Master thesis

Non-verbal communication with the focus on manipulating objects

Università di Genova

Anna Hauschild

Department of Informatics, Bioengineering, Robotics and System Engineering (DIBRIS)

University of Genova

Supervisor

Francesco Rea, Alessandra, Sciutti, Fulvio Mastrogiovanni, Nicholetta Noceti

In partial fulfillment of the requirements for the degree of

European Master on Advanced Robotics

August 28, 2024

Acknowledgements

This thesis is the result of six months work in the Italian Institute of Technology under the supervision of Alessandra Sciutti, Francesco Rea, Nicoletta Noceti and Fulvio Mastrogiovanni. This work was included in the European Master on Advanced Robotics during which I had a lot of support from Leslie Cubizolles and Giulia Repetto. As first practical experience in research I would like to thank Alessandra Sciutti, Francesco Rea, Linda Lastrico and Luca Garello for their constant support and help, for giving me the freedom to explore my research field and providing me a guideline when necessary. This work would not have been possible without Federico Figari Tomenotti and Nuno Ferreira Duarte, who generously took the time to explain parts of their work to me and made it possible to me using some of their functionalities and ideas in this project. I would also like to thank the entire Human Technology Team for warmly welcoming me and supporting me throughout this journey.

I dedicate this thesis to my friends. They have been my dearest supporters, constant guides, and cherished companions, both within and outside of university life. I also dedicate this work to all those who devote themselves to responsible technology, embracing it with care and caution, ensuring that it serves humanity wisely and compassionately.

Abstract

Social robotics aims to bridge the gap between humans and machines, making interactions more intuitive, natural, and beneficial for both parties. The primary goal in creating them is to model their behavior and perception as closely as possible to ours. By doing so, interaction behaviors between humans are investigated, modeled and replicated on the robot. Enhancing the naturalness and efficiency of human-robot cooperation helps to overcome existing barriers. As humans, we perceive cues unconsciously, without needing explicit communication. In this thesis we focus on non-verbal communication cues during a natural transportation movement with the goal of obtaining information about the manipulated object. We investigate the gaze of a person through the integrated iCub camera and eye-tracker glasses to predict the carefulness level of the person. Our data analyses have led us to conclude that the image processing on the iCub system does not meet the performance requirements for this detail-oriented task. Nevertheless, we were able to demonstrate with the groundtruth data that head position serves as a good approximation for gaze detection and can accurately classify carefulness with a true positive rate of over 70% within 0.65 seconds.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivations

As humanoid robots advance, they will increasingly be able to understand and replicate human emotions, gestures, and expressions. This capability will enable them to communicate and collaborate with humans in an intuitive and seamless manner. Such nuance in human interaction is crucial for creating robots that can integrate effortlessly into our daily lives. Current research in computer-based gaze estimation continues to improve, allowing gaze to be integrated into robots as an important component. This approach provides an alternative and valuable contribution to challenging object manipulation tasks. Traditional object detection methods can be extremely complex, such as determining the water level in a transparent glass solely through computer vision. Instead, by analyzing human behavior, movements, and gaze patterns in various situations, robots can infer important information about the objects being manipulated. For example, the way a person handles an object, their careful movements, and their focus can reveal characteristics about the object, such as its weight, fragility, or content level.

By shifting the focus from direct object analysis to understanding human interaction with objects, robots can better adapt to and assist in tasks that are otherwise difficult for pure computer vision systems. This method leverages the natural human ability to interpret non-verbal cues and apply it to robotic systems, making them more capable of operating in dynamic and complex environments. This not only enhances the robots' functional capabilities but also brings them closer to human-like perception and interaction.

1.2 Context of the Study

In the interdisciplinary research field of cognitive science and human-robot interaction, the perception and action of body language play a significant role. Humans are highly skilled at communicating through non-verbal cues during interactions. Unconsciously perceived signals are naturally understood and facilitate our ability to react and collaborate with one another. We have spent our entire lives learning how to interact with others, and our expectations help us predict another person's actions.

The concept of mirror neurons supports this understanding: we perceive another person's actions in the same way we would execute those actions ourselves. This neural mechanism not only enhances our ability to understand others but also enables us to anticipate and respond to their movements, making our interactions more fluid and effective [1]. For example, when a person lifts a bottle, the movement and velocity can indicate how full it is, even without seeing inside. This information helps someone receiving the bottle to prepare and apply the right amount of force to take it over smoothly.

To exploit these capabilities, humanoid robots are designed to resemble humans in their behavior. If robots look somewhat human-like, it sets the expecta-

tion that they will behave in a human-like manner. To avoid the uncanny valley effect, that describes the uneasy feeling of an almost human like presence, it is common practice to design robots with human-like behavior but less human-like features, such as realistic skin or hair. This approach helps maintain user comfort while ensuring the robots' actions and interactions are as natural and intuitive as possible.

1.3 Objectives and Contributions

Since we are trained in perceiving and interpreting human cues, we naturally expect robots to act in a similar manner. When robots exhibit human-like behaviors, we can apply the same social rules and expectations, facilitating smoother and more natural interactions.

Several studies have investigated how humans accept robots and whether their behavior changes when interacting with a person versus a robot. The research indicates that the efficiency of collaborative tasks can be comparable whether the interaction is between two humans or between a human and a robot. For instance, the smoothness of object handover tasks improves significantly when the robot moves with the same velocity profile as its human counterpart.

Individuals, both human and robots, capable of interpreting the gaze of another can enhance collaborative performance. For effective teamwork, it is important that both parties understand and predict each other's actions. Robots that can both read behavior cues and reproduce human-like behavioral patterns set them apart from traditional machines. This bidirectional capability ensures that humans can also interpret the robot's focus and level of carefulness, further enhancing the collaborative process.

From our experience and behavior, we know that when a person is moving

slowly and their gaze is consistently focused on an object, they are likely being careful. Such a behaviour can be measured and used as an input feature for a machine learning model. We aim to find a carefulness model based on gaze cues to deduce object properties. Our research explores how robots perceive and demonstrate these social cues during careful transportation tasks. This is important for ensuring that robots can accurately interpret and respond to human social signals in real-world situations. A key part of our work involves integrating robotic sensors that mimic human senses, such as sight, hearing, and touch. We are specifically testing the iCub camera, built into the robot's head, to see if it works well enough for our needs compared to eye-tracking glasses. While eye-tracking glasses are effective, they are often uncomfortable and less suitable for natural interactions. By focusing on sensor integration, we aim to make our approach more practical for real-world use. Additionally, understanding non-verbal cues is crucial for robots to perform well in teamwork. Our research looks at how humans and robots work together, focusing on how they share tasks and achieve common goals. The insights from our study help develop robots that can interact with humans more naturally and effectively, making them better partners in various situations.

1.4 Overview of the Thesis

In this thesis, we designed an experiment to model a human-robot interaction involving the transportation of an object, specifically a glass of water, between a human and the iCub robot from the Italian Institute of Technology (IIT). The experiment includes the simulation of handover and pick-and-place movements. The cups can be empty, half-full, or full. The movements can be either a handover action or a pick-and-place action.

We evaluated a new head pose estimation program using iCub's integrated cameras to assess if the camera quality and the software together could accurately track a person's head position during human-robot interactions. After validating the program, it was used to gather head position data throughout the experiment. This data was processed to extract relevant features for detecting the level of carefulness in the subject's movements. We developed a classification model to discern whether actions were careful or not. This model was then optimized and tested for real-time implementation on the iCub robot, enabling it to analyze and adapt to the carefulness in real time. This capability forms part of a series of experiments aimed at enabling effective cooperation between humans and robots in dynamic environments.

Chapter 2

State of the Art

2.1 Introduction

As in almost all technical fields, the inspiration for the design and development of social robotics comes from nature. Humanoid robots are digital replicas of humans. To achieve the acceptance and integration of these systems in our society, the goal is to educate machines based on our knowledge of human psychological behavior.

2.1.1 iCub

Robots like iCub bring artificial intelligence into the physical world. The name "iCub" stands for "cognitive universal body", highlighting its role in testing the embodied cognition theory. This theory suggests that interacting physically with the environment is crucial for cognitive development. Designed to resemble a human child, iCub uses its humanoid form to learn and adapt by engaging with people and its surroundings in ways similar to how a child does. It is equipped with a range of sensors and modules to perceive the world as we do. Among its features are a distributed sensorized skin that provides tactile feedback, micro-

phones for audio input positioned similarly to human ears, and two RGB cameras installed in its eyes for visual perception. On the software side, in its cognitive architecture, iCub includes many modules that allow it to perceive the world in a way similar to humans. Log-polar mapping, for instance is an image processing technique that mimics the way the human eye filters information by focusing only on a single point rather than the entire visual field. Modelling this human perception can be done by remapping the distribution of the pixels density in a perceived image. Enhancing the resolution in the center and compressing it at the periphery results in a sharp focus and less detailed but still preserved wide-field information. This reduces the computational load needed to analyze images because less data is processed. Consequently the robot can process information more quickly and is able to respond faster to their environment. It is advantageous for navigation and object recognition, where it is essential to quickly identify central objects while maintaining a broader understanding of the surroundings.

The log polar mapping for visual perception forms just one of the numerous modules inside the YARP libary, which contains all the software components developed for the robot. YARP (Yet Another Robot Platform) is the middleware framework that combines various elements and manages the robot's control [2]. Through the Yarp-manager sensors and motors are linked to the software components. The availability of more modules allows for greater versatility in combining them to enhance the robot's overall perception and action capabilities. Each module represents a building block that can be activated within iCub's behavior, contributing to its lifelike and adaptable functionalities.

2.1.2 Biological Perception System

Biological perception systems, such as the eyes and ears, provide exemplary models for efficiently, flexibly, and accurately handling a wide range of complex tasks in real life. For example, considering the human visual system. When we look at a crowded scene, our brains automatically prioritize and process important details, such as recognizing faces or detecting motion, while filtering out irrelevant background information. This selective attention allows us to navigate complex environments efficiently without being overwhelmed by sensory input. Another example is our auditory system, which can focus on a single conversation in a noisy room, a phenomenon known as the "cocktail party" effect. Our brains can isolate and follow the voice of the person we are speaking with, while tuning out other noises. This ability demonstrates how our perceptual systems efficiently identify and process relevant information.

Computational models inspired by human perceptual systems are pursued to enhance robots and other technologies, enabling them to interact more naturally and effectively in human environments. By replicating the human ability to prioritize and process sensory information efficiently, these models can help robots discern what aspects of their environment are most relevant, which is crucial for tasks like navigation, interaction, or complex problem-solving. For example, in robotics, such models could lead to better performance in dynamic settings like homes, workplaces, or urban areas, where robots must operate safely and interactively among people. They can improve the robots' ability to understand and respond to human actions or spoken commands even in noisy or visually cluttered situations. This capability not only serves to make robots more functional but also enhances their ability to coexist with humans as helpers, companions, or team members across a wide range of activities. To develop computational models capable of perceiving their surroundings, it is logical to draw inspiration from

the mechanisms underlying human motion perception. This involves studying how humans perceive and interpret movements and subsequently applying this knowledge to computational systems of robots.

2.2 Modelling Motion

Various studies focus on the modeling of movements which aims to link observations from cognitive science with computational models to develop systems that support natural human-robot interactions [3]. When considering how to describe and measure motion, we arrive at the key aspects of force, time, space, and flow. Rudolf Laban, a prominent choreographer, developed the Laban notation to systematically record dance and movements. Each movement can be described by these attributes, with each movement executed with varying speed and force, and each having a direction and trajectory. Flow refers to the continuity or interruption of movement and is an integral part of Laban notation to capture the dynamics and fluidity of a movement. This element helps describe the degree of relaxation or tension in movements, which is particularly useful for understanding the emotional expressiveness and stylistic elements of a dance performance [4]. Dance serves as a fascinating medium for modeling movement and non verbal communication in robotics because it involves not only precise but often exaggeratedly clear movements that can aid in understanding and replicating the complexity of human motions. Dance typically unfolds within a narrative or emotional context, where each movement carries specific meaning or emotion.

2.2.1 Motion Recognition

In their research, Noceti et al. developed a method to distinguish biological motion using a binary classifier that assesses whether movements are of biolog-

ical origin. This capability, evident in newborns soon after birth, aids them in identifying potential interaction partners [5].

To detect biological motion within a video stream, the method initially identifies areas of movement using optical flow. Optical flow analyzes changes in pixel brightness across consecutive frames to identify dynamic regions in the video. Subsequently, the method extracts a set of low-level features inspired by the Two-Thirds Power Law, a principle that correlates the speed of a movement with the curvature of its trajectory [6]. This law was explored in the 1980s by researchers including Viviane Desmurget and Michael I. Jordan.

Notably, the approach in these studies presents an alternative to traditional person detection methods. Instead of merely identifying individuals, the focus is on recognizing biological motion patterns, which can be a more nuanced and effective way of detecting human presence.

