Towards Proactive Support for Human-Robot Collaboration

Ayesha Jena

Licentiate Thesis, 2025

Department of Computer Science
Lund University

ISBN 978-91-8104-656-4 (electronic version)
ISBN 978-91-8104-655-7 (print version)
ISSN: 1652-4691
Licentiate Thesis 4, 2025

Department of Computer Science
Lund University
Box 118
SE-221 00 Lund
Sweden

Email: `ayesha.jena@cs.lth.se`

Typeset using LaTeX
Printed in Sweden by Tryckeriet i E-huset, Lund, 2025

ABSTRACT

As robots increasingly integrate into our social environments, from factories to social spaces, there is a growing need to find ways to effectively collaborate in these dynamic environments. However, current robotics research is mostly aimed towards task and environment specific programming. Even the state-of-the-art collaborative robotics technologies lack or have a very rudimentary understanding of the multimodal methods used by human teammates to communicate in real-time. This leads to an increased workload for human operators and becomes a critical problem that limits robots to operate in dynamic environments. We focus on one such dynamic setting of search and rescue (SAR) scenarios. In order to achieve effective collaboration between humans and robots in this scenario, there is a need for robots to naturally understand human intentions through the interpretation of multimodal communication cues such as gaze, gesture, and contextual signals in real-time. This research aims towards achieving mixed-initiative interaction by addressing the gap of robots proactively collaborating with humans through a two step process. The first part of the thesis, following a Design Science approach, explores the use and integration of non-verbal communication cues to conduct collaborative tasks in a SAR environment. Through designing the human-in-the-loop collaboration system CueSense and testing different collaboration strategies, we investigate when and how humans and robots can dynamically share control during missions. This modular system is capable of tracking gaze to predict task focus and gesture inputs for nuanced intent interpretation. This is validated through user studies where participants work alongside the system in different collaborative settings for a simulated search-and-rescue scenario. The results show that the system successfully assists users in the task and improves task efficiency, performance, and reduces mental workload. The second part of the thesis focuses on intention recognition as a foundation of proactive support and mixed-initiative interaction in human robot collaboration. We present an extensive review of the literature on intention recognition and identify gaps and challenges in implementing robust intention recognition systems in robotics. In summary, we focus our research on communication modalities, interfaces, and intention recognition for mixed-initiative interaction to allow efficient and seamless human robot collaboration in dynamic high-stakes scenarios.

CONTRIBUTION STATEMENT

The following papers are included in this dissertation:

Paper I Ayesha Jena, and Elin Anna Topp. “Chaos to control: human assisted scene inspection.” In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction. 2023. Stockholm, Sweden* DOI: 10.1145/3568294.3580133.

Paper II Ayesha Jena, and Elin Anna Topp. “Towards Understanding the Role of Humans in Collaborative Tasks”. In *7th International Workshop on Virtual, Augmented, and Mixed-Reality for Human-Robot Interactions at HRI 2024. Boulder, CO, USA*

Paper III Ayesha Jena, Stefan Reitmann, and Elin Anna Topp. "Impact of Gaze-Based Interaction and Augmentation on Human-Robot Collaboration in Critical Tasks." In *2025 International Conference on Social Robotics + AI. Springer Nature. Naples, Italy.* Accepted and to be published.

Paper IV Ayesha Jena, and Elin Anna Topp. “Bridging Minds and Machines: A Comprehensive Review of Intention Recognition in Human-Robot Interaction”. Submitted to the ACM Transactions on Human-Robot Interaction.

Most of the work in the above mentioned papers was contributed by Ayesha Jena, with discussions about concepts, implementations, and evaluation methods with her supervisors. The 3D template of the real-world location employed in the user study is provided by the WASP Research Arena for Public Safety (WARA-PS).

ACKNOWLEDGEMENTS

First of all, I would like to thank my supervisor Elin Anna Topp and my co-supervisors Jacek Malec and Björn Olofsson, who have been supporting, guiding, and recalibrating me throughout the whole process. Thank you, Elin, for your help in writing the papers and making posters, your insights, encouragement, and support, which not only helped me shape the research but also supported my professional growth. I am especially grateful for your patience with my abundance of ideas and for helping me channel them into a clear research direction. I would also like to acknowledge my late co-supervisor, Anders Robertsson. His inputs and encouragement will be warmly remembered.

Furthermore, I would like to thank Stefan and Rohith for the discussions during our joint meetings, as well as everyone in the RSS group and RobotLab for the insightful talks and lunch sessions. A special thanks to my project partners at Linköping University, Mariusz and Piotr, for their valuable inputs. I would also like to thank the WASP Research Arena for Public Safety (WARA-PS) for providing the necessary resources for conducting the user study. I am grateful to the technical and administrative group at the Department of Computer Science for their support in making many aspects of the journey smoother. I would also like to extend my thanks to the broader Computer Science Department for the many enjoyable and wonderfully unconventional discussions around the lunch table.

I would like to thank my family and friends for their unwavering support and encouragement.

This work was funded by ELLIIT - the Excellence Center at Linköping University and Lund University for Information Technology, and partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

CONTENTS

Included Papers 45

INTRODUCTION

1 Motivation

Traditionally, robots were kept in factories and within defined safety boundaries while working on predefined, repetitive tasks. With the emergence of Industry 5.0 came a more human-centric approach in collaboration and automation [DDS19]. This transition put robots right in the middle of human workspaces, which poses a significant challenge in terms of collaboration. Humans not only started sharing workspaces with robots, but the workspaces themselves evolved from fixed stations in factories to more outdoor and social environments, such as public walkways, hospitals, and so on. This shift also facilitated the use of robots in more critical and high-risk environments such as architectural inspection, nuclear decommissioning, and search and rescue scenarios, amongst others. Such collaborations in critical environments would not only require robots to have seamless communication but also adaptable behavior. Robots have to go beyond the role of just tools and function as adaptive agents that can coordinate through flexible decision-making to ensure efficient goal execution. These collaborative scenarios emphasize the need for mixed-initiative interactions, where control is dynamically shared between humans and robots and each member is allowed to intervene and seize control of it [JA15].

Over the last decade, the field of Human Robot Interaction (HRI) has made significant progress at facilitating effective interaction through a broad array of research aimed towards providing support for collaborative tasks [Bar+24]. However, there is still a substantial gap in terms of adaptive, bidirectional communication and anticipatory support in the search and rescue domain. Robots often follow pre-programmed responses to assigned tasks. This leads to humans being limited to the actions they can perform for robots to understand them intuitively. As a result, humans have to adapt their actions based on the robot's understanding, rather than the other way around. This problem becomes critical in time-sensitive and high stakes scenarios where efficiency and adaptability are essential.

This thesis is motivated by the challenges identified within the WARA public safety arena [And+21], where critical goal executions require effective bidirectional human-robot collaboration. Addressing these collaboration challenges requires advancing beyond traditional reactive human-robot interaction towards more adaptive, mixed-initiative interaction. To overcome these limitations in bidirectional communication and adaptability, this thesis focuses on the search and rescue domain with an aim towards achieving mixed-initiative interaction through proactive support. For such collaboration, robots need the ability to interpret multimodal human communicative cues dynamically. Robots can interpret human cues, such as gaze, gesture, and intent, to enable seamless collaboration by anticipating actions and combining them with environmental cues, thereby providing context-aware support. With this mixed-initiative understanding, robots can support humans to act as proactive teammates, capable of recognizing the user's intentions and providing dynamic support.

This thesis addresses the research challenge of enabling natural collaboration with robots in a dynamic environment through the following research questions:

1. How to design robotic systems for efficient and natural collaboration in real-time?

2. What are the key factors and behavioural cues that can be used to determine when and with whom to communicate during interactions?

3. How to model multimodal signals such as gaze, gestures, and full-body cues to infer human intentions during task execution in dynamic and complex environments?

To address these research questions, this work begins by identifying the relevant modalities for effective human-robot collaboration. The first of which is gaze. A robotic interaction system is designed and validated through a user study in which participants engage in a simulated search and rescue task using a test environment developed from the 3D template provided by the WASP Research Arena for Public Safety (WARA-PS) [WP24]. This study evaluates the influence of head gaze-based control on task performance during critical tasks, highlighting the complex relationship between human intention, gaze direction, and system design in collaborative settings. Building on these insights, an extensive review of existing frameworks, architectures, and models for intention recognition is conducted. This review examines the complexity of intention recognition and provides a foundation for more systematic and practical approaches to enabling seamless collaboration. This facilitates an understanding of intentions on various levels. With intention termed as as an agent's goal, along with an established dynamic action plan within a given environment, intention recognition becomes the key that allows us to understand the desired state and the sequence of actions necessary to achieve this goal. Having robots equipped with the capabilities of inferring human intent during tasks, would facilitate seamless collaboration and uninterrupted support. This would facilitate seamless collaboration and mixed-initiative interaction between humans and robots in high-stakes, dynamic environments. In the next sections, we go through the background of the work and detail the work done and explain how it relates to answering the research questions formulated earlier.

1.1 Thesis Outline

This thesis consists of 6 chapters, including this one. The following is the brief outline of each section:

- In Section 1, we present our motivation behind the work done.
- In Section 2, we discuss the background on human robot collaboration, proactive support, mixed-initiative interaction, associated communication modalities, and interface systems.
- In Section 3, we talk about our system and its implementation, the user study conducted with our system, the results from the study, and intention recognition which is a core part of providing proactive support.
- In Section 4, we conclude our work and discuss plans for future work.
- In Section 5, we talk about our contributions from each of our paper presented in this thesis.
- In Section 6, we mention some other papers authored by us which were not added to this thesis.

2 Background

In this section, we review and discuss the significant developments in the field of Human Robot Interaction with a focus on the progress made towards human-in-the-loop systems, proactive collaboration, mixed-initiative interaction, and multimodal communication. We then identify the gaps in the existing research to inform the work made in this thesis towards furthering the research in the field of human-robot collaboration.

2.1 Humans as the centre of Robotics

With the advancement of robotics over the years and the transformation of the Industrial Revolution, humans have become central to robotics research. Although the field of Human Robot Interaction has been around for some decades, the integration of humans into automated systems in Industry 5.0 has redefined this relation [DDS19, Xu+21].

Early robotics development was primarily focused on industrial robots, programmed for single tasks with a lack of sensors to detect or avoid humans in the vicinity. The field has evolved over the years, diversifying not only in the roles robots play in automation but also in the roles humans play. This shift in robotics, often termed as Collaborative robots [Vic21], led to robots working side by side with humans in diverse contexts. Some use cases of robots in social spaces are the Amazon warehouse robots for fetching items to humans [Dha20], BMW KUKA LBR iiwa to assist workers with heavy lifting [TMK16], Healthcare robots like moxi for delivering supplies and assisting nurses [Ayd23], and Boston Dynamics SPOT robot used for industrial inspection [CP24], construction site monitoring, and search and rescue operations, among others.

As we see from the discussion above, there are three main components of human-robot interaction: the human, the robot, and the interaction that bridges them based on the collaborative task:

- Humans: With humans forming a central part of the process of designing interactions, it is also crucial to interpret the role of humans in the robotic design process. One way of addressing this could be looking at it from a user-centered design perspective [Bar+24, Geb+24]. This involves separating users into three categories: primary users (those with direct interaction, such as nurses using a delivery robot or a teleoperated robot), secondary users (those with occasional interactions), and tertiary users (those indirectly affected by the design process) [Bar+24].

- Robots: The focus in case of robots lies in their capabilities, limitations, and design considerations that directly influence interaction quality. This includes both their physical appearance (shape, manipulators, etc.) and the software (perception, decision-making, learning, etc.).

- Interactions: Although interaction in HRI can be a broad multidisciplinary field, the interaction modalities through which humans can communicate with robots can be broadly classified into verbal and non-verbal methods. While verbal methods constitute the ways in which we explicitly state the tasks to a robot, non-verbal refers to interpreting gaze, gesture, body language, and the meaning behind such actions [Bar+24].

While the field has diversified over time with robots like SPOT [CP24], Unitree G1 [Ghi+24], NASA Valkyrie [Rad+15], and other robots working alongside humans in industrial, healthcare, social, and service domains, there is still a significant gap to allow humans to safely and seamlessly collaborate with robots on every frontier. The continued efforts towards human-centric automation in robotics and the shift from an isolated understanding of robotic task implementation to synergetic human-robot interaction in diverse settings have thus become important [Geb+24].

However, over the last decade or so, the research in the field has been pushed towards more autonomous robot behaviors, leading to the introduction of a subfield called Proactive Human Robot Interaction or Proactive HRI [DBM24]. This field has been increasingly researched with a focus on topics such as human-in-the-loop solutions and human-centered solutions, which are aimed towards more natural and intuitive human-robot interaction [DBM24, Top17]. These interactions are inspired by how humans interact with one another, leading to mixed-initiative interaction. This is in contrast to reactive robotics, which focuses on robots responding to human actions and changes in the environment as they occur without anticipating future needs or creating any long-term plans (for example, reactive obstacle avoidance) [DBM24].

2.2 Proactive Support for Mixed-Initiative Interaction

While a broad range of definitions have been used to define proactive in the case of HRI, the common idea focuses on going beyond the work in reactive behavior. Pandey et al [PAA13] refer to this as “taking the initiative whenever necessary to support the ongoing interaction/task”, van Den Broek et al. [DBM24] provide a broader overview on the topic, stating that Proactive HRI is a subfield where robot could anticipate future states or show control over the situation in some way.

Building on these definitions, the field of human-robot interaction has evolved over the years to include a wide range of paradigms in which interaction is a central and integral part of robot interfaces and control systems. These include collaborative interaction [Mar+20], social interaction [LPT22], multi-agent interaction [Dah+23], mixed-initiative interaction [JA15], supervisory control [RSR16], assistive interaction [CTS21], and teleoperation [CHB07], among others.

In each of these interaction paradigms, the human plays a crucial role, with the nature of that role varying depending on the context, task complexity, and system goals. Understanding how the roles of humans and robots shift across these settings is essential to exploring the associated dynamics, challenges, and outcomes. This would enable human-centered operation to be safer, efficient, intuitive, and natural. It would also enable robots to provide proactive support to humans and contribute towards mixed-initiative interaction for seamless collaboration.

There have been many definitions proposed for mixed-initiative interactions. This concept was originally rooted in the Human-Computer Interaction (HCI) domain as referenced by Carbonell in 1970 [Car70], and was used to create intelligent conversational agents. The approach was later extended to human-robot teams in 1997 by Kortenkamp et al. [Kor+97], and was used as a novel planning perspective for the tasks involved in shared control scenarios. Jiang et al. [JA15] provided a comprehensive definition which explains Mixed-Initiative HRI "A collaboration strategy for human-robot teams where humans and robots opportunistically seize (relinquish) initiative from (to) each other as a mission is being executed, where initiative is an element of the mission that can range from low-level motion control of the robot to high-level specification of mission goals, and the initiative is mixed only when each member is authorized to intervene and seize control of it."

