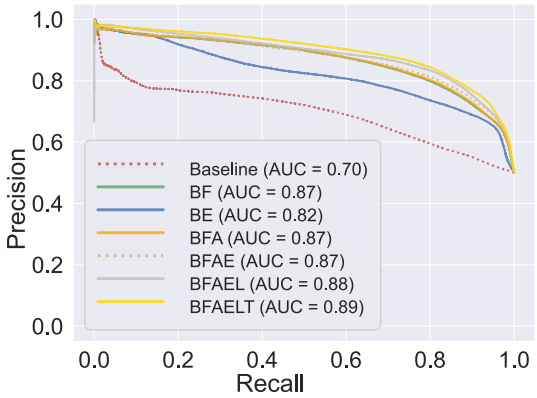
- Baseline (3D head pose and gaze features)
- Baseline + FAU (BF)
- Baseline + Emotions (BE)
- Baseline + FAU + current robot Actions (BFA)
- Baseline + FAU + current robot Actions + Emotion (BFAE)

- Baseline + FAU + current robot Actions + Emotion + Last action (BFAEL)
- Baseline + FAU + current robot Actions + Emotion + Last action + Time since last action (BFAELT)

6.5.4.C Ablation study of the error detector



(a) Receiver operating characteristic curves for error detectors using distinct sets of features.



(b) Precision/recall curves for error detectors using distinct sets of features.

Figure 6.8: ROC average (±1 standard deviation) and Precision/Recall curves of the error detector for distinct sets of features. The greatest the area under the curve (AUC) the better the performance.

Table 6.6: Descriptive statistics and statistical analysis of Accuracy and F1 score metrics for distinct sets of features for the Random Forest error detector.

(a) Error detector statistics for Accuracy and F1 score with distinct combinations of features.

Features sets	Accuracy		F1	
	μ	σ	μ	σ
Baseline	65.2 %	3.28 %	64.3 %	3.59 %
BF	79.1 %	3.22 %	78.9 %	3.22 %
BE	74.7 %	3.69 %	74.6 %	3.81 %
BFA	79.3 %	3.27 %	79.1 %	3.38 %
BFAE	79.6 %	3.47 %	79.5 %	3.55 %
BFAEL	81.4 %	3.46 %	81.3 %	3.46 %
BFAELT	82.0 %	3.44 %	81.9 %	3.53 %

(b) Hypothesis tests for error detection using Random Forest with multiple sets of features. **** means that $p < 0.0001$. For normal data we used Cohen's d to measure the effect size, while for non-normal data we used the rank-biserial correlation r.

Feature sets	Accuracy			F1 score		
	t-test			t-test		
	t (29)	p	d	t (29)	p	d
Baseline v.s. BF	24.17	****	4.33	23.22	****	4.30
Baseline v.s. BE	16.88	****	2.72	16.44	****	2.77
BE v.s. BF	12.40	****	1.28	12.15	****	1.25
BFA v.s. BF	13.23	****	1.29	12.86	****	1.27
BFAE v.s. BFA	2.86	0.008	0.10	2.82	0.009	0.10
BFAEL v.s. BFAE	7.72	****	0.51	7.76	****	0.51
	Wilcoxon			Wilcoxon		
	W	p	r	W	p	r
BFAEL v.s. BFAELT	121	0.022	0.48	126	0.03	0.46

We evaluated the performance of the Random Forest error detector (the pipeline's first module) with distinct sets of proposed features. First, we compared the performance using the baseline features against the baseline with action units. The results are depicted in Figure 6.8 and Table 6.6. We used a paired-samples t-test when the result followed a normal distribution and a Wilcoxon signed rank test otherwise.

As we can see, all features significantly improved the results over the baseline, and adding more features makes the algorithm more accurate. We can see that adding FAU was more enhancing than

Table 6.7: Study of the impact of the Emotion feature in 25 experimental sessions of error detection. Each *#SBR* (Number of Significantly Better Runs) column represents the number of runs of the condition that had a significantly higher performance than the opposite condition (according to McNemar's hypothesis test). Overall, the presence of emotions slightly improves the performance of the error detector.

Session with 25 *runs*	With Emotions (BFAELT)			No emotions (BFALT)			t-test		
	μ	σ	#SBR	μ	σ	#SBR	$t(24)$	p	d
1	78.9 %	2.84 %	13	78.6 %	2.89 %	5	2.91	0.008	0.14
2	80.3 %	3.53 %	12	79.9 %	3.71 %	4	2.01	0.056	0.09
3	79.0 %	3.50 %	9	78.82 %	3.78 %	4	1.58	0.128	0.06
4	80.3 %	3.61 %	10	80.1 %	3.83 %	10	0.84	0.408	0.04
5	80.5 %	3.02 %	15	80.1 %	3.24 %	4	2.40	0.025	0.11
6	80.7 %	2.64 %	13	80.5 %	3.04 %	7	1.04	0.307	0.06
7	79.1 %	2.54 %	12	78.9 %	2.59 %	6	1.85	0.076	0.111
8	81.0 %	2.75 %	14	80.8 %	2.85 %	3	2.62	0.015	0.088
9	80.8 %	3.81 %	10	80.6 %	3.58 %	3	1.48	0.152	0.057
							Wilcoxon test		
							W	p	r
10	80.2 %	3.41 %	12	80.2 %	3.54 %	3	129.0	0.38	0.21

emotions. Additionally, the small effect size of the BFAE v.s. BFA test (Table 6.6b) suggests that Emotions may not yield a very positive overall contribution. Since the slightly positive result may have happened due to a lucky set of training/evaluation splits, one of our concerns was whether emotion features could actually degrade the results. To answer this question, we performed an additional study comparing the detector with all features against a version of it without emotions. This experiment was composed of 10 training/evaluation sessions with 25 *runs* each. Each *run* consists of (i) randomly sampling 75 % of the videos for training and 25 % for the test set; (ii) computing the accuracy of each condition on the test set; (iii) performing the McNemar's test between conditions; and (iv) testing the accuracy of each condition for statistically significant differences over all runs. The McNemar hypothesis test tells us that one algorithm makes more mistakes than the other if the p-value is below 0.05. Otherwise, the algorithms fail similarly. As seen in Table 6.7, when the detector uses emotions, it generally performs significantly better (according to McNemar's test) in more runs than without emotions. Although the average accuracy was only significantly better with emotion features in 3 of the 10 runs, adding the emotion feature did never significantly degrade the average accuracy.

6.5.4.D Ablation study of the error classifier

Similar to the previous experiments, we also evaluated the performance of the Random Forest error classifier with distinct features. Since this classifier assumes that an error occurred, we only used data samples that contain them. Error classification is a multilabel problem since the output can

Table 6.8: Descriptive statistics and statistical analysis of Accuracy and Hamming Loss for distinct sets of features for the Random Forest error classifier.

(a) Error classifier statistics for Accuracy and Hamming Loss with distinct combinations of features.

Features sets	Accuracy		Hamming Loss	
	μ	σ	μ	σ
Baseline	58.7 %	10.3 %	0.27	0.08
BF	62.9 %	9.47 %	0.24	0.07
BE	60.7 %	9.88 %	0.26	0.07
BFA	67.1 %	9.57 %	0.20	0.06
BFAE	67.3 %	9.59 %	0.20	0.06
BFAEL	71.1 %	8.83 %	0.16	0.05
BFAELT	74.6 %	8.60 %	0.14	0.05

(b) Hypothesis tests for error classification using Random Forest with multiple sets of features.

Feature sets	Accuracy			Hamming Loss		
	Wilcoxon			Wilcoxon		
	W	p	r	W	p	r
BF v.s. Baseline	11.0	****	0.95	23.0	****	0.90
BE v.s. Baseline	13.0	****	0.94	25.0	****	0.89
BFA v.s. BE	0.0	****	1.0			
BFAE v.s. BFA				215	0.73	0.08
BFAEL v.s. BFAE				0.0	****	1.0
	t-test			t-test		
	t (29)	p	d	t (29)	p	d
BE v.s. BF	4.21	***	0.23	3.56	0.001	0.27
BFA v.s. BE				8.90	****	0.92
BFAE v.s. BFA	1.06	0.29	0.025			
BFAEL v.s. BFAE	8.87	****	0.40			
BFAELT v.s. BFAEL	11.84	****	0.40	11.78	****	0.52

be SNV, TF, or SNV&TF. To deal with this problem with a different number of samples per label, we used the hamming loss in addition to the balanced accuracy. We show the results in Tables 6.8. When FAUs were present, all features except emotions significantly improved the algorithm's performance.

Since the effect of adding emotion features is not clear, we performed a similar experiment as we did for the error detector. We performed 10 training/evaluation sessions with 25 *runs* each. Each *run* consists of (i) randomly sampling 75 % of the error samples for training and 25 % for the test set; (ii) computing the accuracy of each condition on the test set; and (iii) testing the accuracy of each condition for statistically significant differences over all runs. As shown in Table 6.9, emotions frequently degrade the results of error classification.

6.6 Discussion and conclusions

In this Chapter, we put our efforts into making robots autonomously detect their failures during Human–Robot Interactions. It followed two distinct paths: (i) using the knowledge of the expected outcome of the interaction and (ii) detecting people's non-verbal social feedback on the robot's actions.

For the first part, we focused on creating a hand/no hand binary detector that used the data from the robot's hand's tactile sensors. With this ability, the robot can detect during a close salutation handshake if it was successful or if it made a mistake and grasped an object distinct from the human's hand (or nothing at all).

The second part of the Chapter dealt with detecting people's social feedback when robots' do not respect social scripts or people's expectations of social norms. To address this problem, we created a data-collection procedure. Although with significant limitations, it still allowed us to collect a

Table 6.9: Study of the impact of the Emotion feature in 25 experimental sessions of error classification. Overall, there is a tendency for the presence of emotions to degrade the accuracy of error classification.

Session with 25 *runs*	With Emotions (E)		No Emotions (NE)		t-test		
	μ	σ	μ	σ	$t(24)$	p	d
1	72.4 %	9.51 %	72.8 %	9.43 %	1.83	0.09	0.04
2	74.5 %	10.2 %	74.9 %	9.67 %	1.32	0.20	0.03
3	75.3 %	9.77 %	75.6 %	9.59 %	1.13	0.27	0.03
4	75.0 %	9.88 %	75.2 %	9.83 %	1.11	0.28	0.02
5	74.7 %	10.1 %	74.8 %	9.83 %	0.66	0.51	0.01
6	75.1 %	9.94 %	75.6 %	9.78 %	1.64	0.11	0.05
7	76.3 %	10.2 %	76.5 %	10.1 %	0.95	0.35	0.02
8	73.9 %	9.42 %	74.25 %	9.13 %	1.32	0.20	0.03
					Wilcoxon		
					W	p	r
9	75.1 %	7.66 %	75.4 %	7.63 %	109.0	0.16	0.33
10	74.4 %	8.16 %	74.78 %	8.19 %	90.0	0.05	0.45

dataset of people's reactions to Vizzy's actions that we used to develop an automatic interaction error detector and classifier.

We then proposed an algorithm that detects and classifies HRI errors during one-to-one interactions. Our goal was to build a system that only relied on non-verbal features that the robot could extract with its sensors (for possible in-the-wild non-invasive interactions). We used a baseline of features from previous automatic error detectors and classifiers that met these requirements and showed that the proposed set of features improved the results. Then, we broke down the system with ablation studies to study the contribution of the additional features.

Our results showed that FAUs and robot context features significantly improve error classification results. Although previous works show links between these social signals and interaction errors, as far as we know, this was the first attempt to create a working automatic error detector and classifier that uses them. Emotions, however, did not significantly enhance the results. First, they provided minor improvements to the Random Forest error detector. Nonetheless, since they are computationally fast to compute with the *AverageAU* method, we used them for error detection. However, for the Random Forest error classifier, emotions actually degraded the results. A possible reason for minor improvements is that the classifier intrinsically captures the structural knowledge that relates FAUs to emotions. Thus, it may not need emotion detection algorithms to have the necessary information. The deterioration of the results of the error classifier may also result from the accumulation of errors in FAU estimation that get amplified with a wrongly estimated emotion. This issue needs further clarification, possibly by labeling people's emotions and using the ground truth as a feature.

Part II

Human–Robot Interaction studies

7

ROBOTIC FIRST ENCOUNTERS WITH ELDERLY CARE CENTER USERS

As stated before, the scope of this thesis is to have a mobile robot open the interaction with healthy adults during possible first encounters and potentially propose a task. One of the possible scenarios is having a visiting robot coach suggest activities in elderly care centers. Since this thesis follows a Human-Centered Design, we needed to study people's expectations and perceptions of social robots, assess the Vizzy robot's fitness for the interaction paradigm, and test whether the dependent measures are sensible under these conditions. In other words, we needed to get acquainted with the end-user environment to trace our research path. Thus, we designed an empirical WoZ study where we tested elderly care center residents' social perceptions of human versus robotic coaches in the context of an active and healthy aging program. Kendon's greeting protocol is a model of HHI where two social actors exchange social cues to signal their interest in mutual interaction. Hence, we believe that evaluating if people perceive the robot to have a social presence and how they attempt to interact is relevant to assessing whether adapting Kendon's greeting model to robotics is a promising research path.

We visited five elderly care centers, where we could observe the environment dynamics, how people interact with the caretakers, and how people interact with the robot. The interaction paradigm was that of a coach (human or robotic) that invites seniors to play a game in the Portable Exergames Platform for the Elderly (PEPE) [189]. PEPE is an augmented reality gaming platform that projects games on the floor and captures users' movements using a Kinect sensor. The chosen game for our experiments was ExerPong [190].

This Chapter is an extended version of our previously published papers at the Personal Robots for Exercising and Coaching workshop of the ACM/IEEE International Conference on Human–Robot Interaction and the International Journal of Social Robotics:

J. Avelino, H. Simão, R. Ribeiro, P. Moreno, R. Figueiredo, N. Duarte, *et al.*, "Experiments with vizzy as a coach for elderly exercise", in *ACM/IEEE International Conference on Human–Robot Interaction – Workshop on Personal Robots for Exercising and Coaching (PREC)*, Chicago, Illinois, USA, Mar.

2018.

M. Čaić, J. Avelino, D. Mahr, G. Odekerken-Schröder, and A. Bernardino, "Robotic versus human coaches for active aging: An automated social presence perspective", *International Journal of Social Robotics*, vol. 12, no. 4, pp. 867–882, Jul. 2019, ISSN: 1875-4805. DOI: 10.1007/s12369-018-0507-2.

The contributions of this Chapter are twofold. First, since this was our initial user study with a WoZ paradigm, we developed a series of tools to make the interaction smooth. Here we describe the software applications developed for this and future teleoperation-based studies. Second, we analyzed elderly care residents' perceptions of the Vizzy robot's social presence and compared the perceived warmth and competence of the robot with a human in the same task.

7.1 Hypotheses development

To test the interaction paradigm and Vizzy's appropriateness for this role and set a baseline with basic robot skills, we focus on the concept of *automated social presence*. This concept is defined as "the extent to which machines (e.g., robots) make consumers feel that they are in the company of another social entity" [48](p. 44). It is reasonable to assume that Vizzy, a robot coach, should be perceived to have an automatic social presence. Since people tend to treat non-human actors as social beings [46], [47], knowing whether Vizzy can elicit the perception of having an automated social presence should support our idea of applying Kendon's greeting model, a model of HHI.

Hypothesis 1.1 (H1.1)**.** *Vizzy is perceived to have an automated social presence (i.e., people feel they are in the company of a social entity).*

We are also interested in checking differences regarding perceived warmth and perceived competence between a human and a robot coach. Therefore, we defined the following hypothesis:

Hypothesis 1.2 (H1.2)**.** *There is a difference in the (a) perceived warmth and (b) perceived competence of human versus robotic coaches.*

7.2 Wizard-Of-Oz interfaces

Safe and smooth teleoperation control of the Vizzy robot required us to develop and adapt appropriate tools. Our goal was to have a remote system that gave us quick and secure control over the robot's actions, smooth reactions to users' interactions and utterances, and guarantee their safety. These requirements were especially relevant since the robot was interacting with seniors, who it may harm (for example, due to collisions) due to their physical vulnerability. Thus, we developed two sets of tools: a pack of *Rviz* plugins to control the robot's motor functions and a web interface for dialogue control.

7.2.1 Motor control interface

Rviz is the ROS ecosystem's default visualization and control application. It contains plugins for camera image visualization, navigation, robot motor state visualization, and plugins for multiple

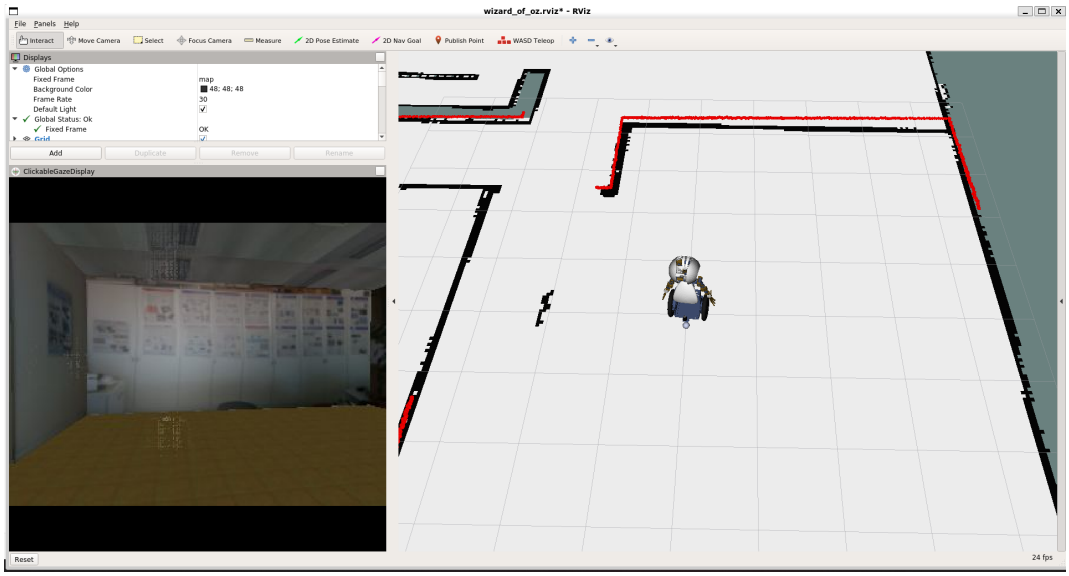


Figure 7.1: The robot control window (Rviz) with our custom plugins (*ClickableGazeDisplay* and *WASDTeleop*). Base control can be achieved with a planner or manually with the keyboard and gaze by clicking on the camera image. The right part of the screen shows the map, the robot and obstacles, enabling the "wizard" safely control the robot.

sensors. Since it is open-source and well documented, the community keeps expanding *Rviz*'s capabilities through new plugins. Given these amenities, we saw *Rviz* as our base application to control Vizzy remotely. However, we identified some limitations during early experiments.

The first limitation was the lack of fine control over the robot's wheels. Even though *Rviz* allows us to control the robot's navigation, this control is based on goal poses with trajectories generated by a navigation planner. Thus, it becomes challenging to manually control the robot's steering for fine positioning of the robot's posture with respect to people. For that purpose, we implemented a new plugin that lets the "wizard" control the robot's velocities directly from Rviz with the W, A, S, and D keys, with Shift as a speed booster (*WASDTeleop tool*).

Easy control of the robot's head is an essential function of a WoZ GUI. Accurate gaze control is crucial to let users know that the robot is indeed talking to them instead of another person nearby. We implemented this functionality as the *ClickableGazeDisplay* view. The clickable view allows the "wizard" to simultaneously view the scene from one of the robot's cameras and choose a point for the robot to look at with a mouse click on the image.

For the remaining non-speech-related visualization and control, the default *Rviz* plugins sufficed. Figure 7.1 shows the final view configuration. It allowed us to view the scene through the robot's cameras, detect obstacles through laser scanner readings, view the robot's estimated location on the care center map, and use the robot's autonomous capabilities.

7.2.2 Dialogue control interface

We developed the dialogue control interface with cross-platform compatibility in mind and to be natural to use in touch screen devices. Our approach was to create the dialogue control GUI as a web app, written with HTML and JavaScript and hosted in the Vizzy robot. ROS communication was

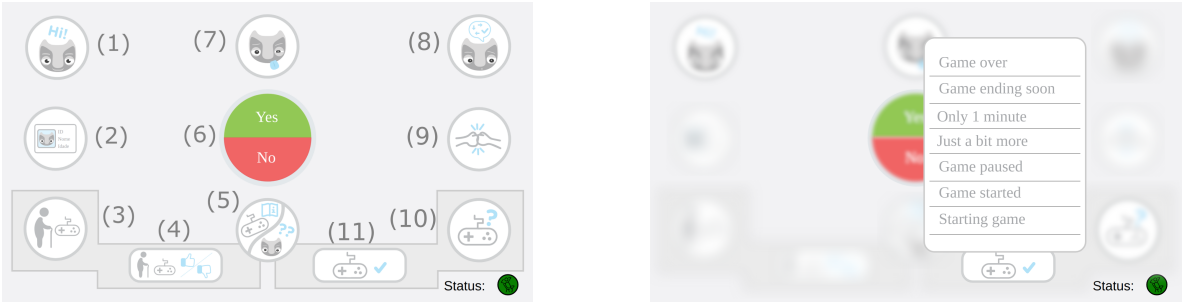
possible with the *roslibjs* library. This interface communicates with ROS via the *rosbridge_websocket* communication server and uses *actionlib* to request a speech action from a node (Speech synthesis server). During this work, we used NUANCE SPEECH's trial API for speech synthesis (the free service is no longer available). The robot used a female voice since all caretakers were female, thus avoiding possible gender effects. Speech actions contain the desired utterance, its language, and the desired voice. This way, we can easily change the voice and language as needed. To allow the "wizard" to listen to what users say, we used an available ROS package (*audio_common*) that streams the microphone's inputs over the network.

Figure 7.2 shows the dialogue interface web application. From the control menu, the "wizard" can select sentences from a predefined list hierarchically organized in dialogue categories as shown in Table 7.1. For instance, if the "wizards" wished to inform users that the game was over, they would press the game events button (button number 11). Upon selecting the "Game Over" verbal intention, the system would randomly select and speak one of the following utterances: (i) "Game Over! Well Done!"; (ii) "Nicely done, the game has ended."; (iii) "Game finished. Good job!"; and (iv) "It's game over. Nice job!".

Table 7.1: Dialogue interface buttons, categories and stages. Button numbers correspond to those depicted in Figure 7.2a.

Button number	Dialogue category	Stage
1	Salutation	Preceding game
2	Presentation	
3	Invitation	
4	Persuasion	
5	FAQ's	Before and during game
6	Yes/No	
7	Humor	
8	Coaching	During and after game
9	Positive Reinforcement	
10	Query user condition	
11	Game event	

The organization of these dialogue category buttons obeys a segmentation based on the various stages of interaction with the person (see Table 7.1). We made this choice so that wizards' choices would be more intuitive and quick. Moreover, we designed the appearance of each button so that their role was visual. The most recent update on this interface allows the "wizard" to create custom utterances in real-time and switch languages.



(a) (b)

Figure 7.2: The dialogue control GUI (a) is composed by a set of buttons grouping several verbal intentions into categories. When pressed, the buttons will expand (b) presenting the available verbal intentions. Button icons were designed by Hugo Simão.

7.3 Materials

7.3.1 Vizzy's skills and control

During these experiments, the robot had two base control modalities: (i) semi-autonomous via-points navigation and (ii) direct wheel velocity control with *WASD* keyboard keys. We used modality (i) when Vizzy moved between rooms, either for fetching a new participant or guiding them to the game room. Then, for the remaining time (approaching people, parking near the game area), we used modality (ii). A researcher controlled the robot's gaze behaviors during an interaction. Vizzy kept switching its gaze between people's faces and the game. Another researcher used a tablet with the dialog GUI to control Vizzy's speech behaviors.

7.3.2 Portable Exergames Platform for Elderly

We used the Portable Exergames Platform for Elderly [189] to perform the exergames with seniors. PEPE is an augmented reality gaming platform that projects exergames on the floor, while a Kinect sensor captures the person's movements to control the game elements. Hence, users can play games on PEPE without any wearable sensor or controller, thus minimizing the burden and complexity for older adults. Two touchscreens on the top of the platform show additional information to staff. For now, the games need to be initiated and terminated by a person, but in the future, the social robot will be able to control the game flow. We chose ExerPong [190] for people to play. In this game, the player controls the green paddle (see Figure 7.3b) with body movements. The objective is to hit the yellow ball with the green paddle, ensuring it does not leave the game area. Colored boxes populate the gaming area, which the yellow ball can destroy on contact, yielding points to the player. The player completes the level when the yellow ball destroys all boxes. One box reapers every time the player misses the yellow ball. Audiovisual stimuli give performance information to the player as red visual feedback and success and failure sounds during the game. The players can either control the paddle by walking sideways or with horizontal arm movements, allowing people with distinct degrees of mobility to play ExergPong.