The study from Vignolo et al. implements a model on the humanoid robot iCub and directly links the perception of biological motion with the robot's actions and attention. The implementation not only recognizes biological movements but also utilizes robotic oculomotor mechanisms to steer the robot's attention toward potential interacting partners in the scene. This approach allows the robot to focus on the activities of human partners. By adjusting its gaze, the robot can intuitively communicate its focus and indicate its willingness to interact and giving human partners clues about what it is currently focusing on. The method used is based exclusively on movement, neither prior knowledge of the human form or skeleton nor the recognition of faces and hands is required [7].

2.2.2 Action Recognition

As infants reach several months of age, they not only maintain their ability to distinguish between biological and non-biological motion but also begin to enhance

their perceptual skills to recognize and differentiate between various biological actions. This developmental progression represents a critical milestone, transitioning from general motion detection to the specific identification of actions such as walking, running, and jumping. These advanced perceptual capabilities are essential for the development of more complex cognitive and social interactions. In the paper of Sigala et al., a neural network is employed to learn specific motion features to effectively recognize biological movements. Instead of focusing on static object recognition like traditional computer vision algorithms, the neural network in the paper is designed to capture and analyze dynamic motion patterns. Specifically, it learns features that are essential for recognizing movements such as walking. The objective is to extract robust motion data features that enable reliable recognition even under variable and challenging conditions, such as with background motion. Background motion refers to any unrelated movement within the visual scene that can potentially interfere with the recognition process, making it difficult to isolate the primary movement of interest [8].

A proven method for recognizing actions is the Gaussian Mixture Model. It involves effectively capturing and modeling temporal structures in videos. These models use a combination of Gaussian distributions to describe the temporal sequences within a video, where each Gaussian filter serves to identify and highlight important moments within the video [9].

The focus of the work of Coppola et al. is on recognizing social activities between two or more people, as opposed to actions performed by a single individual. This approach is particularly relevant for applications where human interaction is central, such as in assistive technologies, monitoring public spaces, or in care giving, where recognizing group interactions can be crucial. By analyzing the dynamics between individuals, the system can better understand the context in which certain actions occur, and respond or assist accordingly [10].

To gain more context, Coppola et al. do not focus on interactions between individuals but rather on interactions between humans and objects, also known as affordances. Instead of merely recognizing actions, they take a step further and predict activities.

To model activities and object affordances, it is necessary to capture the rich context and anticipate the distribution over a large number of possible future human activities. In their work, they represent each possible future scenario with an anticipatory temporal conditional random field, which models the spatio-temporal relationships through object affordances.

For example, when a robot observes a person moving their hand towards a glass of water, the glass could be brought to various places, such as to the mouth, to the dish washer, or to another spot. If a robot can anticipate this, it would not start pouring more water but would instead maybe helpfully open the dishwasher. Activities often have a hierarchical structure, where an activity is composed of a sequence of sub-activities and involves interactions with specific objects. For instance, a glass is used in the drinking activity, which is composed of the sub-activities of reaching, moving, and drinking. Therefore, we can anticipate the future by observing the sub-activities performed in the past and reasoning about the structure of activities and the functionality of the objects being used [11].

2.3 Implicit Interaction in HRI

2.3.1 Vitality Forms

Vitality Forms is a concept from psychology and developmental psychology originally developed by Daniel Stern to describe the dynamics and "life spirit" of human movements and interactions [12]. These forms relate to the way actions are performed—for example, gently or energetically, calm or aggressive. In robotics,

this concept is used to impart more human-like, expressive, and intuitive behaviors to robots. Advancing beyond mere action detection to the recognition and replication of expressive behaviors represents a nuanced and significant challenge. This step involves not only enabling robots to interpret human expressions accurately but also empowering them to express their own reactions in a manner that humans can understand and relate to. However, the field of robotics largely remains focused on mastering fundamental tasks such as walking and grasping. These basic capabilities form the essential groundwork upon which more sophisticated functions can be built. The integration of human-like expressive behaviors and nuanced interactions is still viewed as a future goal, contingent on the development of robots with robust technical, hardware, and motor skills. Once these foundational elements are firmly established, the models for more empathetic and responsive interactions can be effectively implemented, pushing the boundaries of what robots can achieve in human environments.

In a series of experiments, it was demonstrated that a robot can be perceived as aggressive or friendly based on its behavior [13]. The same action, for example passing by an object, can convey different meanings, including a sense of urgency or importance, when performed with varying degrees of vitality. Recent neurophysiological findings have revealed that different vitality forms trigger distinct activations in the dorso-central insula of the brain. Specifically, this brain region distinguishes between rude and gentle behaviors, both in observing actions and in perceiving variations in tone of voice. This suggests a sophisticated level of interpretation by the human brain, which can significantly influence human-robot interactions [14].

In the described experiment, human movements were directly transferred to the joints of the robot to replicate human motions as accurately as possible. This approach allows the robot to perform the movements with the same timing

precision and natural communicative aspects as exhibited by the human actor. Employing this method in a condition characterized as aggressive aims to investigate whether the robot can generate the same expressiveness and emotional impact that a human would show in similar movements. The goal, therefore, is to determine whether the robot can not only technically replicate the movements but also convey the associated social and emotional signals.

Imitating human movements in robots encounters challenges due to fundamental differences in the physical structure and material capabilities between humans and robots. Human joints and the associated movements are extremely complex, capable of fluid and flexible motions due to the arrangement of bones, muscles, tendons, and ligaments. This complexity allows for a wide range of movements and subtle variations that are difficult to replicate with mechanical joints [15].

It is not enough to represent a rude movement quickly and a gentle movement slowly. The precise replication of human actions by the robot, including kinematic parameters such as peak and velocity profile, led to the activation of the dorso-central insula. This indicates that accurate kinematic parameters are crucial for conveying vitality forms and triggering the corresponding brain activity [16].

2.3.2 Object Manipulation

As children age, their ability to gather information not only about other people but also about the objects being manipulated expands. Observing people manipulate objects conveys hidden characteristics of these objects. These characteristics arise from the combination of the human's manipulation and the inherent properties of the object, which would not be apparent from the object alone. For example, children aged 5 to 7 can estimate the weight of a wooden block simply by observing another person lift it [17]. Additionally, research has shown

that size estimation can be optimized merely through the observation of grasping movements [18].

Research by Sciutti et al. has demonstrated that humanoid robots can also convey such information. In their experiments, a humanoid robot lifts bottles whose contents are not visible to the participants. iCub was programmed to perform lifting actions under two different conditions: Standard and Proportional. In the Standard condition, the robot's kinematics were planned independently of the object's weight. In the Proportional condition, the movement kinematics varied according to the object's weight. Specifically, a smaller vertical velocity was associated with an increase in weight. This programming choice was motivated by previous research on human behavior during lifting of different weights. Participants were able to accurately estimate the weight of the bottles based solely on the observed movement and trajectory [19]. These findings illustrate that the ability to infer hidden properties of objects through the observation of manipulation is not limited to human interactions but can also be replicated by humanoid robots.

2.3.3 Carefulness Expressed in Motion

Carefulness in object manipulation encompasses various aspects depending on the nature of the object and the personal or practical reasons for caution. This may involve being cautious due to the object's fragility, high monetary or personal value, or when handling dangerous objects such as knives.

To detect carefulness in a transportation movement, Garello et al. developed a velocity profile similar to the idea of classifing Vitality forms due to velocity. What sets their approach apart is that instead of replicating exact human movements, they utilized generative adversarial networks (GANs) to enable autonomous velocity modeling [20]. By using generative models like GANs, new

and consistent motion patterns without needing a large number of recorded human demonstrations can be generated, allowing the robot to learn and adapt its actions independently and reducing dependency on human input.

It would be interesting to explore the classifications between similar velocity profiles, for instance between a non-careful and a rude velocity profile. The question comes up if the velocity profile provides enough information for a real world situation. When handing over an object it can differ only in subtle nuances if the handover would be rude, hectic, non careful or annoyed. These distinctions may require additional context or a combination of more features. The question we address in this thesis is whether carefulness can be detected and expressed solely based on gaze. For future research gaze might be integrated as one of several features to characterize careful behavior.

Multisensory Integration describes the combination and coordination of information from different senses. The brain processes various sensory inputs and coordinates them to perform specific actions. For example, when transporting a full glass of water, visual information helps ensure that nothing spills, while motor senses provide feedback regarding the speed and balance required to handle the glass safely.

By integrating information from multiple senses, the brain creates a more accurate and reliable perception of the environment, enabling coordinated and effective interactions with the surroundings.

In the carefulness-experiment from Lastrico et al., a multimodal system was used to obtain synchronized and comparable data from various multi modal tracking systems. This included external camera data, where optical flow was applied, a motion capture system with infrared markers placed on key points of the participants' hands and arms, and Inertial Measurement Unit data, worn on the wrist to measure linear acceleration and angular velocity. These all provided a rich dataset

for comparing different techniques and pushes from body-mounted measurement devices toward external cameras, making the procedure more comfortable for the participant [21]. In this study, we aim to analyze data captured by the cameras in the eyes of the iCub robot to generate behavior as realistic as possible without using external devices for realistic situations outside the laboratory.

2.4 Modelling Gaze

With our eyes, we observe the environment and take in information about what we see. This is the process in which we gather visual impressions. At the same time, our eyes are also active in sending information. For instance, the direction of our gaze or the eye contact we make can signal to others what we are focusing on or what interests us. The eye movements provide again subtle non-verbal cues and on top we rely on it due to our lifelong experience. Humans have high expectations for human-like behavior in interactions. Any unnatural motion can leave us feeling unsettled and disrupt the smooth flow of these interactions. Eye contact and gaze are crucial in this context. They significantly contribute to communication and mutual understanding. When a humanoid robot fails to replicate natural eye movements, it can interrupt the seamless interaction and make the experience feel less intuitive and more mechanical.

2.4.1 Social Gaze

Social Gaze refers to the way people use eye movements and eye contact to communicate and interact with others in social contexts. It encompasses various types of eye behavior that convey information, regulate social interactions, and help build and maintain relationships, while providing cues about characteristics, emotional state, and attention. Attention refers to the ability to focus on specific

stimuli or activities while filtering out irrelevant information. People's attention can be inferred by observing their gaze direction, which indicates what they are looking at. Observing someone's gaze direction helps to understand their interests or intentions, thereby aiding in non-verbal communication. There are different types of gaze.

- Mutual gaze, known as eye contact, occurs when two people look into each other's eyes at the same time. It is fundamental in establishing a connection, indicating attention and interest, and is often used to express emotions and intentions.

- Joint attention involves two or more individuals focusing on the same object or event simultaneously, aware of each other's focus. This behavior is crucial for shared experiences and cooperative activities. It helps in developing social cognition and understanding the intentions and actions of others.

- Gaze aversion is the act of deliberately looking away from another person's gaze. It can be used to regulate the flow of conversation, indicate discomfort or disinterest, and manage social interactions.

- Referential gaze refers to looking directly at an object or a specific location. This type of gaze typically happens alongside verbal cues or indications towards the object. For example, when someone says, "Look at that bird," they usually direct their gaze towards the bird while speaking [22].

Gaze reading enhances interaction efficiency by providing parallel information to other communication forms like speech or gesture, thereby reducing the complexity of information transfer. For instance, gaze can clarify speech by grounding ambiguous references to objects. In dialogues, gaze can convey emotional states or intentions without interrupting the verbal flow. We are adept at interpreting

the gaze of our conversation partners and react to them. When persons look up, it is understood they think and the reaction could be waiting a bit longer for an answer. If they glance at their phone, we recognize they might be distracted. In a group setting, we can precisely discern when we are being addressed [23].

Gaze is essential when it comes to object reference and manipulation. People often look at objects before naming them, and can predict their partner's actions based on referential gaze. People respond faster to their partner when they can see their referential gaze on the object of interest. Mutual gaze also carries important information during object manipulation between people.