For a successful implementation of a mixed-initiative system, the robot must detect, understand, and interpret cues such as gaze behavior and body posture, not just in isolation but also in relation to the task. This requires an understanding of the human's intention and the semantics of the scene. These may include verbal communication to explicitly instruct a robot about a task, non-verbal demonstration through gestures, motion, or actions, as well as haptic and tactile sensing methods. Dani et al. demonstrate the non-verbal aspects of communication in their work about human-in-the-loop for human-robot collaboration [Dan+20]. They utilize human motion intentions to generate safe and suitable robot trajectories that aid humans in collaborative tasks. Additionally, devices such as VR headsets or augmented reality headsets like the Meta Quest or Oculus can be used to assist and control the robot based on operational needs through teleoperation [Wal+23].

In Human Robot Collaboration, communication modalities form the foundation of intention-aware systems, enabling robots to support humans dynamically. These modalities range from using gaze cues to full-body communicative gestures to infer the actions. These can be further supported by using devices and systems that enable the robots to directly or indirectly infer these cues. Consider the case of teleoperation, where robots can be directly controlled through interfaces or AR devices like Oculus Quest, enabling direct manipulation of the robot arms to perform tasks such as assembly or inspection [Wal+23]. These are cases of direct inference of intent through controlling the robot using hardware input. In the case of indirect inference, a robot would be able to infer the intentions of humans and provide support autonomously. For example, in manufacturing, understanding the intentions of the workers could help synchronize the robot's actions for efficient goal execution. For instance, a dual-arm YUMI could assist a worker with obtaining boxes triggered by eye contact for in-time component delivery. Similarly, a Boston Dynamics SPOT robot could follow a firefighter's gaze and hand directions to prioritize search locations in areas that are dangerous for humans to enter. These use cases could also be applied to humanoid robots like NASA's Valkyrie in a post-disaster scenario, to map traversal routes for humans and clear debris in the predicted paths.

Achieving effective collaboration of this nature would require an analysis of various aspects of human-robot interaction, encompassing both robots and humans, which are essential for proactively supporting collaborative tasks. The following sections discuss these essential components critical to bridging the gap in achieving seamless collaboration.

Communication Modalities

The examples above highlight the pivotal role of nonverbal communication in enabling intuitive and efficient human-robot collaboration in different scenarios. This is especially true for scenarios where verbal instructions may be impractical due to environmental noise, safety requirements, or time-critical operations. Nonverbal communication methods such as gaze, facial expression, gestures, posture, body language, and proxemics provide suitable communication channels between humans and robots in such cases [Bar+24].

- Gaze: Among these, gaze is one of the most powerful and intuitive channels for communication and intention signaling. Robots can use gaze to interpret human intention and provide timely and context-aware support without explicit verbal commands. Huang et al. have shown how gaze is used to predict target ingredients of customers around 1.8s before they even express it verbally [Hua+15a]. Admoni and Scassellati review social eye-gaze in human-robot interaction, focusing on how people respond to gaze and how robot gaze behavior improves interaction [AS17]. Belcamino et al. demonstrate how eye gaze can be used to infer human intentions and allow robots to collaborate in assembly tasks [Bel+24].

- Gesture: Apart from gaze, gesture also plays a crucial role in channeling communication and intention for enabling proactive support in HRC. While gaze is mainly focused on either eye or head gaze, gesture can be represented by a combination of multiple pose points in the body. One of the most distinct classifications of gestures given by McNeill differentiates between pointing/deictic, metaphoric, iconic, and beat gestures [McN92]. Other gesture classifications are presented in the works by Nehaniv et al., where the authors categorised them as: Irrelevant/Manipulative gestures, Side effect of expressive behavior, Symbolic gestures, Interactional gestures, and Pointing gestures. Not all gestures represent interaction intent. Their work also explains how there is a need not just to consider the kinematics of the gesture but also the interactional context of it. Additionally, it aids in classifying gestures, enabling robots to respond appropriately when needed [Neh+05]. Recognition of gestures is invaluable in environments with high noise levels, such as on a shop floor in a manufacturing facility [Net+19, LW18]. Different gestures map naturally to various classes of tasks. For example, static hand gestures can be used to stop a robot, whereas dynamic gestures could be used to move the robot to a given position [Net+19]. Apart from these, there are also upper-body gestures that combine hands and head gesture combinations, which predominantly allow for the interpretation of engagement or contextual cues [Xia+14]. Full-body gestures, such as posture shifts and spatial orientation, also play a significant role in providing environmental context, which can allow robots to perform safe maneuvers around the human workers [TBG23].

- Haptics and Physical interaction: Apart from gaze and gesture modalities, it is also essential to understand the role of haptics and physical support as modalities used for proactive support in HRC. Based on the levels of engagement with the robot, HRI can be categorised into remote and proximate interaction [FC24]. While remote interaction involves humans and robots being separated in space or time, proximate interaction refers to being colocated in the same space [FC24]. In both cases, physical interaction is possible either through direct (contact-based) or indirect interaction [FC24]. Direct or contact-based physical interaction involves modalities such as touch, grasp, and kinesthetic teaching, among others, whereas indirect physical interaction occurs through interaction through the objects as intermediary [FC24]. Physical cues, such as touch, guiding forces, or contact pressure variations, also help convey user intent and prompt support [Liu+22]. For example, in a collaborative manipulation task, force changes can be used to signal the direction of intended motion or request the release of the object. Timely and seamless interpretation of the above-discussed cues in real time would help robots anticipate human actions, to provide context-based, appropriate support, leading to intuitive, seamless, and safe collaboration.

- Multimodal interaction: The above modalities are singular in nature. This could, however, limit the robustness and adaptability in complex and high-stakes environments. In addition, humans often combine multiple modalities such as gaze, gesture, touch, and speech to convey intent. Gao et al. [GLJ21], Yongda et al. [YFH18], Li et al [LLT22], Cid et al. [CMN15], and Zlatintsi et al. [Zla+18] have used fusion techniques to highlight the importance of multimodal information. These include hand gesture identification, voice and gesture combination to infer instructions for the robot, auditory and visual modalities for perception tracking, visual and auditory signals to identify different states of emotion, voice and gestures to provide support for elderly healthcare [GLJ21, YFH18, LLT22, CMN15, Zla+18].

> While multi-modal techniques are the closest to interpreting context-aware behaviours, they are also more resilient to failures compared to single modality approaches.

These multimodal fusion techniques discussed above naturally extend to the need of designing interfaces where it is possible to fuse information from multiple sources. These could operate as operational layers or systems that fuse communication channels, enabling robots to sense, capture, interpret, and seamlessly provide feedback. These interfaces synchronize multiple modalities to enhance user immersion and awareness, contributing to more effective human-robot collaboration [Tri+20]. In the context of virtual environments, Wonsick et al. [WP20] give a categorization of interfaces based on robot control models for different levels of autonomy and interaction modalities. This categorization includes human inputs, robot movements, and virtual systems and identifies three main types of interfaces in HRI, namely, direct, cyber-physical, and homuncular. In addition to this, research in the field shows that multi-modal interfaces play a critical role in reducing cognitive workload and improving task performance [Tri+20].

The work done highlights the considerable progress that has been made from collaborative systems and multimodal communication to proactive support to facilitate seamless and efficient Human Robot Collaboration. With the immense potential and the use case of multimodal interactions, there still lies the challenge of integrating multiple input methods into a system. This will not only be computationally expensive but also cause processing delays in real-time. Additionally, while these modalities present rich interaction potentials, they often increase cognitive workload and hinder situational awareness, modularity, scalability, and reusability. In this work, I investigate and address this gap and work towards frameworks and recognition systems that allow natural, seamless, and proactive human-robot collaboration. With a focus on high-stakes scenarios of search and rescue, this work is aimed towards achieving mixed-initiative interaction, wherein humans and robots can dynamically seize (relinquish) initiative during task execution.

3 Implementation and Evaluation

The thesis consists of two sub-projects. The first sub-project, following a Design Science approach, aimed to methodologically explore the use and integration of non-verbal communication cues to conduct collaborative tasks. This resulted in the development of a system that incorporates non-verbal inputs into a human-robot collaboration task in a search and rescue scenario. The system is further validated through a user study of different collaboration strategies, with results providing insights for future applications. The system, along with the results from the user study, helps us answer our first research question entirely, and our second and third research questions partially. The second sub-project works as a foundation to bridge the gap in incorporating multi-modal cues with context-aware environment information to equip robots with the capabilities of intent recognition. Both sub-projects are aimed towards achieving mixed-initiative interaction in a search and rescue scenario, with a focus on critical task execution in a high-risk environment.

3.1 CueSense System: Technical Implementation

In this section, we discuss the robot interaction system that was designed for this work and the user study that was used to evaluate the system. Our system comprises different modules - gaze detection, input mapping, robot command generation, visual feedback, and an augmentation system. An overview of the entire architecture can be seen in figure 1. Images of the system can be seen in Paper II and Paper III.

The system integrates gaze and gesture recognition to enable intuitive, real-time human input to operate a robot in a dynamic environment. It is novel in its design and integration of head-gaze based interaction and real-time foveation-based visual augmentation for collaboration in a critical environment. *"Foveation"* is traditionally used as a technique in graphics performance optimization. However, we use it here to show areas of interest unblurred and other regions as blurred. This is done as an *"augmentation technique"* meaning enhancing the method in which the information is presented to the user. In simple terms, the user's non-verbal inputs and robot inputs are integrated into a system that works as both an interface and a system to help execute the mission specifications. An RGB camera is used to capture the user inputs and process them using Mediapipe's [Goo22] pretrained models. The model identifies the 3D facial and hand landmarks with an average precision of 98.61% and 95.7% respectively. The gaze vectors and gesture States are then detected, encoded, and transmitted using a low-latency UDP protocol to the control interface developed in Unreal Engine. With an instantaneous maximum latency of 10 ms within the same devices, the latency was within the admissible limits needed to transmit user input into the control system. This control interface is what the users see during robot control, providing inputs through gaze and gesture commands, and other inputs to the scenarios designed for testing the

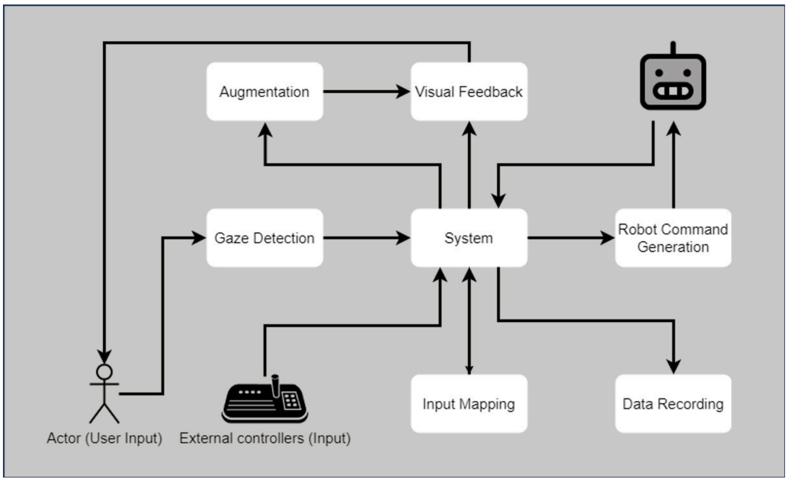


Figure 1: Overview of the system

usability of the interface. The interface consists of a customizable 3x3 screen grid, which receives visual input from the robot camera either in simulation or real-time. Other sections of the screen include the following functions: Task Allocator, Items found, Restore, and End Task. Task Allocator helps the user to select the priority of different locations in the scene after their analysis through robot navigation. These can be used to send information to Search and Rescue operators to further analyze important locations. Items found are used to indicate the number of objects of interest detected by the human during exploration through the robot, such as potential victims. The Restore function is used to record the last known tagged location, and End Task is used to conclude the current operation, log outcomes, and save the data for post-operation analysis. Users use these to provide inputs to the system apart from using gaze-based selection and gesture-based confirmation for robot navigation, allowing natural command input.

The system detects and translates these cues to control commands based on the below:

Gaze detection module It uses the egocentric perspective of the user to acquire the direction of the gaze. This is achieved using the Mediapipe library [Goo22]. At the same time, alternatives such as Tobii eye tracker [Fun+16] and other virtual reality-based devices [CKK19] can also be used based on the situation. Additionally, this module can be modified to acquire a full body pose to supplement the eye gaze as well.

Algorithm 1 : Gaze Detection

```
Process image for face.
if face landmarks detected then
    Extract specific landmarks.
    Calculate 2D and 3D coordinates.
    Define camera and distortion matrices.
    Solve PnP for rotation and translation vectors.
    Convert to rotation matrix, get Euler angles.
    Calculate gaze direction.
end if
```

Input Mapping The detected gaze direction is used in a confirmation procedure to light up the screen where the user is looking. This marks the pre-confirmation phase, where the user verifies the action they want the robot to perform. The user then confirms their intent to navigate the robot according to that particular command using the respective keyboard keys. The screens not only work as pre-confirmation interfaces, but also provide visual feedback of the scene through a live feed of the robot's camera.

Algorithm 2 : Input Mapping: User to System

```
Perform a line trace from the eye's position based on gaze direction
Determine where the user is looking at
Set the actor hit by the line trace as the Gazed Screen
Highlight the Gazed Screen orange
if Key press is same as highlighted screen then
    Highlight the Gazed Screen green
    Set confirmation message as Gazed direction
end if
Send confirmed Gazed direction
```

Robot Command Generation The confirmed commands from the user are converted to appropriate velocity commands for the robot in the simulation. Presently the system generates holonomic velocity commands which can be mapped to non-holonomic robots based on the requirements.