(a) (b)

Figure 7.3: Both experiment conditions: b shows scenario 1, where a human coaches the elderly user while playing the exergame using PEPE, and a shows scenario 2, where Vizzy performs the coaching role.

7.4 User study

7.4.1 Experimental setup

We performed a between-subjects design experiment where a robotic or a human caretaker coach (the independent variable) guided elderly participants during a session. We installed the exergame platform in the main activities room of each care institution. To keep both experimental scenarios as constant as possible, we developed a standardized set of steps for each actor (i.e., human/robot) needed to follow:

1. Human/robot approached a senior who satisfied the inclusion criteria.
2. Human/robot introduced the new activity at the center (i.e., exergames) and invited the elderly subject to join the game.
3. If the elderly subject accepted the invitation, the human/robot escorted the elderly to the gaming area.
4. The human/robot gave instructions on how to play the game.
5. The human/robot motivated the participant with words of encouragement and feedback on the game progression during the game.
6. Human/robot asks the participant whether to continue or terminate the game (Figure 7.4).

7.4.2 Quantitative data collection

After the activity, the participants spent 10 to 15 minutes discussing their overall experience. They went through guided surveys. We evaluated automated social presence with four items ($\alpha = 0.7$) adapted from [191] and [192]. We measured all items on a five-point Likert-like scale. We adopted a three-item scale ($\alpha = 0.76$) to measure the perceived warmth and a three-item scale ($\alpha = 0.88$) to measure the perceived competence. Both scales were adopted from [193]. Finally, we also gathered items related to the robot's physical attributes. Our publications [9], [10] have a full description of all scales.

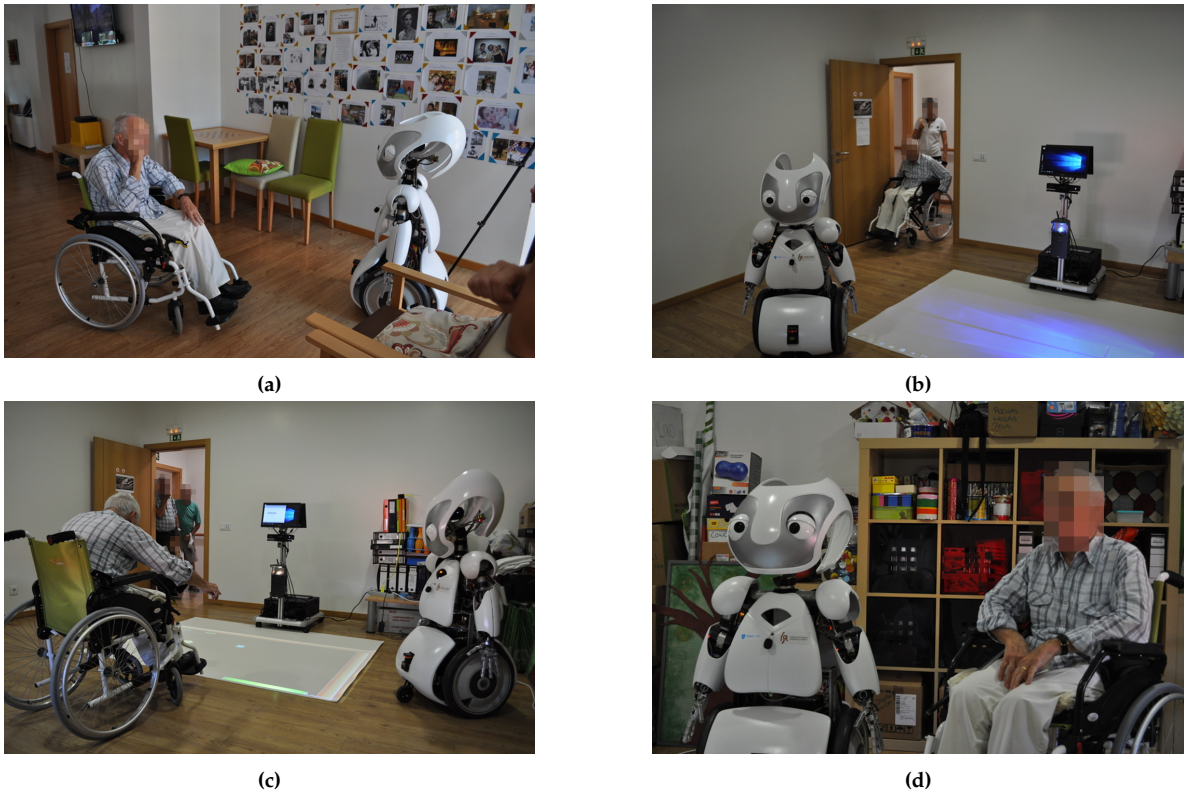


Figure 7.4: Experimental protocol. The robot invites the person a and guides the person to the gaming area b. Upon arrival the robot explains how to play the game and motivates the person based on the performance c. Posing for a photo before the questionnaire d.

7.4.3 Qualitative data collection

We chose to employ a convergent parallel mixed-methods design [194] in which we collected both quantitative and qualitative data simultaneously, intending to analyze whether the findings from both types of data will support each other. During the surveying process, the researcher would further probe particular items from the questionnaire either through 'tell me more' probes [195] or laddering technique [196]. Additionally, we observed how people attempted to interact with the robot and took note of interaction problems.

7.4.4 Participants

A total of 58 elderly persons (42 female and 16 male) participated in the study in five different locations. Their ages varied between 48 and 94 years old ($\mu = 78.79, \sigma = 9.85$). Elderly persons suffering from severe physical (e.g., complete physical immobility) or mental (e.g., dementia) health problems did not participate as well those incapable of consenting. The target population comprised elderly persons living autonomously (e.g., in their own home) or in a nursing home, where they accept services from the care institution. In particular, the main inclusion criterion was that the experiment participants were users of the elderly care centers, which offer a wide range of senior-tailored activities (e.g., card games, fitness, or handcrafts). We further randomly divided the total sample into two experimental groups: i) 22 participants in the *human coach group* (16 female and 6 male), aged between 65 and 94 years old ($\mu = 75.46$, $\sigma = 12.79$); and ii) 36 participants in the

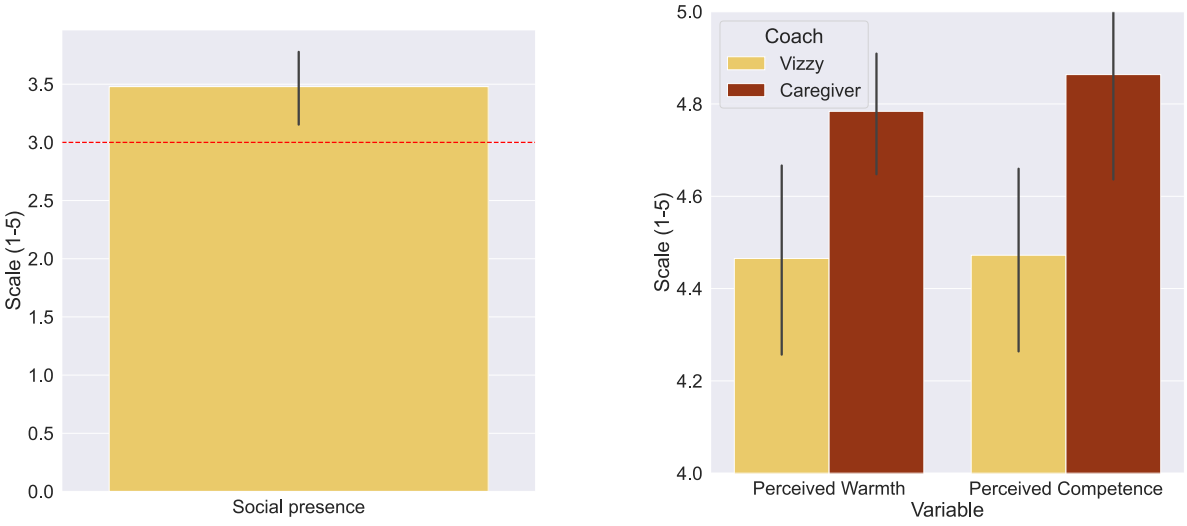
robotic coach group (26 female and 10 male), aged between 48 and 91 years old ($\mu = 80.83$, $\sigma = 6.98$). The human coach experiments were carried out at two locations (Location A: 10 and Location B: 12 elderly participants), while the robotic coach experiments were carried out at three locations (Location C: 11, Location D: 10, and Location E: 15 elderly participants).

7.5 Results

7.5.1 Quantitative results

To investigate H1.1, we ran a one-sample t-test to determine whether the mean value of automated social presence is statistically different from a neutral value, in our case, distinct from a mid-point ($= 3$) on a 5-point Likert scale. We found support for H1.1, with the mean score of 3.5 ($\sigma = 0.97$) which is significantly different from 3, $t(35) = 2.97$, $p = 0.005$, Cohen's $d = 0.49$ (Figure 7.5a). Hence, there is evidence that elderly people experience some sort of automated social presence when interacting with the robot.

To test H1.2, we ran an independent samples t test, to determine whether the means of two groups (human vs. robot) differ statistically. We found a significant difference in the mean scores for perceived warmth when the coach is a human ($\mu = 4.82$, $\sigma = 0.37$) and when the coach is a robot ($\mu = 4.45$, $\sigma = 0.73$), $t(54.40) = 2.517$, $p < 0.05$, Cohen's $d = 0.57$. Similar follows for the perceived competence of the coach: human ($\mu = 4.84$, $\sigma = 0.47$) and robot ($\mu = 4.50$, $\sigma = 0.67$), $t(54.88) = 2.431$, $p < 0.05$, Cohen's $d = 0.67$ (Figure 7.5b). These results support H1.2, the difference for both perceived warmth and perceived competence is statistically significant. The elderly people perceive the robot to be quite friendly, well-intentioned, and understanding of their needs, but to a lesser extent than the human coach.



(a) Automated social presence.

(b) Perceptions of warmth and competence

Figure 7.5: Questionnaire results.

7.5.2 Qualitative results

Informal observations allowed us to collect interaction information that we could use to fine-tune the interfaces and guide our research path. Some people reported they were anxious about the interaction since that was their first encounter with a robot. Some people were frozen, gazing at the robot for some time without knowing if they should really interact with Vizzy or not. However, after this initial state, people started to speak and interact more confidently with the robot.

We noticed that one of the central elements for interaction was the gaze direction while talking. A participant would not interact with Vizzy if its gaze direction was ambiguous or did not match expectations (e.g., Vizzy's gaze direction pointing to a different person). During these situations, the participant would ask for confirmation from the caregivers. These events happened during the first experiment, where the WoZ gaze interface was keyboard-based (i.e., up/down/left/right directions controlled using a keystroke), thus not precise enough. To address this issue, we improved the GUI WoZ interface with the *ClickableGazeDisplay*. Head movements also seemed to have some effect when trying to persuade a person to play. We observed that when the robot nodded up and down while asking them to play, it was easier to convince them to participate. This effect needs further investigation and quantitative data.

Finally, several people wanted to handshake the robot. On one occasion, we had some free time for free interaction with the seniors after the experiment. Given the handshake interest, we directly controlled the robot's motors to stretch its arm, intending to see whether people would try to handshake it. And they did, as exemplified in Figure. 7.6.

Figure 7.6: An elderly lady handshakes Vizzy.

7.5.3 Participant's reports

From interviews with participants, we identified the emergence of a thematic paradox that we label "We know it's a machine, but it feels like a human". Participant's reports show that they experienced conflicting thoughts, as reported in the following quotes:

> "It's almost the same thing [as talking with a real person]"
>
> Participant #G4.12

> "No, I mean... we can see that it's a robot, right? But we can see that it's something intelligent... but it's not a person... but it's as if it is a person."
>
> Participant #V5.3

> "It's a robot but... I imagined it as a person there talking to me."
>
> Participant #V5.5

> "It felt like talking to an adult person. It's like the grown-ups."
>
> Participant #E3.9

These discourses support the hypothesis that people perceive the robot to have an automated social presence.

7.6 Discussion and conclusions

These field experiments allowed us to gather support for using our social robot, Vizzy, as the central research tool for this thesis. During these experimental sessions with older (but mentally healthy) adults, we found support for H1.1 via quantitative results and people's reports. The questionnaire results showed us that people already see the robot as a social entity but that it can improve (3.5/5). In addition, their reports told us that even though it was clear to them that they interacted with a machine, they saw it as a social actor. This evidence is relevant in two ways. First, we claim that if people see the robot as a social actor, it may be appropriate/expected from the robot to behave according to social norms when engaging with someone. Thus, applying a HHI model, like Kendon's greeting model, to open the interactions may be appropriate. People's eagerness to handshake the robot, which is part of the close salutation phase of Kendon's model, further supports this idea. Second, both people's answers and reactions showed room for improvements in automated social presence and the robot's social skills. We noticed some participants had trouble interacting with the robot during the first encounter. They did not know what to expect from the robot and what the robot expected them to do. This difficulty was even more troubling due to people's hearing problems, which made verbal instructions ineffective. Thus, these observations highlight the importance of social norms and gestures in mutual nonverbal understanding.

Perceived warmth and competence also have room for improvement since seniors see the robot as significantly less warm and competent than a human caretaker. Although we expected these conclusions, they showed us that questionnaires measuring *perceived warmth* and *perceived competence* could capture differences in people's perceptions of this robot for this interaction paradigm. Thus, we believe they are a promising tool to evaluate social models for HRI in subsequent user studies.

Finally, we found the teleoperation interfaces developed for this task efficient for HRI studies. We believe they can be a baseline for future studies with the Vizzy robot (or other robots running ROS) with or without additional modules.

8

EFFECTS OF HANDSHAKES IN HUMAN–ROBOT FIRST ENCOUNTERS

The handshake is a physical interaction gesture widely used in western civilizations when opening the interaction, frequently the first form of interaction between people. It is a powerful non-verbal behavior that can influence how individuals perceive social interaction partners and even their interest in future interactions [197]. Studies [198] have shown that people make personality judgments based on handshakes and that the way one performs a handshake has an impact on the perceived employment suitability [199] in recruitment tasks. Other studies [200] also claimed that handshakes influence negotiation outcomes and promote cooperative behavior. As Kendon claims, in his observations, it is part of the Close Salutation phase in western cultures. While the occurrence of other stages of Kendon's greeting model depends on the physical constraints of the scenario or people's acquaintance, the Close Salutation is where people fully acknowledge their willingness to interact. Thus, it is bound to happen in most encounters.

In this Chapter, we study the impact of opening the interaction with handshakes by the Vizzy robot (Figure 4.1) in a task-based scenario. To our knowledge, this is the first investigation of how a physical close salutation may influence user perceptions and acceptance of a mobile humanoid social robot. This Chapter is an extended version of the previously published conference paper:

J. Avelino, F. Correia, J. Catarino, P. Ribeiro, P. Moreno, A. Bernardino, *et al.*, "The power of a hand-shake in human–robot interactions", in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, Madrid, Spain, Oct. 2018.

This user study measures the impact of handshakes by the Vizzy robot (Figure 4.1) in a task-based scenario. We make this investigation in a context where a human and a robot share the same place and task, but the human does not need the robot's help to finish it. Thus, we analyze the helping pro-social behaviour [201], which is not mandatory for the success of the person's task. We believe this is the first attempt at studying the effects of a robot handshake in a situation where

people do not need to cooperate with the robot to succeed. Moreover, we consider that its insights are informative about the relevance of using a physical close salutation during the greeting protocol when people create perceptions about the robot during first encounters.

8.1 Related work

While studying the impact of a close salutation gesture, the handshake, this experiment deals with the constructs of people's willingness to help robots and touch in HRI. We briefly overview work related to these concepts over the following subsections.

8.1.1 Willingness to help a robot

Although robots can achieve superhuman performance in specific tasks and structured environments, social robots often act in unstructured and dynamic scenarios. With current hardware and software, their ability to complete even simple goals (like navigating to a point in space) can fail in unpredicted scenarios. The concept of "*symbiotic relationships*" [202] states that humans and robots can support each other since their strengths and weaknesses complement each other. Thus, if humans are willing to help robots overcome ongoing challenges, both can benefit from this mutual support. As the general interaction paradigm of this thesis is that of a mobile social robot engaging with someone to perform a task, we believe it is relevant to review which variables influence people's willingness to help a robot.

Several works study humans' willingness to help robots. Veloso and colleagues [203] had a receptionist robot lead people through buildings and bring them coffee. If it had a low probability of completing the task (like the need to use a lift), the robot would ask for help from people. The study states three variables that affect the willingness to help: interruption of busy people, the frequency of requests to the same person, and if someone nearby is available. Kuhnlenz et al. [204] made their robot mimic the participants' emotional states. Their results show that adapting to users' emotional states increases their willingness to help the robot. In Srinivasan et al.'s study [205], participants were significantly more willing to help with small requests when they were more familiar with the robot and when the robot was more polite.

8.1.2 Handshakes and the role of touch in Human–Robot Interaction

Besides signaling a mutual recognition between social agents during the greeting protocol, the touch component of human-robot handshakes is a powerful component that has been shown two increase trust and affection, as well as affect physiological responses. Similar effects are also part of other physical close salutations like hugs. For instance, the study of Nakata et al. [206] weakly suggests that when a robot hugged participants, they donated more money than when it did not. Regarding handshakes, Tsalamlal and colleagues [207] showed that human-robot handshakes affect perceived arousal and dominance. A study from Nakanishi et al. **nakanishi2014remote**, where they evaluated remote handshakes through a telepresence device, showed a significantly stronger feeling of closeness and friendliness when people handshake than when they did not. Bevan et al. [208]

found that, during a negotiation task with a teleoperated *Nao* robot, shaking hands resulted in increased cooperation and better economic results for both human and robot.

While these studies address touch and human-robot handshakes, to our knowledge, no study exists addressing how a human-robot handshake before a task affects the participants' perceptions of the robot's social and physical attributes, their help behavior, and their willingness to help in the future.

8.2 Hypotheses development

Given the insights of subsection 8.1.2, we argue that robot handshakes elicit positive emotional responses from people. Studies from cognitive neuroscience have shown that people have different neural responses to interactions preceded by a handshake compared to those without one [197]. Moreover, people also evaluated these interactions more positively. Thus, we have hypothesized a similar effect in Human–Robot Interactions:

Hypothesis 2.1 (H2.1)**.** *Participants will have a more positive perception of a robot that greets them with a handshake.*

Besides, touch behaviors have relevant effects on interpersonal relationships at a sociological level, including pro-social behaviors [209]. In addition to the HRI references in subsection 8.1.2, findings showed a simple touch could increase compliance with different types of requests [210], [211], revealing its positive effect on altruistic behaviors. Since the literature on the touch of a human-robot handshake is still scarce, we were motivated to expand it with this work. Consequently, we have hypothesized that

Hypothesis 2.2 (H2.2)**.** *Participants will be more willing to help a robot that greets them with a handshake.*

8.3 User study

We conducted a user study to analyze the impact of a handshake from a social robot during a collaborative interaction. We have manipulated how the robot introduced itself to participants, with or without a handshake, in a between-subjects design. Due to implementation challenges at the study's time, we opted for a WoZ approach, where a researcher controlled the robot's head and body movements.

8.3.1 Robot handshake

In this work, we used the Vizzy robot and one of the handshake strategies developed in Chapter 4. The handgrip uses the finger position-based controller ("Fixed HS" from section 4.5) since the force sensors were not available. Since the previous handgrip-based study (see section 4.8) did not show significant differences between controller preferences and we recruited people with similar characteristics and demographics, we did not expect this design choice to harm the experiment. The robot's hand had a 3D palm cover to improve the comfort of the handgrip. The shaking motion was

Figure 8.1: Setup of the user study. A - Task instructions; B - Initial Position; C - Target picture with geometric shapes; D - Two obstacles for the robot, a box and a chair; E - Researcher controlling the robot

an open-loop signal with the following characteristics. The wrist had a mean position of 0.84 m, an amplitude of 2.0 cm and a frequency of 1.7 Hz.

8.3.2 Procedure and task

The experiment took place in a large "L" shaped open-space room. Opposing edges of the room were not mutually visible or audible. We used the area of the first edge, which consisted of a simulated living room, to perform the task with the robot. The briefing, questionnaire, and debriefing occurred in the secondary area. Thus, to optimize the experiment time, as soon as a participant started to fill out the questionnaire, we allowed the next one to start the experiment. We warned people working in the open space not to stare or come closer to participants. We ensured that participants filling up the questionnaire and reading the experiment briefing/informed consent did not share information. Each participant started by reading the consent form in the secondary area and signing the consent form, while a researcher initiated the video recording in the living room area. Afterward, the researcher accompanied the participant to the living room area (Figure. 8.1) and introduced Vizzy. The robot's behaviors at that instant depended on the condition:

Handshake: greet the person with a gaze and a handshake.
No handshake: greet the person with just a gaze.

Afterward, the researcher pointed to the sheet with the task instructions and asked the participant to return to the secondary area after finishing the task. Then, the researcher left the participant alone. The experiment ended with the final questionnaire and a debriefing.

The task was composed of four phases:

1. Stand in the initial position and say out loud the voice command - "I am going to start."
2. Move to the target position where a picture with several geometric shapes is.

3. Count how many triangles the picture contains.
4. Return to the initial position and say out loud the sentence - "I saw [N] triangles" - where [N] is the number of triangles the participant counted.

The instructions sheet also mentioned the robot would perform the task in parallel. However, it did not say that obstacles in the robot's way would prevent it from completing the task.

8.3.3 Robot's behaviors

During the experiment, the teleoperator used the set of *Rviz* plugins and speech interfaces developed in Chapter 7 for Vizzy's visits to elderly care centers. The robot only uses speech if it succeeds in counting the triangles, reporting, in the end, the number of triangles it saw.

Through *Rviz* and our custom plugins, the teleoperator could see through one of the robot's cameras and choose fixation points by clicking on the image, controlling the robot's gaze. We defined rules for gaze patterns during the experiment. First, the robot gazed at the participants' faces when it greeted them. While navigating, the robot did not move its head, continuously looking forward. Upon successfully arriving at the objective, the robot would move its head down to simulate the counting of triangles on the picture. The teleoperator sent direct velocity commands to the robot's base using the *WASD* keyboard keys. We developed a gesture panel with buttons for *Rviz* to execute the handshake.

8.3.3.A Robot handshake

During the "Handshake" condition, the robot's handshake was composed of three sequential primitives activated by the teleoperator:

Stretch arm: the robot stretches its arm in the participant's direction with its fingers slightly flexed.

Handshake: after receiving the handshake command from the teleoperator, the robot closes its fingers in an attempt to grab the user's hand. When finger joints achieve the predefined handshake values, the robot performs the shaking motion by oscillating three times, releasing the user's hand afterward.

Home position: the robot's arm returns to its home position (arm pointing down side-by-side with the robot torso).

8.3.3.B Inderect help request

While doing its own task and encountering obstacles, the robot performed an indirect ask for help. To maximize the probability that participants would notice that the robot was struggling, we devised a three-phase behavior for this situation:

Phase 1: the robot moved back, forth, and sideways near the obstacles, simulating it was trying to pass through them.

Phase 2: if the participant did not help the robot, it stretched its arm forward in the direction of the obstacle while moving back, forth, and sideways near it.

Phase 3: the robot's arm returned to its home position, and the robot repeated phase 1. If the participant did not help the robot, it returned to the initial position.

8.3.4 Dependent measures

As our hypotheses involve perceptions of the robot and help behaviors, we use the following dependent measures:

1. **Robotic Social Attribute Scale Questionnaire [74]** using its three dimensions of Warmth (e.g., "feeling"), Competence (e.g., "capable"), and Discomfort (e.g., "awkward") in a scale from 1 ("Definitely not associated") to 7 ("Definitely associated");
2. **Godspeed Questionnaire [73]** using the dimensions of Anthropomorphism (e.g., "fake/natural"), Animacy (e.g., "stagnant/lively"), and Likeability (e.g., "unpleasant/pleasant") in a 7-point semantic differential;
3. **Perceived Closeness** based on [75], using a 7-point scale;
4. **Help behavior** was assessed through objective video analysis and confirmed with the questions "During the task, did you help Vizzy?" ("Yes/No" answer) and "Why?";
5. **The perception that the robot needed help** using the single-item question "During the task, did you feel Vizzy needed help?" and a "Yes/No" answer;
6. **Willingness to help the robot in the future** using the single-item question "In a hypothetical future interaction with Vizzy, in which it needed help, how willing would you be to help it?" and the same five possible answers of [205].