2.4.2 Approaches in Social Robotics

For a robot to perceive eye movement information, different parameters can be measured. The highest precision is provided by a head-mounted eye tracking device, such as eye-tracking glasses worn by the participant, which accurately tracks both the person's field of view and their eyes. It offers extensive insights into a person's visual behavior and physiological responses. They track precise the gaze points in 2D and 3D (fixations), showing exactly where a person is looking, and the gaze path (saccades), which reveals the sequence and pattern of visual exploration. They also measure gaze duration, indicating how long specific areas of interest are viewed. Additionally, they monitor pupil diameter and changes in pupil size (dilation and constriction) to infer arousal levels and emotional states. Other key metrics include blink rate and duration, head movement, eye vergence for depth perception, and fixation distribution, which helps identify the most engaging aspects of a visual scene. One limitation of the glasses is that they do not provide tracking of smooth pursuit movements. This is a type of eye movement where the eyes move smoothly to follow a moving object allowing the eyes to keep a moving target in focus. The eyes move at the same speed

as the moving object, maintaining a stable image on the retina. To detect the smooth pursuit with the eye-tracking glasses additional computer vision model is needed. Eye-tracking glasses are effective tools for tracking eye movement data and building initial models of visual behavior. They offer precise measurements and are often used for ground truth validation. However, they do not accurately represent the robot's natural perspective when perceiving human gaze. Consequently, while these glasses are invaluable for collecting and validating data, they are not intended as the final perception method for robots in human-robot interaction scenarios. Instead, robots should rely on their own sensors and cameras, which process visual information differently to perceive and interpret human gaze during interactions. During the approach In Palinko's approach, the accuracy of iCub's cameras for gaze detection was tested. In a scenario where iCub holds an object in each hand, the participant is supposed to communicate only with their gaze which of the two objects they want, and iCub hands it to them. It turned out that the quality is sufficient for this task, but limitations were also noted, such as the need for a close distance and the fact that the task primarily involved distinguishing between left and right [23].

If we aim to make the behavior of robots as biologically and human-like as possible, then deep learning models, often referred to as "black boxes," may not be the ideal approach. The primary reason is that deep learning models are typically very complex and lack transparency, making them difficult to understand. These models make decisions based on patterns in data that are not always interpretable by humans. In contrast, a biologically inspired approach seeks to mimic the precise mechanisms of the human brain and psychology, ensuring that models not only produce accurate results but also operate in a manner consistent with natural human processes. For social robots, it is crucial that decisions are transparent and align with human behavior to be predictable and understandable. Models based

on clear rules and principles may offer advantages over deep learning models in this regard. While deep learning has achieved impressive results in many areas, there remains a strong interest in methods that provide more transparency and explainability. Research focusing on replicating biological processes could eventually lead to models that are both effective and comprehensible. Therefore, choosing an approach that is not only powerful but also understandable and interpretable is essential for accurately mimicking human behavior.

A systematic approach to capturing and analyzing human attention can be supported through the use of tools for gaze detection. This technology serves as an information source for analyzing gaze behavior by utilizing physical capabilities to accurately track and measure eye movements. The actual recognition of where someone is looking is largely based on physical processes (visual perception and neural processing) while the recognized information is psychologically interpreted to derive social and communicative meanings. The gaze detection enables the collection of valuable data that can provide insights into cognitive processes and attention patterns. By integrating this data into rule-based models, behaviors can be developed that mimic human perceptual and decision-making processes, leading to an authentic and comprehensible replication of human attentiveness. Tomenotti et al. developed a model to predict a person's head pose in a video with a worst-case accuracy loss of only about 2 degrees. This addresses previous concerns that head pose estimation might not be adequate as an alternative to eye tracking [24]. The model is based on Head Pose Prediction Net (HHP-net) and aims to estimate the direction of a person's head in single images by using a minimal number of key points on the head. The network predicts the corresponding Euler angles - yaw, pitch, and roll - for each frame [25]. This approach facilitates a more accurate approximation of perception and enhances our ability to detect gaze direction, making it a valuable tool being integrated in

the overall behavioral model and will later help us detect gaze direction.

2.4.3 Carefulness Expressed in Gaze

To accurately measure carefulness, which is a reflection of psychological behavior, various metrics can be employed. These include the duration and direction of gaze, smooth pursuit movements and saccades. Each of these measurements provides valuable insights into attentional focus and cognitive processes, allowing for a comprehensive analysis of how attentiveness and careful behavior are exhibited.

In the work of Duarte et al., defined areas of interest were used to measure the duration and order of gaze falling into these areas. From the sequence of gaze movements, a Hidden Markov Model was derived to interpret the behavior of the person [26]. This model was employed to distinguish between handover situations and pick-and-place movements, allowing the counterpart to recognize in time that they were being addressed and to initiate a response. This approach could also be applied to measure carefulness using the same input data.

In the same experiment, where water glasses of varying fullness were transported through pick-and-place and handover movements, carefulness was also assessed. They utilized an Echo State Network, an effective solution for processing time-dependent data, to measure the level of attentiveness and care in handling the glasses [27].

This published dataset is one of the few available datasets focused on carefulness [27][28]. All existing gaze datasets in this domain are based on head-mounted eye trackers. To date, there is no dataset that provides gaze data from an external camera in relation to carefulness.

2.5 Interaction and Coordination

When faced with the dual task of manipulating an object and conveying our intent to interact, our gaze simultaneously focuses on two aspects: the precise control of the object and the communicative connection with the other person.

An interesting discovery emerged from the work discussed in the previous section. While detecting carefulness in the two situations of handover and pick-and-place, the accuracy during a pick-and-place was significantly lower than during a handover. The parameters and gaze information transmitted during the handover had a greater impact on the social communication aspect. As the water level in the cup increases, the risk of spilling also rises. Consequently, more time is spent on focusing on the glass rather than gaze communication. This finding indicates and underscores the importance and advantages of social communication in conveying information about an object. Unlike the mere observation of a person, social interaction allows for the utilization of more information and additional features to obtain comprehensive and accurate details about the object. Visuomotor control focuses on ensuring a secure grasp and safe transportation of the cup to prevent spilling, while visual communication is aimed at conveying the intent to hand over the cup to others. In a situation of cooperation there may occur both.

2.5.1 Visuomotor Control

Visuomotor control refers to the coordination of visual perception and motor actions. It's the process by which the brain uses visual information to guide physical movements. This concept is crucial in many activities requiring precise hand-eye coordination, like threading a needle. When tasks demand cognitive effort, visual feedback is essential for adjusting and guiding motor actions. The

brain processes visual input and integrates it with proprioception. Based on this input, the brain plans and executes motor actions. When a person's gaze is fixed on an object, it often signifies a higher level of concentration and cognitive load. Fixations—periods during which the gaze remains steady—allow for detailed processing of visual information. This typically occurs during tasks that require focused attention and deep cognitive engagement. In contrast, saccades, which are rapid eye movements between fixations, are associated with the search for new information rather than in-depth processing. Studies have shown that cognitive load can be measured by analyzing eye movement patterns, including the duration of fixations and the frequency of saccades. Longer fixations often indicate that the individual is engaged in complex cognitive tasks that require sustained attention, while frequent saccades may suggest that the task is less demanding or that the individual is scanning for relevant information rather than deeply processing it [29][30].

In the experiment described above, the analysis indicates that the fuller the cup of water, the longer the fixation on it to prevent spilling.

2.5.2 Visual Communication Balance

When a complex task is coupled with the need to communicate, there is a dynamic shift between visuomotor control and visual communication. This transition depends on the object characteristics and required attention. The way a person can engage in both task execution and interaction provides information about the object. Once the object involved in the task is securely held—indicating successful visuomotor coordination—the person's gaze can then shift towards communicative targets. Simpler tasks generally require less cognitive and motor resources, which leaves more room for visual communication. In a handover situation the duration of time the person looks at the other person in comparison to look on

the cup might give information how full the glass is only by measuring the visual communication part. However, a notable observation from the above experiment was the relatively short trajectories involved in the movements. Specifically, during 'pick-and-place' tasks within the participant's visual field, carefulness was maintained as the cup remained within sight while focusing on the final position. Conversely, during handover tasks, where individuals had to shift their gaze over a larger distance to address another person, the differentiation between 'careful' and 'non-careful' actions became more pronounced.

However, it is noticeable that the trajectories in the experiment are relatively short. During the transportation in the pick-and-place task, the cup remained inside the field of view, especially when the person was already focused on the final position. In contrast, during the handover, the person had to look up, covering a greater distance both in terms of arm movement and gaze. This could be an additional explanation for why there was a more significant difference in the classification between careful and less careful behavior during the handover compared to the pick-and-place task.

Chapter 3

Methodology

3.1 Data Collection

In this study, we collect in parallel two datasets with the goal of determining if they can provide the same information regarding gaze behavior. One dataset is captured from the iCub's perspective, while the other is obtained through eye-tracking glasses—a proven method for accurately capturing gaze data. We will analyze the data, transform features to suit our problem, and compare the datasets to validate their equivalence. The data will be prepared for the use in machine learning models to further assess and quantify levels of carefulness. To this end, an experiment was designed and conducted at the IIT, simulating human-robot interaction.

The final dataset will include six distinct categories: three levels of carefulness and two types of actions. The carefulness levels are categorized as full, half-full, and empty, while the actions include handover and pick-and-place movements. These categories align with previous experimental setups and offer a basis for future comparisons. Although not all classes are the focus of the current study, they are included to support future research that may explore these variations

in greater detail. For this initial approach, our primary focus is on handover situations involving full and empty containers.

3.1.1 Gaze Tracking

Eye tracking is a technology that measures and records the position and movement of the eyes to determine where or what a person is looking at, known as the point of gaze. It has a wide range of applications, including research, accessibility, driving safety, and enhancing virtual reality and gaming experiences. Eye trackers typically use near-infrared light to illuminate the eyes, creating reflection patterns that are captured by camera sensors. These reflections, specifically from the cornea and pupil, are analyzed to estimate the point of gaze using advanced image-processing algorithms. The glasses contain a front-scene camera that records what the user is looking at, providing valuable first-person insights. Before each recording, a calibration process is required to ensure accurate eye-tracking data. Calibration involves aligning the eye tracker with the participant's unique eye characteristics, typically by having them focus on a single point or follow a series of visual targets on a screen.

One dataset is collected using Tobii Pro Lab 2 glasses worn by the participant to track the first person view during the action. It captures images at a resolution of 1920x1080 pixels within a frame rate of 100 Hz. This method has proven to be accurate and reliable in previous experiments, including in detecting carefulness. However, the camera's field of view does not capture objects close to the body—such as a glass held about 20 cm away—unless the head is specifically moved. This dataset serves as a ground truth for the new dataset.

The second dataset is captured from the iCub perspective. The standard built-in camera in iCub's head capture images at a resolution of 640x480 pixels. To extract information from these images, such as the gaze direction of a person

appearing in iCub's field of view, the Head Pose Estimation (HPE) module, introduced in 2.4.2, processes the images. The HPE program operates within a YARP module, connecting to the camera output and display input via YARP ports. It was tested and validated that it tracks the gaze as required even when the person is wearing the eye tracker glasses. On the iCub robot it was implemented for the first time, and the camera resolution has proven sufficient for accurate program performance. However, to achieve a higher frame rate, the program should be executed on a computer equipped with a GPU. In comparison to the Tobii Lab frame rate of 100 Hz, the HPE program on iCub runs only with 7 Hz what is a huge loss of information. The other aspect to consider when comparing to the glasses is the observable field. While between glasses and the persons eyes is no obstacle and the tracking in an ideal situation is not interrupted the HPE can not capture the gaze when the face is not seen, just like in a natural situation. For our experiment we have to estimate the values once the person turns away from the camera. Inside the yaw range the head pose orientation covers around 80° accurately, beyond that range while the face is still visible the degrees decrease almost in the same way the person would turn to the front. For our purpose it is still valuable because as we will later see we focus on the dynamic aspects instead the actual orientation. During the experiment, the output data Yaw, Pitch, Roll, and the X and Y positions in the image, are extracted. The Tobii glasses have a various amount of features while we focus on the comparable output data of Gyro X, Gyro Y and Gyro Z which measure the rotation around the three axis in degrees per second and samples at 100Hz. Both the HPE data and the Tobii glasses output data are recorded and synchronized with an event message sent by the State Machine module, which controls and coordinates the entire process. Based on the timestamp we tried to validate the HPE Yaw output in a comparison to Gyro Y. Due to the 93% difference of frame rate it was challenging finding a

one to one comparison.