Algorithm 3 : Robot Command Generation

```
Receive confirmed Gazed direction
if Gazed Direction then
    Move the robot towards corresponding direction at the required velocity until the command received
end if
```

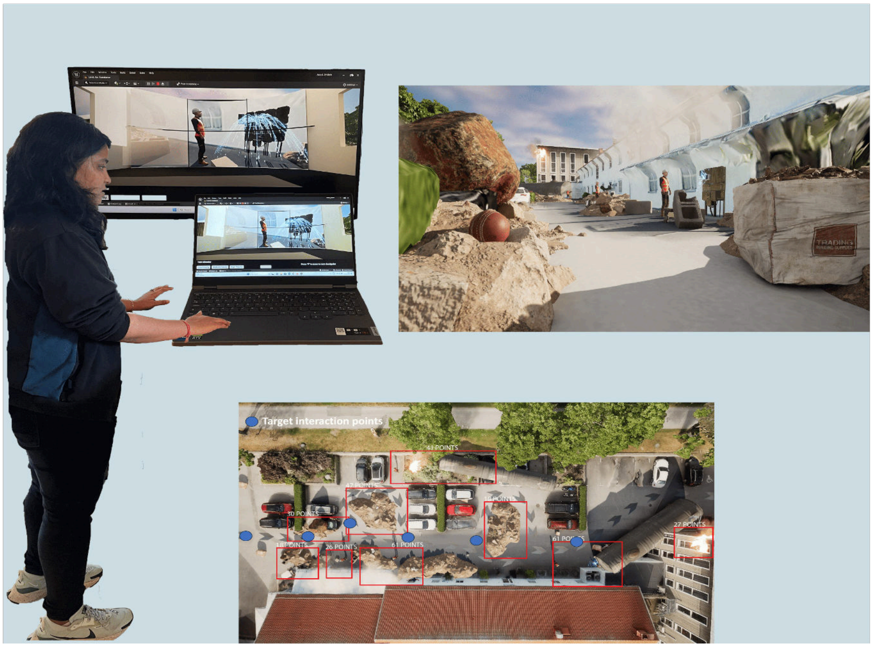
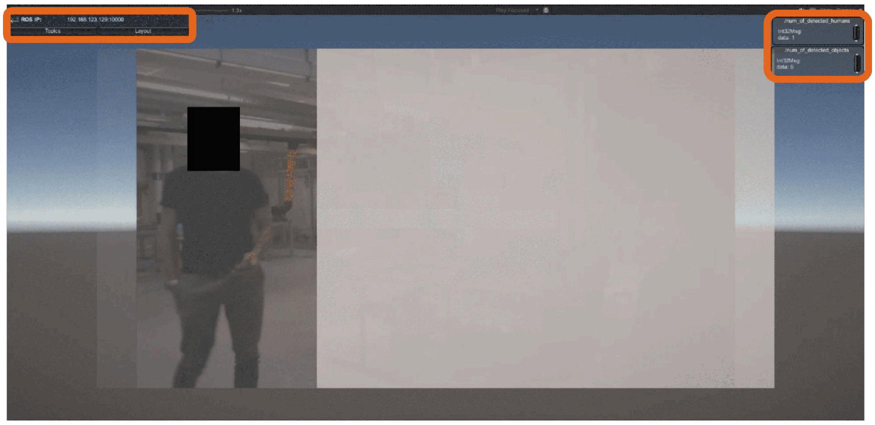


Figure 2: Experimental setup showing a human operator leveraging their intuition to guide the robot in search and rescue virtual scene using "gaze and hand signaling controls" while identifying areas of interest.

For the user study, we set up a post-disaster environment in a 28 * 83 m simulation using a test-bed environment provided by WARA-PS [WP24]. The motivation behind the search and rescue environment was to replicate challenging and high-stakes conditions for collaboration in a controlled and safe setup. Additionally, the environment includes simulated SAR personnel to replicate realistic operational constraints and potential coordination efforts in real-time rescue missions. The robot used in the simulation for the study is a mobile robot platform MiR200 [Rob22] with a UR5e robotic arm [Rob08] and a Schunk gripper [Sch].

During the experiment, participants were placed in a simulated control room. This control room is designed for real-world SAR operation centers and interacts with the virtual disaster site by navigating a mobile robot fitted with cameras and sensors. During each mission, they encountered survivors, some of whom were deliberately hidden in difficult-to-reach spots to increase the challenge.

The system has undergone several iterations and changes for integration of various modules, improvements, and ease of use. The current version of the system is reimplemented in Unity [Uni22] for validating its modular reproducability across platforms. Additional functionalities have been included: foveated augmentation



ROS IP: 192.168.123.129:10000

/num_of_detected_humans
Int32Msg:
data: 1

/num_of_detected_objects
Int32Msg:
data: 6

[redacted]

Figure 3: System: Unity version

based on human detection, real-time video input from robot, dynamic recognition of objects and humans in the scene, and display functionality for the number of detections. Figure 3 shows the user view in the Unity interface. The ROS connection happens through ROS TCP Connector [Tec22].

Study Setting

The study compares two implementations of the system to devise intuitive human-in-the-loop robotic systems for collaborative scenarios: Human-Assisted (HA) and System-Assisted (SA). Both implementations were tested in a search and rescue test environment to study the effects of different levels of user support and system design on performance and decision-making. A detailed study setting with images is described in Paper II and Paper III.

Human-Assisted (HA) Scenario: In this scenario, participants were given more control over the robot's movements. They navigated the virtual SAR environment by combining head-gaze detection with keyboard confirmations, effectively steering the robot to locate and assess the post-disaster area. This dual-confirmation input is used to mitigate the "Midas-touch" problem [VSU97] associated with using gaze-based inputs. Head-gaze-based input is confirmed via a mapped keyboard input, allowing participants to interact more directly with the system and make real-time decisions about where to move next.

During navigation, participants continuously monitored live visual feedback from the robot, which helped them identify Areas of interest(AOI) such as potential victims, debris, and other hazards. Once an AOI was identified, participants provided a number for how many important objects were in that location and assigned them each a priority level from Low, Medium, or High. This approach demands more natural and active human intervention, emphasising direct user control communicated through head-gaze inputs that conveyed participants' intentions in real-time.

System-Assisted (SA) Scenario: The second scenario employs a higher degree of system automation to guide participants in identifying and prioritizing AOIs. Using a predefined point-based technique for AOIs in the virtual SAR scene, the system clusters and highlights regions (rubble + humans + fire, electrical failure + rubble, fire + smoke + human) based on their priority scores derived from proximity to threats, human presence, and structural urgency. The scene view, captured via the robot's front-facing camera, is streamed to a 2D screen partitioned into six dynamically adaptive sections, each mapped to spatial zones in the robot's field of view and optimized to align with human visual working memory limits (5–9 chunks) [Mil56]. The foveation on these sections of the 2D screen is updated in real-time to reflect AOI density and criticality in regions with overlapping hazards, such as smoke near trapped humans. These regions are accentuated using foveated rendering [Su+23], which prioritized high-detail rendering for critical zones while gradually reducing visual fidelity in peripheral sections. Although the system guides participants to prioritized regions, participants remain the final decision-makers, confirming or adjusting the importance levels of each AOI (Low, Medium, High) through a simplified input interface. This hybrid workflow is adopted to balance automation with human oversight, to reduce cognitive load while retaining situational control.

This design and iterative development of the CueSense system directly addresses RQ1 and RQ3 by demonstrating how multimodal non-verbal cues such as gaze and gesture can be modeled for efficient, real-time collaboration in dynamic environments.

User Study

The experiment included two primary tasks: 1) an areas of interest (AOI) identification task, where participants located and prioritized relevant AOIs within the virtual environment, and 2) a decision-making task, which measured how head-gaze augmentation influences cognitive workload and efficiency in task execution. In the SA condition, the system dynamically adjusts foveation based on real-time head-gaze input, whereas in the HA condition, participants rely solely on manual head-gaze exploration without system assistance.

Before beginning the experiment, user demographic information was collected, and ethical considerations were addressed following the university's guidelines [LU23], which did not require a formal committee review for this study type.

Each experiment session began with a system functionality check. Participants were then briefed about the overall objectives of the study, signed consent forms, and completed a set of demographic questionnaires. Following this preliminary phase, participants received a short training session to help familiarize themselves with both the head-gaze-based interaction system and the virtual environment. After training, participants proceeded to the main part of the study, which consisted of two different scenarios presented in random order. The order of these scenarios was counterbalanced across participants to control for potential learning effects. Each participant performed both scenarios in sequence, with the total session lasting approximately 60 minutes per participant. Using the provided reference sheet in appendix 9.1, participants assigned priority markers on the screen to each point of interest they discovered, ensuring systematic identification of urgent rescue needs or potential hazards. Their ability to identify and categorize these points of interest was recorded, forming the basis for post-task performance assessments. The recorded metrics are discussed in detail in the next section.

While the CueSense system addresses RQ1 and RQ3, results from the user study partially address RQ2 and RQ3. The user study explores the use of gaze and gesture for communication and to deduce direct intentions and improve collaboration in high-stakes scenarios.

Study Validation

The validity and effectiveness of the system were assessed using the evaluation metrics mentioned in the Table below.

Metrics Used	**Description**
Demographics information	Pre-experiment; data to correlate participants experience with controllers such as AR/VR devices and disaster relief scenarios to the final analyzed results
Total time taken	During-experiment; time taken by participants from start of experiment till end task button is clicked on the screen
Total humans saved	During-experiment; calculated using marked regions by all participants
Average humans saved	During-experiment; calculated using marked regions by individual participants
NASA TLX Score	Post-experiment; the average subjective mental workload reported by participants after the task
System Usability Scale	Post-experiment; the average subjective usability of the system reported by participants after the task
System Assisted Search Questionnaire	Post-experiment; used to collect feedback on user experience, trust, and challenges
Human Assisted Search Questionnaire	Post-experiment; used to collect feedback on user experience, user strategies, decision making, and challenges
End of Experiment Questionnaire	Post-experiment; used to collect feedback on comparative search efficiency, accuracy, and user strategies. User inputs were collected for operator handover and shared control scenarios
Head-gaze values	During-experiment; head-gaze direction as 3D rotation values (Rx, Ry, Rz) to compute participants' specific regions of interest and correlations
Gesture inputs	During-experiment; dual confirmation key presses indicating action selection
Robot State transformation	During-experiment; Trajectory data was computed from the location and rotation value of the robot in the map
Marked regions	During-experiment; areas of interest marked by participants
Items located	During-experiment; total number of items found in each of the various areas of interest

Table 1: Evaluation metrics for two scenarios

Results

Participants Overview: The total sample recruited for the user study consisted of 18 participants. There were 13 males, 4 females, and 1 unspecified, with a mean age of 31.29 years (SD = 9.78), excluding 1 participant who declined to report their age. Out of the 18 participants, 7 had some level of vision impairment mostly corrected with eye-glasses. Since the task involves a search and rescue scenario, the use of multiple interfaces, and virtual scenarios, we were also concerned about the experience of the participants in those aspects. Only 4 participants had experience in providing disaster relief. Participants also reported varying levels of experience across different domains: with robots (M = 2.72, SD = 1.7), with any form of virtual, augmented, or mixed reality system (M = 1.83, SD = 1.79), and with using controllers (M = 3.83, SD = 1.2). In order to eliminate any order and learning effects, half of the participants (N = 9) started the study with HA scenario while the other half started with the SA scenario.

Objective and Subjective Analysis: As seen from Table1, users completed the task faster and with higher accuracy in the SA scenario. In terms of performance metrics, SA outperformed HA. Additionally, in terms of identifying trapped humans, 47 out of 54 total possible instances were identified correctly in foveated augmentation as compared to 20 correct identifications out of 54 in the other scenario. The results in case of the subjective measurements also support the performance outcomes.

With 38% lower perceived workload in SA, and system usability ratings well above the benchmark score of 68, users found the system more intuitive to use. Based on the subjective questionnaires participants found the gaze-driven interface intuitive but noted limitations such as slow turning and restricted peripheral vision. In the SA scenario, the participants wanted greater explainability of system decisions. Head-gaze analysis of the participants highlighted the difference in behavior of participants in both the scenarios. For comparing the gaze data across two scenarios, we collected a baseline to establish the thresholds for each section of the screen. These thresholds allowed us to map the gaze direction. Baseline regions based on thresholds from baseline data are shown in figure 4. This also served as the method for removing outliers in the data collected. Across the two scenarios, we used the view counts and the variance of the gaze position to compare where the participants were focusing during the experiments.

Scenario	Total Time Taken (in s)	Humans Saved	NASA TLX Score	SUS
HA	678.88 ± 233.98	1.11 ± 0.75	53.85 ± 18.06	58.61 ± 14.80
SA	$*$**274.41 ± 52.95**	$*$**2.61 ± 0.69**	$*$**33.4 ± 15.24**	$*$**80.13 ± 16.30**

Table 2: Objective and Subjective performance results for Human-Assisted (HA) and System-Assisted (SA). Results show that SA outperformed HA significantly in all the evaluation metrics with $*$ indicating significance according to the independent t-test with $p < 0.05$.

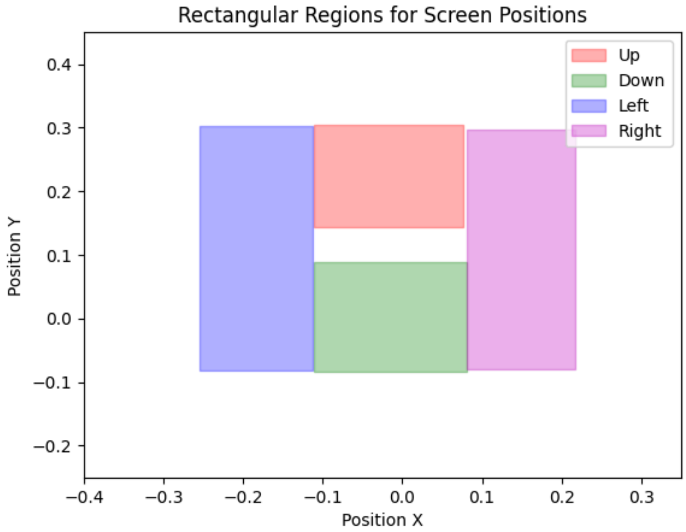


Figure 4: Baseline Regions based on thresholds from baseline data

The view counts and variance of the gaze position for each of the 9 sections of the system-assisted scenario were compared as can be seen in Table 3 and figure 5.

Section	Foveation	View Counts (Left)	View Counts (Right)	View Counts (Up)	View Counts (Down)	Average Gaze Position
Section 1	Down	88	162	48	3872	X: -0.01 ± 0.06, Y: 0.03 ± 0.06
Section 2	Left	277	81	166	3615	X: -0.03 ± 0.04, Y: 0.03 ± 0.07
Section 3	Up, Down, Right	290	48	173	5162	X: -0.03 ± 0.04, Y: 0.03 ± 0.06
Section 4	Down, Right	132	309	144	5617	X: -0.01 ± 0.05, Y: 0.03 ± 0.06
Section 5	Up, Left	233	113	148	3274	X: -0.03 ± 0.06, Y: 0.03 ± 0.07
Section 6	Up, Down	185	68	262	3423	X: -0.04 ± 0.04, Y: 0.03 ± 0.07
Section 7	Down, Left	185	45	489	5918	X: -0.03 ± 0.06, Y: 0.02 ± 0.08
Section 8	Down, Left	228	139	413	4346	X: -0.04 ± 0.07, Y: 0.03 ± 0.07
Section 9	Up	321	21	757	5537	X: -0.04 ± 0.06, Y: 0.02 ± 0.07

Table 3: Results of view counts and average gaze position for different sections of the System-Assisted scenario

Aggregate results show that, while using SA, users preferred to look predominantly towards the lower part of the screen (67.82% views, $p < 0.05$) as can be seen in Fig. 3. In the horizontal direction, users had a higher inclination of looking towards the left (3.4% views, $p < 0.05$) compared to the right side. While using HA, users had a more spread-out view count across the regions. In the horizontal direction, the users looked equally towards the left and the right (17.85% left views, 16.99% right views, $p < 0.05$). However, unlike the predominantly downward gaze in the case of SA, users in HA preferred to look more towards the upper part of the screen (33.7% views, $p < 0.05$).