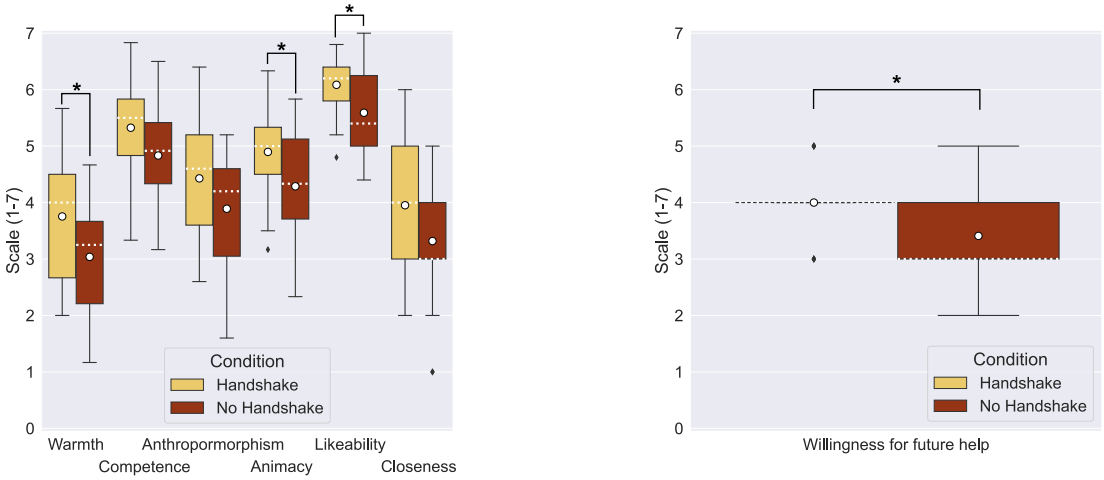
We used measures 1 to 3 to gather support for hypothesis H2.1 and measures 4 and 6 to support hypothesis H2.2. We used measurement number 5 as a manipulation check.

8.3.5 Sample

We recruited 45 university students. However, we excluded 2 participants that did not touch the robot's hand during the Handshake condition. We made this decision since we view the handshake as a touch modality. Thus, our final sample was of 43 participants (23 female and 20 male), with ages from 18 to 27 years old ($\mu = 19.86, \sigma = 1.54$). The Handshake and No Handshake conditions had 21 and 22 participants, respectively.

8.4 Results

After conducting a normality analysis using the Shapiro-Wilk test, we used the parametric Student's t test for dependent variables with normal distributions and the nonparametric Mann-Whitney U test otherwise. We now present the results for participants' perception of the robot and willingness to help it.



(a) Averages and standard deviations per condition for Godspeed, RoSAS and closeness measures. *p < 0.05

(b) Averages and standard deviations per condition for the willingness for future help. *p < 0.05

Figure 8.2: Questionnaire results.

8.4.1 Participants' perceptions of the robot

Within the three dimensions of the RoSAS Questionnaire (Figure 8.2a), we did not use the Discomfort as it presented an extremely low internal consistency (Cronbach's $\alpha = 0.455$). A possible explanation may be the inaccurate translation as the questionnaire was validated in English and applied in Portuguese, the native language of the participants. The Warmth and Competence dimensions revealed good internal consistencies ($\alpha = 0.867$ and $\alpha = 0.835$, respectively). Participants in the Handshake condition attributed significantly higher levels of Warmth to the robot ($\mu = 3.734, \sigma = 1.124$) compared to participants in the No Handshake condition ($\mu = 3.038, \sigma = 1.063$), $t(41) = 2.148, p = 0.038, r = 0.311$. However, there was a non-significant difference between the levels of Competence attributed to the robot in both conditions, $t(41) = 1.733, p = 0.091, r = 0.255$.

Regarding the three dimensions of the Godspeed Questionnaire (Figure 8.2a), Anthropomorphism and Animacy revealed a good internal consistency ($\alpha = 0.838$ and $\alpha = 0.838$, respectively), while Likeability showed only an acceptable internal consistency ($\alpha = 0.790$). There was a non-significant difference between the levels of Anthropomorphism, $t(41) = 1.72, p = 0.093, r = 0.254$, but the difference between the levels of Animacy and Likeability were statistically significant, $t(41) = 2.163, p = 0.036, r = 0.314$ and $t(41) = 2.464, p = 0.018, r = 0.353$ respectively. Participants in the Handshake condition rated the robot with higher values of Animacy ($\mu = 4.897, \sigma = 0.814$) compared to the No Handshake condition ($\mu = 4.288, \sigma = 1.016$). Similarly, they rated the robot as more likeable in the Handshake condition ($\mu = 6.086, \sigma = 0.578$) compared to the values attributed in the No Handshake condition ($\mu = 5.591, \sigma = 0.726$).

The difference between the levels of Perceived Closeness attributed to the robot in both conditions was not statistically significant (Figure 8.2a), $U = 172, p = 0.139, r = -0.225$.

8.4.2 Willingness to help

The first measure related to the willingness to help the robot was the objective helping behavior during the task, which we evaluated in video analysis. In a previous pilot study, we found out that people would assist the robot in various ways (e.g., remove one of the obstacles, inform the robot out loud of the number of triangles, or show the picture to the robot). However, during this study, the participants only assisted the robot by removing one of the obstacles. Moreover, we double-checked the objective analysis with the subjective single-item question - "During the experiment, did you help the robot?" - which matched all participants except one. He considered saying the final command as helping the robot, which was not as it was part of the task, and all the remaining participants did it as well.

There was no statistically significant association between the condition (Handshake or No Handshake) and the helping behavior, $\chi^2(1) = 1.865, p = 0.172, r = 0.208$. Although non-significant, the tendency suggests that more participants helped the robot when it greeted with a handshake (57.1 %) compared to when it did not greet with a handshake (36.4 %).

Additionally, there was no statistically significant association between the condition (Handshake or No Handshake) and the perception that the robot needed help, $\chi^2(1) = 2.751, p = 0.097, r = 0.253$. Although non-significant, the tendency suggests more participants in the Handshake condition understood the help request (85.6 %) than in the No Handshake condition (63.6 %).

Furthermore, among the 32 participants that understood the robot was in trouble, we also analyzed the association between the condition (Handshake or No Handshake) and their helping behavior, which was not statistically significant, $\chi^2(1) = 3.030, p = 0.082, r = 0.308$. Again, the tendency suggests when the people perceived the robot as needing help, participants in the Handshake condition helped it more (12 out of 18, 66.7 %) than participants in the No handshake condition (5 out of 14, 35.7 %).

Finally, there was a statistically significant difference between conditions in the willingness for future help (Figure 8.2b), $U = 138, p = 0.015, r = -0.369$. When asked about a hypothetical future situation where Vizzy was in need of help, participants in the Handshake condition reported significantly higher values ($\mu = 4.00, \sigma = 0.154$, "4 - Yes, I would help even if I was busy") than participants in the No Handshake condition ($\mu = 3.409, \sigma = 0.170$, "3 - Yes, I would help even if I was somewhat busy").

8.5 Discussion

Our results support H2.1, which predicted that a robot greeting participants with a handshake would be perceived more positively. We found that the handshake increased participants' perceptions of Warmth, Animacy, and Likeability. Although we cannot claim a similar effect on the remaining measures used to assess the robot's perception, i.e., Competence, Anthropomorphism, and Perceived Closeness, we believe we cannot ignore their considerable effect sizes and tendencies.

According to H2.2, we expected the handshake would have positively influenced the partici-

pants' willingness to help. Our results partially support this hypothesis as we can only claim the handshake positively affected the participants' willingness for future help. The pro-social behavior of helping the robot during the task was not statistically significant between conditions. However, the considerable effect sizes and tendencies seemed to suggest the handshake might have had a small impact, especially among participants who understood the robot needed help.

8.6 Discussion and conclusions

This Chapter explored the impact of the handshake - a physical close salutation gesture - on people's perceptions of the Vizzy robot and their pro-social behavior toward it. The interaction design accounted that the robot's skills match the challenge of the task. Results show that people greeted with a robot handshake improve their perception of the robot and willingness to help it. The relevance of these results for our thesis is twofold. First, they contribute to our goal of improving people's perceptions of a mobile social robot when opening the interaction in a possible first encounter. Second, even though H2.2 was only partially supported, it hints that a physical close salutation gesture may play a relevant role when attempting to accomplish regular and symbiotic collaboration. Given our scope of having a mobile robot open the interaction with someone in a first encounter with the intention of providing a service or asking for help, these results are highly encouraging.

Nonetheless, the present study has some limitations. The first one is the still robotic handshake behavior. Currently, our colleagues are implementing significant improvements that provide a more comfortable and warmer handshake, which we believe will further improve how people perceive the robot and their willingness to help it. However, we warn that during future similar studies, if the robot displays a highly elaborate and lifelike handshake, participants will not expect it to get stuck in minor tasks (like navigation). There would be a significant discrepancy between the sensed handshake behavior and the expected robot's skills. Thus, researchers should design future experiments with care to avoid breaking people's expectations. Finally, all the participants are from western countries, where handshakes are a standard greeting behavior, share similar cultural backgrounds, and are from the same age group. A more diverse sample is needed to generalize the results.

9

EVALUATION OF GREETING MODELS IN A PUBLIC PLACE

Researchers in past works used several approaches to create engaging mobile robots with varying sensor modalities and engagement behaviors. As highlighted in Chapter 3, we identified that the social scripts followed by these robots were mainly subsets of Kendon's greeting model and only used limited social signals. To our knowledge, no comparison exists between the literature's greeting approaches in first encounters under the same conditions. In addition, we proposed a perceptual model (Chapter 5) that uses multimodal information to estimate people's mental states according to Kendon's model. Does this extra information and uncertainty tracking improve how engaged people are with the robot?

This Chapter addresses some gaps highlighted in subsection 3.1.6: (i) the lack of comparison between greeting models under the same conditions; and (ii) the evaluation of a greeting model that acts based on Kendon's phase estimation. Thus, we propose an experiment in a public place where we evaluate how Kendon's greeting model improves people's engagement during a first encounter. Thus, we experimented in a public place where we evaluated whether two components of greeting models improve people's engagement during a first encounter. These are: (i) related to the robot's capability to mimic the behaviors of Kendon's model; and (ii) related to the robot's abilities to perceive and act according to the social signals described in Kendon's works.

We evaluate the impacts of three greeting models on multiple dimensions of the robot's engagement success (behavioral, emotional, and cognitive engagement). The first model uses a subset of behaviors (three greeting phases) and social signals inspired by the works of Satake et al. [11], [12], Shi et al. [63], Brščić et al. [86], and Kato et al. [87] (previously described in Chapter 3). The second model commands the robot to mimic the six phases of Kendon's greeting model (complete model) as a response to the same social signals of the first model (position, body orientation, and gaze). Our implementation is inspired by the work of Heenan et al. [90], with the addition of the Head Dip phase and adaptations to our robot, Vizzy. The third model uses the same behaviors as the second but responds to more of the social signals described by Kendon. Comparing the first and second

Figure 9.1: Vizzy navigating through the art exhibition during inauguration day.

models, we can assess whether behaving more closely to Kendon's model improves the success of a mobile robot's engagement while reacting to a small set of signals. Comparisons between the second and third models evaluate if using additional social signals to track people's greeting state further improves engagement.

The experiments described in this Chapter occurred during a real-life art exhibition of the digital painting of Professor Andreas Wichert at the Civil Engineering building of Instituto Superior Técnico - University of Lisbon [1]. Our experimental setup was composed of two scenarios: (i) a controlled experiment with invited participants and (ii) an in-the-wild experiment, where the robot was free to interact with people passing through the exhibition.

9.1 Hypothesis development

9.1.1 Behavioral engagement with the robot

The crucial goal of this thesis is to improve the chances of a mobile social robot being able to open interaction with people during first encounters. Even though we can measure the success of the engagement in distinct ways, in our opinion, there are two critical questions. Was the robot successful in opening the interaction and fulfilling its interaction goals? If not, why did the robot fail? As illustrated in Figure 1.2 in Chapter 1, and proposed by Satake and colleagues [11], [12], a mobile social robot may fail at distinct stages. They report a sequential model of failures that may happen, which we adapted and illustrated in the diagram of Figure 9.2. The first type of failure (Unreachable) occurs if the robot cannot intercept the target human to open the interaction. Since we did not focus on path planning for this thesis, we did not evaluate this type of failure during this experiment. Even though people are reachable, they may fail to notice the robot, which Satake and colleagues call the Unaware failure type. When people are aware of the robot's presence and gestures, they may still not understand its intentions, leading to an Unsure error. Finally, even if humans fully

[1] `https://bit.ly/ist_exhibit`

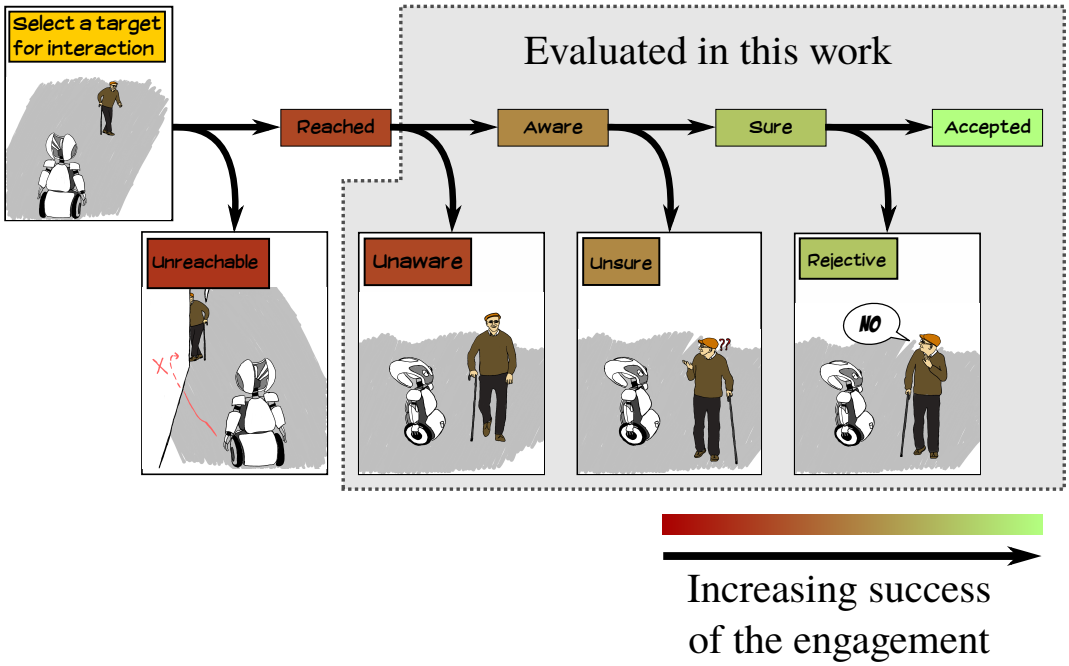


Figure 9.2: According to Satake and colleagues [11], [12], when mobile robots attempt to engage with someone, there is a sequence of problems that may prevent them from succeeding. In this Figure, we represent their sequential model, illustrating it with our robot.

understand the robot's intentions to interact, they may still reject the attempt to interact and leave, a Rejective error. This error not be solely the robot's fault (like using inappropriate discourse or gestures) but also caused by people's current mental state (e.g., being in a hurry, not being in the mood). Finally, if everything goes well, the person accepts the robot's attempts and interacts with it. Thus, following this sequence, the closer the outcome is to people accepting to interact, the more successful we consider the interaction attempt to be. We consider this aspect of interaction the main form of behavioral engagement.

Since we claim that Kendon's greeting model can improve the way robots engage with people, we define our first set of behavioral engagement hypotheses:

Hypothesis 1.1 (H1.1)**.** *The robot's engagement success increases if the robot:*

Hypothesis 1.1a (H1.1a)**.** *increases the number of Kendon's greeting phases that it follows.*

Hypothesis 1.1b (H1.1b)**.** *acts according to the full Kendon greeting model reacting to estimates of people's current greeting phase.*

Additionally, observed human behaviors provide additional information about people's mental states, demonstrating their commitment to the interaction. These social signals allow humans to gather certainty about the state of interactions, as reported by Kendon's studies, and as Kendon's and Goffman's [56] argued, more complex greeting scripts occur during first encounters. Thus, we expect that if people see a robot displaying more social signals described in Kendon's greeting model, they will be more likely to respond with such gestures. We, therefore, set the following behavioral engagement hypothesis:

Hypothesis 1.2 (H1.2)**.** *People will behave more closely to Kendon's greeting model when the robot:*

Hypothesis 1.2a (H1.2a)**.** *increases the number of Kendon's greeting phases that it follows.*

Hypothesis 1.2b (H1.2b)**.** *acts according to the full Kendon greeting model reacting to estimates of people's current greeting phase.*

Since the remaining hypothesis follow the same sub-hypothesis structure (action - perception) as the previous ones, from now on, we will use a single hypothesis that encompasses both.

9.1.2 Emotional engagement with the robot

Not only do we want the robot to complete the engagement attempt by initiating the interaction with people but to do it pleasantly. The way people interact with the robot during a first encounter can have repercussions on their willingness to help the robot in the future, as observed in Chapter 8. Additionally, our past work (in Chapter 7 and Čaić et al. [9]) suggests that constructs like perceived warmth and perceived competence of the robot have an impact on users' experience related to a human-robot task (in that particular case, playing exergames).

To address this challenge, we once again investigate a set of hypothesis based on past literature on social sciences and HRI. A relevant aspect is how the robot performs gaze patterns. Excessive staring in human-human interactions can be threatening [212], and gaze increases with familiarity [213], thus being lower in first encounters. Additionally, people felt significantly more nervous with a constant gaze from a robotic head [214], and women were less likely to be persuaded by a staring robot [215]. Thus, we expect people to feel more emotionally engaged with a robot that follows the full Kendon greeting model more closely when they first meet since that would mimic human behavior, with gaze and gaze aversion phases, in first encounters. These arguments raise the following hypothesis:

Hypothesis 1.3 (H1.3)**.** *The closer a robot follows Kendon's greeting model during first encounters, the more emotionally engaged people will be in the interaction.*

9.1.3 Behavioral and cognitive engagement with the task

Finally, and inspired by the insights of Chapter 7 we expect people to be more dedicated to the task after greetings that resemble Kendon's greeting model more. As such, we raise our two final hypotheses:

Hypothesis 1.4 (H1.4)**.** *The closer a robot follows Kendon's greeting model, the greater the:*

Hypothesis 1.4a (H1.4a)**.** *behavioral engagement with the task.*

Hypothesis 1.4b (H1.4b)**.** *cognitive engagement with the task.*

9.2 Greeting models / conditions

We manipulated how our robot, Vizzy, greets people during the experiment to test our hypothesis and evaluate whether following Kendon's greeting model more closely improves engagement

success. Thus, we implemented the following three greeting models with distinct levels of similarity to Kendon's greeting model.

Model 1 (M1)**.** Based on behaviors and inputs used by Kanda and Ishiguro's group [11], [12], [63], [86], [87]. Even though each work had a tailored engagement strategy for a specific task, they had a baseline of shared behaviors that we classify as three of Kendons phases (section 3.1): (i) initiation of the approach; (ii) final approach; and (iii) close salutation. This model uses distance, body orientation, and gaze. We highlight that the two main behaviors where this model deviates from Kendon's full model are the absence of distance salutation and lack of gaze aversion when the robot is approaching people. The model's state machine is depicted in Figure 9.3 and an illustrative video with the interaction with a researcher is available online[2].

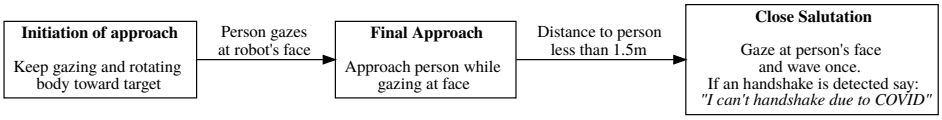


Figure 9.3: Model 1 state-machine

Model 2 (M2)**.** This model implements the behaviors of all six phases of Kendon's greeting model and reacts to gaze, distance, and body orientation signals like M1. Compared to Model 1, this model adds distance salutation, head dip, and gaze aversion during approach behaviors. We based this implementation on Heenan et al.'s [90] work unlocking each phase sequentially. Thus, the robot cannot retrocede to a previous greeting stage after entering a subsequent one. Unlike Heenan and colleagues, we also added the Head Dip phase, only used the robot's onboard cameras, and adapted the implementation's characteristics to the Vizzy robot's size and features, which differ significantly from their NAO robot. As far as we know, this model was never compared with other greeting models, nor was it evaluated outside of the lab or quantitatively. The model is illustrated as a state-machine in Figure 9.4 and exemplified in an online video [3].

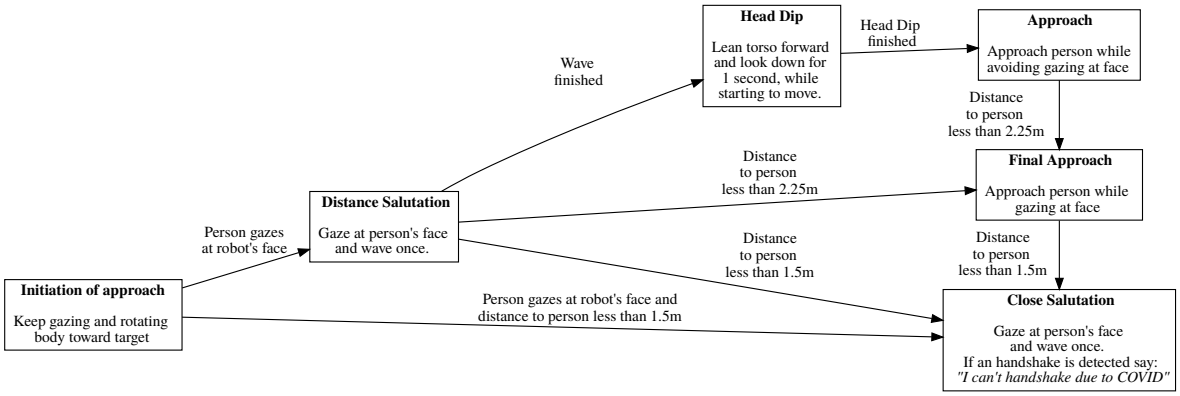


Figure 9.4: Model 2 state-machine

[2] https://bit.ly/greet_m1
[3] https://bit.ly/greeting_m2

Model 3 (M3)**.** The third model of this work uses the implementation of the phases of Kendon's greeting model of M2, but distinct social signals and reasoning to switch between phases. This model uses an HMM that estimates the human's current greeting phase according to Kendon's model, using distance, orientation, velocity, gaze detection, and gesture signals. The model uses all this information to choose a greeting phase to execute or preempt at each instant. Unlike previous models, it allows the robot to repeat greeting phases since they are reactions to the robot's estimate of people's greeting stage and controls the robot to act according to a more extensive set of social cues. According to Schiffrin [50], humans also repeat greeting behaviors when they are uncertain if others have observed them, which supports this model. To simplify the representation for most people, who may not be familiar with BT, we describe this model using a diagram in Figure 9.5. An online [4] video exemplifies the execution of this model.

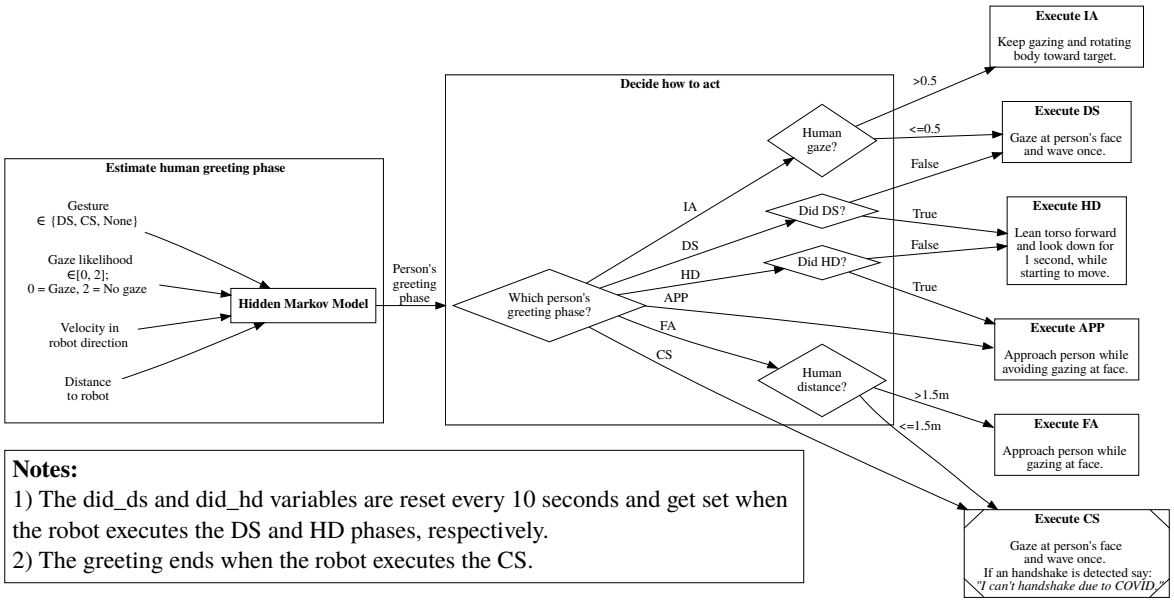


Figure 9.5: Model 3 diagram

Every greeting model ends successfully if the robot executes the close salutation phase. We set the greeting as failed if the robot fails to enter this phase for 60 seconds or stops detecting the target for 10 seconds.