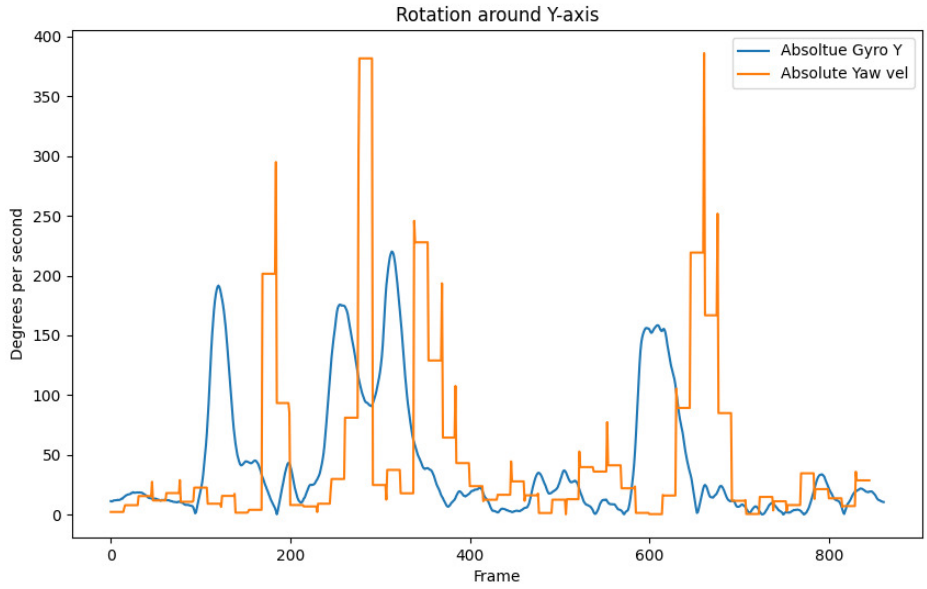

Figure 3.1: Validation of Yaw compared as |ΔYaw| with Gyro Y

So we analyzed two scenarios: during movement and in an idle state, comparing the average absolute rotation between the Tobii and HPE outputs. In a nearly idle state, where the person's movement ranged within 10° for approximately 5.5 seconds, we used 559 frames of gyroscopic data around the y-axis and 35 frames of yaw values for our dataset. The average absolute velocity on the y-axis, denoted as |ΔYaw|, was measured at 10°/s with a standard deviation of 9.82°/s. Tobii recorded an average velocity of 9.19°/s with a standard deviation of 4.41°/s. In a moving example, seen in Fig.3.1, the person moves in the range of 94°. The yaw average velocity was 44°/s with a standard deviation of 71.03°/s, whereas Tobii showed an average of 42.86°/s and a standard deviation of 53.31°/s. In the plot we see high fluctuation in the Yaw values and a slight delay due to latency in the image processing system of the robot. The overall trend is the same.

3.1.2 Experimental Setup

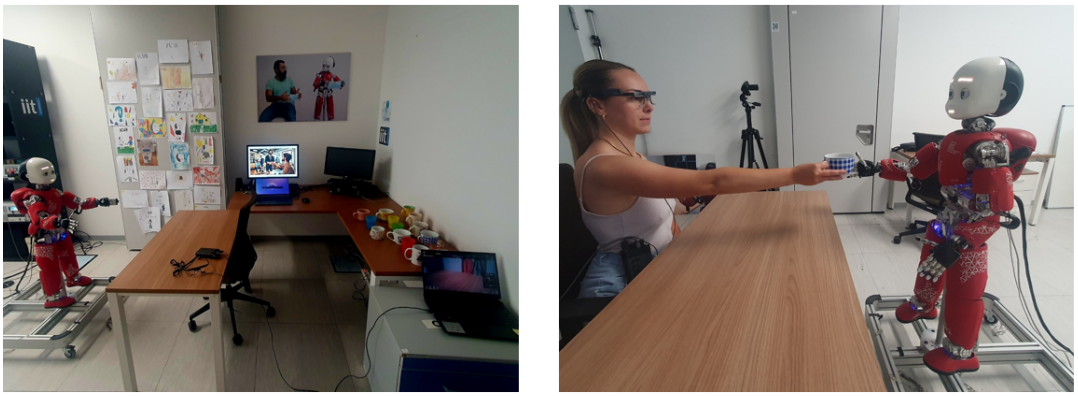

Figure 3.2: Experimental setup, on left the overall scene, on the right a participant during a handover

The experimental setup, shown in Fig. 3.2, includes a screen, a keyboard, the iCub robot with a built-in monocular camera, Tobii Pro Lab 2 glasses, and 15 cups filled with varying amounts of water. The screen is used to display images of natural scenarios where individuals perform handover or pick-and-place movements, which instruct the participant on the situations they need to replicate. Specifically for this purpose, we used ChatGPT-4's DALL·E to generate uniform yet distinct images of various cup handover and placing situations. Examples of these images are shown in Fig. 3.3. In the experimental setup iCub stands in front of a table, unable to reach the cups on the second table to the right of the scene. The participant is seated between iCub and the cups, positioned within reach of the cups and close enough to cooperate with iCub in completing the tasks. This setup was chosen to emphasize carefulness along the transport trajectory, challenging participants to maintain careful handling, especially compared to scenarios with shorter paths. The scene from above is shown in Fig. 3.4.

Figure 3.3: Example pictures of situation. Left Handovers, right Placing

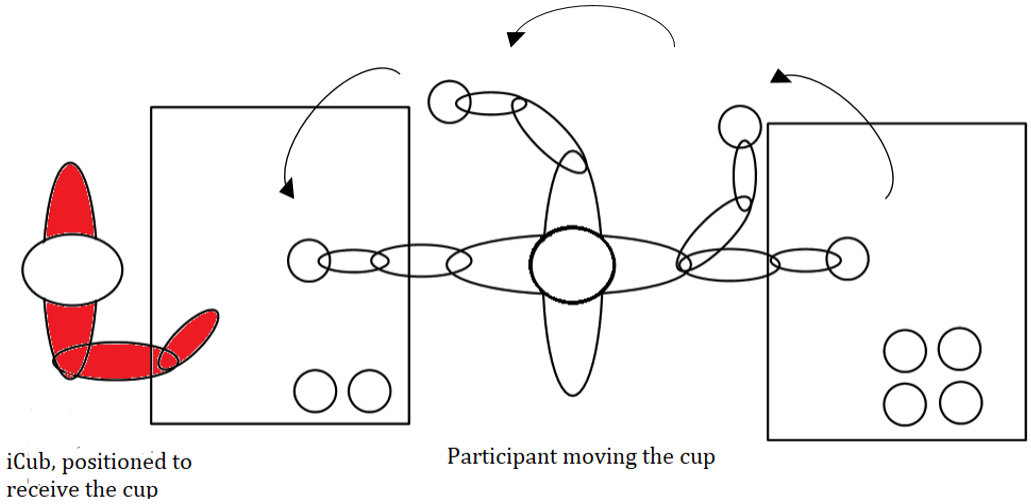

Figure 3.4: Experimental set viewed from above

3.1.3 Participants

In total, thirty adults (11 males and 19 females, in between 20 and 60 years) participated in the experiment. While they were naive to the field of robotics,

some had prior experience with experimental studies conducted by IIT. Of the participants, 29 successfully wore the Tobii eye-tracking glasses, for one participant, calibration issues arose. Participants were informed in advance about wearing the eye-tracking glasses and were asked to possibly use contact lenses instead of regular glasses if they required vision correction. This precaution was taken because the eye-tracking glasses can have difficulty tracking when worn over regular glasses. Each participant received a compensation of 10 euros for their participation in the experiment.

3.1.4 Experiment Procedure

The procedure involved transporting cups filled with varying amounts of water - full, half-full and empty - across three rounds. During the transportation the participant's eyes were tracked by the Tobii glasses and the HPE module was running on the iCub left eye. iCub did not move but was positioned as if to theoretically accept the cup. The task for the participants was to open an image on the screen depicting one of two scenarios—handover or placing—and then replicate the scenario themselves. For each scenario, there was a designated key on the keyboard that participants pressed to send the corresponding YARP event and start the recording. In the handover scenario, participants transported the cup from the right table towards iCub, simulating a handover, and after a brief pause of 1-2 seconds, placed the cup on the table. They were informed that a complete handover could not be performed due to the robotic hand's fragility, which prevented it from grasping the cup. In the placing scenario, participants directly placed the cup on the table. They were further instructed to begin placing the cups on the left side of the table, keeping the area clear for the remaining cups, which totaled 15 across all rounds. At the end of each action, participants pressed another key to stop the recording. The first five trials of the initial round, which

involved half-full cups, were conducted to allow the participant to become familiar and comfortable with the task, after which they continued independently without further contact with the experimenter. Rounds two and three were completed with empty and full cups. The entire experiment was recorded with an external camera to provide an overview and allow for a clear understanding and review of the situation.

Fig. 3.5 illustrates iCub's view with the HPE program running. The five key points, both eyes and ears and the nose, on the participant's head are detected to calculate the head pose which is represented in the image by the red vector. In the first row, three sequential images show the participant handing over an empty cup. The second row displays three sequential images capturing the handover of a full cup.

Figure 3.5: iCub perspective of a handover situation. Above with an empty cup, below the transportation of a full cup

Figure 3.6 illustrates the same situation from the participant's view through the glasses. The circle indicates the point of gaze, and the red line represents the

Figure 3.6: Participants perspective of a handover situation. Above with an empty cup, below the transportation of a full cup

saccade in the image.

3.2 Preprocessing

To precisely capture the segment of the data where the actual transportation of the cup takes place , the data is initially manually filtered. Each trial begins at the moment the cup is grasped and concludes at the potential handover position. This also has the reason that the participants face is not seen during the grasping.

A summary of the dataset details is presented in Table 3.1. It comprises 191 valid trials for both full and empty cups from the iCub perspective, referred to here as the HPE datase. Additionally, the Tobii eye-tracking glasses recorded 225 valid trials with full cups and 229 with empty cups, briefly referred to as the Tobii dataset. The Tobii eye-tracker glasses operate at a frame rate of 100 Hz, with an average of 458 frames per full cup trial and 284 frames per empty cup trial. In contrast, the HPE dataset records at a frame rate of 7 Hz, with an

average of 29 frames for full cups and 17 frames for empty cups per trial. The average duration of full cup transportation is 4.5 seconds, and for empty cups, it is 2.8 seconds. The angular velocity averages 49 degrees per second for full cups and 70 degrees per second for empty cups.

Parameter	**Full**	**Empty**
Tobii		
Number of Trials	225	229
Average Frames per Trial	458	284
HPE		
Number of Trials	191	191
Average Frames per Trial	29	17
General Metrics		
Average Duration (s)	4.5	2.8
Angular Velocity (degrees/s)	49	70

Table 3.1: Summary of Gaze Data from Tobii and HPE Datasets

3.2.1 Feature Transformation

The output data from the HPE system include Yaw, Pitch, Roll, and the X and Y positional coordinates for each frame. However, simply analyzing head positions does not provide insights into levels of carefulness. Previous studies have established that carefulness is dynamically measured in time-dependent units. To better capture these dynamics, we transform the Yaw, Pitch, Roll, X, and Y coordinates into their respective differences:

- $|\Delta\text{Yaw}| = |\text{Yaw}(t+1) - \text{Yaw}(t)|$,
- $|\Delta\text{Pitch}| = |\text{Pitch}(t+1) - \text{Pitch}(t)|$,
- $|\Delta\text{Roll}| = |\text{Roll}(t+1) - \text{Roll}(t)|$,

- $|\Delta X| = |X(t+1) - X(t)|$,
- $|\Delta Y| = |Y(t+1) - Y(t)|$.

These changes between consecutive frames serve as features for our analysis.

3.2.2 Data Analysis

In our study, we assume that the |ΔYaw| and |ΔPitch| values are crucial for the movements performed during the experiment. Visualization of this data from individual test subjects shows that in the case of an empty cup, the movements remain more stable but occasionally exhibit strong peaks, while in the case of a full cup, smaller, quicker changes are observable. These observations are illustrated in Figures 3.7 and 3.8. In Fig. 3.7 , |ΔPitch| and |ΔYaw| are plotted together in one plot, with the top showing the full cup condition and the bottom showing the empty cup condition for four test subjects evenly selected from the dataset. In the second plot shown in Fig. 3.8, the same values are presented differently; here, both full and empty conditions are in a single plot, with |ΔPitch| at the top and |ΔYaw| at the bottom, again for the same four individuals.

However, the individual trials from different test subjects vary greatly, and it is nearly impossible to find a uniform measure. The large peaks are interpreted as strong saccades during a head turn, the quiet sections as fixations, and the fluctuations as small saccades during a smooth pursuit. Typically, a person performs 3-5 saccades per second, depending on the cognitive demand. The amplitude of a saccade varies depending on the nature of the task.

The interpolation across all trials in Fig. 3.9 attempts to provide an overview, yet it reveals only a slight upward trend for the empty cups. Towards the end, in both cases — full and empty — we observe a smooth decrease in the change of position, indicating that at this moment, the subject was facing the robot.