There was also a high alignment (67%) of the gazed and foveated regions in the SA scenario, with the deviations occurring due to extra foveal attention capture

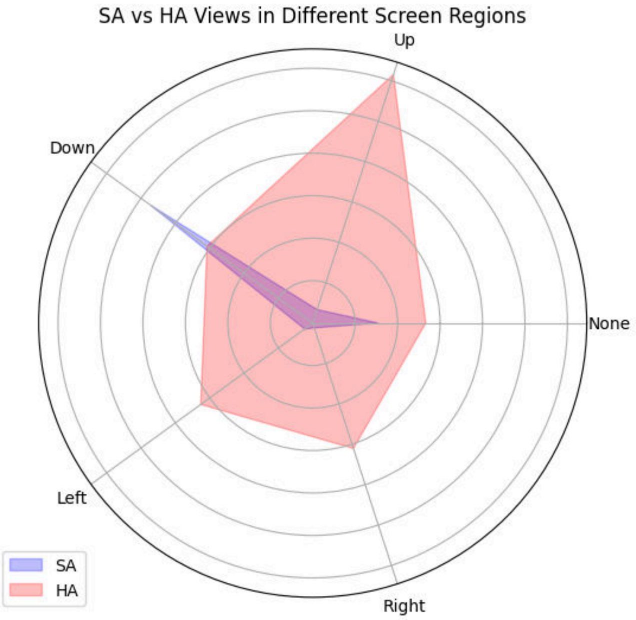


Figure 5: View Percentages in SA and HA implementations. Here, None refers to the central white region in figure 4.

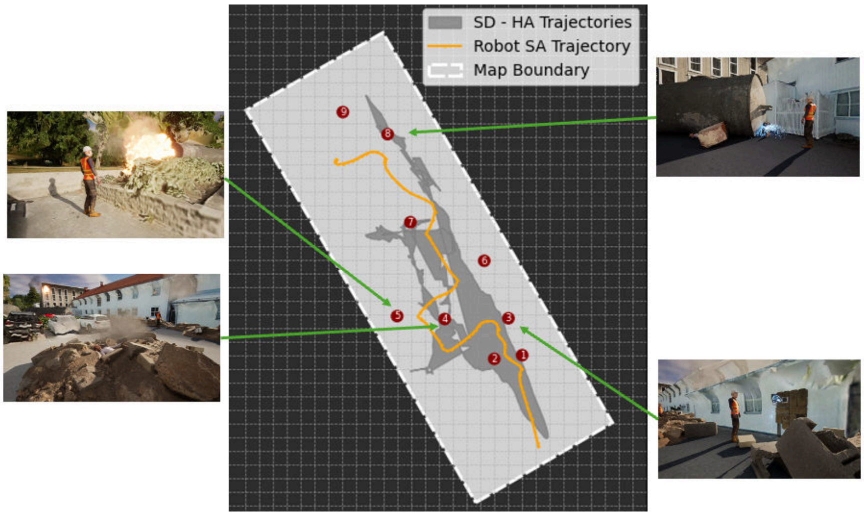


Figure 6: Mapped trajectories in HA vs SA approaches.

due to high-priority items outside the immediate field of view of the participants. Extrafoveal effects were also observed in HA scenario where participants showed interest in distant AOIs over the near ones.

A comparison of the trajectories of the robot in both the scenarios can be seen in figure 6. A detailed analysis of the results for subjective questionnaires can be found in Paper II and gaze analysis in Paper III.

3.2 Intention Recognition

This part of the thesis focuses on using the understanding taken from the work done in the previous section. The primary objective is to understand how to equip robots with the ability to interpret multiple modalities from humans. Here, using the review as a foundation, our analysis provides an in-depth understanding to allow integration of additional cues and helps bridge the critical gaps in incorporating multi-modal cues to robotics systems to provide support during collaborative task executions.

Intention recognition serves as an essential aspect that enables the integration of multimodal cues with environmental factors, which allows robots to understand and predict human intentions. Traditionally, intention recognition uses a combination of sensors, data analysis techniques, machine learning, and deep learning approaches to interpret behavior through emotions, gestures, gaze, speech, and motion. However, as robots have evolved from working solely on factory floors to operating within social environments, it has become essential to integrate context-based information that could be inferred and learned to interpret dynamic and unpredictable scenarios. Thus, it is crucial to make robotic systems adaptable and resilient. Intention could be defined as an agent's goal with a dynamic action plan within a defined environment and herein intention recognition then becomes the key that allows us to understand the desired State and the sequence of actions necessary to achieve this goal IV.

The work in Paper IV focuses on categorization of intentions into high-level, low-level, and robot intention recognition. The previous categorizations such as distal (D-intentions), proximal (P-intentions), motor (M-intentions), active, passive, have successfully attempted to capture certain aspects of intention recognition such as future intentions, intentions during planning and decision making, intentions inferred from observing others, amongst others [Pac08, Per22, Bra87, Omo+08]. However, the previous categorizations lacked the breadth needed to encapsulate the research done in the field. Our review also includes all the different methods in which intention is referred to and understood during execution phase with robots.

Figure 7 summarizes the important aspects of intention recognition.

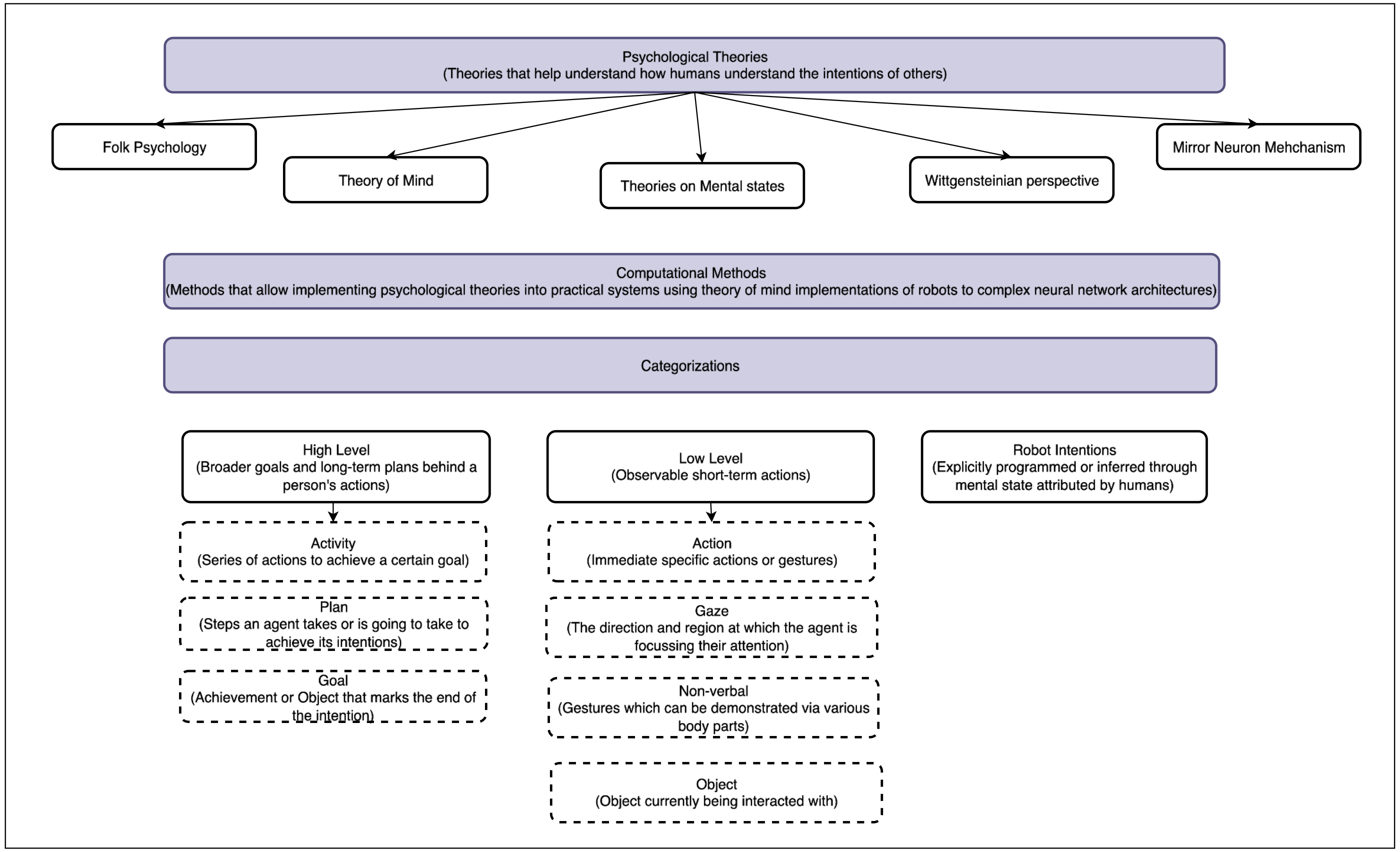

Psychological Theories
(Theories that help understand how humans understand the intentions of others)
Folk Psychology
Theory of Mind
Theories on Mental states
Wittgensteinian perspective
Mirror Neuron Mehchanism
Computational Methods
(Methods that allow implementing psychological theories into practical systems using theory of mind implementations of robots to complex neural network architectures)
Categorizations
High Level
(Broader goals and long-term plans behind a person's actions)
Activity
(Series of actions to achieve a certain goal)
Plan
(Steps an agent takes or is going to take to achieve its intentions)
Goal
(Achievement or Object that marks the end of the intention)
Low Level
(Observable short-term actions)
Action
(Immediate specific actions or gestures)
Gaze
(The direction and region at which the agent is focussing their attention)
Non-verbal
(Gestures which can be demonstrated via various body parts)
Object
(Object currently being interacted with)
Robot Intentions
(Explicitly programmed or inferred through mental state attributed by humans)

Figure 7: Intention Recognition

4 Conclusions and Future Work

Robotics research is progressing at a fast pace. However, we have a long way to catch up to the maturity and breadth achieved in artificial intelligence in recent times. With the continued efforts towards mixed-initiative interaction in robotics, the shift from an isolated understanding of robotic task implementation to synergistic human-robot interaction across diverse settings has become crucial. To enable this integration, it is essential to develop methods that equip robots with multimodal inference capabilities and adaptive behaviour. This would allow robots to understand and respond appropriately in complex and dynamic environments.

This thesis addresses the research challenge by designing real-time systems for natural collaboration, enabling effective strategies to balance humans' situational awareness with robots' automation, and undertaking comprehensive reviews to better understand effective methods of interaction through intention recognition techniques. All of this aims toward achieving mixed-initiative interaction. The research questions in the thesis have been addressed as follows:

1. How to design robotic systems for efficient and natural collaboration in real-time?

   The CueSense system demonstrates the feasibility of using nonverbal communication cues to carry out collaborative tasks. This was done by leveraging gaze input to assist users in executing tasks within search and rescue scenarios. The integration of additional modalities such as foveated augmentation and navigation assistance resulted in a 60% reduction in task completion time, which has been discussed previously. This highlights the need for methodological approaches and more studies into using human-in-the-loop mixed-initiative systems to support critical tasks, leading to better goal execution.

2. What are the key factors and behavioural cues that can be used to determine when and with whom to communicate during interactions?

   We use gaze and gesture as the primary modality in the system. The results from the study show that gaze can be used to determine where the attention is focused. However, users demonstrate extrafoveal attention captures in high-risk environments, highlighting the need to explore such mixed-initiative collaborative settings with robotic systems further. This will enable robots to handle attention shifts better while working in collaborative workspaces. Building on this understanding further, we suggest that Intention recognition in robots should be the key towards determining support protocols during task execution. The techniques in Intention recognition should utilize a combination of multimodal interaction cues, task context, and environmental conditions to provide timely support.

3. How to model multimodal signals such as gaze, gestures, and full-body cues to infer human intentions during task execution in dynamic and complex environments?

   To understand how non-verbal communication cues could be incorporated into a complex system, we began by examining individual cues. In addition to this, we employ a direct approach to utilize them in a collaborative task. This consists of an iterative process involving designing and redesigning the system and examining ways in which gaze and gesture can be directly used to infer intent. A search and rescue scenario is used for this purpose. This is not only representative of an uncontrolled, high-risk, and dynamic environment but also makes it a challenging task to balance human supervision with system automation. Further, a deeper understanding of different aspects of effective use of non-verbal communication cues, mixed-initiative interaction, and an uncertain environment was needed to design the system and formulate studies that efficiently address the challenges discussed previously. The next step involved identifying a multi-modal approach that integrates understanding from direct system-based methods with gaze, gesture, full-body cues, and scene information into a more context-aware framework. Through a comprehensive review of the methods and techniques in the field of human-robot collaboration, we conclude that intention recognition for robots provides the right strategies and direction to anticipate the goals behind human actions and adapt proactively in complex environments.

This thesis advances the field of collaborative robotics by bridging the gap in establishing the effectiveness of human-in-the-loop mixed-initiative systems and multimodal communication cues in human-robot collaboration and providing both theoretical and empirical evidence to support the claims.

While much work still needs to be done to bridge the gap in human-centered collaboration, achieving smooth and seamless collaboration is a key objective. We believe the work done in this thesis and the results obtained would serve as a valuable contribution to the field of mixed-initiative interaction and proactive human-robot collaboration that makes robots intelligent, explainable, and trustworthy.

In future work, we propose to build a comprehensive system for next-step human intention recognition in collaborative environments. The aim is to develop a multi-stage pipeline starting with real-time pose estimation of humans using tools like OpenPose or MediaPipe to extract keypoint features such as gaze, lower-limb and upper-limb positions, and torso orientation. Simultaneously, the environmental context sensing part of the pipeline will use an RGB-D camera and vision, along with pretrained models such as CogVLM [Wan+24c], to identify and localize surrounding objects in space. The multimodal data of humans will be input into a machine learning based intent classifier, which will infer one of the three main intentions: observation, interaction, or navigation. These predicted intentions, along with raw sensor inputs, data of the environment, and with normalized task-conditioned relevance weights for each object obtained from an LLM-based scoring function, will be integrated into a Bayesian inference framework, enabling structured probabilistic reasoning about the human's following most likely actions which are then used to trigger robot policies for proactive assistance.