Model 1 and Model 2 share the same input signals and differ in the possible behaviors that the robot can use. We claim that, by comparing both, we are evaluating the effects of using more phases of Kendon's greeting model. Model 2 and Model 3 share the same possible behaviors for execution but differ in the input signals and the way the robot uses them for decision-making.

As a note, even though Model 1 and Model 2 are conceptually state-machines, we implemented them as behavior trees to leverage previous behavior implementations and avoid low-level behavioral differences with Model 3. Nonetheless, these two implementations are equivalent to a state machine from an execution perspective.

[4] `https://bit.ly/greeting_m3`

9.3 Methodology and experimental setup

To compare the effects of the three greeting models, we designed two experiments in a public place with the Vizzy robot. The experimental site consists of an art exhibition hosted in the civil engineering building at Instituto Superior Técnico (IST). The first type of experiment consisted of having the robot interact in the wild when there were no scheduled participants. During the second type of experiment, we invited people (one person at a time) to enter the painting exhibition area as a visitor would. Participants were recruited both on-site and through an online schedule. We performed this experiment during hours when the exhibition place was unlikely to be crowded.

The interaction paradigm was the same for both cases and is described in subsection 9.3.1.

9.3.1 Interaction paradigm

The experiments occurred in the central area of the painting exhibition. Since the sides of the building are passage and access areas to rooms and cafeterias, and not necessarily appreciating the paintings, the robot does not select people in those areas for interaction. During the robot's idle state, it stands in the position depicted in Figure 9.6 and looks around randomly. This behavior occurs under two conditions: (i) when the robot is not in someone's line of sight (outside a visual cone of 120°) and (ii) when the robot is in the "not willing mode," which occurs for 20 seconds after a previous interaction. When none of the conditions hold, the robot selects the closest person that has it in line of sight. Afterward, the robot executes the greeting defined for the session (possible greetings described in section 9.2). If the greeting is successful, the robot presents itself using the following discourse:

> » Hi! Can you help me? Please, please, please!
>
> **1 second pause**
>
> » My name is Vizzy and I'm learning about art. I need your interpretation of some parts of these artworks.
>
> **1 second pause**
>
> » Unfortunately I cannot recnognize speech.
>
> **1 second pause**
>
> » To interact with me, access the website in my chest with your smartphone.
>
> **19 seconds pause**
>
> » Then show the QR on your phone's screen to my chest camera.
>
> **2 second pause**
>
> » If you are having trouble, put your phone 15 centimeters away from the camera, as exemplified.
>
> **After user shows the QR code**
>
> » Interface unlocked, check your smartphone.

The web interface interaction is summarized in Figure 9.7 and described in subsection 9.3.2. The

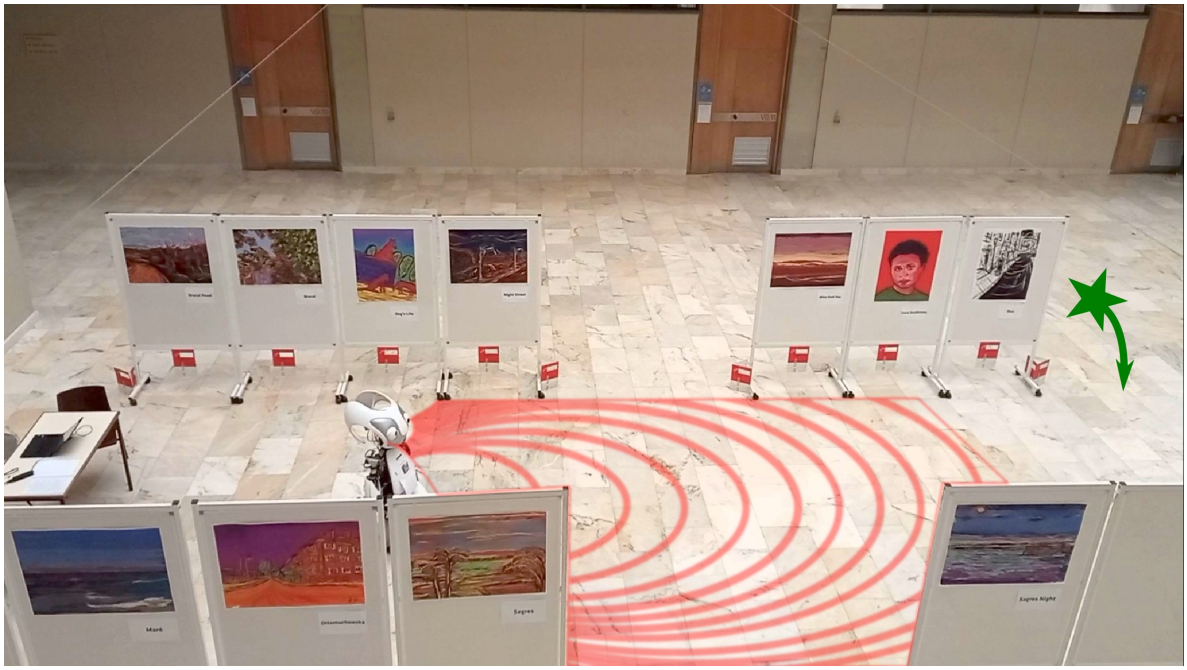


Figure 9.6: Initial conditions before interaction. The robot detects people through its left-eye camera. The red patterned area in the picture represents the robot's detection field relative to the robot's position. People moving on the sideways corridors were ignored. The green arrow and star represent the starting point and movement direction for invited participants during the controlled experiment sessions. During in-the-wild sessions, there were no constraints on people's movements.

robot reacts to people's answers in the interface with voice and gazes at participants through the whole interaction. Since participants need to get close to the robot to unlock the web interface, the robot moves back to respect proxemics as soon as people finish the unlocking task. This behavior has two functions. On the one side, it should make people more comfortable during the interaction since the robot respects their personal space. On the other side, this distance improves the reliability of the robot's onboard sensors (mainly its eye cameras) since it can capture more information about people's bodies' positions and movements. This idea was inspired in past literature works like ones of Mead and Matarić [216] where they simultaneously control the robot to respect the proxemics distance and minimize sensor errors.

The people detection setup using the robot's onboard cameras can make detections up to around 4.0 m (rough estimate). Given the fact that we programmed the robot to not engage with people in the side areas, the initial perception area is depicted in red in Figure 9.6. This area moves with the robot since it performs all perception with its onboard sensors. The robot is free to navigate the whole exhibition area. Once the robot finished an interaction (either successfully or not), it returned to the initial pose and to the idle state.

9.3.2 Interaction interface

When the robot successfully opens the interaction, it asks the participant to interact through a smartphone by accessing a web application, which we now describe (Figure 9.7 and Figure 9.8). The application requires proof of presence to be used, preventing unintended access from external

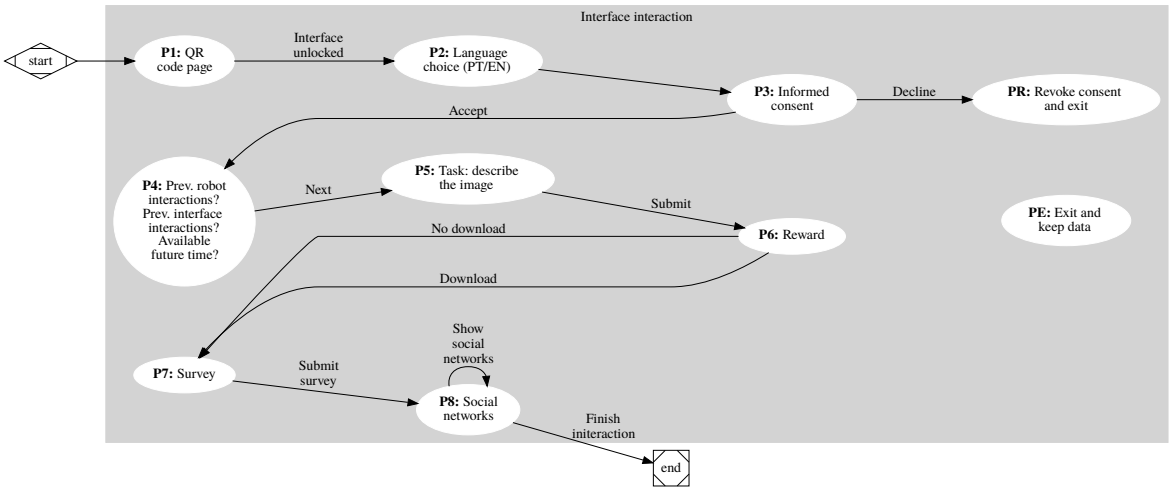


Figure 9.7: Diagram depicting the web interface that participants use to interact with the robot. Each white ellipsis represents a distinct interface page, and arrows represent transitions between pages. Users can access the **PR** and **PE** pages at all times, being able to revoke their consent and delete their data or exit the experiment and keep their data.

users. Thus, the participant must show a QR Code (P1) to the robot's chest camera to unlock the web interface for that device. Then, the participant chooses the preferred language (P2) and needs to accept the informed consent to proceed (P3). Henceforward, the robot speech and the web interface will use the user-selected language. On page P4, the interface asks the number of times the user interacted with the robot and the web interface before and how much time the participant would be willing to spend to help the robot in a future encounter (between 0 and 120 minutes). Page P5 depicts an artwork that Vizzy asks the participant to describe. Afterward, the robot offers a wallpaper as a reward (P6), which the user can download or not. Then P7 shows a questionnaire, and the robot asks the participant to fill it. Before the final page, the robot informs the participant that it has social networks. At the same time, the web interface presents a button to show some screenshots of Vizzy's Facebook and Instagram pages (P8) and another to finish the interaction.

The participant can end the interaction anytime, either keeping or revoking the shared data with the robot, through two buttons on the interface bottom.

9.3.3 Collected data

We collected anonymized data through the robot and the web interface during this experiment. From the perspective of the robot, we collected:

- The number of engagement attempts
- The initial distance of the target (right before starting the greeting)
- Detected gestures
- Blurred RGB images with people masked with Openpose keypoints and landmarks
- A 2D point cloud from the robot's base laser range finders.

The web application collected the following data:

- The number of interface interactions and respective engagement attempts

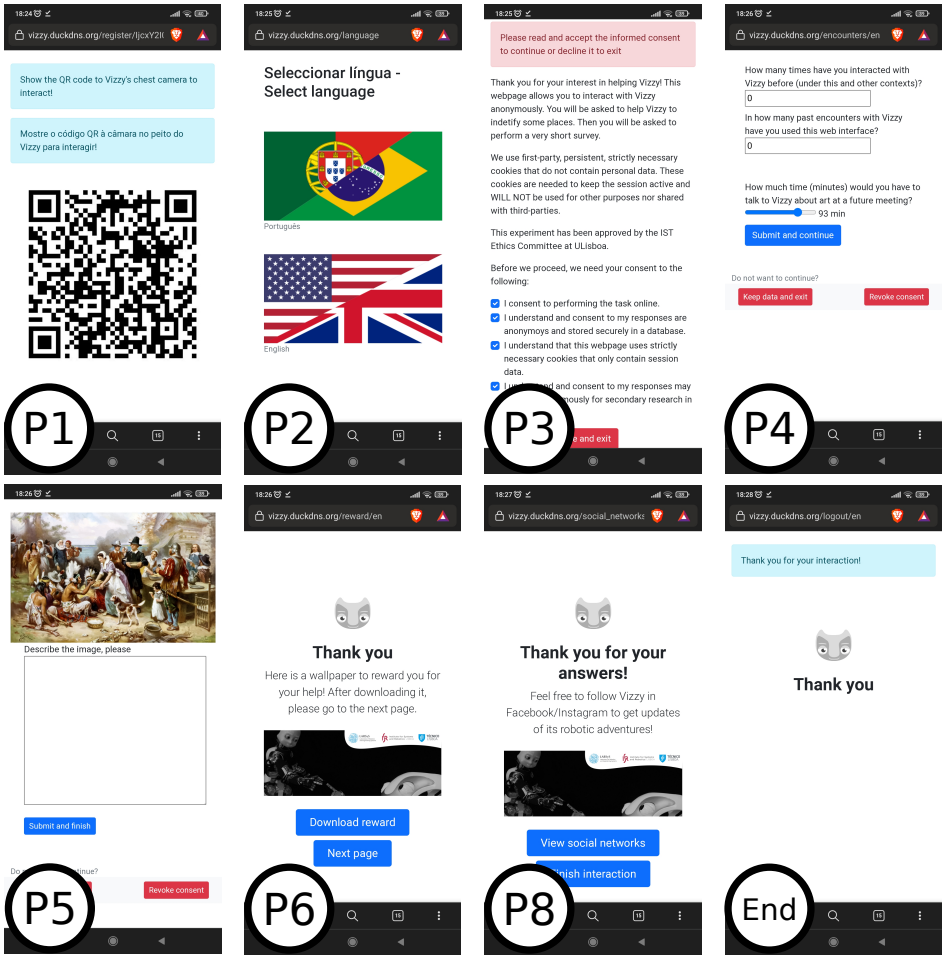


Figure 9.8: Interface screenshots of the web interface illustrated by Figure 9.7 (except the survey pages).

- If participants accepted informed consents
- Self-reported number of past robot and interface interactions
- Self-reported time to spend teaching the robot about art in a future encounter
- The description of the artwork on **P5**
- If the person downloaded the reward
- Survey answers:
  - Gender, age group, geographic area, self-reported past interactions with robots, and self-reported experience in robotics development
  - Perceived warmth, competence, and discomfort [74]
  - Additional 1 to 6 Likert-like items asking if: (i) the robot's movements were natural, (ii) if the robot recognizes when they pay attention to it, (iii) if it knows their position, and finally (iv) if the robot knows how to greet people.
- If the person clicked on the "Show social networks" button.

9.3.4 Measures

To test our hypothesis, we needed several metrics that could capture people's behavioral and emotional engagement with the robot and their behavioral and cognitive engagement with the robot's proposed task. We now present the chosen metrics and argue how they can support each hypothesis. Table 9.1 summarizes the information in this section.

9.3.4.A Behavioral engagement with the robot

Our main goal is to increase the success of the robot's approach to engaging with people from the perspective of the outcome of each attempt (Figure 9.2). We formalized our proposal to achieve this goal with hypothesis H1.1. To measure the outcomes of each interaction, we needed separate metrics for in-the-wild and invited participant scenarios.

For experiments in-the-wild, we labeled the outcomes of each interaction to measure the effects of our conditions following the sequence illustrated in Figure 9.2. Then, we converted them to an ordinal scale from 1 - *Unaware* to 4 - *Success*.

We cannot apply the same metric when interacting with invited participants. We argue that that metric is not fit for such scenarios since people would not leave the experimental area if the robot started talking with them (they would think it was part of the experiment). The bias would make the *Rejective* outcome unlikely to occur under this setup. In addition, since people's movements were bounded to the exhibition area (stopping and looking around), it would be too challenging to discern between the *Unaware* and *Unsure* outcomes. Due to these challenges, we opted to use a binary metric, measuring whether people initiated the interaction with the robot or not, to analyze the results of invited participants.

In addition, we also want to study whether people act more according to Kendon's greeting model toward social robots that follow it more closely. We formalized this goal through hypothesis H1.2 and used the frequency of people demonstrating behaviors reported in Kendon's works to

gather support for the hypothesis. More specifically, we counted distance salutations, smiles, head dips, gaze aversion, close salutations, and whether people approached the robot or waited for it.

9.3.4.B Emotional engagement with the robot

Our second goal is to make the robot's engagement attempt pleasant for people, creating a positive first impression. Measurements of emotional engagement often include the estimation of affect through facial expressions [64], [70]. However, these data are difficult to use for our experiment since (i) there are still some people wearing masks, (ii) the acquired data is processed to anonymize people on videos, and (iii) face images captured with the robot's camera lack the necessary quality for reliable facial expression analysis in the wild (due to motion blur and reduced resolution for farther away people). Thus, we opted for two types of measurements to gather support for H1.3. As observational data, we used the number of people (not wearing masks) who smiled. From our experience, checking whether someone smiled or not during the experiment was feasible under the constraints of the experiment. In addition, we used self-reports from the web interface's questionnaire with items measuring the constructs of *perceived warmth*, *perceived discomfort*, and *perceived intelligence* taken from [74]. Some of these items showed sensibility to the manipulations of previous studies related to parts of this thesis, making us argue that they are appropriate.

9.3.4.C Behavioral and cognitive engagement with the task

To test whether people's behavioral engagement with the robot's proposed task was affected by how the robot initiated the encounter (formulated with hypothesis H1.4a), we analyzed how participants used the web interface. Thus, we checked for: (i) unique interface interactions, (ii) if people downloaded the reward, and (iii) if people demonstrated interest in the robot's social networks by clicking the *Show social networks* button.

Assessment of whether people were more cognitively engaged with the proposed task depending on the greeting model (formulated as hypothesis H1.4b) measured the number of unique words people used to describe the painting on **P5**.

9.3.5 Implementation details

We performed this experiment aiming at a fully autonomous system. Although a WoZ approach could potentially simplify the implementation, our goals are not solely to study the effects of the proposed models on human behaviors and perceptions but to have models that social robots can autonomously execute in practice. Ultimately, the system implemented in this experiment is the culmination of previous contributions in Part I of this thesis.

The social signals extracted by the perception pipeline feed the HMM that estimates and tracks people's mental states according to Kendon's greeting model, as proposed in section 5.7. There, we described and compared two versions: (i) a handcrafted model based on Kendon's observations of an outdoor birthday party and (ii) a data-driven model trained with indoor greetings. In this experiment, we used the handcrafted HMM model since the scenario was an open space, unlike the

Table 9.1: Summary of robot and task engagement hypothesis and associated metrics used in this work.

Engagement target	Engagement type	Hypothesis	Metrics
Robot	Behavioral engagement	H1.1: *The robot's engagement success increases if the robot (a) increases the number of Kendon's greeting phases that it follows, and (b) acts according to the full Kendon greeting model reacting to estimates of people's current greeting phase.*	- **In-the-wild:** Ordinal scale of interaction outcomes: 1 - Unaware to 4 - Success - **Invited:** Binary scale - Interacted/Did not interact
		H1.2: *People will behave more closely to Kendon's greeting model when the robot (a) increases the number of Kendon's greeting phases that it follows, and (b) acts according to the full Kendon greeting model reacting to estimates of people's current greeting phase.*	- **Observed:** Frequency of people demonstrating behaviors reported by Kendon's greeting model toward the robot.
	Emotional engagement	H1.3: *The closer a robot follows Kendon's greeting model during first encounters, the more emotionally engaged people will be in the interaction.*	- **Observed:** Frequency of people who smiled - **Self-reported:** Perceived warmth, discomfort, intelligence, and individual items
Task	Behavioral engagement	H1.4a: *The closer a robot follows Kendon's greeting model, the greater the behavioral engagement with the task.*	- **Interface logs:** Unique interface interactions, reward downloads, interest in Vizzy's social networks
	Cognitive engagement	H1.4b: *The closer a robot follows Kendon's greeting model, the greater the cognitive engagement with the task.*	- **Interface logs:** Number of unique words to describe painting (P5)

training data for the data-driven method. Preliminary tests supported this decision.

Due to a temporary hardware incompatibility in the Vizzy robot, it could not use an onboard GPU to perform the social signal processing from images. Thus, we had to rely on a local server to process the initial algorithms of the perception pipeline, specifically *Openpose*, *OpenHeadPose*, and 3D pose estimations. All the remaining code ran on Vizzy. The robot and the local server communicated via Wi-Fi.

9.4 Experiment in-the-wild

9.4.1 Methodology

The in-the-wild experiments consisted of having Vizzy freely interact with passersby. Thus, we used the slack time between invited participants to run the same greeting system without interruptions. The robot's initial conditions were the the ones described in section 9.3.1. After detecting someone with the robot in their field of view, the robot would attempt to initiate the interaction with that person using the currently active greeting model. If people left or the greeting failed, the robot would return to its initial pose and idle state. During each in-the-wild session, the robot executed the same model so that present people would not see the robot with distinct models, possibly acting contrary to their expectations. We only changed models after some shutdown time (for charging), attempting to ensure that the people sharing the space were different from those of the previous condition.

9.4.2 Sample

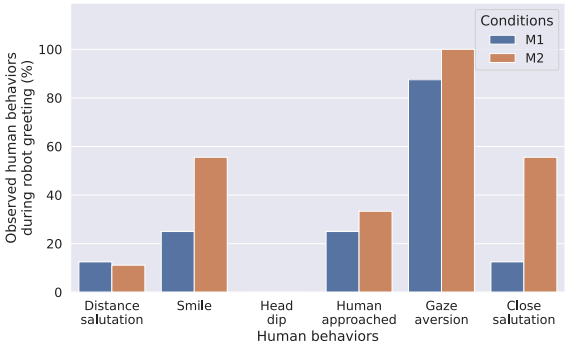
After the experiment, we performed a data cleansing process to remove samples with severe TFs that made it impossible to execute the greeting behaviors according to the plan. The most notable examples included the saturation of the robot's cameras due to extreme light conditions caused by the building's skylight (inhibiting the detection of social signals) and lost communications with the social signal processing server. Since these errors would lead to robot behaviors that do not follow the models under testing, we discarded them. We ended up with the following distribution of samples:

M1: 8 subjects
M2: 9 subjects
M3: 3 subjects

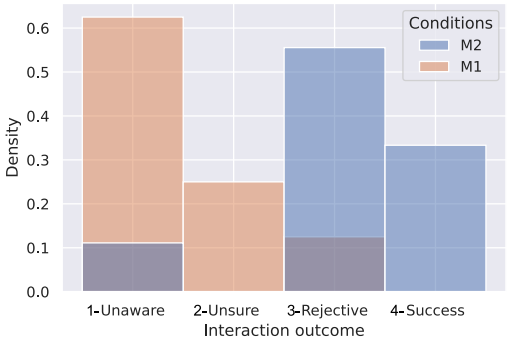
9.4.3 Results

We now present the results of the in-the-wild experiments. Our central focus is on two measures: (i) people's behaviors (according to Kendon's model) observed during the interaction attempts and (ii) the outcome of the robot's attempt on a one-to-four scale according to Figure 9.2.

In these experiments, we only obtained four interactions with the web interface (3 for M2 and 1 for M3). All of them were reported as first encounters. Of these, only one participant (M2) completed



(a) Human behaviors observed during in-the-wild encounters. Pearson's χ^2 tests did not find significant differences between M1 and M2. However, the plot shows a tendency of increased frequency in these behaviors for M2 over M1.



(b) Distribution of engagement outcomes. A one-tailed Mann-Whitney U test ($U = 64.0$, $p = 0.0027$, $RBC = -0.778$) shows that the outcome of robot attempts using Model M2 (robot could execute 6 greeting phases) are closer to success than the outcomes of Model M1 (robot could only execute 3 greeting phases).

Figure 9.9: Observed behaviors and interaction outcomes for in-the-wild participants.

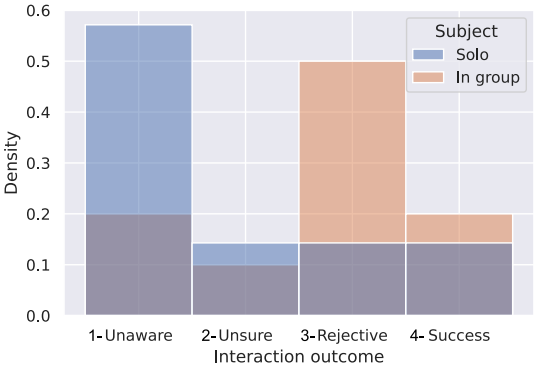
all steps of the web interface. Another participant in the M2 left the interaction after page **P3**. The two remaining participants (1 for M2 and 1 for M3) left the web interface after downloading the reward on Page **P6**. There were no web interface interactions in the M1. Due to this lack of samples, we cannot perform a more thorough analysis of web interface-related data.

Figure 9.9a shows the frequency of observed human behaviors for conditions M1 and M2. We performed a Pearson's χ^2 for each behavior, finding no statistically significant differences in frequencies for any. However, we hypothesize that this lack of significance occurred due to insufficient data since the plot depicts a tendency for more *smiles*, *humans who approach the robot*, *gaze aversion*, and especially the frequency of *close salutations* for the M2 condition. It is also worth noting that not all of people's *close salutation* gestures were the same as the robot's. Besides the *waves* people also performed: 4 high fives while the robot was executing the wave gesture and 2 attempted handshakes before the robot started waving / said it could not handshake due to COVID-19.