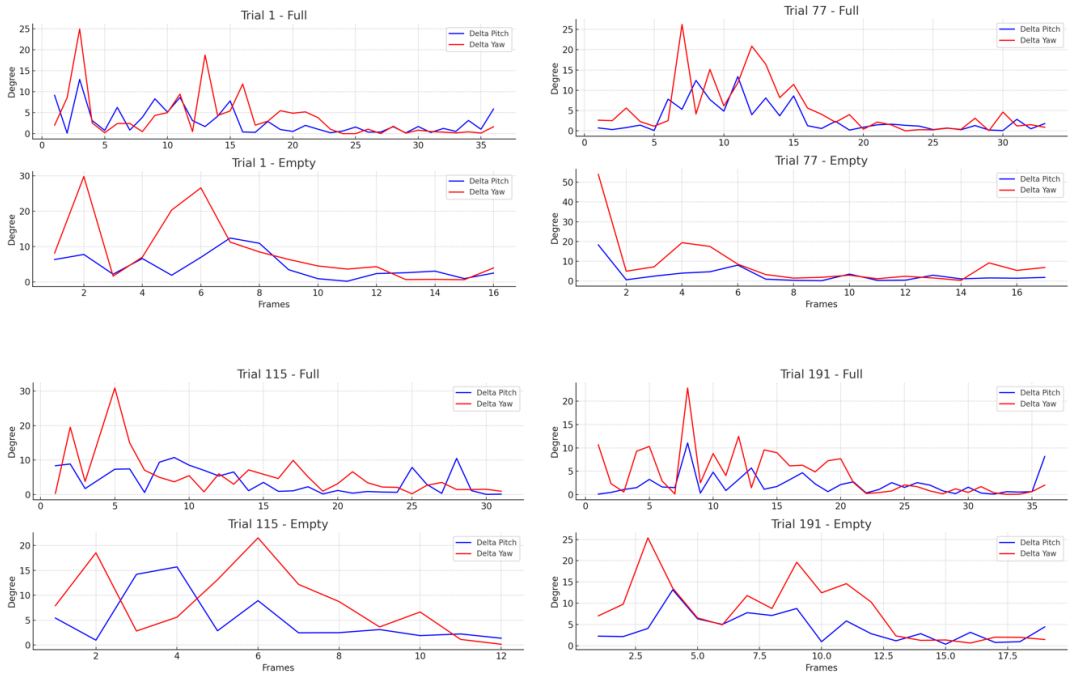

Figure 3.7: Four participants head movements $|\Delta\text{Pitch}|$ and $|\Delta\text{Yaw}|$ in the same plot. Above with a full cup, below the transportation of a empty cup

To determine when the person is looking at the face of iCub based on the HPE values, manual measurements were conducted, and the following intervals were established.

$$\text{face}_{|\Delta\text{Yaw}|} = [-18, 19.23] \quad \text{face}_{|\Delta\text{Pitch}|} = [-9.01, 3.09]$$

From various positions in the room, always facing the camera, the mean values and standard deviations for Yaw and Pitch were calculated; the standard deviations serve as intervals. This corresponds to the theoretical ideal value: when a person looks directly into the camera, Yaw, Pitch, and Roll should be zero. In the current study, the average percentage of time after the initial mutual gaze following the first observation of the robot's face was 72.78% in the Empty condition and 77.27% in the Full condition. The mean duration of mutual gaze was

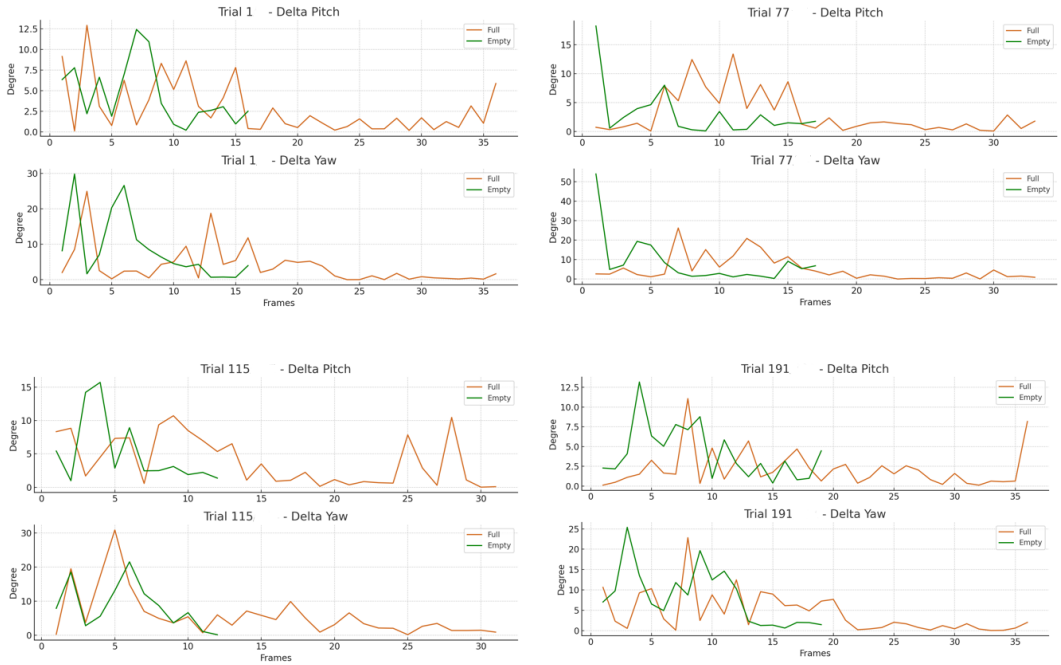

Figure 3.8: Four participants head movements full and empty in the same plot. Above with for $|\Delta\text{Pitch}|$, below for $|\Delta\text{Yaw}|$

21.30% in the Empty condition and 19.43% in the Full condition. When these values are compared to the previously observed results in the dataset [27] gained from a dynamic human-human scenario, where the Empty condition reported 25% and the Full condition approximately 15%, our findings fall between these ranges and do not exhibit a significant difference. In our case, we can cautiously hypothesize that the visual gaze is similar in both scenarios and therefore should not contribute to the classification of carefulness, at least in a Human-Robot Interaction (HRI) scenario. This observation could also be attributed to the robot's lack of movement, speech, or interaction. It is possible that in a more dynamic experiment, the results might differ.

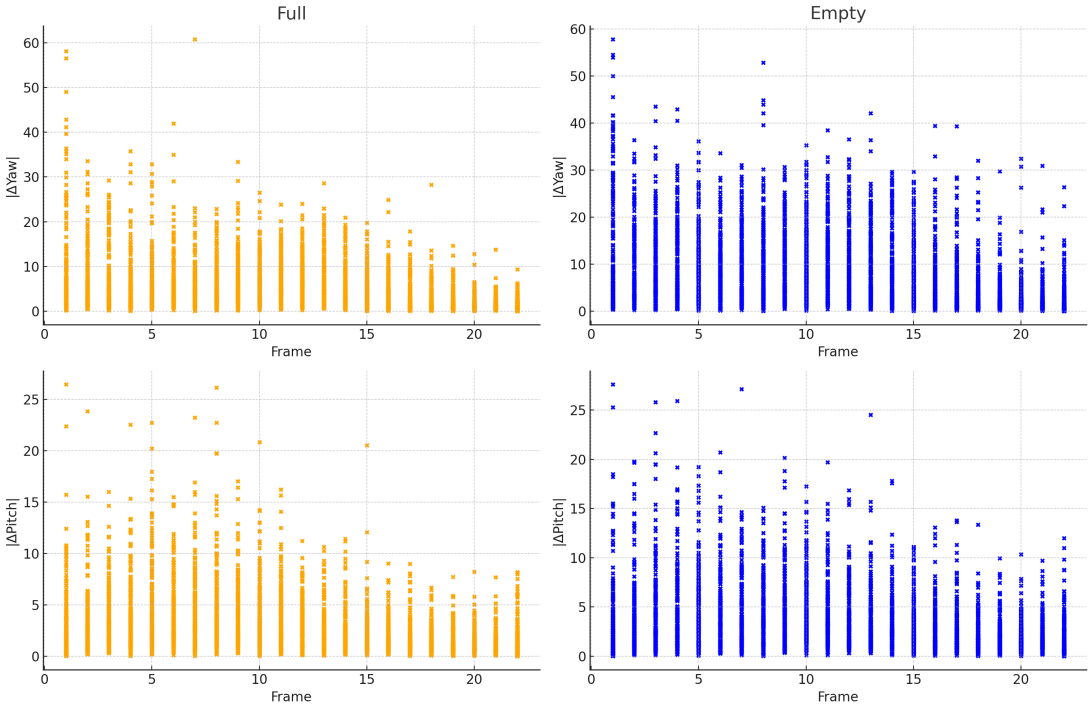

Figure 3.9: Interpolation over all trials, on the left for full cups, on the right for empty cups

3.2.3 Groundtruth Comparison

When comparing the findings from the analysis in the previous chapter with the results from the Tobii datasets, we can observe similarities. One of the provided outputs recorded by the Tobii glasses includes the rotation around the X-axis, Gyro X, and the rotation around the Y-axis, Gyro Y, which we use as approximations for Pitch and Yaw. They measure the same qualitative patterns and are plotted in Fig. 3.10. A bell shape emerges for the full dataset in yellow and a half-bell shape for the empty dataset in blue. As mentioned earlier, the dataset is filtered so that the trial begins when the person has already grasped the cup. In our experimental scenario, the person knows they need to turn in the direction of the robot. Therefore, they fixate on the cup, ensuring that their hand reaches it,

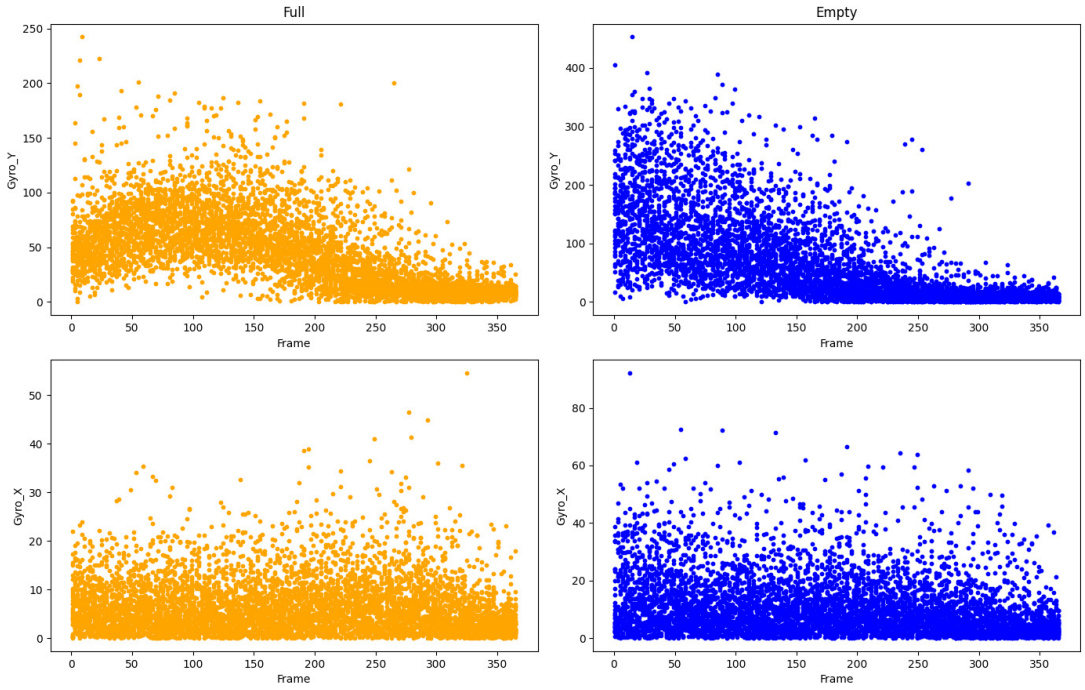

Figure 3.10: Tobii Interpolation data over all trials, on the left for full cups, on the right for empty cups

and then turn their gaze away from the cup even before completing the grasping movement. Different from a handover where both persons are turned towards each other, it would be unusual to first look at the other person and then retrieve the cup. Consequently, in the Tobii data, we observe due to sample wise manually measurements, as well as in the HPE data, a mutual gaze of about 20% in both cases—whether the cup is full or empty. In Fig. 3.11, we see the overall distribution of Fixation, Saccade, and Unclassified movements. Movements with significant head turns were not captured by Tobii and explain some of the Other eye movement which have a higher percentage for empty handovers indicating the head turn towards iCub as an exagerated saccade, without following the glas. We see 58.07% of Fixation for the empty data and 70.26% for the full data set,

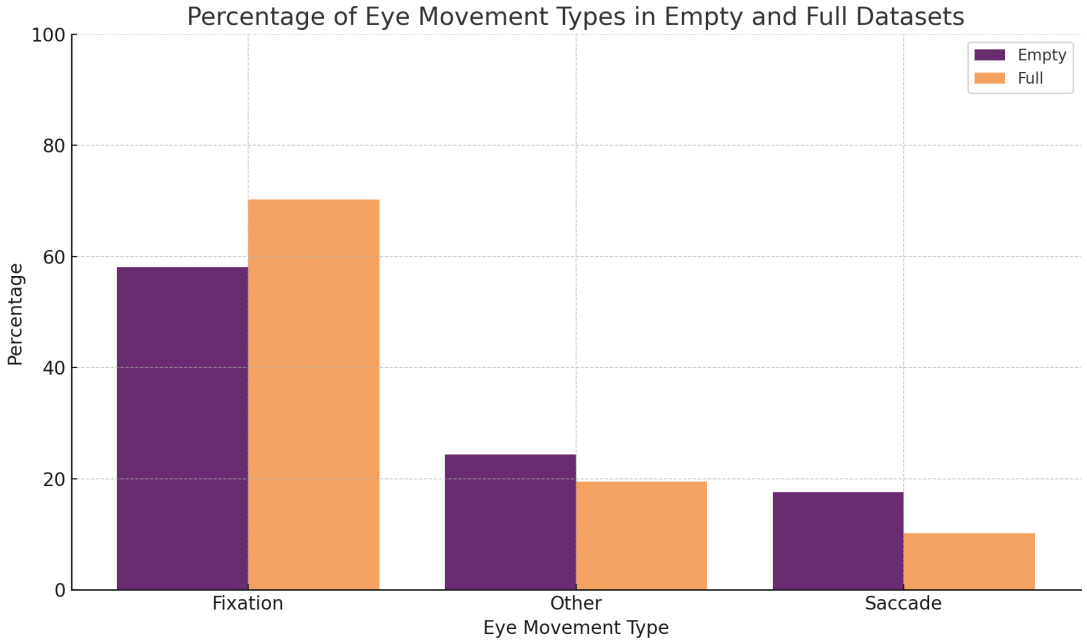

Figure 3.11: Percentage average Eye Movement Types in Empty and Full Datasets

17.53% Saccades for empty and 10.24% for full handovers and 24.39% empty and 19.50% Unclassified eye movements.