This Bayesian model will run in an iterative inference loop, updating predictions in real time as soon as new observations are collected. Each module of this pipeline remains to be implemented and evaluated, and a key part of the future work will be designing a robust Bayesian network structure (static or dynamic) that can handle uncertainty and incorporate dependencies across modalities (for example, combining gaze and hand gestures for highly accurate intent about interaction with an object). Additionally, along with object-task relevance scores, we also plan to explore how peripersonal relevance can influence priors within this network to guide inference. This will be followed by ablation studies to understand the impact of each component and module in the pipeline. The goal is to enable a robot to anticipate human intentions in real-time and respond appropriately, supporting more intuitive and seamless collaboration in shared environments.

5 Contributions

This section describes the contributions of this thesis.

In **Paper I**, we present a mixed-reality interface for human-assisted robotic scene inspection in search and rescue operations. The existing algorithms and control schemes for autonomous systems have not matured enough to allow for safe operations in high-risk, unstructured, and dynamic environments. Fully autonomous systems are susceptible to failures in unpredictable environments, and manual control can also be slow and cognitively demanding. Our method introduces human-in-the-loop by using high-level human input through intuitive gaze and gesture recognition to guide the robot. This approach works effectively while using humans' situational awareness with robots spatial and navigational capabilities in unstructured environments with minimal information and no communication loss. This allows the system to be useful for robots in time-sensitive and high-risk environments to carry out operations seamlessly.

In **Paper II and III**, we further develop the system presented in Paper I. In addition to the system, a comparative user study is carried out with two interaction designs within this system: Human-Assisted and System-Assisted. While several techniques in the past mostly focused on speech and gesture to support seamless collaboration, the use of these explicit commands limits their effectiveness in dynamic real-world scenarios. Our work is novel in its design and integration of non-verbal cues such as head-gaze based interaction and real-time foveation-based visual augmentation for task collaboration. Paper II mostly focuses on the experiment setup and preliminary results analyzed from the study. In Paper III detailed analysis and correlations in the data obtained from the gaze analysis is also presented. The study demonstrates that the head-gaze based foveated augmentation improves performance and user experience with significant improvement in task performance and decrease in cognitive workload during collaboration.

In **Paper IV**, we conduct a review of intention recognition in human-robot interaction. While the concept of intention in case of human-human communication is well defined, its meaning and application in the field of robotics is quite varied. We aimed to bridge this gap of fragmented approach to provide a meaning and classification of intention recognition in human robot interaction. On the basis of the state-of-the-art review of the different psychological theories, computational models, methods, terminologies, approaches, and applications we categorize it as low-level, high-level, and robot intentions with further clusters within these categories. In the paper, we discuss about the advancements in the field and identify gaps in the research. We also propose future research directions through discussion of methods in which these gaps could be addressed paving the way towards seamless collaborations between humans and robots.

6 Other Contributions

The following papers are related, but not included in this thesis.

Virtual, augmented, and mixed reality for human-robot interaction (vam-hri) MK Wozniak, M Pascher, B Ikeda, MB Luebbers, **A Jena** Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction, Boulder, Colorado, USA. DOI: 10.1145/3610978.3638158

Framework for assessing situational awareness in Beyond Visual Line of Sight UAV operations RP Maben, **A Jena**, S Reitmann, EA Topp 2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Melbourne, Australia. DOI: 10.1109/HRI61500.2025.10973804

Impact of Gaze-Based Interaction and Augmentation on Human-Robot Collaboration in Critical Tasks A Jena, S Reitmann, EA Topp Accepted as a Late Breaking Report to the 2025 34th IEEE International Conference on Robot and Human Interactive Communication.

References

[AS17] Henny Admoni and Brian Scassellati. “Social eye gaze in human-robot interaction: a review”. In: *Journal of Human-Robot Interaction* 6.1 (2017), pp. 25–63.

[And+21] Olov Andersson et al. “WARA-PS: a research arena for public safety demonstrations and autonomous collaborative rescue robotics experimentation”. In: *Autonomous Intelligent Systems* 1.1 (2021), p. 9.

[Ayd23] Ezgi Uzel Aydınocak. “Robotics systems and healthcare logistics”. In: *Health 4.0 and medical supply chain*. Springer, 2023, pp. 79–96.

[Bar+24] Christoph Bartneck et al. *Human-robot interaction: An introduction*. Cambridge University Press, 2024.

[Bel+24] Valerio Belcamino et al. “Gaze-based intention recognition for human-robot collaboration”. In: *Proceedings of the 2024 International Conference on Advanced Visual Interfaces*. 2024, pp. 1–5.

[Bra87] Michael Bratman. “Intention, plans, and practical reason”. In: (1987).

[Car70] Jaime R Carbonell. “AI in CAI: An artificial-intelligence approach to computer-assisted instruction”. In: *IEEE transactions on man-machine systems* 11.4 (1970), pp. 190–202.

[CP24] Jiho Chang and Jeongho Park. “Design of a Robot System for Surveillance and Anomaly Detection in Industrial Environments”. In: *2024 15th International Conference on Information and Communication Technology Convergence (ICTC)*. IEEE. 2024, pp. 2229–2234.

[CHB07] Jessie YC Chen, Ellen C Haas, and Michael J Barnes. “Human performance issues and user interface design for teleoperated robots”. In: *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 37.6 (2007), pp. 1231–1245.

[CTS21] Meia Chita-Tegmark and Matthias Scheutz. “Assistive robots for the social management of health: a framework for robot design and human–robot interaction research”. In: *International Journal of Social Robotics* 13.2 (2021), pp. 197–217.

[CMN15] Felipe Cid, Luis J Manso, and Pedro Núnez. “A novel multimodal emotion recognition approach for affective human robot interaction”. In: *Proceedings of fine* (2015), pp. 1–9.

[CKK19] Viviane Clay, Peter König, and Sabine Koenig. “Eye tracking in virtual reality”. In: *Journal of eye movement research* 12.1 (2019).

[Dah+23] Abhinav Dahiya et al. “A survey of multi-agent Human–Robot Interaction systems”. In: *Robotics and Autonomous Systems* 161 (2023), p. 104335.

[Dan+20] Ashwin P Dani et al. “Human-in-the-loop robot control for human-robot collaboration: Human intention estimation and safe trajectory tracking control for collaborative tasks”. In: *IEEE Control Systems Magazine* 40.6 (2020), pp. 29–56.

[DDS19] Kadir Alpaslan Demir, Gözde Döven, and Bülent Sezen. “Industry 5.0 and human-robot co-working”. In: *Procedia computer science* 158 (2019), pp. 688–695.

[DBM24] Marike Koch van Den Broek and Thomas B Moeslund. “What is proactive human-robot interaction?-a review of a progressive field and its definitions”. In: *ACM Transactions on Human-Robot Interaction* 13.4 (2024), pp. 1–30.

[Dha20] Amandeep Dhaliwal. “The rise of automation and robotics in warehouse management”. In: *Transforming management using artificial intelligence techniques*. CRC Press, 2020, pp. 63–72.

[FC24] Mohammad Farajtabar and Marie Charbonneau. “The path towards contact-based physical human–robot interaction”. In: *Robotics and Autonomous Systems* 182 (2024), p. 104829.

[Fun+16] Gregory Funke et al. “Which eye tracker is right for your research? performance evaluation of several cost variant eye trackers”. In: *Proceedings of the Human Factors and Ergonomics Society annual meeting*. Vol. 60. 1. SAGE Publications Sage CA: Los Angeles, CA. 2016, pp. 1240–1244.

[GLJ21] Qing Gao, Jinguo Liu, and Zhaojie Ju. “Hand gesture recognition using multimodal data fusion and multiscale parallel convolutional neural network for human–robot interaction”. In: *Expert Systems* 38.5 (2021), e12490.

[Geb+24] Ferran Gebellí et al. “Co-designing explainable robots: a participatory design approach for HRI”. In: *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*. IEEE. 2024, pp. 1564–1570.

[Ghi+24] Bianca Ghinoiu et al. “From Cobot to Dog Intelligent Robot GO2 and Humanoid Intelligent Robots G1/H1. Educational and Economic Applications”. In: *International Conference on Innovation of Emerging Information and Communication Technology*. Springer. 2024, pp. 125–135.

[Goo22] Google. *Mediapipe*. 2022. URL: `https://google.github.io/mediapipe/`.

[Hua+15a] Chien-Ming Huang et al. “Using gaze patterns to predict task intent in collaboration”. In: *Frontiers in psychology* 6 (2015), p. 1049.

[JA15] Shu Jiang and Ronald C Arkin. “Mixed-initiative human-robot interaction: definition, taxonomy, and survey”. In: *2015 IEEE International conference on systems, man, and cybernetics*. IEEE. 2015, pp. 954–961.

[Kor+97] David Kortenkamp et al. “Traded control with autonomous robots as mixed initiative interaction”. In: *AAAI Symposium on Mixed Initiative Interaction*. Vol. 97. 4. 1997, pp. 89–94.

[LU23] LU. *Ethical review — staff.lu.se*. `https://www.staff.lu.se/research-and-education/research-support/research-ethics-and-animal-testing-ethics/ethical-review`. [Accessed 10-02-2024]. 2023.

[LLT22] Yidi Li, Hong Liu, and Hao Tang. “Multi-modal perception attention network with self-supervised learning for audio-visual speaker tracking”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 2. 2022, pp. 1456–1463.

[LW18] Hongyi Liu and Lihui Wang. “Gesture recognition for human-robot collaboration: A review”. In: *International Journal of Industrial Ergonomics* 68 (2018), pp. 355–367.

[Liu+22] Yiming Liu et al. “The role of haptic communication in dyadic collaborative object manipulation tasks”. In: *arXiv preprint arXiv:2203.01287* (2022).

[LPT22] Birgit Lugrin, Catherine Pelachaud, and David Traum. *The Handbook on Socially Interactive Agents: 20 Years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 2: Interactivity, Platforms, Application*. ACM, 2022.

[Mar+20] Jeremy A Marvel et al. “Towards effective interface designs for collaborative HRI in manufacturing: Metrics and measures”. In: *ACM Transactions on Human-Robot Interaction (THRI)* 9.4 (2020), pp. 1–55.

[McN92] David McNeill. *Hand and mind: What gestures reveal about thought*. University of Chicago press, 1992.

[Mil56] George A Miller. “The magical number seven, plus or minus two: Some limits on our capacity for processing information.” In: *Psychological review* 63.2 (1956).

[Neh+05] Chrystopher L Nehaniv et al. “A methodological approach relating the classification of gesture to identification of human intent in the context of human-robot interaction”. In: *ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication, 2005.* IEEE. 2005, pp. 371–377.

[Net+19] Pedro Neto et al. “Gesture-based human-robot interaction for human assistance in manufacturing”. In: *The International Journal of Advanced Manufacturing Technology* 101.1 (2019), pp. 119–135.

[Omo+08] Takashi Omori et al. “Computational modeling of human-robot interaction based on active intention estimation”. In: *Neural Information Processing: 14th International Conference, ICONIP 2007, Kitakyushu, Japan, November 13-16, 2007, Revised Selected Papers, Part II 14*. Springer. 2008, pp. 185–192.

[Pac08] Elisabeth Pacherie. “The phenomenology of action: A conceptual framework”. In: *Cognition* 107.1 (2008), pp. 179–217.

[PAA13] Amit Kumar Pandey, Muhammad Ali, and Rachid Alami. “Towards a task-aware proactive sociable robot based on multi-state perspective-taking”. In: *International Journal of Social Robotics* 5.2 (2013), pp. 215–236.

[Per22] Michele Persiani. “Expressing and recognizing intentions”. PhD thesis. Umeå University, 2022.

[Rad+15] Nicolaus A Radford et al. “Valkyrie: Nasa’s first bipedal humanoid robot”. In: *Journal of Field Robotics* 32.3 (2015), pp. 397–419.

[Rob22] Mobile Industrial Robot. *MIR*. 2022.

[Rob08] Universal Robots. *UR*. 2008. URL: `https://www.universal-robots.com/products/ur5-robot/`.

[RSR16] Alessandra Rossi, Mariacarla Staffa, and Silvia Rossi. “Supervisory control of multiple robots through group communication”. In: *IEEE Transactions on Cognitive and Developmental Systems* 9.1 (2016), pp. 56–67.

[Sch] Schunk. *Schunk gripper*. URL: `https://schunk.partcommunity.com/3d-cad-models/gripping-systems-schunk?info=schunk%2Fgreifsysteme_neu&cwid=6782`.

[Su+23] Yun-Peng Su et al. "Integrating virtual, mixed, and augmented reality into remote robotic applications: a brief review of extended reality-enhanced robotic systems for intuitive telemanipulation and telemanufacturing tasks in hazardous conditions". In: *Applied Sciences* 13.22 (2023).

[Tec22] Unity Technologies. *ROS TCP Connector*. 2022. URL: `https://github.com/Unity-Technologies/ROS-TCP-Connector`.

[TBG23] Matteo Terreran, Leonardo Barcellona, and Stefano Ghidoni. "A general skeleton-based action and gesture recognition framework for human–robot collaboration". In: *Robotics and Autonomous Systems* 170 (2023), p. 104523.

[TMK16] Carsten Thomas, Bjoern Matthias, and Bernd Kuhlenkötter. "Human-robot collaboration–new applications in industrial robotics". In: *International conference on competitive manufacturing*. 2016, pp. 293–299.

[Top17] Elin Anna Topp. "Interaction patterns in human augmented mapping". In: *Advanced Robotics* 31.5 (2017), pp. 258–267.

[Tri+20] Eleftherios Triantafyllidis et al. "Study of multimodal interfaces and the improvements on teleoperation". In: *IEEE Access* 8 (2020), pp. 78213–78227.

[Uni22] Unity Technologies. *Unity*. Version 2022.3.48f1. 2022.

[VSU97] Boris Velichkovsky, Andreas Sprenger, and Pieter Unema. "Towards gaze-mediated interaction: Collecting solutions of the "Midas touch problem"". In: *Human-Computer Interaction INTERACT'97: IFIP TC13 International Conference on Human-Computer Interaction, 14th–18th July 1997, Sydney, Australia*. Springer. 1997.

[Vic21] Federico Vicentini. "Collaborative robotics: a survey". In: *Journal of Mechanical Design* 143.4 (2021), p. 040802.

[WP24] WARA-PS. *GitHub - wara-ps/cesium-unreal — github.com*. `https://github.com/wara-ps/cesium-unreal`. [Accessed 10-02-2024]. 2024.