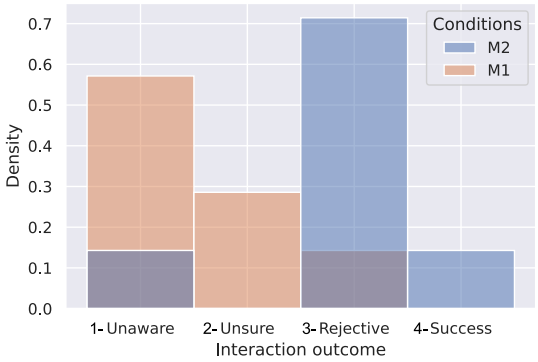
The distribution of engagement outcomes is depicted in Figure 9.9b. A one-tailed Mann-Whitney U test ($U = 64.0$, $p = 0.0027$) shows that model M2 is significantly more successful than model M1. We can see that while most outcomes of M1 end up in *Unaware* and *Unsure* outcomes, those of M2. The data has a rank biserial correlation of -0.778, which reveals a very large effect size. Additionally, we computed the common language effect size $CLES = 0.889$, which tells us that the probability of having a more successful outcome with M2 over M1 is of 88.9 %.

9.4.4 Investigation of group bias in interaction outcomes during in-the-wild encounters

During in-the-wild experiments, we cannot control whether people interact with the robot individually or in a group. While labeling the data, we noticed that many of the robot's interaction targets were part of a group of people. The results in subsection 9.4.3 do not take this information into account. Thus, we hypothesize that they may lead to misleading conclusions if people were more willing to interact with the robot while belonging to a group and if these events did not occur



(a) Distribution of robot approach outcomes for individuals and groups. We can see that the interaction outcome tends to be worse when the robot attempts to engage people who are alone, during the in-the-wild scenario.



(b) Distribution of engagement outcomes after balancing the samples (subsubsection 9.4.4.A). A one-tailed Mann-Whitney U test ($U = 41.5$, $p = 0.012$, $RBC = -0.693$) shows that the differences are still significant, with Model M2 still being closer to successful outcomes than Model M1.

Figure 9.10: Outcomes of robot attempted interactions.

equally among conditions.

Figure 9.10a shows us that interaction attempts with people in groups had a more successful outcome than those with alone people. Thus, we need to perform an extra analysis that considers this information.

9.4.4.A Worst-case data balance

Our new analysis balances the data with the worst-case scenario for our hypothesis. To do this, we removed samples from both conditions, making them equal in number and having the same ratio of solo and in-group interactions while discarding the minimum amount of data. Thus, balancing the data for the worst-case scenario consisted of removing the two most successful in-group samples from the M2 condition and the most unsuccessful solo sample from the M1. With this process, we ended with 7 samples per condition, where 3 samples were interaction attempts with an isolated person and 4 samples were interaction attempts with someone in a group.

9.4.4.B Results

We performed a one-tailed Mann-Whitney U test ($U = 41.5$, $p = 0.012$) that still confirms the interaction outcome is statistically significantly closer to success for the M2 condition than for the M1 with a rank biserial correlation of -0.693 (very large effect size) and the common language effect size $CLES = 0.847$ tells us that model M2 has a 84.7 % chance of achieving a better interaction outcome than M1. Thus, interaction with people in groups did not have a significant impact on the conclusions of subsection 9.4.3.

9.5 Experiment with invited participants

9.5.1 Methodology

In this experiment, we used the general interaction paradigm, asking participants to enter the experimental area from the point marked with a green star in Figure 9.6. If participants appeared in groups (for instance, friends in the study rooms), we kindly asked them to wait outside while only one was participating in the experiment so that people would not see each other's participation. When instructing participants, we used the following briefing text:

> "In this experiment, we are studying how humans and robots can coexist and interact in public places like museums. Thus, we ask you to be a participant in this exhibition, behaving as you would behave in a museum or art exhibition that has a robot present. Please, enjoy the experience and do not restrain your actions. Act as you would act in a real museum. To start, please enter the exhibition area at the second entrance. Ignore the paintings on the outside of the central area."
>
> Experiment briefing

9.5.2 Sample

After filtering samples with the same issues reported in section 9.4.2, we ended up with the following distribution of samples:

M1: 16 participants (13 male, 2 female, and 1 unknown)
M2: 16 participants (9 male and 7 female)
M3: 10 participants (6 male and 4 female)

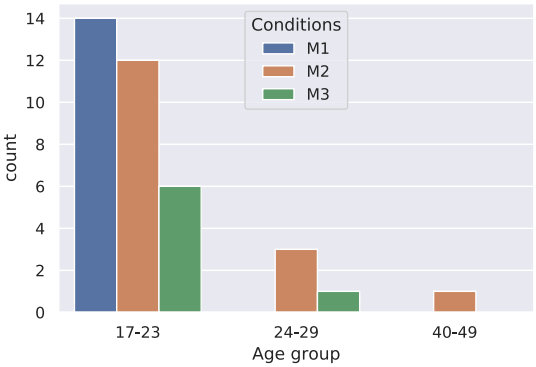
Then, we filtered our people who had previously interacted with Vizzy since our focus was to study interactions during first encounters. The remaining sample distribuition is the following:

M1: 14 participants (12 male, 1 female, and 1 unknown)
M2: 16 participants (9 male and 7 female)
M3: 7 participants (4 male and 3 female)

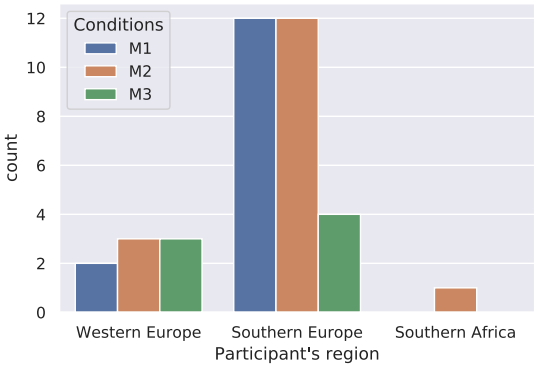
The age and geographic distributions of first encounter participants are depicted in Figure 9.11.

We note that the number of participants in condition M3 was lower than in the two remaining ones. Since this condition relied on a more extensive set of social signals, it was more sensitive to the TFs of the cameras and network. For instance, when the camera was saturated, it was still possible to detect static signals (people's positions, orientation, and gaze direction) with a lower frame rate. However, it became impossible to estimate and detect social data dependent on time, like gestures and velocity towards the robot, which are relevant for the HMM. Thus, the M3 condition was limited by environmental factors and time contraints.

Our methods to recruit participants were fourfold:



(a) Distribution of age groups reported by invited participants.



(b) Distribution of regional areas reported by invited participants.

Figure 9.11: Invited people's demographics.

- Advertising the experiment in the campus with posters at IST (with a link for enrollment).
- Advertising with panflets at IST and the Faculty of Psychology.
- Advertising the experiment online (with a link for enrollment).
- Personally inviting people from nearby study rooms.

9.5.3 Quantitative results - behaviors and cognitive engagement

9.5.3.A Interaction with the robot

During the experiment with invited participants, we obtained the following distribution of people with who the robot engaged successfully without being ignored:

M1: 8/14

M2: 7/16

M3: 5/7

We did not find statistically significant differences between the M1 and M2 conditions with a Pearson's χ^2 test ($\chi^2 = 0.134$, $p = 0.71$). We also did not find significant differences between the success of M2 and M3 with a Fisher's exact test ($p = 0.37$).

We present the observed human behaviors in Figure 9.12. We performed statistical analysis for each item, finding a statistically significant difference in the execution of a distance salutation between M2 and M3 through Fisher's exact test ($p = 0.005$). We note that the annotation of smiles depended on whether people were wearing masks (one masked in M1 and one masked in M3). Even though we could not get statistically significant differences in the frequencies of *close salutations*, we can see that there is a tendency for people to perform them more, the more similar the robot behaves to Kendon's greeting model.

The average time participants were willing to spend discussing art with Vizzy in a future encounter was $\mu = 9.64\,\text{min}$ ($\sigma = 1.33\,\text{min}$) for M1, $\mu = 10.0\,\text{min}$ ($\sigma = 1.86\,\text{min}$) for M2, and $\mu = 10.42\,\text{min}$ ($\sigma = 4.61\,\text{min}$) for M3. We did not find statistically significant differences between M1 and M2 nor between M2 and M3.

Figure 9.12: Observed human behaviors while the robot performed the greeting protocol of each condition. We only compared conditions that differ either the robot actuaction possibilities (M1 v.s. M2) or in the robot's signal sensing capabilities (M2 v.s. M3). The number of participants on each condition M1, M2, and M3 was of 14, 16, and 7, respectively. We note that we could only detect smiles of people not wearing masks (there were two people wearing masks: one in the M1 and one in the M3 conditions).

9.5.3.B Interaction with the web interface

Regarding the interaction with the interface, all participants submitted a description of the image on **P5**. The average number of unique words used to describe the painting was $\mu = 13.357$ ($\sigma = 7.967$) for M1, $\mu = 15.687$ ($\sigma = 15.395$) for M2, and $\mu = 18.14$ ($\sigma = 10.991$) for M3 (Figure 9.13a). Since the data did not meet the normality criteria, we used the non-parametric one-tailed Mann-Whitney U test between M1 and M2 ($U = 126.0$, $p = 0.287$) and between M2 and M3 ($U = 43.5$, $p = 0.210$), finding no statiscally significant differences.

The frequency of participants that downloaded the reward of **P6** was 7/14 for M1, 9/16 for M2, and 5/7 for M3 (Figure 9.13b). No significant results were found with Person's χ^2 test M1 and M2 ($\chi^2 = 0.001$, $p = 0.98$) nor with Fisher's exact test between M2 and M3 ($p = 0.642$).

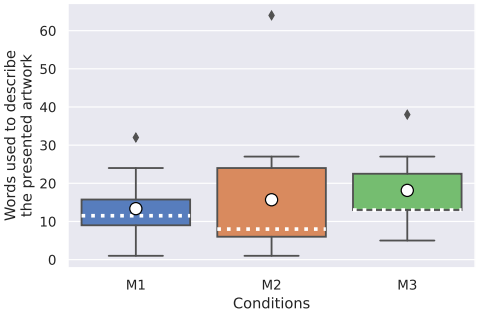
Participants demonstrated interest in Vizzy's social networks (through a click on *Show social networks* in **P8**) with a frequency of 3/14, 3/16, and 2/7 for M1, M2, and M3, respectively (Figure 9.13c). No significant differences were found with Fisher's exact test when comparing M1 with M2 ($p \approx 1.0$) nor between M2 and M3 ($p = 0.621$).

As shown in Figure 9.13d, the overall number of unique interactions within the web application also did not significantly differ among conditions.

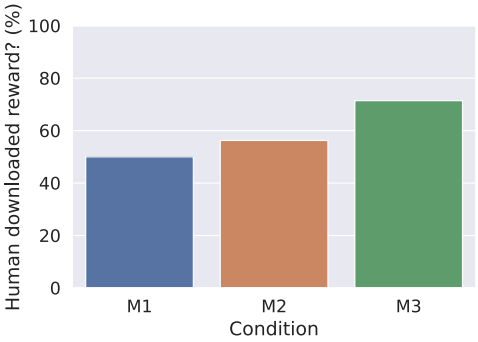
9.5.4 Quantitative results - questionnaire

The questionnaire focused mailing on the items of people's *perceived warmth*, *discomfort*, and *perceived intelligence*. First, we analyzed each item. Then, we aggregated them to study each dimension.

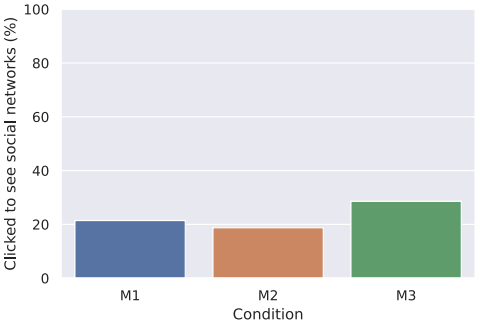
As shown in Figure 9.14a most *warmth* related items did not differ statistically between conditions, apart from the *compassionate* item. A one-tailed Mann-Whitney U test showed that peo-



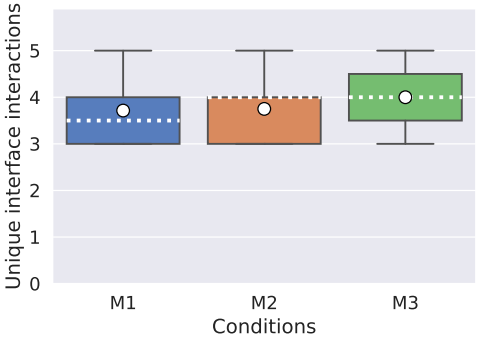
(a) Number of words the participants used to describe the painting presented in P5 (see Figure 9.8).



(b) Percentage of participants that downloaded the reward presented P6 (Figures 9.7 and 9.8).



(c) Percentage of participants that wanted to see more about Vizzy's social networks by clicking the *View social networks* button on P8 (see Figure 9.8)

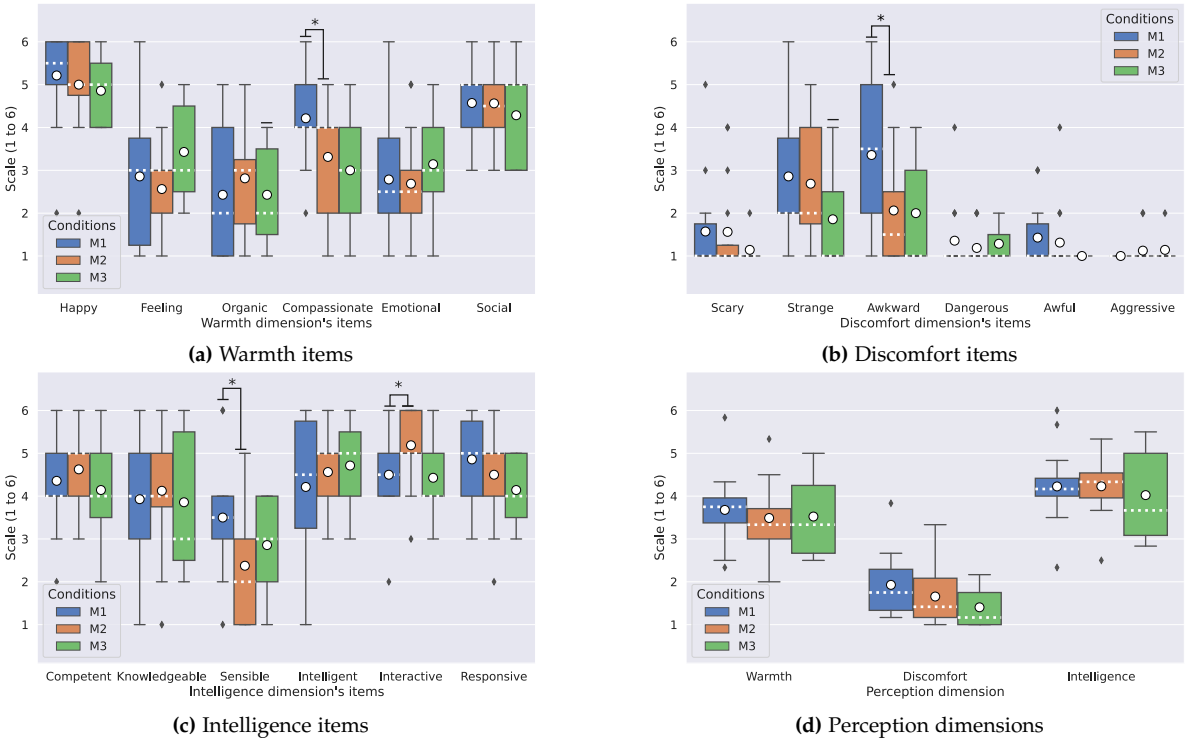


(d) Unique interactions with the web interface in the following transitions (see Figure 9.7): P3-P4, P5-P6, P6-P7 (*Download*), P7-P8, and P8-P8 (*Show social networks*).

Figure 9.13: Results of measuring participants' behavioral and cognitive engagement with the web interface.

ple consider Vizzy with the M1 model significantly more compassionate than when following M2 ($U = 155.5$, $p = 0.03$). A rank biserial correlation of -0.778, reveals a large effect size, while the common language effect size $CLES = 0.694$ tells us that the probability of people perceiving M1 as more *compassionate* than M2 is of 69.4 %. On the contrary, we found a non-significant tendency for people to find the robot that followed M3 more *feeling* that M2 (one-tailed Mann-Whitney U teste, $U = 35.5$, $p = 0.08$). We computed the *perceived warmth* dimension with the average of all items from Figure 9.14a ($\alpha = 0.762$), as shown in Figure 9.14d. The normality assumption was assured with a Shapiro-Wilke test for all conditions ($W_{M1} = 0.908$, $p_{M1} = 0.148$, $W_{M2} = 0.943$, $p_{M2} = 0.382$, $W_{M3} = 0.908$, $p_{M3} = 0.385$) as well as the homoscedasticity assumption (Levene test, $W = 0.515$, $p = 0.602$). Thus, we used two two-sided independent samples *t-tests* to check for significant differences between M1 and M2 ($t(28) = 0.63$, $p = 0.53$) and between M2 and M3 ($t(21) = -0.09$, $p = 0.93$). We could not find statistically significant differences.

Regarding *discomfort*, we found that people reported the robot following M1 to be significantly more *awkward* than when executing M2 (one-tailed Mann-Whitney U test - $U = 161$, $p = 0.019$). The data as a large effect size (rank biserial correlation of -0.4375) and the common language effect size $CLES = 0.719$ tells us that the probability of people perceiving M1 as more *awkward* than M2 is of 71.9 %. In addition, there was also a tendency to find M3 less *strange* than M2, although the results were not statistically significant. None of the other items had statistically significant results using a



(a) Warmth items

(b) Discomfort items

(c) Intelligence items

(d) Perception dimensions

Figure 9.14: Results for invited participants' reponses to the Warmth, Discomfort, and Intelligence dimensions and items on the web application questionnaire as boxplots. White circles represent the mean value and white dashes represent the median.

Mann-Whitney U test. Computing the mean value of these items resulted in the *discomfort* dimension results in Figure 9.14d ($\alpha = 0.726$). Since Shapiro-Wilke tests rejected the normality assumption, we used a Mann-Whitney U test to check for statistically significant differences, finding none. However, we note a slight tendency for people to report that the robot using M2 causes them less *discomfort* than M1, and that M3 generates less *discomfort* than M2.

Figure 9.14c shows the results for items of the *intelligence* dimension. While most items showed non-significant differences using a Mann-Whitney U test, we found two conflicting statistically significant results. First, people considered when the robot followed M1, it was more sensible than when using M2 (one-tailed Mann-Whitney U test: $U = 162.5$, $p = 0.02$, rank biserial correlation$= -0.451$, common language effect size $= 0.725$). However, they also reported that the robot was significantly more interactive when it used M2 than M1 (one-tailed Mann-Whitney U test: $U = 67.0$, $p = 0.026$, rank biserial correlation$= 0.402$, common language effect size $= 0.7$). Computing the *perceived intelligence* dimension from these items ($\alpha = 0.78$) resulted in the third boxplot group in Figure 9.14d. The data Shapiro-Wilkes tests did not reject the normality assumption for all conditions ($W_{M1} = 0.909$, $p_{M1} = 0.154$, $W_{M2} = 0.940$, $p_{M2} = 0.154$, $W_{M3} = 0.872$, $p_{M3} = 0.194$) and Levene's test also did not reject the homoscedasticity assumtption ($W = 1.592$, $p = 0.218$). Thus, we used an independent samples *t-test* to compare M1 ($t(28) = -0.01$, $p = 0.991$) with M2 and M2 with M3 ($t(21) = 0.553$, $p = 0.586$), which revealed no statistically significant differences.

9.5.5 Qualitative results - participants' feedback

Although questionnaires and observations provide quantitative data that we can use with statistical analysis tools, people's reports can give us additional insights we did not think of when preparing the experiment. Additionally, we can gather information to improve the proposed models in future works. In this section, we overview some of the participants' pos-experiment feedback. We collected these opinions from invited people who did not interact with the robot in its first attempt to engage with them. Unlike the remaining results in this thesis, we also considered the reports of people who had met Vizzy before. We observed three recurring topics in people's feedback: (i) the unawareness of the robot's movements or pose; (ii) the inappropriateness or ineffectiveness of behaviors; and (iii) being unnatural to interact socially with robots.

9.5.5.A Unawareness of the robot's movements and pose

These reports occurred with participants who experienced the M1 condition. People were surprised and somewhat scared to have the robot appear next to them without warning.

> "It's a bit scary to have the robot appear right next to us without noticing it."
>
> Participant the in M1 condition

> "I only noticed the robot when it was close to me."
>
> Participant in the M1 condition

These reports agree with our theory that social robots should exchange social signals to capture people's attention before starting to approach them and state intentions clearly. These behaviors belong to the first part of Kendon's greeting model, which M1 lacks.

9.5.5.B Ineffective or inappropriate behaviors

The following reports suggest that people did not interact with the robot because its behaviors did not correctly express the robot's intentions. Just using the *wave* during the *distance salutation* was not an obvious interaction intent cue for some people.

> "I didn't interact at first because I thought the robot's gestures were just used to decorate the scene. Maybe if the robot had walked toward me, the interaction would have been more natural. But it is very, very cute!"
>
> Participant in the M3 condition

> "I thought he was just saying *goodbye* [during distance salutation] and did not want to interact. During a normal situation I would also not interact with it. Maybe using the voice [to call for people's attention] would be more effective."
>
> Participant in the M2 condition

> "First, it looked like it was just saying *goodbye*. I suggest it to move toward people and starting talking more quickly. Getting closer is the most important thing to demonstrate the robot's intention to interact."
>
> Participant in the M3 condition

Unlike M2, in M3, the robot waited for people to demonstrate signals that they were willing to interact (thus, making the HMM transit to a subsequent state). People who experienced condition M3 suggested that the robot's intentions would become clear if it started moving toward them instead of just waving. The participant in the M1 condition also suggested adding a vocal callout to get people's attention and state its intentions would be effective. Kendon's greeting model also reports these behaviors, making this suggestion match the existing theoretical model.

Other participants complained about the inappropriateness of the *wave* behavior to close the interaction and its implementation.

> "I don't think the wave is the more appropriate to close the greeting. And I don't think people would like a handshake but a fist bump. So the robot should extend its arm, and we would touch it to finish the greeting."
>
> Participant in the M2 condition

> "It's interesting to have the robot look at us before starting to move, and it would be a funnier idea to have it explain each part of the exhibition. The wave movements are too monotonous. If the robot is close to people, it would make sense to warn them [with speech]. Otherwise, maybe not."
>
> Participant in the M1 condition

We agree that adding a handshake would probably clarify that the robot was greeting and intending to open the interaction with people. It was actually our intention to integrate the handshake into the greeting pipeline. However, we opted not to do it since we followed a zero-contact approach due to the COVID-19 pandemic. People's feedback also tells us to improve the *wave* gesture.

9.5.5.C Unnatural to interact socially with robots

Some people also told us that, even though they were aware of the robot's movements, they do not feel it to be natural to interact socially with robots.

> "I noticed what the robot was doing even though I was not looking at it. However, I think that, in a real scenario, I would also not interact with the robot making a response to the gesture... I don't feel it to be natural to interact socially with robots."
>
> Participant in the M2 condition

> "I think that, in a real-life situation, I would maybe perform the gesture [wave], but I would not try to talk with the robot. When I approach a robot, I'm not expecting to have a conversation with it."
>
> Participant in the M3 condition

These cases are challenging to address since people's behaviors depend upon their preconceived beliefs of robots as social entities. Unfortunately, we did not measure this dimension in the questionnaire, and thus we can only hypothesize. We believe that, in these cases, it is even more important for social robots to follow social norms and act in ways that make people perceive them as social actors during their first encounters.

9.6 Discussion and conclusions

In this experiment, we tested the greeting theories and some of the methods developed along with this thesis. We prepared a HRI study in a public place using two distinct but complementary approaches: (a) experiments in the wild and (b) experiments with invited participants. While the first is more ecologically plausible, the second gave us a structured control of all variables that strengthens the conclusions from statistical analysis. In-the-wild sessions mimic real-life applications more closely and let people interact with the robot more naturally. We will now discuss our results, starting with the most supported hypothesis and ending with the least supported.

Starting with our hypothesis, H1.1. Although the experiments with invited participants did not give us evidence on this matter, we discovered statistically significant differences when measuring the engagement success during in-the-wild sessions (Figures 9.9b and 9.10b). As we can see, the model that implements more of Kendon's greeting phases (M2) had a greater probability of being more successful than the other (M1). We made efforts to eliminate possible bias related to group/solo interactions, still achieving significant results. Thus we argue that we found support for hypothesis H1.1 for in-the-wild experiments. Unfortunately, we could not test M3 since it was impossible to gather samples with the necessary social signals with the robot sensors under the experimental site conditions in due time.

In hypothesis H1.2, we considered that people would act more like Kendon's greeting model the more similar the robot followed that model. We found some support for this hypothesis, as depicted in Figures 9.12 and 9.9. The most obvious differences were in the *distance salutation* and *close salutation*-related behaviors. We obtained significantly more distance salutations ($p < 0.01$) when following M3 than M2 during sessions with invited participants, as well as strong tendencies for more close salutation gestures when increasing the robot's similarity to Kendon's greeting model. A strong tendency was also observed for close salutation gestures during in-the-wild sessions between M1 and M2, although to a non-significant level. It is also worth noting that, unlike described in Kendon's greeting model, only one person performed the head dip gesture during the whole experiment. Whether the absence of head dips can be an important cue of how people perceive the

robot's social attributes remains an open question.