The Tobii system, with its higher frame rate and various informative output, delivers more precise results, enhancing our understanding of the HPE data. For the overall interpretation it suggests that the HPE data might also show a bell shape pattern. And for the individual plots it delivers some explanation about the interpretation for saccade and fixation. To further analyze the fixation distribution more focus could be placed on identifying several areas of interest. Even the fixations are really precise defined, every movement between a saccade is considered a fixation, additional information is necessary to fully understand the targets of these fixations. This work was done by Duarte et al. [27]. One aspect not adressed in this study, which would extend this work involves extracting additional body key point, such as the hand, provided by the HPE. This would

enable us to explore the relationship between gaze direction and hand movements, particularly when a person is looking at their hand while transporting an object. We conclude this analysis with the hypothesis that that the increased number of fixations in cases where the cup is full may be due to the need to monitor the cup closely to prevent spilling. Conversely, the higher frequency of saccades and unclassified eye movements in the empty handover case could indicate faster head movements and saccades. These findings will serve as feature input for further machine learning analysis.

3.3 Classification

3.3.1 Window-Based Feature Extraction

After an initial exploration of the data, it becomes evident that the complexity make it difficult to discern clear patterns or rules through simple visual inspection. While some indications of saccades and fixations can be observed, the exact differences, are not easily distinguishable by the human eye. To systematically analyze the data we will use machine learning models. To capture the timing within the data, we apply a windowing technique that groups samples together based on the chosen window size. The window in our study is an overlapping window, meaning it moves across the dataset incrementally, overlapping with the previous window at each step. For each window the mean, minimum, maximum and standard deviation are calculated for both $|\Delta\text{Pitch}|$ and $|\Delta\text{Yaw}|$. These statistics are then used as features for the machine learning models Random Forest, Support Vector Machine, and Linear Regression.

The windowing technique offers practical benefits for real-time applications. This approach allows the model to update its predictions continuously as new data becomes available. We implemented a real-time simulation to test the best

model by dividing the original dataset into an 80/20 split. The 80% was used to build the model, while the remaining 20% served as new and unseen input. In this setup, the inputs were the raw yaw and pitch data, which were converted in real-time to velocities. These velocities were then collected into windows, from which, when the desired window size was reached, the statistical features were calculated and applied to the model. Different window sizes were tested.

3.3.2 Temporal Modeling with LSTM

In our random forest approach, we concentrated on examining individual values extracted from a data window. Now, we aim to shift our focus back to the sequence of data and take into account the temporal relationships between data points. To this end, we applied an LSTM model to the absolute changes in yaw, pitch, roll, X, and Y positions. Long Short-Term Memory (LSTM) networks are a specialized type of Recurrent Neural Network (RNN) specifically designed to process long input sequences and learn long-term dependencies in data. Unlike traditional RNNs, which tend to lose performance during training due to the "Vanishing Gradient Problem," LSTMs effectively circumvent this issue.

Each data point in a sequence serves as an input for the LSTM, which identifies and learns patterns within the temporal variations of the data and classifies them into careful or non-areful. Our simple LSTM is structured with three layers. The Input Layer receives sequences of Delta values and positions, LSTM Layer processes inputs through the input, forget and output gates gates, maintaining an internal state that retains relevant information about the sequence. And the Output Layer which performs the final classification, using a Softmax activation to yield a probability distribution over the classes.

Chapter 4

Experimental Results

4.1 Model Comparison

In this section, the three machine learning models—Random Forest (RF), Logistic Regression (Log. Reg.), and Support Vector Machine (SVM)—are evaluated. The input data for these models consisted of statistical features extracted from an overlapping window that moves across the dataset. The initial tests were conducted using a sequence length of 10. This length was chosen as it was assumed to be long enough to capture the movement before it ends, but not so short that important data is missed. Hyper parameters for each model were optimized using GridSearchCV. The following hyper parameters were selected:

- RF: max_depth = None, min_samples_leaf = 1, n_estimators = 200
- Log. Reg.: C = 100, solver = lbfgs
- SVM: C = 150, gamma = auto, kernel = rbf

The performance of each model was evaluated with the following results:

- RF: Training Accuracy: 1.0, Testing Accuracy: 0.917

- Log. Reg.: Training Accuracy: 0.768, Testing Accuracy: 0.779
- SVM: Training Accuracy: 0.844, Testing Accuracy: 0.825

It is worth noting that the SVM model took significantly longer to train. The detailed classification report for each model can be seen in Tab. 4.1. Subsequent tests were conducted with different window sizes, 5, 15, and 20, to evaluate the impact of window size on model performance, seen in Fig. 4.1.

Model	Precision	Recall	F1-Score	Support
Random Forest				
Non-careful	0.92	0.83	0.88	581
Careful	0.91	0.96	0.94	1084
Accuracy			0.92	1665
Macro Avg.	0.92	0.90	0.91	1665
Weighted Avg.	0.92	0.92	0.92	1665
Log. Reg.				
Non-careful	0.82	0.64	0.72	581
Careful	0.83	0.92	0.87	1084
Accuracy			0.83	1665
Macro Avg.	0.82	0.78	0.80	1665
Weighted Avg.	0.82	0.83	0.82	1665
SVM				
Non-careful	0.79	0.71	0.74	440
Careful	0.85	0.90	0.87	1084
Accuracy			0.83	1665
Macro Avg.	0.82	0.80	0.81	1665
Weighted Avg.	0.83	0.83	0.83	1665

Table 4.1: Classification Report for RF, Log. Reg. and SVM

We observed that the Random Forest consistently outperformed the other models across all tested sequence lengths. The improvement in accuracy with larger windows was initially thought to be due to the model capturing more temporal context. However, it turns out that this was mainly because the $|\Delta\text{Yaw}|$ mean feature became more prominent, not because the model was better at understanding the time-dependent structure of the data. The LSTM model was

Sequence Length 5

Model	Training	Testing
RF	1.0	0.804
Log Reg	0.717	0.717
SVM	0.765	0.725

Sequence Length 15

Model	Training	Testing
RF	1.0	0.955
Log Reg	0.817	0.825
SVM	0.910	0.881

Sequence Length 20

Model	Training	Testing
RF	1.0	0.980
Log Reg	0.820	0.845
SVM	0.938	0.918

Sequence Length 25

Model	Training	Testing
RF	1.0	0.995
Log Reg	0.850	0.860
SVM	0.950	0.950

Figure 4.1: Performances for window size 5, 15, 20 and 25

employed to account for sequential time dependencies, the results, seen in Fig. 4.2, indicate a consistent accuracy across different sequence lengths. This consistency could be a sign of being able to classify behaviors effectively, independent of the sequence length. This indicates that the model's ability to identify patterns does not rely heavily on the length of the sequence, allowing it to maintain stable performance across various scenarios.

Seq. Length	Loss	Accuracy	Validation Loss	Validation Acc.
5	0.510	0.747	0.542	0.749
10	0.347	0.845	0.471	0.811
15	0.261	0.916	0.698	0.802
20	0.102	0.963	0.780	0.829
25	0.041	0.970	0.160	0.820

Table 4.2: LSTM Training and Validation Metrics Across Sequence Lengths

4.2 Feature Importance

The advantage of using a Random Forest model lies in its transparency and the insight it provides into the importance of different features. Tab. 4.3 shows the

results of the Random Forest feature analysis during the training of the HPE data with a window size of 10 frames. In the experiment, we observed a pronounced motion around the yaw axis due to the back-and-forth turning between the cups and the robot. It was assumed speed would play a significant role. Seeing the maximum value in second place confirms our hypothesis that outliers—specifically saccades—are particularly important in shaping the structure. The standard deviation ranking third supports the idea of chaos and the lack of calm phases, even during fixations. $|\Delta\text{Yaw}|$ minimum indicates the idle state when the person's gaze has reached the robot's face, as well as the maximum pitch movement when looking up to the face. Pitch minimum could be interpreted as spending more time looking at the cup, but it does not appear to be the critical factor distinguishing our two cases. With $|\Delta\text{Pitch}|$ mean and standard deviation ranking last, it suggests that pitch remains quite consistent overall, as also illustrated in Fig. 3.9.

Feature Importance	
1. $\|\Delta\text{Yaw}\|$ mean	0.195
2. $\|\Delta\text{Yaw}\|$ max	0.174
3. $\|\Delta\text{Yaw}\|$ std	0.158
4. $\|\Delta\text{Yaw}\|$ min	0.109
5. $\|\Delta\text{Pitch}\|$ max	0.102
6. $\|\Delta\text{Pitch}\|$ min	0.094
7. $\|\Delta\text{Pitch}\|$ mean	0.086
8. $\|\Delta\text{Pitch}\|$ std	0.082

Table 4.3: Feature Importance result from RF analysis

4.3 Real Time Testing

We proceeded to test the Random Forest model in a real-time simulation. For this purpose, the original dataset was divided into 80% for training and 20% for

testing. The model was built using the training data, and we observed that the performance remained consistently high, with an accuracy of 0.91.

In the simulation, the input data is fed into the program continuously, piece by piece, as if the data were being received in real-time. This simulates a scenario where data is being recorded live and immediately processed by the model, rather than being processed all at once. As soon as the window size was reached, the model continuously made predictions with each new input. The confusion matrix, displaying the true positives, true negatives, false positives, and false negatives, is shown in Tab. Tab. 4.4.

		Predicted Class	
		non-careful	**careful**
True Class	**non-careful**	468	292
	careful	264	1036

Table 4.4: Confusion Matrix for window size 10

The confusion matrix illustrates the performance of our model. When testing with data representing an empty glass (labeled as "non-careful"), the model correctly identifies this behavior 62% of the time. Conversely, when the input data represents a full glass (labeled as "careful"), the model achieves a higher accuracy, correctly classifying the behavior with a true positive rate of 80%. This difference in true positive rates can be attributed to the larger amount of data available for the "careful". Having more examples likely helped the model to learn and recognize careful behavior better than non-careful behavior. When we test across different sequence lengths, we observe an increasing trend in the performance for the non-carefulness classification, while the performance for carefulness remains relatively stable. During the training of the models, it became evident that the importance of the yaw mean feature consistently increased as the sequence length grew. This suggests that as the sequence length increases,

yaw mean becomes a more critical factor in distinguishing between non-careful and careful behavior. In Fig. 4.2, the lower plot shows the development of true