[Wal+23] Michael Walker et al. "Virtual, augmented, and mixed reality for human-robot interaction: A survey and virtual design element taxonomy". In: *ACM Transactions on Human-Robot Interaction* 12.4 (2023), pp. 1–39.

[Wan+24c] Weihan Wang et al. "Cogvlm: Visual expert for pretrained language models". In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 121475–121499.

[WP20] Murphy Wonsick and Taskin Padir. “A systematic review of virtual reality interfaces for controlling and interacting with robots”. In: *Applied Sciences* 10.24 (2020), p. 9051.

[Xia+14] Yang Xiao et al. “Human–robot interaction by understanding upper body gestures”. In: *Presence* 23.2 (2014), pp. 133–154.

[Xu+21] Xun Xu et al. “Industry 4.0 and Industry 5.0—Inception, conception and perception”. In: *Journal of manufacturing systems* 61 (2021).

[YFH18] Deng Yongda, Li Fang, and Xin Huang. “Research on multimodal human-robot interaction based on speech and gesture”. In: *Computers & Electrical Engineering* 72 (2018), pp. 443–454.

[Zla+18] Athanasia Zlatintsi et al. “Multimodal signal processing and learning aspects of human-robot interaction for an assistive bathing robot”. In: *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE. 2018, pp. 3171–3175.

INCLUDED PAPERS

Chaos to Control: Human Assisted Scene Inspection

1 Abstract

We are working towards a mixed reality-based human-robot collaboration interface using gaze and gesture as methods of communicating intent in a search and rescue scenario to optimize the operation. The lack of mature algorithms and control schemes for autonomous systems makes it still difficult for them to operate safely in high-risk environments. We are approaching the problem through symbiosis while utilizing humans' intuition of the environment and robots' capability to travel through unknown environments for optimal performance in a given time.

2 Introduction

Autonomous systems have come a long way in what they are capable of achieving. This has allowed robots to permeate various spheres of society to fulfil small and large needs ranging from the mundane to the critical.

In disaster response scenarios, time is of crucial importance. The goal is to potentially carry out a rapid and recursive search, identification, and response to

Ayesha Jena, and Elin Anna Topp. "Chaos to control: human assisted scene inspection." In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction. 2023. Stockholm, Sweden*
DOI: 10.1145/3568294.3580133.

minimize further loss of life and property. However, caution must be maintained while carrying out these tasks so that additional damage or risk does not occur. Robots can prove to be of crucial assistance in this regard by directly helping the affected victims, providing support to structures, or aiding in additional support activities [SKK08].

Additionally, robots can be of help in aspects of Search and Rescue (SAR) like concentrated search, wider reconnaissance and mapping, rubble removal, skeletal inspection of the damaged structures, in-situ medical aiding or assistance, and expanding network ranges in the area of operations using beacons, among other tasks [SKK08]. In order to carry out such operations autonomously, robots need to be equipped with a number of functionalities in navigation, perception, manipulation and reasoning. However, integrating the above functionalities in a robot comes with its own set of challenges such as power supply, computing resources, space and cost. Furthermore, the algorithms and control schemes for autonomous systems have not matured enough to allow for safe operation in such high-risk environments. A major disadvantage here is the lack of bi-directional communication between the human and the robot for effective collaboration [Wal+22]. We assume that the challenges faced by an autonomous system can be mitigated by falling back to a semi-autonomous one. This allows the system to continue functioning while relying on human expertise to address the issue at hand. Thus, in this paper, we present a human-in-the-loop mixed-reality based robot control system for the teleoperation of a mobile robot in unknown scenarios to facilitate a natural and intuitive user interface for SAR operations. The components of the system involve: (i) a virtual system with reconfigurable screens, (ii) real-time visual reproduction of the target environment, and (iii) use of gaze and gesture for robot control.

3 Related Work

Teleoperation in virtual, augmented, and mixed reality for human-robot interactions makes use of compatible hardware (Video displays, Tablets, Projectors, Cave Automatic Virtual Environments (CAVEs), Head Mounted Displays (HMDs) [Wal+22]) and graphics engines (Unreal Engine, Unity3D, etc. [Nac+19]) to control an agent from a remote location. Besides legacy input devices such as controllers and joysticks, nonverbal cues from humans can also be used as input to directly control the robot. Lipton *et al.* have defined three mapping models to categorize various teleoperation interfaces, namely: direct, cyber-physical, and homunculus [LFR17]. These categories define the amount of information being exchanged between the user and the robot's space during teleoperation. Hirschmanner *et al.* used a Leap Motion Controller to get the teleoperator's arm pose and used algorithms to directly calculate the robot's joint angles from the same [Hir+19]. The entire environment was presented virtually to the user with

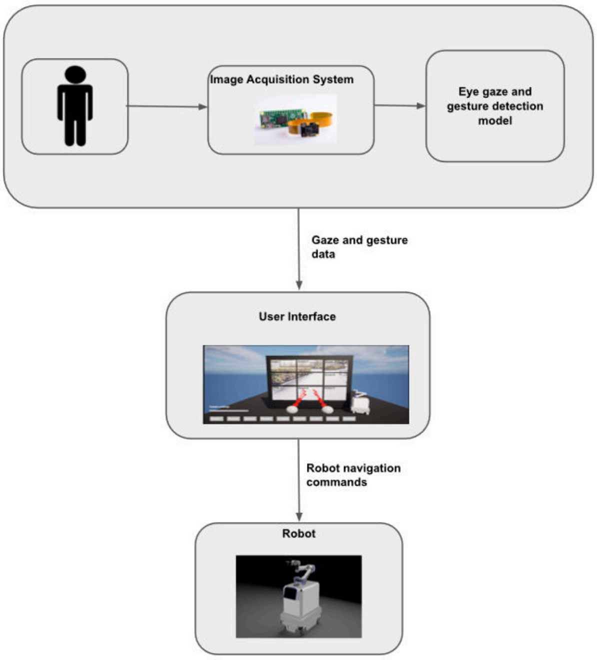


Figure 1: User interface in a top-down approach

the help of an Oculus Rift VR headset. Gaurav *et al.* proposed a machine-learning approach to estimate correspondence for robotic teleoperation from the operator's pose [Gau+19]. A Microsoft Kinect depth camera is used to perceive the robot's workspace while an HTC Vive is used to visualize the workspace and provide 3D control to the operator. Sun *et al.* developed an interaction method based on the homunculus model to individually control the position and orientation of the robot independent of each other [Sun+20]. This was made possible by using a virtual sphere as the controller with which the operator interacted to control the two robot modes. Comparing this to the work done by Whitney *et al.* [Whi+20], where both the robot's position and orientation are controlled simultaneously, the method developed by Sun *et al.* achieved a 93.75% success rate compared to 25% success rate of the first mentioned system. Lager *et al.* have compared different graphical user interfaces and found that a VR setting has benefits over traditional Graphical User Interface regarding situation awareness [LTM18, LTM19]. Based on these findings, they further proposed a VR-based Graphical User Interface to allow a remote operator to support an unmanned vessel performing GPS-free navigation [LTM20].

4 Methods

Our proposed system can be categorized under the homunculus model as mentioned in the previous section [LFR17], and is shown in Figure 1. The gaze and gesture inputs are decoupled to control the orientation and motion of the robot in real-time through a virtual space. The mapped user inputs from the virtual space are sent to the robot in the real world via ROS.

4.1 User Input

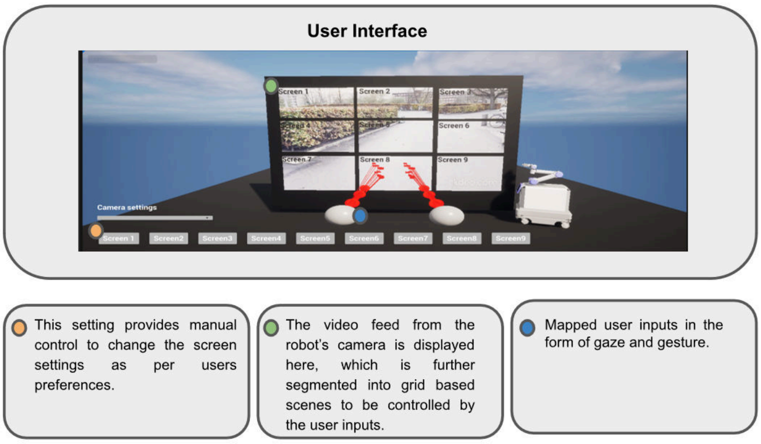
In order to detect the user's gesture and gaze, we use an RGB camera to capture real-time video input and use OpenCV to feed the frames to pre-trained models from mediapipe [Goo22]. These models allow us to detect 21 3D landmarks across the user's hand and 468 3D landmarks across the user's face for highly accurate tracking. The 21 landmarks of the hand and the rotation values of the gaze vector are calculated from the face landmarks and combined into a string and encoded in the form of bytes. These bytes are then sent across the network in the form of UDP messages to the virtual system.

4.2 Network

To allow efficient and real-time data transfer between the user input system and the virtual system, a UDP server and client connection is created on the local network. We are using UDP messaging instead of the standard TCP-IP in order to reduce latency between the delivered packets. The UDP connection is established using the sockets library in python and UDP-Unreal plugin [get22] in the virtual system.

4.3 Virtual System

The virtual system shown in Figure 2 works as a virtual space which receives visual input from the robot's camera and control inputs from the user. The system also ensures reliable delivery of mapped user input to the robot for precise control. The virtual system is developed using Unreal Engine in order to leverage the software's physics system and rendering capabilities. Within the virtual space, the bytes from the user input are parsed back into relevant transform values which are then used to control a virtual representation of the gaze and gesture. Additionally, 9 screens in the form of a 3x3 grid are presented to the user in the virtual space. The video feed from the robot's camera is segmented into these 9 screens. These individual screens can be turned on or off based on the user's preference. A ray cast from the gaze vector is used to select one of these 9 screens and another ray cast from the gesture is used to confirm the choice. These screens are mapped to different robot navigation commands. Once the user has confirmed their choice, the virtual system relays this information to the robot via ROS.



● This setting provides manual control to change the screen settings as per users preferences.

● The video feed from the robot's camera is displayed here, which is further segmented into grid based scenes to be controlled by the user inputs.

● Mapped user inputs in the form of gaze and gesture.

Figure 2: Egocentric view of the user interface. The colour is used to indicate the system components and functions.

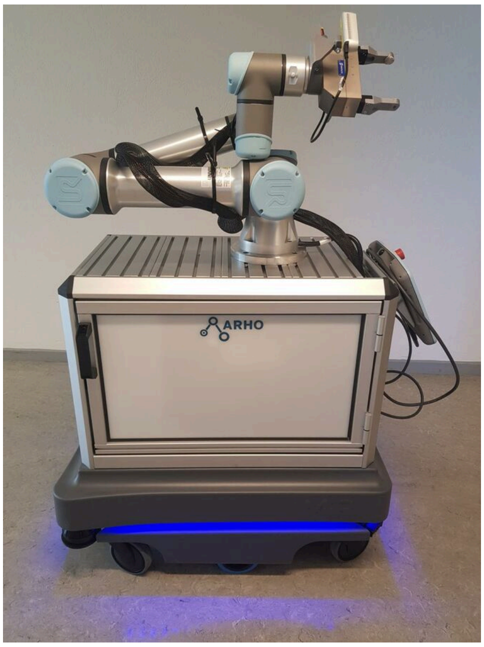


Figure 3: Robot used for the purpose of the system design and operation. Universal Robot arm UR5e mounted on a MiR200.

4.4 Robot

The mobile robot used for this setup is a combination of a Mobile robot platform MiR200 (Mobile Industrial Robots) [Rob22], a Universal Robots collaborative arm UR5e [Rob08], attached with an Intel Realsense D435 RGBD camera [Int18] on the plate of Schunk WSG-50 parallel gripper [Sch], as shown in the Figure 3. MiR200 is an autonomous mobile robot with a load-carrying capacity of 200 kg. MiR200 comes with a web-based interface, which can be accessed by a browser on any device. The Universal Robot arm UR5e, is a 6 DOF collaborative robot arm with a 5 kg payload capacity. Intel Realsense D435 is an RGBD camera, with an effective depth range of 0.1-0.8m. This camera input is used as a real-time feed into the screens in virtual space.

5 Technical Results

The core components are the gaze vector from the face detection model, the hand detection model and the user input information to the virtual space. The facial detection model used in the system has an average precision of 98.61% with an inference time of 0.6 ms on a high-end device [Baz+19], and the hand detection model in the system achieves an average precision of 95.7% [Goo22]. Further, the data communication latency between user input and virtual space is 10 ms within the same system. This value is the instantaneous maximum latency value and was obtained by comparing the clocks of the user input system and the virtual system over a period of time. LAN is used for transmitting the sensor data. This is done keeping in mind the sensors used. For UDP, there is no insurance in place for packet loss happening. However, this is also of no concern for this version since the data is collected and transmitted at a frequency of more than 60 Hz which does not affect the performance even if there is some amount of packet loss.

6 Discussion and Future Work

We proposed a user interface system to assist operators in real-time to control a robot using gaze and gesture signals. The preliminary results from testing the system show that the models are able to accurately identify the presence of faces and hands in the input images, while the low inference time suggests that the models can process the images quickly. The network latency value for information exchange puts the system ahead of the required threshold of 80 ms, above which the user experience gets affected [Cho+12]. This is useful in tasks requiring speed and accuracy.

The current system is designed keeping in mind that using human intuition to guide a mobile robot in SAR scenarios could lead to faster search and locate times, as human intuition is often based on experience and pattern recognition, which can be valuable in quickly identifying potential areas of interest. We assume that a limited field of view can help to maximize rescue efforts in chaotic situations by allowing rescuers to focus on specific tasks or areas, rather than becoming overwhelmed by a large and chaotic environment. Additionally, using a limited field of view while controlling a robot towards a selected goal could lead to quicker goal execution. By focusing on a limited area, the operator can more easily identify potential obstacles and plan a path to the goal without being overwhelmed by information from a wider field of view. This can allow the operator to make quicker and more efficient decisions. The future work includes evaluating the user interface with the robot to carry out user experiments in order to confirm these two assumptions.

Acknowledgements

This work was funded by ELLIIT - Excellence Center at Linköping-Lund on Information Technology.

References

[Baz+19] Valentin Bazarevsky et al. *BlazeFace: Sub-millisecond Neural Face Detection on Mobile GPUs*. 2019.

[Cho+12] Sharon Choy et al. “The brewing storm in cloud gaming: A measurement study on cloud to end-user latency”. In: *2012 11th Annual Workshop on Network and Systems Support for Games (NetGames)*. IEEE. 2012, pp. 1–6.

[Gau+19] Sanket Gaurav et al. “Deep correspondence learning for effective robotic teleoperation using virtual reality”. In: *2019 IEEE-RAS 19th International Conference on Humanoid Robots (Humanoids)*. IEEE. 2019, pp. 477–483.