We now focus on the emotional engagement with the robot. On the one side, people of in-the-wild sessions had a noticeable (although non-significant) tendency to smile more at the robot when it implemented more phases of Kendon's greeting model (M1 v.s. M2) as shown in Figure 9.9a. However, this tendency was not observed during sessions with invited participants. On the other side, we found that people perceived the robot to be significantly less awkward and more interactive when behaving according to M2 compared to M1. On the contrary, people also reported that the robot was more sensible and more compassionate under condition M1 over M2. When looking at Figure 9.14d, we can see a minor tendency to find the interaction with the robot less uncomfortable the more similar the models are to Kendon's greeting model (M1 v.s. M2 and M2 v.s. M3). Given these observed tendencies and some significant differences favorable to our hypothesis, we argue that the metrics for emotional engagement with the robot gave us weak support for hypothesis H1.3.

Regarding the behavioral engagement with the task (hypothesis H1.4a), Figure 9.13b shows us that there is a minor tendency for more people to download the reward when the robot acts more closely to Kendon's greeting model (M1 v.s. M2) and also when it uses estimated human greeting states (using the HMM) (M2 v.s. M3). However, these were not statistically significant, not letting us draw further conclusions. The same minor tendency was also observed in the overall number of unique interactions with the interface (Figure 9.13d), but only when comparing M2 against M3 regarding people's interest in social networks (Figure 9.13c). All of these are just tendencies that did not have support from statistical analysis. Whether this lack of statistical significance was due to no effect or a small sample size required further investigation.

The metrics used during experimental sessions with invited participants did not support hypothesis H1.4b. We could not find statistically significant differences between the number of unique words to describe the painting on page **P5** among the different conditions. However, we argue that this metric may be too simple to capture people's cognitive engagement with the task and look forward to a more focused experiment that addresses whether *the closer a robot follows Kendon's greeting model, the greater the cognitive engagement with the task*.

Participants' feedback also goes in line with the description of Kendon's greeting model while providing important hints for future behavioral designs. There were suggestions that the robot should perform vocal callouts and handshakes. Although described in the model, they were not part of the current experiment. People's reports that the robot should perform the *distance salutation* and start moving forward to signal its intentions to interact hint that although relying on the HMM (model M3) and people's social signals may, in theory, decrease the number of rejections, the robot may be missing opportunities to show its intentions in initial stages of the interaction.

Finally, we highlight that deploying a mobile robot capable of perceiving various social signals from its onboard sensors remains a technical challenge, even with out-of-the-box solutions like OpenPose and OpenHeadPose. That became evident when we attempted to test M3. Nonetheless, this work shows that even with a limited amount of information (like people's body and head poses), Kendon's greeting model can still improve the robot's chances of successfully opening the

interaction with people during in-the-wild scenarios.

10

CONCLUSIONS AND FUTURE WORK

In this thesis, we worked to find theories and methods that could enhance how mobile social robots can engage with humans in first encounters, focusing on one-to-one interactions with healthy adults in public places. We aimed to increase engagement success, which we mainly measured with an ordinal scale (1 - Unreachable, 2 - Unnoticed, 3 - Unaware, 4 - Rejective, 5 - Accepted) based on Satake et al.'s [11], [12] sequence of engagement outcomes. Behavioral annotations were also an important tool to gather evidence for our hypotheses. Additionally, we collected qualitative observations and used questionnaires from the HRI and social sciences communities to gather quantitative self-reported measures. We believe these tools appropriately measured people's behavioral, emotional, and cognitive engagement with the robot. Throughout our work, we compiled evidence that we have found a possible answer for our research question: following the social scripts of Kendon's greeting model. To achieve this answer, we developed a set of multidisciplinary works, which we organized into two parts of this thesis: Part I - Methods and Part II - User studies.

10.1 Social agency of the Vizzy robot

Since Kendon's greeting protocol is a model of HHI, we needed assurance that the robot used in this thesis, Vizzy, was seen by the public as a social agent with what the literature defines as "automated social presence" [48]. This way, we would increase the odds of having people interacting socially with it, as they do it with other objects [46], [47]. The study of Chapter 7 was fundamental to assessing this feeling of social agency. Not only did we find that people perceived the Vizzy robot to have an "automated social presence," but we also observed them attempting to interact with it using a gesture reported by Kendon: the handshake. Further, comparing people's perceptions of the robot with those about a human coach showed there was (and there is) room for improvement, which people's reports on their experience with the robot complemented. While they could imagine the robot as a person, they still felt conflicted and saw it as having limited actions. This study confirmed the fitness of the Vizzy robot for our tasks and goals.

10.2 On the previous literature

The literature on this research path was still not homogeneous. Most works where mobile robots engaged with people used *ad-hoc* solutions and did not share a standard theory. As seen in chapters 3, while most models had distinct nomenclatures, we could identify close similarities between the description of their states and those described in Kendon's greeting model. Most of these only implemented behaviors relative to a subset of stages of Kendon's model. However, if not for the heterogeneity in the evaluation methods and nomenclature among these works, the differences in implementation could actually give us insights into whether following the model more closely could improve engagement, be useless, or even harmful (by enacting phenomena similar to the *uncanny valley*). With our taxonomy contribution from Chapter 2 and consequently our published survey [21] we hope to step forward into setting a standard basis for the problem of mobile human-robot engagement during first encounters.

10.3 Performing the close salutation via the handshake

Implementing a comfortable handshake (Chapter 4) for the Vizzy robot was the first step to testing the impacts of one of the scripts described in Kendon's model. While we could choose many other close salutation gestures, this is the default behavior in western civilizations. Moreover, it was actually one of the ways people attempted to interact with the robot during the study of Chapter 7. Our adult test subjects perceived our handshake implementation as safe, comfortable, and firm. It is worth noting that they reported no significant perception differences between a position controlled and force-based controlled handshake, although with a slight tendency to prefer the latter. Both handshakes were positively rated. These results let us conclude that for a population with similar physical characteristics having a more complex hand grasp controller may not be necessary to improve how people perceive the handshake, as long as safety is guaranteed.

We then used this new capability of Vizzy to test the impacts of a handshake during HRI on people's perceptions and behaviors. Through the study of Chapter 8, we showed that when the robot gave a handshake to participants during a first encounter, their willingness to help it in a future task increased. Further, introducing Vizzy with a handshake increased participants' perceptions of warmth, likeability, and animacy. These findings are consistent with our objectives, specifically O3 because this behavior improved people's perceptions of the robot. Thus, we can say that we achieved O3 successfully.

10.4 Perceiving social signals and acting accordingly

On chapters 5 and 6, we pursued the achievement of objectives O1 and O2. While the first focused on building a pipeline that can detect social signals and use them to follow the social scripts of greetings, the second focused on situations where the robot fails to respect social scripts.

To implement a pipeline that allows mobile social robots to engage with people, we studied the

possibilities of off-the-shelf algorithms and complemented them or how they connected as needed. Even though research in perception methods using computer vision is quite hectic, many works seem to forget their application on mobile social robots. For instance, we could not use the Openpose+OpenHeadPose duo to estimate arbitrary head orientations in 3D scenarios directly. The same could be said for other schemes using popular video gaze estimation methods like Gaze360 [103]. We proposed a correction that approximates these estimates to the real solution for arbitrary people in 3D scenarios. Our pipeline could estimate people's 3D body and head poses with sufficient accuracy for our interaction paradigm. The estimation error decreases with the distance to the person, which allows the robot to keep track of people during the approach even if the initial error is quite significant at greater distances.

Our greeting tracking HMM had promising results on greetings provided by the AVDIAR and UoL datasets, with both data-driven and handcrafted modes. However, posterior experiments in Chapter 9 showed that the method is too susceptible to social signal estimation failures that occur during in-the-wild experiments. Nonetheless, it was able to track people's greeting states and react to them during controlled experiments in the same environment.

The problem of being aware of unseen human body parts is fundamental for HRI even from a safety perspective. However, as far as we know, investigation on this topic is still scarce. Our skeleton completion scheme allows robots that use our pipeline to have a plausible estimate of people's unseen limbs. It was crucial to estimate people's 3D position using our "feet on the ground" assumption.

Regarding perception, we also highlight the importance of image suppression mechanisms for highly mobile cameras, like Vizzy's eyes. Without suppressing images, the perception pipeline would fail due to the mismatches between image timestamps and the robot's proprioception during fast head/eye movements. We argue that this issue should not be neglected by the robotics community, especially during in-the-wild conditions.

Regarding decision-making and actuation, our tests with behavior trees allowed us to implement three distinct engagement models in a modular and reactive approach. We were able to create behavior trees intuitively thanks to open-source tools such as the *BehaviorTree.CPP* library and the Groot application. Nonetheless, we needed to adapt the library for our use with ROS and the Vizzy robot, which did not make the process completely out-of-the-box. Approaching people with the ROS navigation stack was impossible, and thus we had to implement a reactive controller. This method made the robot vulnerable to local minima during navigation, causing it to become occasionally stuck during in-the-wild experiments.

When detecting robot failures, we highlight the importance of using the context of robot actions with users' FAU. These were the most prominent features. Unlike our expectations, user emotions did not contribute significantly to the detection of robot failures and could even hinder the performance of the error classifier.

Given the above conclusions, we argue that we partially achieved objectives O1 and O2. Our justification for the partial achievement of O1 is twofold. First, our pipeline allowed our robot to

detect and engage with people in-the-wild although with some limitations when estimating their greeting state. Second, the proposed algorithm of Chapter 6 achieved high accuracy on the test set but was never tested in the real world. Regarding O2, we conclude that Vizzy was able to engage with people in-the-wild, but experienced action failures, especially related to getting stuck in navigation.

10.5 Using Kendon's greeting model during first encounters

In Chapter 9, we could finally fully test whether Kendon's greeting model is an appropriate answer for our research question. In-the-wild results showed that the Vizzy robot was significantly more successful when engaging with people when acting according to the complete Kendon model, even though people rejected further interaction with the robot. This phenomenon, however, is out of the scope of the thesis since people's willingness to take time to interact with the robot depends on other external factors (like how persuasive the robot is or if people are in a hurry). From the thesis point of view, the robot was more successful at displaying its intentions to interact when following the complete model of Kendon. Moreover, we could observe that people displayed higher tendencies to perform distant and close salutations when the robot followed the model. With these results, we claim to have achieved objective O4.

However, this achievement is not without limitations. First, we did not test the model with an automatic handshake for the close salutation phase, which would be preferable to repeating the "wave" gesture. Indeed, some people complained that the robot's movements were too slow and robotic. Second, it was challenging to test whether reacting to more complex social signals (via the HMM) further improves the results or not since we could only gather three samples during the "in-the-wild" sessions, and the differences with invited participants sessions were not significant. Third, we believe this study should have a wider timeframe and be performed in other public places in order to get more participants from a more general population.

10.6 External limitations

Along this work, we faced several external limitations that deeply impacted its development. The most obvious one was the COVID-19 pandemic. Even though lockdowns did not have a significant impact on software development, since we prepared ourselves with simulation environments and datasets, they deeply impacted the experimental setup and recruitment of people. Moreover, mask-wearing made it impossible to collect some social signals, like people's smiles (which Kendon mentions). For safety purposes and to avoid spreading the disease, we also limited the robot's actions to avoid touching people. It also required us to develop a web server for interaction and questionnaires so that people could interact with the robot through their smartphones without touching a shared device. These efforts were quite time-consuming since our alternative for "normal times" would be to attach a tablet to the robot for people to interact and use common platforms for questionnaires (like Google tools).

On the other side, we faced critical hardware issues with the Vizzy robot that delayed our experiments and thesis considerably. Since the robot is an experimental platform, unique in the world, its hardware is not yet robust, especially during in-the-wild scenarios.

10.7 Future work

Given the broad nature of this thesis, the most obvious future work is to invest time in perfecting each of its modules. As discussed before, even though the HRI community asks for more in-the-wild experiments, our work showed that they are still very challenging to perform. Without robust hardware and methods, can we be certain that we are really testing interaction concepts?

Additionally, we finish this thesis willing to test more engagement models with distinct salutation gestures and more robust greeting phase estimation.

During the following subsections, we overview the future work for each contribution of this thesis.

10.7.1 Handshake design

For the time being, our work on handshake design has only focused on the force applied to the contacts during the handshake. To achieve a natural interaction, we must make the robot's arm move in a human-like manner. While previous works [217]–[219] cannot be directly applied to our robot without hardware improvements (i.e., adding torque sensors), the detailed 3D force information captured by the robot's hand sensors may be of use to control the arm movement. An additional line of future work can be research toward personalized handshakes, providing the most comfortable handshake to a particular person or the one that elicits perceptions of the robot's personality that are more appropriate for specific social contexts. This task is, however, very challenging since (i) people's preferences regarding handgrip forces and their distributions seem to be influenced by personality measures [198], (ii) the robot needs to gather information about people's preferences and perceptions (i.e., have appropriate feedback functions to learn), and (iii) it is difficult to capture people's feedback since they may express it with subtle signals. We hypothesize that we might be able to learn and predict individual preferences with the future implementation of a tactile palm that has an array of tactile sensors plus biosignals such as temperature and sweating sensors. These extra sensors might allow us to measure the forces exerted by the person on the robot's hand, which might correlate to the preferable force distribution for that person.

10.7.2 Perception pipeline

Regarding the perception pipeline, we note that even though we could use it in our experiments, its performance depends on the noise in sensing algorithms. We believe that even though recent advances in human pose estimation and tracking show impressive results, most works do not consider rapidly moving cameras in mobile robots and the underlying challenges of motion blur and over/underexposure. We need to find ways to overcome these challenges in future works. In addition, we note that the Markov assumption for our data-driven greeting state estimation model

can be a weakness. We need to test whether using methods that leverage the full knowledge of the sequence of greeting phases and social signals can improve the results. In addition, most works do not consider rapidly moving cameras in mobile robots and the underlying challenges of motion blur and over/underexposure. We need to find ways to overcome these challenges in future work. In addition, we note that the Markov assumption for our data-driven greeting state estimation model can be a weakness. We need to test whether using methods that leverage the full knowledge of the sequence of greeting phases and social signals can improve the results.

10.7.3 Self-perception of HRI failures

The proposed interaction error detection and classification system showed promising results in the human-robot interaction dataset. The natural follow-up for this work is to test it in real-time under the same block assembly collaborative scenario. Then, we could evaluate the system's robustness to in-the-wild error situations. From the user studies point of view, we claim that a study that evaluates how people react to a robot that automatically perceives its own errors is of paramount importance for human-robot interaction during first encounters. In addition, we highlight the importance of studying strategies to recover from failures.

10.7.4 Human-robot first encounter in elderly care centers

Regarding the user studies of human-robot first encounters in elderly care institutions, we propose the following future research. First, we believe we should invest time in developing the Wizard-Of-Oz interface to eliminate the need for two teleoperators. Controlling the robot's movements and speech simultaneously is essential for seamless interaction. A possible research path is to find inspiration in online video game systems that implement functionalities for quick communication (for instance, the Commo Rose from the Battlefield series[1]).

Second, comparing how elderly care center users perceive and engage with a robot that uses our autonomous model against one being teleoperated would benchmark the current status of this research path.

10.7.5 Effects of handshakes in engagement during first encounters

In our opinion, the impact of verbal greetings accompanied by a handshake should be studied in the future. In addition, we feel that handshaking and other types of close salutation gestures (such as waving, fist bumps, and high fives) should be compared. This information is important for roboticists during the behavioral design process.

10.7.6 Evaluation of greeting models in the wild

Future studies should consider additional actions described in Kendon's greeting model. As participants suggested, vocal callouts and handshakes could improve engagement and people's perceptions. It would be interesting to see if having these behaviors in both models further dilutes or

[1] https://battlefield.fandom.com/wiki/Commo_Rose

otherwise increases differences in performance. We also note that we just compared a three-state model against the six-state (full) Kendon's model. It would be interesting (but challenging and time-consuming) to study other combinations of Kendon's greeting phases to measure their importance. Finally, a possible follow-up study would include greeting models focused on engaging groups in first encounters instead of focusing on a single user.

Bibliography

[1] T. Fong, I. Nourbakhsh, and K. Dautenhahn, "A survey of socially interactive robots", *Robotics and Autonomous Systems*, vol. 42, no. 3, pp. 143–166, 2003, Socially Interactive Robots, ISSN: 0921-8890 (cit. on p. 1).

[2] F. Rubio, F. Valero, and C. Llopis-Albert, "A review of mobile robots: Concepts, methods, theoretical framework, and applications", *International Journal of Advanced Robotic Systems*, vol. 16, no. 2, p. 1 729 881 419 839 596, 2019. DOI: 10.1177/1729881419839596 (cit. on p. 1).

[3] A. Powers, S. Kiesler, S. Fussell, and C. Torrey, "Comparing a computer agent with a humanoid robot", in *Proceedings of the ACM/IEEE International Conference on Human–Robot Interaction*, ser. HRI '07, Arlington, Virginia, USA: ACM, 2007, pp. 145–152, ISBN: 978-1-59593-617-2 (cit. on p. 1).

[4] K. Shinozawa, F. Naya, J. Yamato, and K. Kogure, "Differences in effect of robot and screen agent recommendations on human decision-making", *International Journal of Human-Computer Studies*, vol. 62, no. 2, pp. 267–279, Feb. 2005, ISSN: 1071-5819 (cit. on p. 1).

[5] J. Li, "The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents", *International Journal of Human-Computer Studies*, vol. 77, pp. 23–37, 2015, ISSN: 1071-5819 (cit. on pp. 1, 88).

[6] P. Moreno, R. Nunes, R. Figueiredo, R. Ferreira, A. Bernardino, J. Santos-Victor, R. Beira, L. Vargas, D. Aragão, and M. Aragão, "Vizzy: A humanoid on wheels for assistive robotics", in *Robot 2015: Second Iberian Robotics Conference*, doi: 10.1007/978-3-319-27146-0_2, Springer, Cham, 2015, pp. 17–28 (cit. on p. 1).

[7] R. Ventura, M. Basiri, A. Mateus, J. Garcia, P. Miraldo, P. Santos, and P. Lima, "A domestic assistive robot developed through robotic competitions", in *IJCAI 2016 Workshop on Autonomous Mobile Service Robots*, New York, USA, 2016 (cit. on p. 1).

[8] D. Portugal, L. Santos, P. Alvito, J. Dias, G. Samaras, and E. Christodoulou, "Socialrobot: An interactive mobile robot for elderly home care", in *2015 IEEE/SICE International Symposium on System Integration (SII)*, Dec. 2015, pp. 811–816 (cit. on p. 1).

[9] M. Čaić, J. Avelino, D. Mahr, G. Odekerken-Schröder, and A. Bernardino, "Robotic versus human coaches for active aging: An automated social presence perspective", *International Journal of Social Robotics*, vol. 12, no. 4, pp. 867–882, Jul. 2019, ISSN: 1875-4805. DOI: 10.1007/s12369-018-0507-2 (cit. on pp. 1, 6, 9, 103, 107, 126).

[10] J. Avelino, H. Simão, R. Ribeiro, P. Moreno, R. Figueiredo, N. Duarte, R. Nunes, A. Bernardino, M. Čaić, D. Mahr, and O.-S. Gaby, "Experiments with vizzy as a coach for elderly exercise", in *ACM/IEEE International Conference on Human–Robot Interaction – Workshop on Personal Robots for Exercising and Coaching (PREC)*, Chicago, Illinois, USA, Mar. 2018 (cit. on pp. 1, 6, 9, 102, 107).

[11] S. Satake, T. Kanda, D. F. Glas, M. Imai, H. Ishiguro, and N. Hagita, "How to approach humans?-strategies for social robots to initiate interaction-", *Journal of the Robotics Society of Japan*, vol. 28, no. 3, pp. 327–337, 2010. DOI: 10.7210/jrsj.28.327 (cit. on pp. 2, 3, 19, 23, 24, 26–28, 30, 123–125, 127, 149).

[12] S. Satake, T. Kanda, D. F. Glas, M. Imai, H. Ishiguro, and N. Hagita, "A robot that approaches pedestrians", *IEEE Transactions on Robotics*, vol. 29, no. 2, pp. 508–524, Apr. 2013. DOI: 10.1109/tro.2012.2226387 (cit. on pp. 2, 3, 19, 23, 24, 26–28, 30, 123–125, 127, 149).

[13] E. Goffman, *Behavior in public places*. The Free Press, 1963 (cit. on pp. 2, 3, 13).

[14] E. Goffman, *Interaction ritual: Essays in face-to-face behavior*. Anchor Books, 1967 (cit. on p. 2).

[15] A. Kendon, "Goffman's approach to face-to-face interaction.", in *Erving Goffman: exploring the interaction order*, P. Drew and A. Wootton, Eds., Polity Press, 1988, pp. 14–40 (cit. on p. 2).

[16] C. Bassetti, "Social interaction in temporary gatherings: A sociological taxonomy of groups and crowds for computer vision practitioners", in *Group and Crowd Behavior for Computer Vision*, Elsevier, 2017, pp. 15–28 (cit. on pp. 2, 58, 63).

[17] A. Kendon and A. Ferber, "A description of some human greetings", *Comparative ecology and behaviour of primates*, vol. 591, p. 668, 1973 (cit. on pp. 3, 13).

[18] R. Firth, "Verbal and bodily rituals of greeting and parting", *The interpretation of ritual*, vol. 1972, pp. 1–38, 1972 (cit. on pp. 3, 13).

[19] C. L. Sidner, C. Lee, C. D. Kidd, N. Lesh, and C. Rich, "Explorations in engagement for humans and robots", *Artificial Intelligence*, vol. 166, no. 1-2, pp. 140–164, 2005 (cit. on p. 3).

[20] C. Silpasuwanchai, X. Ma, H. Shigemasu, and X. Ren, "Developing a comprehensive engagement framework of gamification for reflective learning", in *ACM Conference on Designing Interactive Systems*, 2016, pp. 459–472 (cit. on p. 3).

[21] J. Avelino, L. Garcia-Marques, R. Ventura, and A. Bernardino, "Break the ice: A survey on socially aware engagement for human–robot first encounters", *International Journal of Social Robotics*, vol. 13, no. 8, pp. 1851–1877, Jan. 2021, ISSN: 1875-4805. DOI: 10.1007/s12369-020-00720-2 (cit. on pp. 5, 8, 11, 23, 34, 150).

[22] J. Avelino, T. Paulino, C. Cardoso, R. Nunes, P. Moreno, and A. Bernardino, "Towards natural handshakes for social robots: Human-aware hand grasps using tactile sensors", *Paladyn, Journal of Behavioral Robotics*, vol. 9, no. 1, pp. 221–234, Aug. 2018. DOI: 10.1515/pjbr-2018-0017 (cit. on pp. 6, 8, 40, 80).

[23] J. Avelino, A. Gonçalves, R. Ventura, L. Garcia-Marques, and A. Bernardino, "Collecting social signals in constructive and destructive events during human-robot collaborative tasks", in *Companion of the ACM/IEEE International Conference on Human–Robot Interaction*, Cambridge, United Kingdom (Online), Mar. 2020 (cit. on pp. 6, 80).

[24] M. Carvalho*, J. Avelino*, A. Bernardino, R. Ventura, and P. Moreno, "Human-robot greeting: Tracking human greeting mental states and acting accordingly", in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, Prague, Czech Republic (Online), Sep. 2021 (cit. on pp. 6, 8, 55, 74).

[25] F. Loureiro, J. Avelino, P. Moreno, and A. Bernardino, "Detecting human–robot interaction failures through egocentric visual head-face analysis", in *EgoVIP - Egocentric vision for interactive perception, learning, and control, Workshop at IROS 2021*, Prague, Czech Republic (Online), Oct. 2021 (cit. on pp. 6, 80, 94).

[26] J. Avelino, F. Correia, J. Catarino, P. Ribeiro, P. Moreno, A. Bernardino, and A. Paiva, "The power of a hand-shake in human–robot interactions", in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, Madrid, Spain, Oct. 2018 (cit. on pp. 6, 9, 113).

[27] H. Simão, J. Avelino, N. Duarte, and R. Figueiredo, "Geebot: A robotic platform for refugee integration", in *Companion of the 2018 ACM/IEEE International Conference on Human–Robot Interaction*, Chicago, Illinois, USA, Mar. 2018 (cit. on p. 7).

[28] R. Livramento, J. Avelino, and P. Moreno, "Natural data-driven approaching behaviors of humanoid mobile robots for f-formations", in *IEEE International Conference on Autonomous Robot Systems and Competitions*, Ponta Delgada, Portugal, Apr. 2020 (cit. on p. 7).

[29] F. Loureiro, J. Avelino, P. Moreno, and A. Bernardino, "Self-perception of interaction errors through human non-verbal feedback and robot context", in *International Conference on Social Robotics*, Florence, Italy, Dec. 2022 (cit. on pp. 8, 80).