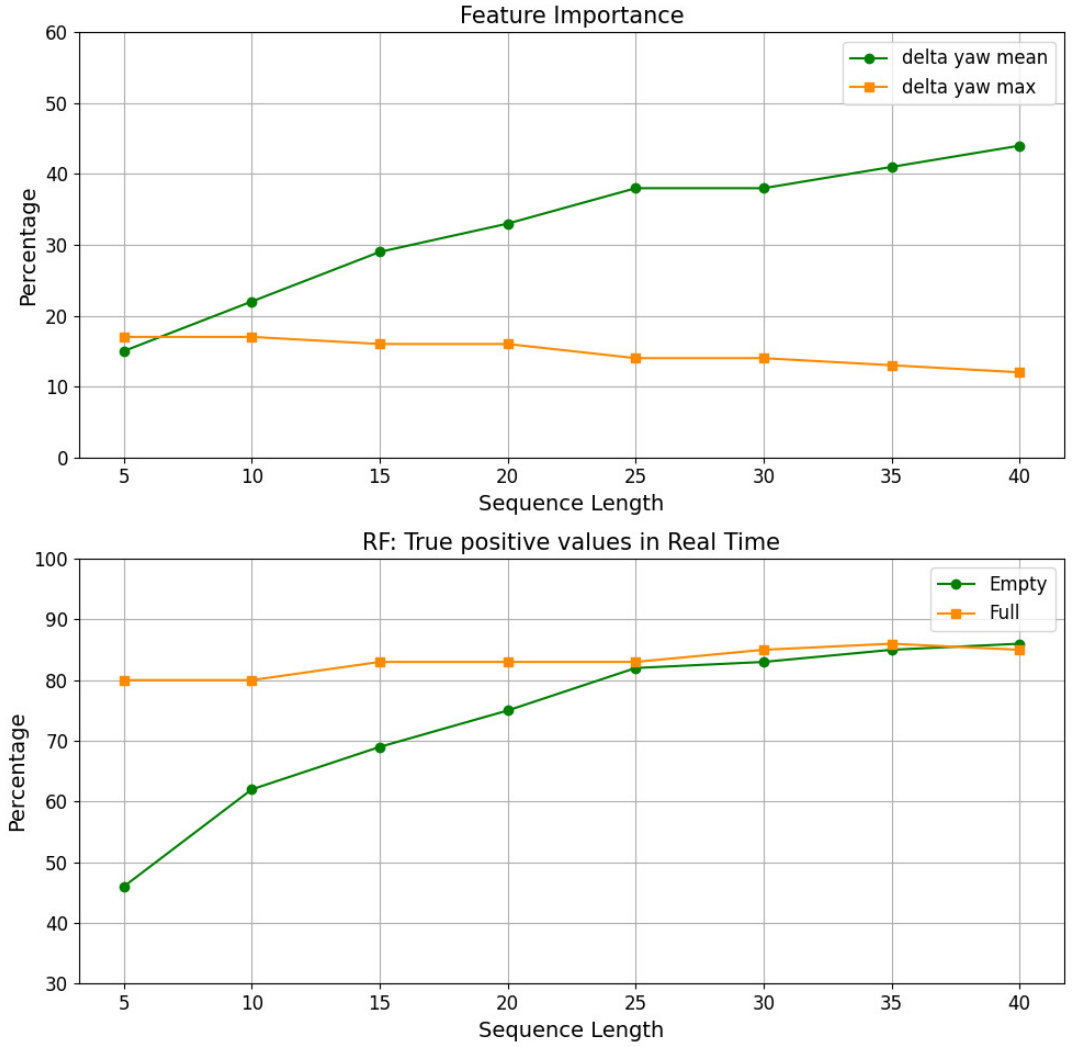

Figure 4.2: Above: Feature Importance, below: Classification full and empty during real time simulation with different window size

positive rates for careful and non-careful classifications during real-time testing, with careful behavior represented in yellow and non-careful behavior in green. Above this, we observe a similar qualitative trend in the development of feature importance, where $|\Delta\text{Yaw}|$ mean is shown in green. For comparison, the second-

ranked feature in the importance ranking (4.3) is represented in yellow. That could suggest that the model is primarily capturing speed or motion, potentially because the other features cannot be effectively extracted due to the noise in the dataset. When a model increasingly relies on a feature like $|\Delta\text{Yaw}|$ mean as sequence length grows, it might indicate that this feature is capturing a more straightforward or "trivial" aspect of the data. If the dataset is imprecise, it can obscure more subtle or complex features, making them harder for the model to detect and utilize effectively. In such cases, the model might default to relying on features that are more resilient to noise, like overall speed. This can happen because these features remain relatively consistent and interpretable even in the presence of noise, while more nuanced features might become too distorted to provide reliable information. So, the increased importance of yaw mean could be a sign that the model is focusing on more robust, but potentially less informative, features due to the difficulty in extracting and leveraging other features in a noisy dataset.

The output of the LSTM in 4.3 provides a similar output, but we cannot derive how the system reached these results, which is a drawback of deep learning models. The almost same curves, the empty one improved about 20%, and the findings just discussed, suggests the prediction might only be based on velocity. The good results might be misleading and could perform well in a real-time scenario where only carefulness is measured. However, once applied to a more complex environment, errors are likely to occur. For example, a movement involving a heavy object, which is likely slow, could be misclassified as careful. The transparency of the random forest model helped us recognize that we should be cautious in valuing this approach, as appearances can be misleading.

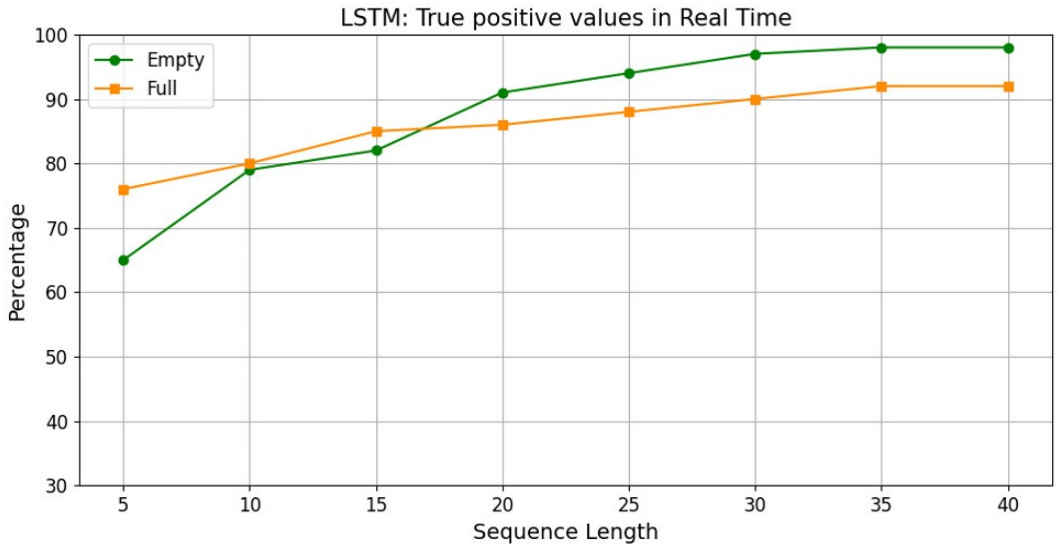

Figure 4.3: True positive output LSTM

4.4 Ground Truth Comparison

When testing the Tobii dataset, we utilized the features gyro X and gyro Y to approximate the HPE values. Additionally, instead of using window sizes of 5, 10, 15 ect. we multiplied these values by 14. This adjustment was done because the Tobii dataset was approximately 14 times denser than the HPE dataset. By scaling the window sizes in this manner, we aimed to capture the same duration in seconds, making this approach more appropriate for our analysis and investigating the precision difference. The results are shown in Fig. 4.4. Also the feature importance is differently ranked. Gyro Y maximum with about 30% and Gyro Y standard deviation with constant 25% are the guiding features, visible in Fig. 4.5.

Due to the density of the data and the smoother transitions, there are fewer abrupt changes within the dataset. As a result, sudden movements such as saccades and head turns become more prominent and easier to detect which can be interpreted through the Gyro Y maximum. The Gyro Y standard deviation can

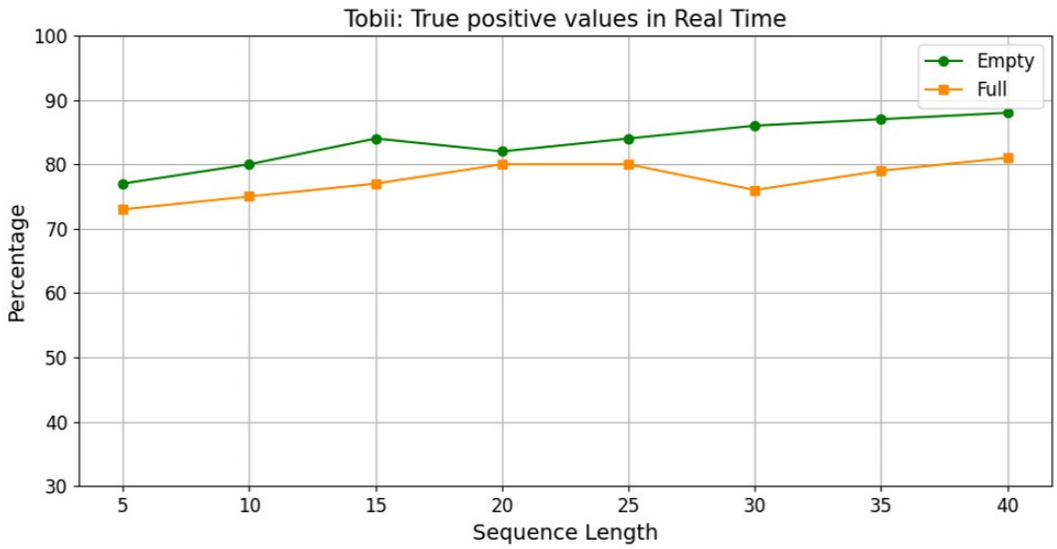

Figure 4.4: Tobii True positive values during real time simulation with different window size

be interpreted as a sign for a fixation versus abrupt movements. The effect of increasing window sizes is less significant compared to the HPE data, meaning smaller windows can still provide accurate measurements. For example, with a window size of 10, we achieve a true positive rate of 80% for empty handovers, compared to 62% with HPE.

4.5 Placing

As mentioned previously, this work focuses on the handover actions involving both full and empty cups. Given that the results from the Tobii data are satisfactory, we now present the same classification applied to another part of the dataset: the transportation of full and empty cups during the placing action. We observed no significant differences in the true positive values of the testing results, as shown

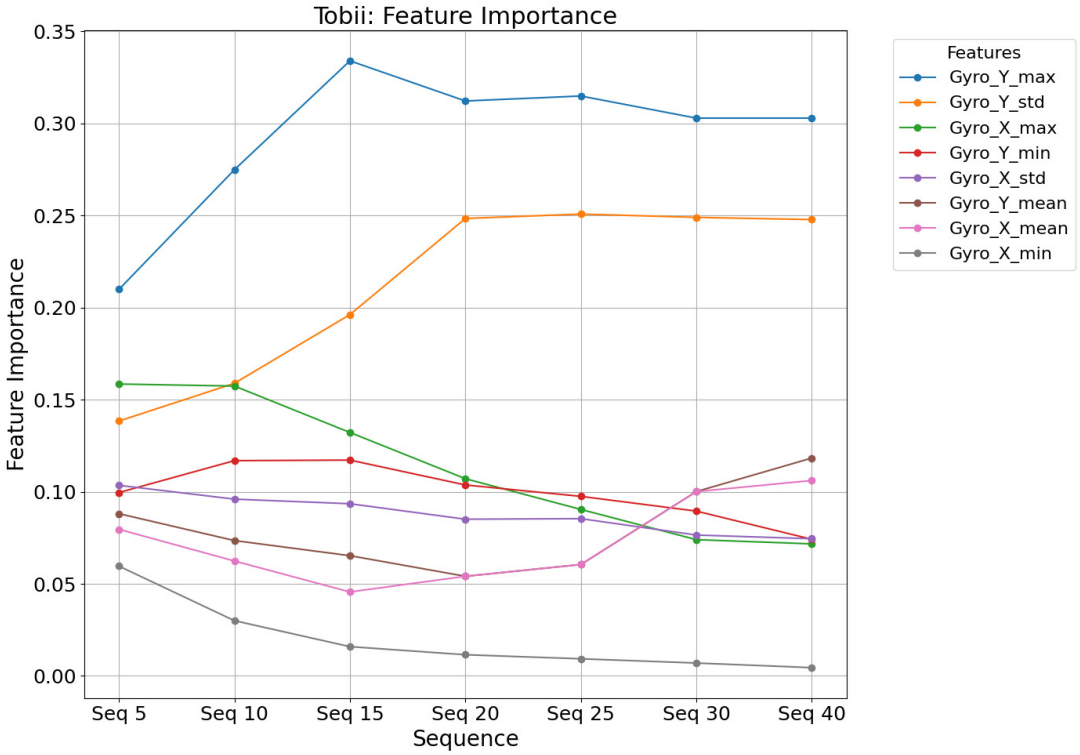

Figure 4.5: Tobii Feature Importance during real time testing with different window size

in Fig 4.6, or in the feature importance, as shown in Fig 4.7.