[Goo22] Google. *Mediapipe*. 2022. URL: `https://google.github.io/mediapipe/`.

[Hir+19] Matthias Hirschmanner et al. “Virtual reality teleoperation of a humanoid robot using markerless human upper body pose imitation”. In: *2019 IEEE-RAS 19th International Conference on Humanoid Robots (Humanoids)*. IEEE. 2019, pp. 259–265.

[Int18] Intel. *Intel depth camera*. 2018. URL: `https://www.intelrealsense.com/depth-camera-d435/`.

[LTM18] Mårten Lager, Elin A Topp, and Jacek Malec. "Remote Operation of Unmanned Surface Vessel through Virtual Reality-a low cognitive load approach". In: *Proceedings of the 1st International Workshop on Virtual, Augmented, and Mixed Reality for HRI (VAM-HRI)*. 2018.

[LTM19] Mårten Lager, Elin A Topp, and Jacek Malec. "Remote supervision of an unmanned surface vessel-a comparison of interfaces". In: *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE. 2019, pp. 546–547.

[LTM20] Mårten Lager, Elin A. Topp, and Jacek Malec. "VR Teleoperation to support a GPS-free Positioning System in a Marine Environment". In: *International Journal on Marine Navigation and Safety of Sea Transportation* 14.4 (2020), pp. 789–798.

[LFR17] Jeffrey I Lipton, Aidan J Fay, and Daniela Rus. "Baxter's homunculus: Virtual reality spaces for teleoperation in manufacturing". In: *IEEE Robotics and Automation Letters* 3.1 (2017), pp. 179–186.

[Nac+19] Abdeldjallil Naceri et al. "Towards a virtual reality interface for remote robotic teleoperation". In: *2019 19th International Conference on Advanced Robotics (ICAR)*. IEEE. 2019, pp. 284–289.

[Rob22] Mobile Industrial Robot. *MIR*. 2022.

[Rob08] Universal Robots. *UR*. 2008. URL: `https://www.universal-robots.com/products/ur5-robot/`.

[Sch] Schunk. *Schunk gripper*. URL: `https://schunk.partcommunity.com/3d-cad-models/gripping-systems-schunk?info=schunk%2Fgreifsysteme_neu&cwid=6782`.

[SKK08] Bruno Siciliano, Oussama Khatib, and Torsten Kröger. *Springer handbook of robotics*. Vol. 200. Springer, 2008.

[Sun+20] Da Sun et al. "A new mixed-reality-based teleoperation system for telepresence and maneuverability enhancement". In: *IEEE Transactions on Human-Machine Systems* 50.1 (2020), pp. 55–67.

[Wal+22] Michael Walker et al. "Virtual, augmented, and mixed reality for human-robot interaction: A survey and virtual design element taxonomy". In: *arXiv preprint arXiv:2202.11249* (2022).

[Whi+20] David Whitney et al. "Comparing robot grasping teleoperation across desktop and virtual reality with ROS reality". In: *Robotics Research*. Springer, 2020, pp. 335–350.

[get22] getnamo. *UDP-Unreal*. 2022. URL: `https://github.com/getnamo/UDP-Unreal`.

TOWARDS UNDERSTANDING THE ROLE OF HUMANS IN COLLABORATIVE TASKS

1 Abstract

This paper explores the dynamics of human-robot collaboration through a comparative study of human-assisted and system-assisted approaches in a search and rescue application. Leveraging virtual environments and mixed-reality interfaces, the study evaluates task performance, workload, usability, and subjective experiences of participants. Results indicate that the system-assisted approach significantly improves task completion time and accuracy in identifying critical elements, and reduces perceived workload compared to human-assisted methods. Subjective assessments reveal valuable insights into user preferences and challenges, informing recommendations for system refinement and protocol development. Findings highlight the potential of human collaboration in enhancing operational effectiveness and promoting seamless collaboration between humans and robots in cluttered and high-risk environments. Interactions aimed at synchronizing goals, task states, and actions can be facilitated through virtual, augmented, and mixed-reality environments providing an intuitive platform for understanding interaction dynamics.

Ayesha Jena, and Elin Anna Topp. “Towards Understanding the Role of Humans in Collaborative Tasks”. In *7th International Workshop on Virtual, Augmented, and Mixed-Reality for Human-Robot Interactions at HRI 2024. Boulder, CO, USA*

2 Introduction

In a time marked by rapid technological advancements, robotics continues to evolve and permeate various facets of society ranging from manufacturing and healthcare to disaster response and space exploration. The synergy between humans and robots holds immense promise across these diverse domains as this paradigm shift transcends the traditional notion of robotics where there was little to no interaction between humans and robots. With the increased integration of robots into social settings, collaborative robots are becoming active participants in our everyday lives forging new frontiers in Human-Robot Interaction (HRI) that have the potential to change the way we interact with the world around us. As we navigate this era of increased collaboration between humans and machines, exploring the dynamics, challenges, and opportunities inherent in this symbiotic relationship becomes important.

While collaborative robots (cobots) are designed to assist humans, they still operate within highly predefined parameters that are constraining. For example, a robot will mostly stop or slow down when working in the periphery of humans, thus, limiting the impact of such collaboration [Nat+23]. In various sectors, fully autonomous cobots function within rigid frameworks, carrying out tasks alongside humans rather than engaging in genuine teamwork. As a result, while they enhance certain aspects of productivity and efficiency, their potential for seamless human-robot collaboration remains largely untapped [Nat+23]. In the case of human-human collaboration, communication is a crucial aspect that leads to successful teamwork and goal completion. Similarly, in human-robot teams, it is essential to have information sharing based on the human supervisory role and the robot's autonomy level. This can be achieved through interactions to synchronize goals, task states, and actions [LHZ97].

Virtual environments and simulations offer a valuable tool for comprehending the dynamics of interaction. They provide an intuitive platform for understanding the mechanics of how interactions would unfold. By leveraging virtual environments, adaptability can be significantly enhanced to cater to the specific requirements of the task space, the user involved, and the capabilities of the robot. These tools also provide the opportunity to evaluate interactions with virtual robots that are restricted by monetary and/or safety concerns in the real world [Wal+23].

Building upon our previous work [JT23], this paper presents a user study to compare *human-assisted* and *system-assisted* methods for human-robot collaboration. Figure 1 gives an overall understanding of the steps involved. The study investigates two collaboration frameworks: one where humans act as teleoperators and scene inspectors for robots (human-assisted, HA), and another where systems suggestions are taken as inputs, with robot teleoperation while humans make final decisions on areas of interest (system-assisted, SA). Objective and subjective measures are analyzed to elucidate factors influencing the development of intelligent and collaborative robots.

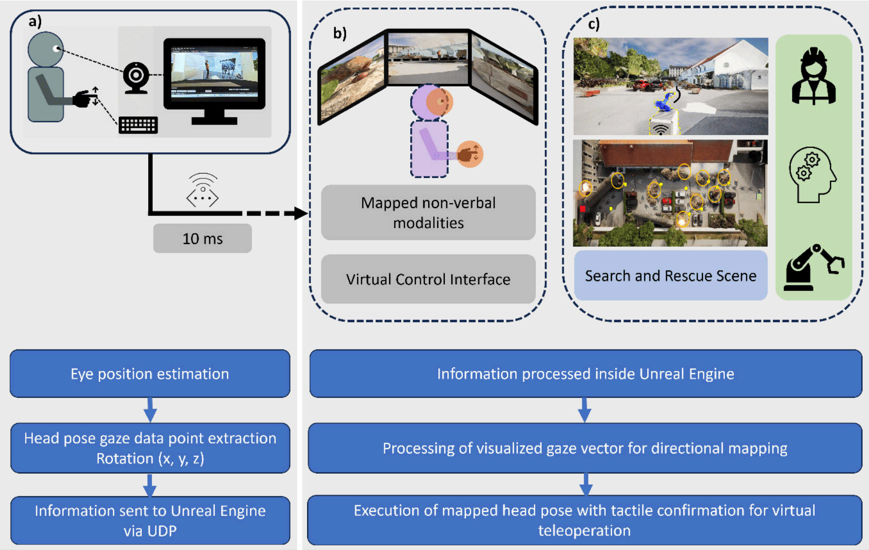

a)
10 ms
b)
Mapped non-verbal modalities
Virtual Control Interface
c)
Search and Rescue Scene
Eye position estimation
Head pose gaze data point extraction Rotation (x, y, z)
Information sent to Unreal Engine via UDP
Information processed inside Unreal Engine
Processing of visualized gaze vector for directional mapping
Execution of mapped head pose with tactile confirmation for virtual teleoperation

Figure 1: An image of the different aspects of mixed-reality based human-in-the-loop robot control system. Here, a) the human operator's gaze and gesture modalities are taken as input using image recognition and tracking, b) the tracked modalities are mapped within the virtual control interface which in turn controls and monitors the agent in the system in human-assisted and system-assisted scenarios respectively, and c) search and rescue scene designed for assessing humans' intuition while performing different tasks.

Two research questions guide this investigation:

- "What level of human-robot collaboration is better at performing search operations in an unknown environment?"
- "How does a limited field of view affect goal execution?"

These questions aim to shed light on the effectiveness of human guidance in various tasks and the impact of cognitive load on goal execution within limited visual contexts.

3 Related Work

3.1 Human-Robot Interaction (HRI) Frameworks and Communication Modalities

Human-robot interaction (HRI) encompasses a spectrum of interaction stages, as categorized by Onnasch et al. [OR21], including bounded autonomy, teleopera-

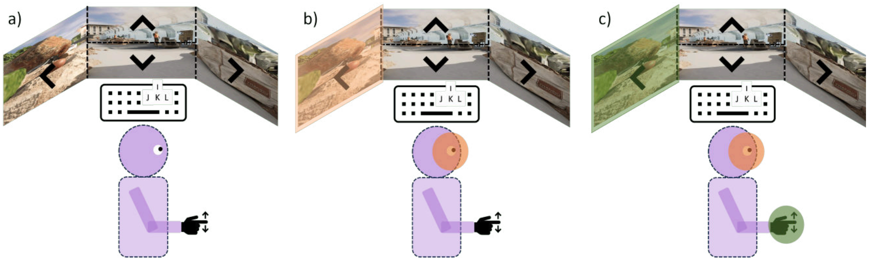


Figure 2: Mapping of head pose and tactile confirmation

tion, supervised autonomy, adaptive autonomy, and virtual symbiosis. However, the practical application of these stages often involves smooth transitions based on human roles and task demands, highlighting the importance of effective communication channels between humans and robots throughout these interactions. Researchers have investigated diverse communication modalities within HRI, surrounding two-way dialogue, natural language, multi-modal communication, and visual messages. While these modalities present rich interaction potentials, they often elevate cognitive workload and present hurdles to situational awareness. In response, discrete and sparse communication channels aimed at preserving human interpretability while strengthening decision-making precision have been suggested [Nat+23]. Gaze, identified as a natural means of interaction, has been leveraged in HRI, either as a primary input signal or in conjunction with other modalities [Plo+22]. However, gaze-only interfaces encounter challenges like the “Midas touch problem”, where deciding when to select input becomes intricate due to the constant nature of gaze [VSU97]. Consequently, separate confirmation mechanisms are necessitated to address these issues [SG19]. Techniques such as Eye & Head Dwell, Eye & Head Convergence, and Eye & Head Pointer have been investigated to enhance stability and efficacy in gaze-based interactions [SG19]. Moreover, head-supported gaze offers greater stability compared to gaze-only approaches [SG19]. Considering that humans utilize their bodies to attend to their environment or convey their attention to others, nonverbal cues like pointing or directing their head and eyes toward objects of interest emerge as natural candidates for further exploration in target selection and manipulation tasks within Extended Reality (XR) contexts [Plo+22].

3.2 Teleoperation Interfaces and Multi-modal Interaction Techniques

Teleoperation offers a bridge between human instinct and robotic capabilities [Zha+19b]. Gesture-based teleoperation systems, utilizing devices like joysticks or motion-tracking devices, enable intuitive control methods for opera-

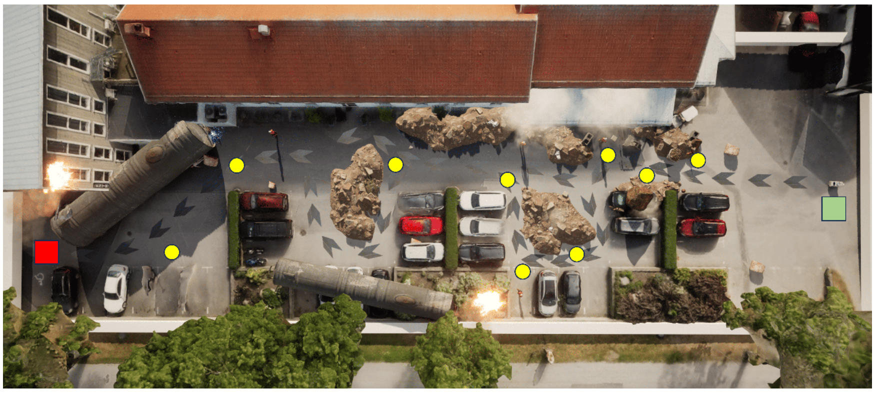


Figure 3: Start (green box), goal (red box), and interaction target points (yellow box) of the robot during experiment in SA scenario

tors [Zha+19b]. Immersive VR teleoperation interfaces replicate natural human motions, although they introduce complexities such as the need for specialized equipment [TRW18]. This, when combined with head-supported gaze, can generate mapped motions in the interface [JT23]. Multi-modal interfaces play a crucial role in reducing cognitive workload and improving task performance in teleoperation scenarios [Tri+20]. These interfaces synchronize multiple modalities to enhance user immersion and awareness, contributing to more effective human-robot collaboration [Tri+20]. In the context of this work where humans need to perform faster searches, foveation methods also offer an interesting way to facilitate search mechanisms in cluttered and cognitively demanding environments [AE17].

In summary, research in the field has explored various communication modalities, teleoperation interfaces, multi-modal interaction techniques, and foveation technologies to enhance human-robot collaboration across different interaction frameworks. These studies provide valuable insights for comparing the effectiveness of HA and SA approaches in HRI scenarios, as investigated in our current research.

4 Experiment

4.1 Aim

In this study, we aim to investigate the effectiveness and efficiency of a mixed reality-based system for improved human-robot collaboration, along with the underlying methods for user support in search and rescue situations involving otherwise autonomous systems.