[30] E. Fehr and U. Fischbacher, "Social norms and human cooperation", *Trends in cognitive sciences*, vol. 8, no. 4, pp. 185–190, 2004 (cit. on p. 11).

[31] B. F. Malle, P. Bello, and M. Scheutz, "Requirements for an artificial agent with norm competence", in *AAAI/ACM Conference on AI, Ethics, and Society*, ser. AIES '19, Honolulu, HI, USA: Association for Computing Machinery, 2019, pp. 21–27, ISBN: 9781450363242. DOI: 10.1145/3306618.3314252 (cit. on p. 11).

[32] R. C. Schank and R. P. Abelson, *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Psychology Press, 2013 (cit. on p. 12).

[33] R. P. Abelson, "Psychological status of the script concept.", *American psychologist*, vol. 36, no. 7, p. 715, 1981 (cit. on p. 12).

[34] N. Hayes, *Foundations of Psychology*, Third. Cengage Learning EMEA, 2000, ISBN: 1861525893 (cit. on p. 12).

[35] A. Kendon, *Conducting Interaction: Patterns of Behavior in Focused Encounters*. Cambridge University Press, 1990, ISBN:978-0521389389 (cit. on pp. 12, 63, 72).

[36] M. L. Knapp, J. A. Hall, and T. G. Horgan, *Nonverbal communication in human interaction*. Cengage Learning, 2013 (cit. on p. 12).

[37] P. Drew, G. Raymond, and D. Weinberg, *Talk and Interaction in Social Research Methods*. SAGE Publications Ltd, 2006. DOI: 10.4135/9781849209991 (cit. on pp. 12, 13).

[38] M. Argyle, *Bodily Communication (2nd edition)*. Methuen, 1988 (cit. on p. 12).

[39] C. C. Bracken, L. W. Jeffres, and K. A. Neuendorf, "Criticism or praise? the impact of verbal versus text-only computer feedback on social presence, intrinsic motivation, and recall", *CyberPsychology & Behavior*, vol. 7, no. 3, pp. 349–357, Jun. 2004. DOI: 10.1089/1094931041291358 (cit. on p. 12).

[40] P. Ekman and W. V. Friesen, "The repertoire of nonverbal behavior: Categories, origins, usage, and coding", *Semiotica*, vol. 1, no. 1, Jan. 1969. DOI: 10.1515/semi.1969.1.1.49 (cit. on p. 12).

[41] R. Hastie and P. A. Kumar, "Person memory: Personality traits as organizing principles in memory for behaviors.", *Journal of Personality and Social Psychology*, vol. 37, no. 1, pp. 25–38, 1979. DOI: 10.1037/0022-3514.37.1.25 (cit. on p. 13).

[42] R. Hastie, *Person memory: The cognitive basis of social perception*. Lawrence Erlbaum Associates, 1980 (cit. on p. 13).

[43] T. K. Srull, "Person memory: Some tests of associative storage and retrieval models.", *Journal of Experimental Psychology: Human Learning & Memory*, vol. 7, no. 6, pp. 440–463, 1981. DOI: 10.1037/0278-7393.7.6.440 (cit. on p. 13).

[44] T. K. Srull, M. Lichtenstein, and M. Rothbart, "Associative storage and retrieval processes in person memory.", *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 11, no. 2, pp. 316–345, 1985. DOI: 10.1037/0278-7393.11.2.316 (cit. on p. 13).

[45] R. Jerónimo, L. Garcia-Marques, M. B. Ferreira, and C. N. Macrae, "When expectancies harm comprehension: Encoding flexibility in impression formation", *Journal of Experimental Social Psychology*, vol. 61, pp. 110–119, 2015, ISSN: 0022-1031. DOI: https://doi.org/10.1016/j.jesp.2015.07.007 (cit. on p. 13).

[46] B. Reeves and C. I. Nass, *The media equation: How people treat computers, television, and new media like real people and places.* Cambridge university press, 1996 (cit. on pp. 13, 103, 149).

[47] M. Heerink, B. Kröse, V. Evers, and B. Wielinga, "Assessing acceptance of assistive social agent technology by older adults: The almere model", *International Journal of Social Robotics*, vol. 2, no. 4, pp. 361–375, 2010 (cit. on pp. 13, 103, 149).

[48] J. Van Doorn, M. Mende, S. M. Noble, J. Hulland, A. L. Ostrom, D. Grewal, and J. A. Petersen, "Domo arigato mr. roboto: Emergence of automated social presence in organizational frontlines and customers' service experiences", *Journal of Service Research*, vol. 20, no. 1, pp. 43–58, 2017 (cit. on pp. 13, 103, 149).

[49] S. Greenspan, "Defining childhood social competence: A proposed working model.", *Advances in special education*, 1981 (cit. on pp. 13, 18).

[50] D. Schiffrin, "Opening encounters", *American Sociological Review*, vol. 42, no. 5, pp. 679–691, 1977, ISSN: 00031224 (cit. on pp. 13, 16, 128).

[51] A. Duranti, "Universal and culture-specific properties of greetings", *Journal of Linguistic Anthropology*, vol. 7, no. 1, pp. 63–97, 1997. DOI: https://doi.org/10.1525/jlin.1997.7.1.63 (cit. on p. 13).

[52] D. Morris, *Peoplewatching: the Desmond Miles guide to body language*. Vintage, Jan. 2002, ISBN: 0099429780 (cit. on p. 13).

[53] G. Yoshioka, T. Sakamoto, and Y. Takeuchi, "Polite approach to engrossing person based on two-dimensional attitude of interaction with other", in *IEEE International Symposium on Robot and Human Interactive Communication*, IEEE, Aug. 2018 (cit. on p. 14).

[54] E. T. Hall, *The hidden dimension*. Anchor Books, 1966 (cit. on p. 16).

[55] J. Åström, "Introductory greeting behaviour: A laboratory investigation of approaching and closing salutation phases", *Perceptual and Motor Skills*, vol. 79, no. 2, pp. 863–897, 1994 (cit. on p. 16).

[56] E. Goffman, *Relations in public: Microstudies of the public order*. Routledge, 2017 (cit. on pp. 16, 72, 125).

[57] B. Skyrms, *Social dynamics*. Oxford University Press, 2014 (cit. on pp. 17, 32, 83).

[58] J. F. Sorce, R. N. Emde, J. J. Campos, and M. D. Klinnert, "Maternal emotional signaling: Its effect on the visual cliff behavior of 1-year-olds.", *Developmental Psychology*, vol. 21, no. 1, pp. 195–200, 1985 (cit. on pp. 17, 83).

[59] T. A. Cavell, "Social adjustment, social performance, and social skills: A tri-component model of social competence", *Journal of clinical child psychology*, vol. 19, no. 2, pp. 111–122, 1990 (cit. on p. 18).

[60] N. R. Crick and K. A. Dodge, "A review and reformulation of social information-processing mechanisms in children's social adjustment.", *Psychological bulletin*, vol. 115, no. 1, p. 74, 1994 (cit. on p. 18).

[61] D. L. DuBois and R. D. Felner, "The quadripartite model of social competence: Theory and applications to clinical intervention.", in *Cognitive therapy with children and adolescents: A casebook for clinical practice.* The Guilford Press, 1996 (cit. on p. 18).

[62] M. R. Goldfried and T. J. D'Zurilla, "A behavioral-analytic model for assessing competence", in *Current Topics in Clinical and Community Psychology*, Elsevier, 1969, pp. 151–196. DOI: 10.1016/b978-1-4831-9972-6.50009-3 (cit. on p. 18).

[63] C. Shi, S. Satake, T. Kanda, and H. Ishiguro, "A robot that distributes flyers to pedestrians in a shopping mall", *International Journal of Social Robotics*, vol. 10, no. 4, pp. 421–437, Nov. 2017. DOI: 10.1007/s12369-017-0442-7 (cit. on pp. 20, 24, 25, 27, 28, 30, 35, 123, 127).

[64] R. Khosla, K. Nguyen, and M.-T. Chu, "Human robot engagement and acceptability in residential aged care", *International Journal of Human–Computer Interaction*, vol. 33, no. 6, pp. 510–522, 2017 (cit. on pp. 20, 134).

[65] S. M. Anzalone, S. Boucenna, S. Ivaldi, and M. Chetouani, "Evaluating the engagement with social robots", *International Journal of Social Robotics*, vol. 7, no. 4, pp. 465–478, 2015 (cit. on p. 20).

[66] Y. Feng, G. Perugia, S. Yu, E. I. Barakova, J. Hu, and G. Rauterberg, "Context-enhanced human–robot interaction: Exploring the role of system interactivity and multimodal stimuli on the engagement of people with dementia", *International Journal of Social Robotics*, pp. 1–20, 2021 (cit. on p. 20).

[67] I. Leite, M. McCoy, M. Lohani, N. Salomons, K. McElvaine, C. Stokes, S. Rivers, and B. Scassellati, "Autonomous disengagement classification and repair in multiparty child-robot interaction", in *IEEE International Symposium on Robot and Human Interactive Communication*, 2016, pp. 525–532 (cit. on p. 20).

[68] C. Kim, D. Kim, J. Yuan, R. B. Hill, P. Doshi, and C. N. Thai, "Robotics to promote elementary education pre-service teachers' stem engagement, learning, and teaching", *Computers & Education*, vol. 91, pp. 14–31, 2015 (cit. on pp. 20, 21).

[69] A. Merkouris, K. Chorianopoulos, and A. Kameas, "Teaching programming in secondary education through embodied computing platforms: Robotics and wearables", *ACM Transactions on Computing Education*, vol. 17, no. 2, May 2017 (cit. on p. 20).

[70] H. Chen, H. W. Park, and C. Breazeal, "Teaching and learning with children: Impact of reciprocal peer learning with a social robot on children's learning and emotive engagement", *Computers & Education*, vol. 150, p. 103 836, 2020, ISSN: 0360-1315 (cit. on pp. 20, 134).

[71] J. A. Fredricks, P. C. Blumenfeld, and A. H. Paris, "School engagement: Potential of the concept, state of the evidence", *Review of Educational Research*, vol. 74, no. 1, pp. 59–109, 2004. DOI: 10.3102/00346543074001059 (cit. on p. 20).

[72] C. Kim, J. Yuan, D. Kim, P. Doshi, C. N. Thai, R. B. Hill, and E. Melias, "Studying the usability of an intervention to promote teachers' use of robotics in stem education", *Journal of Educational Computing Research*, vol. 56, no. 8, pp. 1179–1212, 2019. DOI: 10.1177/0735633117738537 (cit. on p. 20).

[73] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi, "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots", *International Journal of Social Robotics*, vol. 1, no. 1, pp. 71–81, 2009 (cit. on pp. 20, 50, 87, 118).

[74] C. M. Carpinella, A. B. Wyman, M. A. Perez, and S. J. Stroessner, "The robotic social attributes scale (RoSAS)", in *ACM/IEEE International Conference on Human–Robot Interaction*, ACM Press, 2017. DOI: 10.1145/2909824.3020208 (cit. on pp. 20, 87, 118, 133, 134).

[75] A. Aron, E. N. Aron, and D. Smollan, "Inclusion of other in the self scale and the structure of interpersonal closeness.", *Journal of Personality and Social Psychology*, vol. 63, no. 4, pp. 596–612, 1992 (cit. on pp. 20, 87, 118).

[76] T. Nomura and T. Kanda, "Rapport–expectation with a robot scale", *International Journal of Social Robotics*, vol. 8, no. 1, pp. 21–30, 2016 (cit. on p. 20).

[77] B. W. Miller, "Using reading times and eye-movements to measure cognitive engagement", *Educational Psychologist*, vol. 50, no. 1, pp. 31–42, 2015. DOI: 10.1080/00461520.2015.1004068 (cit. on p. 21).

[78] T. Atapattu, M. Thilakaratne, R. Vivian, and K. Falkner, "Detecting cognitive engagement using word embeddings within an online teacher professional development community", *Computers & Education*, vol. 140, p. 103 594, 2019 (cit. on p. 21).

[79] J. Webster and R. T. Watson, "Analyzing the past to prepare for the future: Writing a literature review", *MIS quarterly*, pp. xiii–xxiii, 2002 (cit. on p. 23).

[80] J. vom Brocke, A. Simons, B. Niehaves, K. Riemer, R. Plattfaut, and A. Cleven, "Reconstructing the giant: On the importance of rigour in documenting the literature search process", in *ECIS 2009 Proceedings*, 2009 (cit. on p. 23).

[81] J. vom Brocke, A. Simons, K. Riemer, B. Niehaves, R. Plattfaut, and A. Cleven, "Standing on the shoulders of giants: Challenges and recommendations of literature search in information systems research", *Communications of the Association for Information Systems*, vol. 37, 2015. DOI: 10.17705/1cais.03709 (cit. on p. 23).

[82] X. Truong and T. Ngo, ""to approach humans?": A unified framework for approaching pose prediction and socially aware robot navigation", *IEEE Transactions on Cognitive and Developmental Systems*, vol. 10, no. 3, pp. 557–572, Sep. 2018, ISSN: 2379-8939. DOI: 10.1109/TCDS.2017.2751963 (cit. on pp. 23, 29).

[83] X.-T. Truong and T.-D. Ngo, "Dynamic social zone based mobile robot navigation for human comfortable safety in social environments", *International Journal of Social Robotics*, vol. 8, no. 5, pp. 663–684, May 2016. DOI: 10.1007/s12369-016-0352-0 (cit. on p. 23).

[84] F. Yang and C. Peters, "Appgan: Generative adversarial networks for generating robot approach behaviors into small groups of people", in *IEEE International Symposium on Robot and Human Interactive Communication*, 2019 (cit. on p. 23).

[85] E. Repiso, A. Garrell, and A. Sanfeliu, "Robot approaching and engaging people in a human-robot companion framework", in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, Oct. 2018. DOI: 10.1109/iros.2018.8594149 (cit. on p. 23).

[86] D. Brscic, T. Ikeda, and T. Kanda, "Do you need help? a robot providing information to people who behave atypically", *IEEE Transactions on Robotics*, vol. 33, no. 2, pp. 500–506, Apr. 2017. DOI: 10.1109/tro.2016.2645206 (cit. on pp. 24, 25, 27, 28, 30, 123, 127).

[87] Y. Kato, T. Kanda, and H. Ishiguro, "May i help you? - design of human-like polite approaching behavior", in *ACM/IEEE International Conference on Human–Robot Interaction*, IEEE, 2015, pp. 35–42 (cit. on pp. 24, 25, 27, 28, 30, 123, 127).

[88] M. A. Yousuf, Y. Kobayashi, Y. Kuno, A. Yamazaki, and K. Yamazaki, "How to move towards visitors: A model for museum guide robots to initiate conversation", in *IEEE International Symposium on Robot and Human Interactive Communication*, IEEE, Aug. 2013. DOI: 10.1109/roman.2013.6628543 (cit. on pp. 25, 28, 30).

[89] E. Saad, J. Broekens, M. A. Neerincx, and K. V. Hindriks, "Enthusiastic robots make better contact", in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, Nov. 2019 (cit. on pp. 25, 27, 28, 30).

[90] B. Heenan, S. Greenberg, S. Aghel-Manesh, and E. Sharlin, "Designing social greetings in human robot interaction", in *Conference on Designing interactive systems*, ACM Press, 2014. DOI: 10.1145/2598510.2598513 (cit. on pp. 25–28, 30, 123, 127).

[91] M. E. Foster, R. Alami, O. Gestranius, O. Lemon, M. Niemelä, J.-M. Odobez, and A. K. Pandey, "The MuMMER project: Engaging human–robot interaction in real-world public spaces", in *International Conference on Social Robotics*, Nov. 2016. DOI: 10.1007/978-3-319-47437-3_74 (cit. on pp. 25, 28–30).

[92] M. E. Foster, B. Craenen, A. Deshmukh, O. Lemon, E. Bastianelli, C. Dondrup, I. Papaioannou, A. Vanzo, J.-M. Odobez, O. Canévet, Y. Cao, W. He, A. Martínez-González, P. Motlicek, R. Siegfried, R. Alami, K. Belhassein, G. Buisan, A. Clodic, A. Mayima, Y. Sallami, G. Sarthou, P.-T. Singamaneni, J. Waldhart, A. Mazel, M. Caniot, M. Niemelä, P. Heikkilä, H. Lammi, and A. Tammela, "Mummer: Socially intelligent human–robot interaction in public spaces", in *AAAI Fall Symposium Series*, Arlington, VA, Nov. 2019 (cit. on pp. 25, 27, 28, 30).

[93] M. Zhao, D. Li, Z. Wu, S. Li, X. Zhang, L. Ye, G. Zhou, and D. Guan, "Stepped warm-up–the progressive interaction approach for human–robot interaction in public", in *Design, User Experience, and Usability. User Experience in Advanced Technological Environments. HCII 2019*, Springer International Publishing, 2019, pp. 309–327. DOI: 10.1007/978-3-030-23541-3_23 (cit. on pp. 26, 28, 30).

[94] A. Zaraki, M. Pieroni, D. D. Rossi, D. Mazzei, R. Garofalo, L. Cominelli, and M. B. Dehkordi, "Design and evaluation of a unique social perception system for human–robot interaction", *IEEE Transactions on Cognitive and Developmental Systems*, vol. 9, no. 4, pp. 341–355, Dec. 2017. DOI: 10.1109/tcds.2016.2598423 (cit. on pp. 29, 33).

[95] N. Lazzeri, D. Mazzei, L. Cominelli, A. Cisternino, and D. D. Rossi, "Designing the mind of a social robot", *Applied Sciences*, vol. 8, no. 2, p. 302, Feb. 2018. DOI: 10.3390/app8020302 (cit. on pp. 29, 33).

[96] R. Triebel, K. Arras, R. Alami, L. Beyer, S. Breuers, R. Chatila, M. Chetouani, D. Cremers, V. Evers, M. Fiore, H. Hung, O. A. I. Ramírez, M. Joosse, H. Khambhaita, T. Kucner, B. Leibe, A. J. Lilienthal, T. Linder, M. Lohse, M. Magnusson, B. Okal, L. Palmieri, U. Rafi, M. van Rooij, and L. Zhang, "SPENCER: A socially aware service robot for passenger guidance and help in busy airports", in *Springer Tracts in Advanced Robotics*, Springer International Publishing, 2016, pp. 607–622 (cit. on pp. 29, 33).

[97] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines", in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016 (cit. on pp. 29, 31, 92).

[98] Y. Cao, O. Canévet, and J.-M. Odobez, "Leveraging convolutional pose machines for fast and accurate head pose estimation", in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, Oct. 2018, pp. 1089–1094. DOI: 10.1109/IROS.2018.8594223 (cit. on pp. 29, 31, 33).

[99] S. Sheikhi and J.-M. Odobez, "Combining dynamic head pose–gaze mapping with the robot conversational state for attention recognition in human–robot interactions", *Pattern Recognition Letters*, vol. 66, pp. 81–90, Nov. 2015. DOI: 10.1016/j.patrec.2014.10.002 (cit. on p. 29).

[100] B. Amos, B. Ludwiczuk, and M. Satyanarayanan, "Openface: A general-purpose face recognition library with mobile applications", CMU-CS-16-118, CMU School of Computer Science, Tech. Rep., 2016 (cit. on p. 29).

[101] W. He, P. Motlicek, and J.-M. Odobez, "Deep neural networks for multiple speaker detection and localization", in *IEEE International Conference on Robotics and Automation*, IEEE, May 2018. DOI: 10.1109/icra.2018.8461267 (cit. on p. 29).

[102] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "OpenFace 2.0: Facial behavior analysis toolkit", in *IEEE International Conference on Automatic Face & Gesture Recognition*, IEEE, May 2018. DOI: 10.1109/fg.2018.00019 (cit. on pp. 31, 33, 71, 92).

[103] P. Kellnhofer, A. Recasens, S. Stent, W. Matusik, and A. Torralba, "Gaze360: Physically unconstrained gaze estimation in the wild", in *IEEE International Conference on Computer Vision*, Oct. 2019 (cit. on pp. 31, 33, 151).

[104] R. A. Güler, N. Neverova, and I. Kokkinos, "Densepose: Dense human pose estimation in the wild", in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7297–7306 (cit. on p. 31).

[105] B. Massé, S. Ba, and R. Horaud, "Tracking gaze and visual focus of attention of people involved in social interaction", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 11, pp. 2711–2724, Nov. 2018, ISSN: 1939-3539. DOI: 10.1109/TPAMI.2017.2782819 (cit. on p. 31).

[106] B. Massé, S. Lathuilière, P. Mesejo, and R. Horaud, "Extended gaze following: Detecting objects in videos beyond the camera field of view", in *IEEE International Conference on Automatic Face & Gesture Recognition*, May 2019, pp. 1–8. DOI: 10.1109/FG.2019.8756555 (cit. on p. 31).

[107] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain", *Image and Vision Computing*, vol. 27, no. 12, pp. 1743–1759, 2009 (cit. on p. 32).

[108] A. Nigam and L. D. Riek, "Social context perception for mobile robots", in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, Sep. 2015. DOI: 10.1109/iros.2015.7353883 (cit. on p. 32).

[109] A. H. Qureshi, Y. Nakamura, Y. Yoshikawa, and H. Ishiguro, "Intrinsically motivated reinforcement learning for human–robot interaction in the real-world", *Neural Networks*, vol. 107, pp. 23–33, Nov. 2018. DOI: 10.1016/j.neunet.2018.03.014 (cit. on p. 32).

[110] N. Mitsunaga, C. Smith, T. Kanda, H. Ishiguro, and N. Hagita, "Adapting robot behavior for human–robot interaction", *IEEE Transactions on Robotics*, vol. 24, no. 4, pp. 911–916, Aug. 2008. DOI: 10.1109/tro.2008.926867 (cit. on p. 32).

[111] E. McQuillin, N. Churamani, and H. Gunes, "Learning socially appropriate robo-waiter behaviours through real-time user feedback", in *ACM/IEEE International Conference on Human–Robot Interaction*, ser. HRI '22, Sapporo, Hokkaido, Japan: IEEE Press, 2022, pp. 541–550 (cit. on p. 32).

[112] H. Ritschel, T. Baur, and E. André, "Adapting a robot's linguistic style based on socially-aware reinforcement learning", in *IEEE International Symposium on Robot and Human Interactive Communication*, Aug. 2017, pp. 378–384. DOI: 10.1109/ROMAN.2017.8172330 (cit. on p. 32).

[113] K. Tsiakas, M. Abujelala, and F. Makedon, "Task engagement as personalization feedback for socially-assistive robots and cognitive training", *Technologies*, vol. 6, no. 2, p. 49, May 2018. DOI: 10.3390/technologies6020049 (cit. on p. 32).

[114] M. I. Ahmad, "An emotion and memory model for social robots: A long-term interaction", Ph.D. dissertation, Western Sydney University (Australia), 2018 (cit. on p. 33).

[115] P. Trung, M. Giuliani, M. Miksch, G. Stollnberger, S. Stadler, N. Mirnig, and M. Tscheligi, "Head and shoulders: Automatic error detection in human–robot interaction", in *ACM International Conference on Multimodal Interaction*, ACM Press, 2017. DOI: 10.1145/3136755.3136785 (cit. on pp. 33, 91, 92, 94).

[116] N. Mirnig, G. Stollnberger, M. Miksch, S. Stadler, M. Giuliani, and M. Tscheligi, "To err is robot: How humans assess and act toward an erroneous social robot", *Frontiers in Robotics and Artificial Intelligence*, vol. 4, May 2017. DOI: 10.3389/frobt.2017.00021 (cit. on pp. 33, 91).

[117] J. Avelino, R. Figueiredo, P. Moreno, and A. Bernardino, "On the perceptual advantages of visual suppression mechanisms for dynamic robot systems", *Procedia Computer Science*, vol. 88, pp. 505–511, 2016 (cit. on pp. 34, 71).

[118] S. Honig and T. Oron-Gilad, "Understanding and resolving failures in human–robot interaction: Literature review and model development", *Frontiers in psychology*, vol. 9, p. 861, 2018 (cit. on p. 34).

[119] N. Mirnig, M. Giuliani, G. Stollnberger, S. Stadler, R. Buchner, and M. Tscheligi, "Impact of robot actions on social signals and reaction times in hri error situations", in *International Conference on Social Robotics*, Springer, 2015, pp. 461–471 (cit. on p. 34).

[120] M. Giuliani, N. Mirnig, G. Stollnberger, S. Stadler, R. Buchner, and M. Tscheligi, "Systematic analysis of video data from different human–robot interaction studies: A categorization of social signals during error situations", *Frontiers in psychology*, vol. 6, p. 931, 2015 (cit. on pp. 34, 79, 91, 92).

[121] D. E. Cahya, R. Ramakrishnan, and M. Giuliani, "Static and temporal differences in social signals between error-free and erroneous situations in human-robot collaboration", in *International Conference on Social Robotics*, Springer, 2019, pp. 189–199 (cit. on p. 34).