We conclude from this data that the social visual communication did not influence the carefulness classification in our experiment. This outcome may be attributed to the experimental setup involving a long trajectory

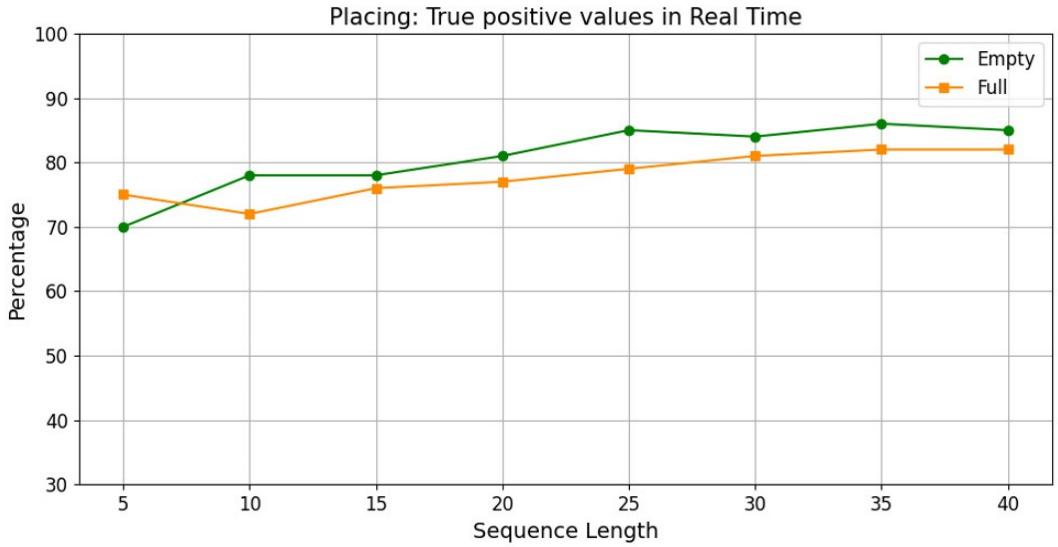

Figure 4.6: True positive values for Carefulness classification on Placing data set

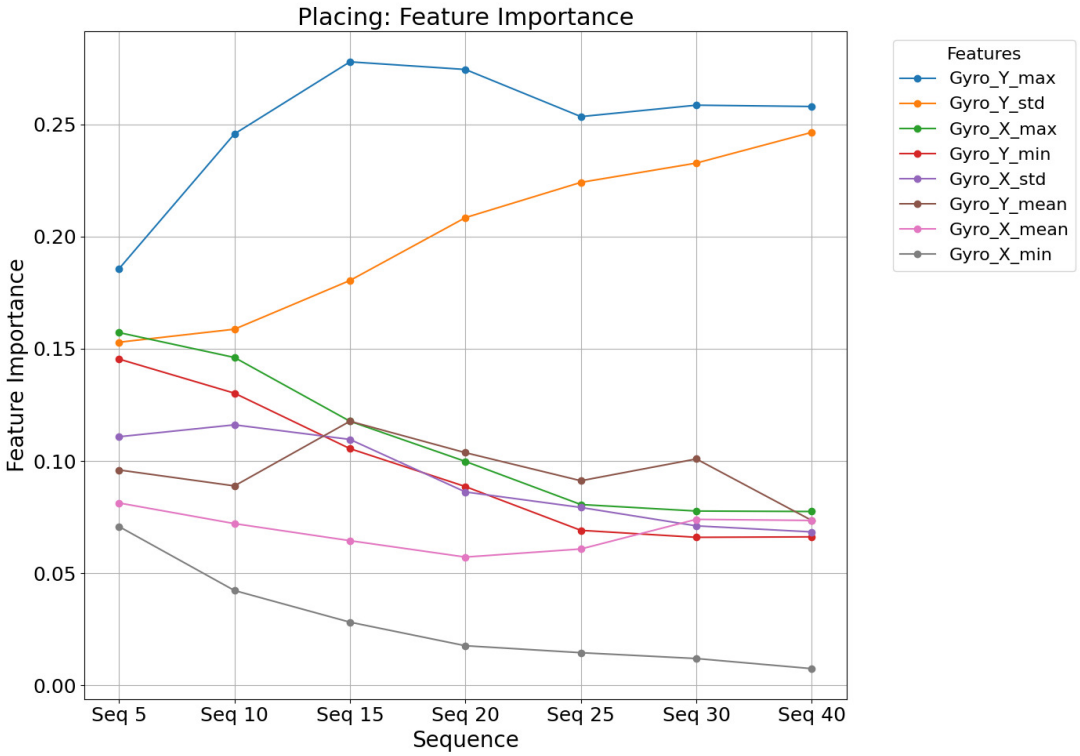

Figure 4.7: Feature Importance for the Placing data set

Chapter 5

Conclusions

In this thesis, we investigated a new program for head pose estimation and attempted to run it on the iCub camera integrated into the robot's head, with the main goal of validating its accuracy against a head-mounted eye tracker in the context of carefulness detection. As previous research had suggested, and our findings confirm, head pose estimation running on the iCub robot cannot fully replace gaze tracking devices and only provides a rough approximation. When it comes to detecting carefulness purely through gaze, small and detailed eye movements become critical features that are difficult to capture. Although the program demonstrated impressive performance, it faced challenges in capturing the subtle features that differentiate careful from non-careful gaze movements when implemented online and embedded on the robot. Instead, the trivial measure of speed emerged as a significant feature in characterizing these behaviors. Thanks to the transparency of the random forest model, we were able to identify this issue. The highly precise eye-tracker glasses, which served as our ground truth, showed that carefulness can be accurately measured by head position alone, with an accuracy of over 70% already after less than 0.65 seconds, as seen with the Tobii dataset using gyro Y and gyro X. Accurate detection of carefulness continues to rely on the precision offered by dedicated eye-tracking systems.

References

[1] G. Rizzolatti and L. Craighero, "The mirror-neuron system," *Annu. Rev. Neurosci.*, vol. 27, no. 1, pp. 169–192, 2004. 2

[2] G. Metta, P. Fitzpatrick, and L. Natale, "Yarp: yet another robot platform," *International Journal of Advanced Robotic Systems*, vol. 3, no. 1, p. 8, 2006. 7

[3] N. Noceti, A. Sciutti, and F. Rea, *Modelling human motion.* Springer, 2020. 9

[4] C. McCoubrey, *Effort observation in movement research: An interobserver reliability study.* PhD thesis, Hahnemann University, 1984. 9

[5] N. Noceti, A. Sciutti, and G. Sandini, "Cognition helps vision: Recognizing biological motion using invariant dynamic cues," in *Image Analysis and Processing—ICIAP 2015: 18th International Conference, Genoa, Italy, September 7-11, 2015, Proceedings, Part II 18*, pp. 676–686, Springer, 2015. 10

[6] N. Noceti, F. Odone, A. Sciutti, and G. Sandini, "Exploring biological motion regularities of human actions: a new perspective on video analysis," *ACM Transactions on Applied Perception (TAP)*, vol. 14, no. 3, pp. 1–20, 2017. 10

[7] A. Vignolo, N. Noceti, F. Rea, A. Sciutti, F. Odone, and G. Sandini, "De-

tecting biological motion for human–robot interaction: A link between perception and action," *Frontiers in Robotics and AI*, vol. 4, p. 14, 2017. 10

[8] R. Sigala, T. Serre, T. Poggio, and M. Giese, "Learning features of intermediate complexity for the recognition of biological motion," in *Artificial Neural Networks: Biological Inspirations–ICANN 2005: 15th International Conference, Warsaw, Poland, September 11-15, 2005. Proceedings, Part I 15*, pp. 241–246, Springer, 2005. 11

[9] Y. Y. Joefrie and M. Aono, "Multi-label multi-class action recognition with deep spatio-temporal layers based on temporal gaussian mixtures," *IEEE Access*, vol. 8, pp. 173566–173575, 2020. 11

[10] C. Coppola, S. Cosar, D. R. Faria, and N. Bellotto, "Social activity recognition on continuous rgb-d video sequences," *International Journal of Social Robotics*, vol. 12, no. 1, pp. 201–215, 2020. 11

[11] H. S. Koppula and A. Saxena, "Anticipating human activities using object affordances for reactive robotic response," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 1, pp. 14–29, 2015. 12

[12] D. N. Stern, *Forms of vitality: Exploring dynamic experience in psychology, the arts, psychotherapy, and development*. Oxford University Press, USA, 2010. 12

[13] F. Vannucci, G. Di Cesare, F. Rea, G. Sandini, and A. Sciutti, "A robot with style: Can robotic attitudes influence human actions?," in *2018 IEEE-RAS 18th International Conference on Humanoid Robots (Humanoids)*, pp. 1–6, IEEE, 2018. 13

[14] G. Di Cesare, M. Marchi, A. Errante, F. Fasano, and G. Rizzolatti, "Mirroring the social aspects of speech and actions: the role of the insula," *Cerebral Cortex*, vol. 28, no. 4, pp. 1348–1357, 2018. 13

[15] G. Sandini, A. Sciutti, F. Rea, A. Goswami, and P. Vadakkepat, "Movement-based communication for humanoid-human interaction," *Humanoid Robotics: A Reference*, pp. 1–29, 2017. 14

[16] G. Di Cesare, F. Vannucci, F. Rea, A. Sciutti, and G. Sandini, "How attitudes generated by humanoid robots shape human brain activity," *Scientific Reports*, vol. 10, no. 1, p. 16928, 2020. 14

[17] M. K. Kaiser and D. R. Proffitt, "The development of sensitivity to causally relevant dynamic information," *Child Development*, pp. 1614–1624, 1984. 14

[18] M. Gori, A. Sciutti, D. Burr, and G. Sandini, "Direct and indirect haptic calibration of visual size judgments," *PLoS One*, vol. 6, no. 10, p. e25599, 2011. 15

[19] A. Sciutti, L. Patane, F. Nori, and G. Sandini, "Understanding object weight from human and humanoid lifting actions," *IEEE Transactions on Autonomous Mental Development*, vol. 6, no. 2, pp. 80–92, 2014. 15

[20] L. Garello, L. Lastrico, F. Rea, F. Mastrogiovanni, N. Noceti, and A. Sciutti, "Property-aware robot object manipulation: a generative approach," in *2021 IEEE International Conference on Development and Learning (ICDL)*, pp. 1–7, IEEE, 2021. 15

[21] L. Lastrico, V. Belcamino, A. Carfì, A. Vignolo, A. Sciutti, F. Mastrogiovanni, and F. Rea, "The effects of selected object features on a pick-and-place task: A human multimodal dataset," *The International Journal of Robotics Research*, vol. 43, no. 1, pp. 98–109, 2024. 17

[22] N. Scarano, *Boundary Conditions for Human Gaze Estimation on A Social Robot: Evaluation of the State-of-the-Art Models and Implementation of Joint Attention Mechanism.* PhD thesis, Politecnico di Torino, 2023. 18

[23] O. Palinko, F. Rea, G. Sandini, and A. Sciutti, "Eye gaze tracking for a humanoid robot," in *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*, pp. 318–324, IEEE, 2015. 19, 20

[24] F. F. Tomenotti, N. Noceti, and F. Odone, "Head pose estimation with uncertainty and an application to dyadic interaction detection," *Computer Vision and Image Understanding*, vol. 243, p. 103999, 2024. 21

[25] G. Cantarini, F. F. Tomenotti, N. Noceti, and F. Odone, "Hhp-net: A light heteroscedastic neural network for head pose estimation with uncertainty," in *Proceedings of the IEEE/CVF Winter Conference on applications of computer vision*, pp. 3521–3530, 2022. 21

[26] M. Raković, N. F. Duarte, J. Marques, A. Billard, and J. Santos-Victor, "The gaze dialogue model: Nonverbal communication in hhi and hri," *IEEE Transactions on Cybernetics*, vol. 54, no. 4, pp. 2026–2039, 2022. 22

[27] N. F. Duarte, M. Raković, and J. Santos-Victor, "Robot learning physical object properties from human visual cues: A novel approach to infer the fullness level in containers," in *2022 International Conference on Robotics and Automation (ICRA)*, pp. 10375–10381, IEEE, 2022. 22, 38, 41

[28] Y. L. Pang, A. Xompero, C. Oh, and A. Cavallaro, "Towards safe human-to-robot handovers of unknown containers," in *2021 30th IEEE International Conference on Robot Human Interactive Communication (RO-MAN)*, pp. 51–58, 2021. 22

[29] B. Mahanama, Y. Jayawardana, S. Rengarajan, G. Jayawardena, L. Chukoskie, J. Snider, and S. Jayarathna, "Eye movement and pupil measures: A review," *frontiers in Computer Science*, vol. 3, p. 733531, 2022. 24

[30] L. Lastrico, N. F. Duarte, A. Carfí, F. Rea, F. Mastrogiovanni, J. Santos-Victor, and A. Sciutti, "Like robots, like humans: Pupil dilation during collaborative object manipulation," in *2023 21st International Conference on Advanced Robotics (ICAR)*, pp. 264–270, IEEE, 2023. 24

[31] B. Rimé and L. Schiaratura, "Gesture and speech in fundamentals of non-verbal behavior," 01 1991.

[32] E. Coronado, J. Villalobos, B. Bruno, and F. Mastrogiovanni, "Gesture-based Robot Control: Design Challenges and Evaluation with Humans," 05 2017.