[122] C. Sirithunge, A. G. B. P. Jayasekara, and D. P. Chandima, "Proactive robots with the perception of nonverbal human behavior: A review", *IEEE Access*, vol. 7, pp. 77 308–77 327, 2019. DOI: 10.1109/access.2019.2921986 (cit. on p. 34).

[123] F. M. Carlucci, L. Nardi, L. Iocchi, and D. Nardi, "Explicit representation of social norms for social robots", in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, Sep. 2015. DOI: 10.1109/iros.2015.7353970 (cit. on pp. 34, 35).

[124] D. Porfirio, A. Sauppé, A. Albarghouthi, and B. Mutlu, "Authoring and verifying human–robot interactions", in *ACM Symposium on User Interface Software and Technology*, ACM Press, 2018. DOI: 10.1145/3242587.3242634 (cit. on p. 35).

[125] M. Jindai and T. Watanabe, "Development of a handshake robot system based on a handshake approaching motion model", in *IEEE/ASME International Conference on Advanced Intelligent Mechatronics*, IEEE, 2007. DOI: 10.1109/aim.2007.4412423 (cit. on p. 36).

[126] M. Jindai and T. Watanabe, "A small-size handshake robot system based on a handshake approaching motion model with a voice greeting", in *IEEE/ASME International Conference on Advanced Intelligent Mechatronics*, IEEE, Jul. 2010. DOI: 10.1109/aim.2010.5695738 (cit. on p. 36).

[127] M. Jindai and T. Watanabe, "Development of a handshake request motion model based on analysis of handshake motion between humans", in *IEEE/ASME International Conference on Advanced Intelligent Mechatronics*, IEEE, Jul. 2011. DOI: 10.1109/aim.2011.6026975 (cit. on p. 36).

[128] S. Ota, M. Jindai, T. Fukuta, and T. Watanabe, "A handshake response motion model during active approach to a human", in *IEEE/SICE International Symposium on System Integration*, IEEE, Dec. 2014. DOI: 10.1109/sii.2014.7028056 (cit. on p. 36).

[129] S. Ota, M. Jindai, T. Sasaki, and Y. Ikemoto, "Handshake response motion model with approaching of human based on an analysis of human handshake motions", in *International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, IEEE, Oct. 2015. DOI: 10.1109/icumt.2015.7382396 (cit. on p. 36).

[130] D. Mura, E. Knoop, M. G. Catalano, G. Grioli, M. Bächer, and A. Bicchi, "On the role of stiffness and synchronization in human–robot handshaking", *International Journal of Robotics Research*, vol. 39, no. 14, pp. 1796–1811, Feb. 2020. DOI: 10.1177/0278364920903792 (cit. on p. 36).

[131] H. B. Amor, D. Vogt, M. Ewerton, E. Berger, B. Jung, and J. Peters, "Learning responsive robot behavior by imitation", in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, Nov. 2013. DOI: 10.1109/iros.2013.6696819 (cit. on p. 36).

[132] H. B. Amor, G. Neumann, S. Kamthe, O. Kroemer, and J. Peters, "Interaction primitives for human-robot cooperation tasks", in *IEEE International Conference on Robotics and Automation*, IEEE, May 2014. DOI: 10.1109/icra.2014.6907265 (cit. on p. 36).

[133] T. Shu, X. Gao, M. S. Ryoo, and S.-C. Zhu, "Learning social affordance grammar from videos: Transferring human interactions to human–robot interactions", in *IEEE International Conference on Robotics and Automation*, IEEE, May 2017. DOI: 10.1109/icra.2017.7989197 (cit. on p. 36).

[134] J. Campbell and K. Yamane, "Learning whole-body human-robot haptic interaction in social contexts", in *IEEE International Conference on Robotics and Automation*, IEEE, 2020, pp. 10 177–10 183 (cit. on p. 36).

[135] 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 *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2010, pp. 1668–1674. DOI: 10.1109/IROS.2010.5650851 (cit. on p. 40).

[136] M. Aragão, P. Moreno, and A. Bernardino, "Middleware interoperability for robotics: A ros–yarp framework", *Frontiers in Robotics and AI*, vol. 3, p. 64, 2016 (cit. on p. 41).

[137] T. Paulino, P. Ribeiro, M. Neto, S. Cardoso, A. Schmitz, J. Santos-Victor, *et al.*, "Low-cost 3-axis soft tactile sensors for the human-friendly robot vizzy", in *IEEE International Conference on Robotics and Automation*, 2017, pp. 966–971 (cit. on p. 42).

[138] J. G. Ziegler and N. B. Nichols, "Optimum settings for automatic controllers", *Journal of Dynamic Systems, Measurement, and Control*, vol. 115, no. 2B, pp. 220–222, 1942 (cit. on p. 49).

[139] Z. Wang, J. Yuan, and M. Buss, "Modelling of human haptic skill: A framework and preliminary results", *IFAC Proceedings Volumes*, vol. 41, no. 2, pp. 14 761–14 766, 2008 (cit. on pp. 49, 50).

[140] W. A. Bainbridge, S. Nozawa, R. Ueda, K. Okada, and M. Inaba, "A methodological outline and utility assessment of sensor-based biosignal measurement in human–robot interaction", *International Journal of Social Robotics*, vol. 4, no. 3, pp. 303–316, Aug. 2012 (cit. on p. 50).

[141] D. Kulić and E. Croft, "Physiological and subjective responses to articulated robot motion", *Robotica*, vol. 25, no. 1, pp. 13–27, 2007. DOI: https://doi.org/10.1017/S0263574706002955 (cit. on p. 50).

[142] H. Sun and P. Zhang, "Causal relationships between perceived enjoyment and perceived ease of use: An alternative approach", *Journal of the Association for Information Systems*, vol. 7, no. 9, p. 24, 2006 (cit. on p. 50).

[143] D. A. Forsyth and J. Ponce, *Computer Vision: A Modern Approach*. Prentice Hall Professional Technical Reference, 2002, ISBN: 0130851981 (cit. on p. 58).

[144] D. Scaramuzza, A. Martinelli, and R. Siegwart, "A toolbox for easily calibrating omnidirectional cameras", in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, Oct. 2006 (cit. on p. 58).

[145] N. R. F. Collaboration *et al.*, "A century of trends in adult human height", *Elife*, vol. 5, e13410, 2016 (cit. on p. 62).

[146] N. Carissimi, P. Rota, C. Beyan, and V. Murino, "Filling the gaps: Predicting missing joints of human poses using denoising autoencoders", in *Computer Vision – ECCV 2018 Workshops*, L. Leal-Taixé and S. Roth, Eds., Cham: Springer International Publishing, 2019, pp. 364–379, ISBN: 978-3-030-11012-3 (cit. on p. 68).

[147] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context", in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., Cham: Springer International Publishing, 2014, pp. 740–755, ISBN: 978-3-319-10602-1 (cit. on p. 68).

[148] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2d human pose estimation: New benchmark and state of the art analysis", in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2014 (cit. on p. 68).

[149] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016 (cit. on p. 70).

[150] H. Strasburger, I. Rentschler, and M. Jüttner, "Peripheral vision and pattern recognition: A review", *Journal of Vision*, vol. 11, no. 5, pp. 13–13, Dec. 2011, ISSN: 1534-7362. DOI: 10.1167/11.5.13 (cit. on p. 71).

[151] D. Jurafsky and J. H. Martin, "Hidden markov models", in *Speech and Language Processing*, 1st, USA: Prentice Hall PTR, 2000, pp. 548–563, ISBN: 0130950696 (cit. on pp. 72, 74).

[152] L. E. Baum and T. Petrie, "Statistical inference for probabilistic functions of finite state markov chains", *The annals of mathematical statistics*, vol. 37, no. 6, pp. 1554–1563, 1966 (cit. on p. 72).

[153] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm", *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977 (cit. on p. 72).

[154] I. D. Gebru, S. Ba, X. Li, and R. Horaud, "Audio-visual speaker diarization based on spatiotemporal bayesian fusion", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 5, pp. 1086–1099, 2018. DOI: 10.1109/TPAMI.2017.2648793 (cit. on p. 73).

[155] C. Coppola, S. Cosar, D. Faria, and N. Bellotto, "Automatic detection of human interactions from RGB-D data for social activity classification", in *IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2017, pp. 871–876 (cit. on p. 73).

[156] M. Carvalho, "Human-robot greeting: A model based on social studies and hidden markov models", M.S. thesis, Instituto Superior Técnico, Universidade de Lisboa, 2021 (cit. on p. 74).

[157] G. D. Forney, "The viterbi algorithm", *Proceedings of the IEEE*, vol. 61, no. 3, pp. 268–278, 1973 (cit. on p. 74).

[158] M. Colledanchise and P. Ogren, *Behavior Trees in Robotics and AI: An Introduction*. CRC Press, 2018, p. 208, ISBN: 9781138593732. DOI: 10.1201/9780429489105 (cit. on p. 76).

[159] A. Marzinotto, M. Colledanchise, C. Smith, and P. Ögren, "Towards a unified behavior trees framework for robot control", in *2014 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2014, pp. 5420–5427 (cit. on p. 76).

[160] P. Ogren, "Increasing modularity of UAV control systems using computer game behavior trees", in *AIAA guidance, navigation, and control conference*, 2012, p. 4458 (cit. on p. 76).

[161] C. Paxton, A. Hundt, F. Jonathan, K. Guerin, and G. D. Hager, "Costar: Instructing collaborative robots with behavior trees and vision", in *2017 IEEE international conference on robotics and automation (ICRA)*, IEEE, 2017, pp. 564–571 (cit. on p. 76).

[162] R. Siegwart, I. R. Nourbakhsh, and D. Scaramuzza, *Introduction to autonomous mobile robots*. MIT press, 2011 (cit. on p. 77).

[163] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie", *Image and Vision Computing*, vol. 28, no. 5, pp. 807–813, 2010, Best of Automatic Face and Gesture Recognition 2008, ISSN: 0262-8856. DOI: https://doi.org/10.1016/j.imavis.2009.08.002. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0262885609001711 (cit. on pp. 80, 84).

[164] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression", in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, IEEE, Jun. 2010. DOI: 10.1109/cvprw.2010.5543262. [Online]. Available: https://doi.org/10.1109%2Fcvprw.2010.5543262 (cit. on pp. 80, 84, 93).

[165] M. Mavadati, P. Sanger, and M. H. Mahoor, "Extended DISFA dataset: Investigating posed and spontaneous facial expressions", in *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, Jun. 2016. DOI: 10.1109/cvprw.2016.182. [Online]. Available: https://doi.org/10.1109%2Fcvprw.2016.182 (cit. on pp. 80, 84).

[166] M. Valstar and M. Pantic, "Induced disgust, happiness and surprise: An addition to the mmi facial expression database", in *Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect*, Paris, France, 2010, p. 65 (cit. on pp. 80, 84).

[167] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard, "Bp4d-spontaneous: A high-resolution spontaneous 3d dynamic facial expression database", *Image and Vision Computing*, vol. 32, no. 10, pp. 692–706, 2014, Best of Automatic Face and Gesture Recognition 2013, ISSN: 0262-8856. DOI: https://doi.org/10.1016/j.imavis.2014.06.002 (cit. on pp. 80, 84).

[168] L. Devillers, S. Rosset, G. D. Duplessis, M. A. Sehili, L. Bechade, A. Delaborde, C. Gossart, V. Letard, F. Yang, Y. Yemez, B. B. Turker, M. Sezgin, K. E. Haddad, S. Dupont, D. Luzzati, Y. Esteve, E. Gilmartin, and N. Campbell, "Multimodal data collection of human-robot humorous interactions in the joker project", in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, IEEE, Sep. 2015. DOI: 10.1109/acii.2015.7344594. [Online]. Available: https://doi.org/10.1109%2Facii.2015.7344594 (cit. on p. 80).

[169] K. Fischer, M. Jung, L. C. Jensen, and M. V. aus der Wieschen, "Emotion expression in HRI – when and why", in *2019 14th ACM/IEEE International Conference on Human–Robot Interaction (HRI)*, IEEE, Mar. 2019. DOI: 10.1109/hri.2019.8673078. [Online]. Available: https://doi.org/10.1109%2Fhri.2019.8673078 (cit. on p. 80).

[170] A. W. Bronkhorst, "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions", *Acta Acustica united with Acustica*, vol. 86, no. 1, pp. 117–128, 2000 (cit. on p. 80).

[171] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series", in *3rd International Conference on Knowledge Discovery and Data Mining*, ser. AAAIWS'94, Seattle, WA, 1994, pp. 359–370 (cit. on p. 81).

[172] H. Ding, G. Trajcevski, P. Scheuermann, X. Wang, and E. Keogh, "Querying and mining of time series data: Experimental comparison of representations and distance measures", *Proceedings of the VLDB Endowment*, vol. 1, no. 2, pp. 1542–1552, 2008 (cit. on p. 81).

[173] C. I. Hooker, "Amygdala response to facial expressions reflects emotional learning", *Journal of Neuroscience*, vol. 26, no. 35, pp. 8915–8922, Aug. 2006. DOI: 10.1523/jneurosci.3048-05.2006 (cit. on p. 83).

[174] C. Ebster and M. Garaus, *Store Design and Visual Merchandising: Creating Store Space That Encourages Buying*. Business Expert Press, Jun. 2011. DOI: 10.4128/9781606490952. [Online]. Available: https://doi.org/10.4128%2F9781606490952 (cit. on p. 86).

[175] A. Edland and O. Svenson, "Judgment and decision making under time pressure", in *Time Pressure and Stress in Human Judgment and Decision Making*, Springer US, 1993, pp. 27–40. DOI: 10.1007/978-1-4757-6846-6_2. [Online]. Available: https://doi.org/10.1007%2F978-1-4757-6846-6_2 (cit. on p. 86).

[176] L. Itti and C. Koch, "Computational modelling of visual attention", *Nature Reviews Neuroscience*, vol. 2, no. 3, pp. 194–203, Mar. 2001. DOI: 10.1038/35058500 (cit. on p. 86).

[177] K. M. Lee, W. Peng, S.-A. Jin, and C. Yan, "Can robots manifest personality?: An empirical test of personality recognition, social responses, and social presence in human–robot interaction", *Journal of Communication*, vol. 56, no. 4, pp. 754–772, Nov. 2006. DOI: 10.1111/j.1460-2466.2006.00318.x (cit. on p. 87).

[178] S. Jeong, C. Breazeal, D. Logan, and P. Weinstock, "Huggable: Impact of embodiment on promoting verbal and physical engagement for young pediatric inpatients", in *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, IEEE, Aug. 2017. DOI: 10.1109/roman.2017.8172290 (cit. on p. 88).

[179] M. Salem, G. Lakatos, F. Amirabdollahian, and K. Dautenhahn, "Would you trust a (faulty) robot?", in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human–Robot Interaction - HRI '15*, ACM Press, 2015 (cit. on p. 91).

[180] C. J. Hayes, M. Moosaei, and L. D. Riek, "Exploring implicit human responses to robot mistakes in a learning from demonstration task", in *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, IEEE, Aug. 2016 (cit. on p. 91).

[181] D. E. Cahya, R. Ramakrishnan, and M. Giuliani, "Static and temporal differences in social signals between error-free and erroneous situations in human-robot collaboration", in *Social Robotics*, Springer International Publishing, 2019, pp. 189–199 (cit. on p. 91).

[182] D. Kontogiorgos, A. Pereira, B. Sahindal, S. van Waveren, and J. Gustafson, "Behavioural responses to robot conversational failures", in *Proceedings of the 2020 ACM/IEEE International Conference on Human–Robot Interaction*, ACM, Mar. 2020. DOI: 10.1145/3319502.3374782 (cit. on pp. 91–94).

[183] T. Ethofer, S. Stegmaier, K. Koch, M. Reinl, B. Kreifelts, L. Schwarz, M. Erb, K. Scheffler, and D. Wildgruber, "Are you laughing at me? neural correlates of social intent attribution to auditory and visual laughter", *Human Brain Mapping*, vol. 41, no. 2, pp. 353–361, 2020. DOI: 10.1002/hbm.24806 (cit. on p. 91).

[184] D. Kontogiorgos, M. Tran, J. Gustafson, and M. Soleymani, "A systematic cross-corpus analysis of human reactions to robot conversational failures", in *Proceedings of the 2021 International Conference on Multimodal Interaction*, 2021, pp. 112–120 (cit. on p. 92).

[185] T. Simon, H. Joo, I. Matthews, and Y. Sheikh, "Hand keypoint detection in single images using multiview bootstrapping", in *CVPR*, 2017 (cit. on p. 92).

[186] P. Ekman, "An argument for basic emotions", *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992 (cit. on p. 92).

[187] P. Ekman and W. V. Friesen, "Measuring facial movement", *Environmental psychology and nonverbal behavior*, vol. 1, no. 1, pp. 56–75, 1976 (cit. on pp. 93, 94).

[188] F. Loureiro, "Detecting interaction failures through emotional feedback and robot context", M.S. thesis, Instituto Superior Técnico, Universidade de Lisboa, 2021 (cit. on pp. 93, 94).

[189] H. Simao and A. Bernardino, "User centered design of an augmented reality gaming platform for active aging in elderly institutions", in *5th International Congress on Sport Sciences Research and Technology Support (icSPORTS 2017)*, Sep. 2017 (cit. on pp. 102, 106).

[190] J. E. M. Cardona, M. S. Cameirao, T. Paulino, S. B. i Badia, and E. Rubio, "Modulation of physiological responses and activity levels during exergame experiences", in *2016 8th International Conference on Games and Virtual Worlds for Serious Applications (VS-GAMES)*, Sep. 2016, pp. 1–8 (cit. on pp. 102, 106).

[191] J. N. Bailenson, J. Blascovich, A. C. Beall, and J. M. Loomis, "Equilibrium theory revisited: Mutual gaze and personal space in virtual environments", *Presence: Teleoperators and Virtual Environments*, vol. 10, no. 6, pp. 583–598, 2001 (cit. on p. 107).

[192] M. Heerink, B. Kröse, V. Evers, and B. Wielinga, "Relating conversational expressiveness to social presence and acceptance of an assistive social robot", *Virtual Reality*, vol. 14, no. 1, pp. 77–84, 2010 (cit. on p. 107).

[193] A. J. Cuddy, S. T. Fiske, and P. Glick, "Warmth and competence as universal dimensions of social perception: The stereotype content model and the bias map", *Advances in Experimental Social Psychology*, vol. 40, pp. 61–149, 2008 (cit. on p. 107).

[194] J. W. Creswell and J. D. Creswell, *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications, 2017 (cit. on p. 108).

[195] H. R. Bernard and H. R. Bernard, *Social research methods: Qualitative and quantitative approaches*. Sage Publications, 2012 (cit. on p. 108).

[196] T. J. Reynolds and J. Gutman, "Laddering theory, method, analysis, and interpretation", *Journal of Advertising Research*, vol. 28, no. 1, pp. 11–31, 1988 (cit. on p. 108).

[197] S. Dolcos, K. Sung, J. J. Argo, S. Flor-Henry, and F. Dolcos, "The power of a handshake: Neural correlates of evaluative judgments in observed social interactions", *J. of Cognitive Neuroscience*, vol. 24, no. 12, pp. 2292–2305, 2012 (cit. on pp. 113, 115).

[198] W. F. Chaplin, J. B. Phillips, J. D. Brown, N. R. Clanton, and J. L. Stein, "Handshaking, gender, personality, and first impressions.", *Journal of personality and social psychology*, vol. 79, no. 1, p. 110, 2000 (cit. on pp. 113, 153).

[199] G. L. Stewart, S. L. Dustin, M. R. Barrick, and T. C. Darnold, "Exploring the handshake in employment interviews.", *J. of Applied Psychology*, vol. 93, no. 5, p. 1139, 2008 (cit. on p. 113).

[200] J. Schroeder, J. Risen, F. Gino, and M. I. Norton, "Handshaking promotes cooperative dealmaking", *Harvard Business School NOM Unit Working Paper No. 14-117; Harvard Business School Marketing Unit Working Paper No. 14-117. DOI: 10.2139/ssrn.2443674*, 2014 (cit. on p. 113).

[201] A. Paiva, F. P. Santos, and F. C. Santos, "Engineering pro-sociality with autonomous agents", in *Proc of AAAI 2018*, 2018 (cit. on p. 113).

[202] S. Rosenthal, J. Biswas, and M. Veloso, "An Effective Personal Mobile Robot Agent Through Symbiotic Human–Robot Interaction", in *International Joint Conference on Autonomous Agents and Multi-Agent Systems*, 2010 (cit. on p. 114).

[203] M. Veloso, J. Biswas, B. Coltin, and S. Rosenthal, "CoBots: Robust Symbiotic Autonomous Mobile Service Robots", in *International Joint Conferemce on Artificial Intelligence*, Jul. 2015 (cit. on p. 114).

[204] B. Kühnlenz, S. Sosnowski, M. Buß, D. Wollherr, K. Kühnlenz, and M. Buss, "Increasing helpfulness towards a robot by emotional adaption to the user", *Int. J. of Social Robotics*, vol. 5, no. 4, pp. 457–476, Nov. 2013, ISSN: 1875-4805 (cit. on p. 114).

[205] V. Srinivasan and L. Takayama, "Help me please: Robot politeness strategies for soliciting help from humans", in *ACM Conf. on Human Factors in Computing Systems*, 2016 (cit. on pp. 114, 118).

[206] A. Nakata, M. Shiomi, M. Kanbara, and N. Hagita, "Does being hugged by a robot encourage prosocial behavior?", in *ACM/IEEE Int. Conf. on Human–Robot Interaction*, 2017 (cit. on p. 114).

[207] M. Y. Tsalamlal, J.-C. Martin, M. Ammi, A. Tapus, and M.-A. Amorim, "Affective handshake with a humanoid robot: How do participants perceive and combine its facial and haptic expressions?", in *Int. Conf. on Affective Computing and Intelligent Interaction*, 2015 (cit. on p. 114).

[208] C. Bevan and D. Stanton Fraser, "Shaking hands and cooperation in tele-present human-robot negotiation", in *ACM/IEEE Int. Conf. on Human–Robot Interaction*, 2015 (cit. on p. 114).

[209] J. B. Van Erp and A. Toet, "Social touch in human–computer interaction", *Frontiers in Digital Humanities*, vol. 2, p. 2, 2015 (cit. on p. 115).

[210] R.-V. Joule and N. Guéguen, "Touch, compliance, and awareness of tactile contact", *Perceptual and Motor Skills*, vol. 104, no. 2, pp. 581–588, 2007 (cit. on p. 115).

[211] N. Guéguen and J. Fischer-Lokou, "Tactile contact and spontaneous help: An evaluation in a natural setting", *The J. of Social Psychology*, vol. 143, no. 6, pp. 785–787, 2003 (cit. on p. 115).

[212] A. Mazur, E. Rosa, M. Faupel, J. Heller, R. Leen, and B. Thurman, "Physiological aspects of communication via mutual gaze", *American Journal of Sociology*, vol. 86, no. 1, pp. 50–74, 1980 (cit. on p. 126).

[213] L. M. Coutts and F. W. Schneider, "Affiliative conflict theory: An investigation of the intimacy equilibrium and compensation hypothesis.", *Journal of Personality and Social Psychology*, vol. 34, no. 6, p. 1135, 1976 (cit. on p. 126).

[214] Y. Zhang, J. Beskow, and H. Kjellström, "Look but don't stare: Mutual gaze interaction in social robots", in *International Conference on Social Robotics*, A. Kheddar, E. Yoshida, S. S. Ge, K. Suzuki, J.-J. Cabibihan, F. Eyssel, and H. He, Eds., Cham: Springer International Publishing, 2017, pp. 556–566, ISBN: 978-3-319-70022-9 (cit. on p. 126).

[215] C. J. Stanton and C. J. Stevens, "Don't stare at me: The impact of a humanoid robot's gaze upon trust during a cooperative human–robot visual task", *International Journal of Social Robotics*, vol. 9, no. 5, pp. 745–753, 2017 (cit. on p. 126).

[216] R. Mead and M. J. Matarić, "Autonomous human–robot proxemics: Socially aware navigation based on interaction potential", *Autonomous Robots*, vol. 41, no. 5, pp. 1189–1201, Jun. 2016. DOI: 10.1007/s10514-016-9572-2 (cit. on p. 130).

[217] T. Kasuga and M. Hashimoto, "Human-robot handshaking using neural oscillators", in *IEEE International Conference on Robotics and Automation*, 2005, pp. 3802–3807 (cit. on p. 153).

[218] Y. Yamato, M. Jindai, and T. Watanabe, "Development of a shake-motion leading model for human-robot handshaking", in *2008 SICE Annual Conference*, 2008, pp. 502–507 (cit. on p. 153).

[219] G. Avraham, I. Nisky, H. L. Fernandes, D. E. Acuna, K. P. Kording, G. E. Loeb, and A. Karniel, "Toward perceiving robots as humans: Three handshake models face the turing-like handshake test", *IEEE Transactions on Haptics*, vol. 5, no. 3, pp. 196–207, 2012 (cit. on p. 153).