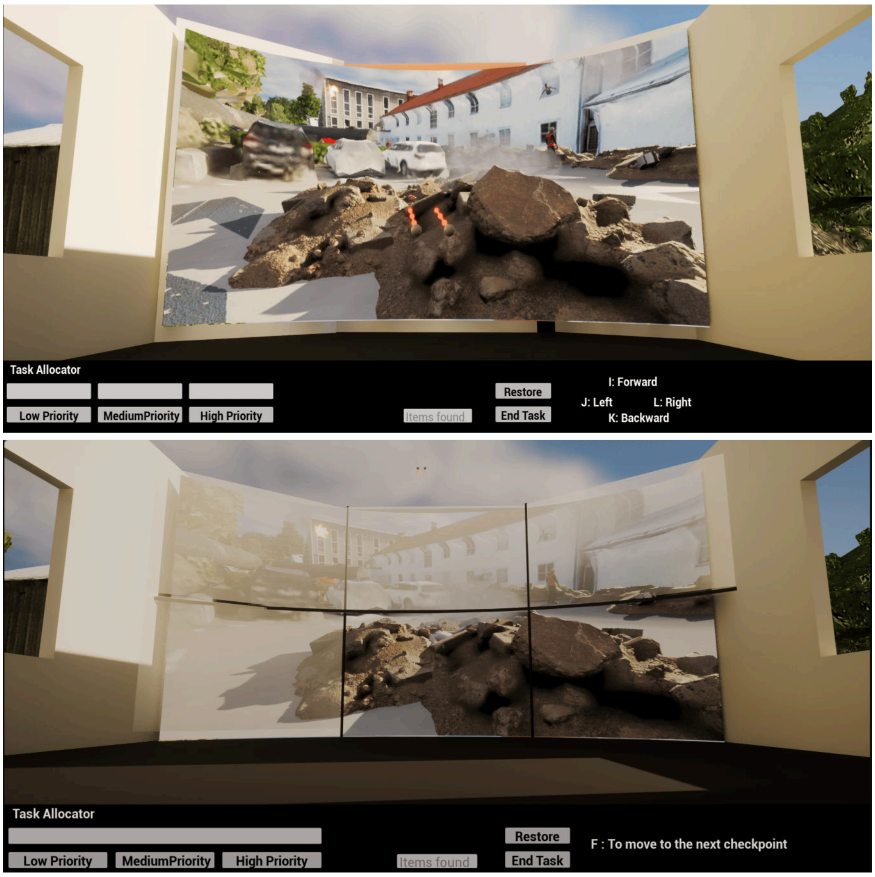


Figure 4: FOVs of the participant in both the scenarios - HA scenario (top) and SA scenario (bottom).

4.2 Test-bed environment

To test in a search and rescue scenario we used an already existing 3D map [WP24] as the virtual environment and made modifications to create a simulated post-disaster scenario. Based on the interface design, similar 3D maps could be integrated at any stage to test other applications.

4.3 Experimental Design and Workflow Steps

Participants engaged in two sequential scenarios presented in random order. The two scenarios were designed with varying degrees of human intervention and decision-making. In the HA scenario, participants directed a simulated robot using eye gaze (left, right, up, down) and corresponding keyboard inputs (J, L, I, K) to traverse a search-and-rescue scene (figure 2), aiming to reach the end of a parking lot. This mode granted participants heightened control over task execution. This can be seen from the top part of figure 4.

In contrast, the SA scenario employed a Wizard-of-Oz technique to cluster identified areas of interest (AOIs) according to assigned importance levels (Low, Medium, High) for objects and humans in the scene. These AOIs could be seen as yellow point marks in figure 3, while green and red indicate start and end locations respectively. The robot autonomously navigated to these AOIs using foveation techniques, thus, reducing the cognitive load. Participants acted as final decision-makers, specifying their priority levels through the interface, utilizing similar importance categories (Low, Medium, High). This can be seen from the bottom part of figure 4.

4.4 Task

The overall goal of the task was to assess the scene and provide information regarding AOIs in a post-disaster scenario. The task for the participants was to count the points of interest they encountered and put corresponding priority markers for each of them. Participants were also provided with a small reference sheet before the experiment started to give a general idea of the importance of various objects in the scene.

4.5 Participant Data, Recorded Information and Ethics

Ethical considerations were taken into account before the experiment. As per the guidelines mentioned in [LU23] the experiment did not require an ethical review process from a committee. Participant demographics and recorded data, including log files, are anonymized, stored, and processed in line with the regulations of the university.

4.6 After Experiment: Analysis

We evaluated task completion time, task accuracy, and the number of identified humans to compare priorities between scenarios. These evaluations were complemented by workload analysis using NASA Task Load Index [HS88], system usability through the System Usability Scale [Bro96], and subjective questionnaires to draw conclusive insights.

5 Results

5.1 Participants

The total sample recruited for the user study consisted of 18 participants. There were 13 males, 4 females, and 1 Other with a mean age of 31.29 years (SD = 9.78) excluding 1 Other participant who refused to report their age. Out of the 18 participants, 7 had some level of vision impairment mostly corrected with eyeglasses. Since the task involved a search and rescue scenario, the use of multiple interfaces, and virtual scenarios, we were also concerned about the experience of the participants in those aspects. Only 4 participants had experience in providing disaster relief. Participants also reported varying levels of experience across different domains: with robots (M = 2.72, SD = 1.7), with any form of virtual, augmented, or mixed reality system (M = 1.83, SD = 1.79), and with using controllers (M = 3.83, SD = 1.2). To eliminate any order and learning effects, half of the participants (N = 9) started the study with the HA scenario while the other half started with the SA scenario.

5.2 Task Timing

This is the first objective performance metric that we use to measure the performance of the participants in the two scenarios. Participants took an average time of 11m 17s (SD = 4m 34s) to complete the HA scenario and an average time of 3m 54s (SD = 53s) to complete the SA scenario. A paired two-sample t-test for two-tail significance of means shows that participants performed significantly better in the SA scenario ($p < 0.001$). This can be seen in figure 5.

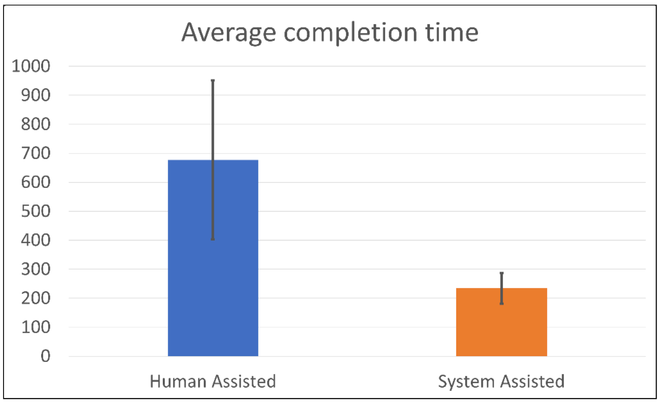


Figure 5: Average completion time in seconds of participants in HA and SA scenarios

5.3 Task Performance

The participants provided different priorities at different locations in the scenarios. Based on these, important locations and total instances were calculated for the identification of correct instances of the number of trapped humans present in the scene. There were 3 trapped humans in each scenario for the participants to locate during the task. In case of the HA scenario, out of N = 54 total instances, 20 (M = 1.11, SD = 0.76) were successfully identified. In case of the SA scenario, 47 (M = 2.6, SD = 0.7) instances were successfully identified. A paired two-sample t-test for two-tail significance of means shows that participants performed significantly better in the SA scenario ($p < 0.001$).

5.4 Workload

The participants answered the NASA TLX questionnaire after each scenario which helped measure the perceived workload for each scenario. The results show a high mean workload of 53.85 (SD = 18.07) in case of the HA scenario as compared to a mean workload of 33.41 (SD = 15.24) in the SA scenario with a paired two-sample t-test for two-tail significance of means showing statistical significance ($p < 0.001$). This can be seen in figure 6.

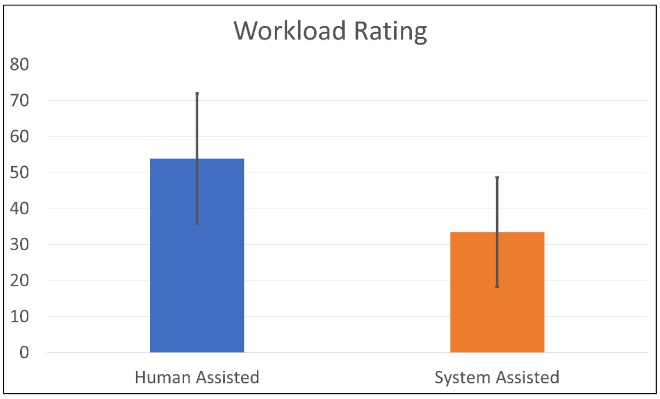


Figure 6: Average workload experienced by participants in HA and SA scenarios

5.5 System Usability

The system usability scale is a quick way to ascertain the usability of systems under scrutiny. Similar to the TLX questionnaire earlier, the participants answered 10 questions from the system usability questionnaire using a Likert scale (1 to 5) to indicate strong disagreement on the leftmost end (1) to strong agreement on the rightmost end (5). After calculating a single value from the responses to all 10 questions, the scores of the participants were averaged to arrive at the presented results. The participants reported an average usability of 58.61 (SD = 14.8) for the HA scenario and an average usability of 80.14 (SD = 16.3) for the SA scenario. Since the system usability score by itself does not represent a percentage, it needs to be normalized and converted to percentile to be interpreted correctly. According to [BKM08], a system usability score of 68 marks the 50th percentile. A paired two-sample t-test for two-tail significance of means shows that participants preferred the SA scenario ($p < 0.001$). This can be seen in figure 7.

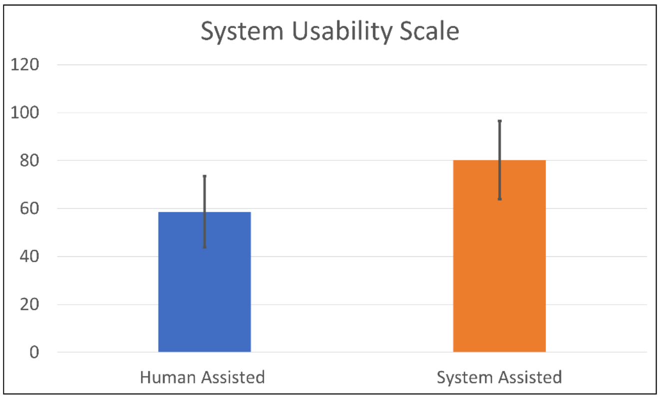


Figure 7: Average system usability score reported by participants in HA and SA scenarios

5.6 Subjective Assessment

The subjective assessment in the form of a questionnaire was presented to participants after the completion of each scenario followed by an end-of-experiment questionnaire. These questionnaires contained both long-answer form and five-point Likert scale-based questions. In the case of HA, the Likert scale-based questions were -

Q1 The non-verbal interactive interface helped me to provide assistance to the robot.

Q2 The non-verbal interface was intuitive and easy to use.

Q3 The robot accurately followed my guidance.

Q4 I am satisfied with the overall outcome of the search task.

Q5 My assistance contributed to the successful completion of the task.

Q6 Human assistance is beneficial for the robot in a search task in a cluttered environment.

The findings based on the responses are presented in the graph shown in figure 8.

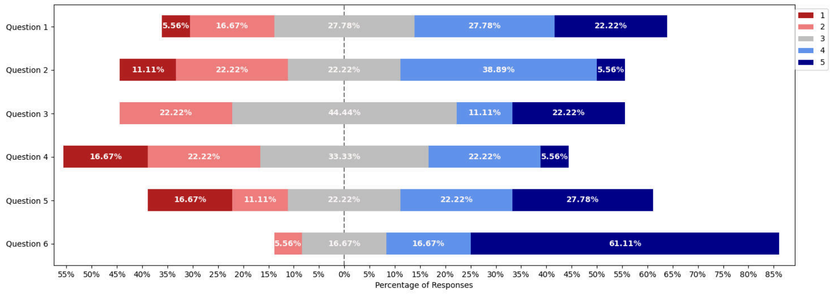


Figure 8: Subjective Likert responses to HA scenario interview questions

Similarly, in the case of SA, the Likert-based questions were -

Q1 I had a good experience with System assisted search for providing assistance to the robot

Q2 The foveated view field improved my experience in finding points of interest and importance in the scene

Q3 I trust the system's understanding of the scene to guide me to particular locations in the scene

Q4 I had a good experience with System assisted search and foveation for providing assistance to the robot in this case

Q5 I am satisfied with the overall outcome of the search task.

Q6 My assistance contributed to the successful completion of the task.

Q7 I am confident in the robot's ability to find the important locations.

Q8 Human assistance is beneficial for the robot in a search task in a cluttered environment.

The findings based on the responses are presented in the graph shown in the figure 9.

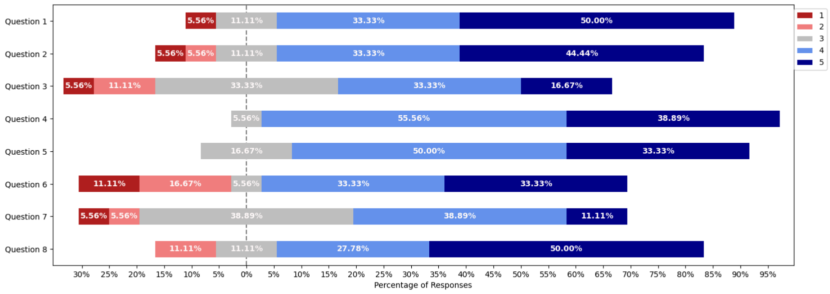


Figure 9: Subjective Likert responses to SA scenario interview questions

6 Discussion

The results presented offer a comprehensive evaluation of the performance, workload, and usability of participants in two different scenarios: HA and SA. These scenarios were designed to assess the effectiveness and efficiency of systems in aiding users in identifying and locating trapped humans within a simulated environment. The analysis and discussion below provide insights into the implications of these findings.

6.1 Task Timing

Participants completed the tasks significantly faster in the SA scenario compared to the HA scenario. The average time to complete the SA scenario was approximately one-third of the time taken to complete the HA scenario. This substantial reduction in task completion time suggests that the SA approach provides a more efficient means of accomplishing the task at hand.

6.2 Task Performance

In terms of task performance, participants demonstrated a higher success rate in identifying instances of trapped humans in the SA scenario compared to the HA scenario. The increased accuracy in identifying trapped humans indicates that the system provides valuable assistance to users, enhancing their ability to detect critical elements within the simulated environment.

6.3 Workload

The perceived workload reported by participants was significantly lower in the SA scenario compared to the HA scenario. This finding suggests that participants experienced reduced mental and physical demands when utilizing the SA approach. A lower perceived workload is desirable as it can lead to improved user satisfaction and overall performance.

6.4 System Usability

Participants rated the SA scenario as significantly more usable compared to the HA scenario. The higher system usability score indicates that participants found the system to be more intuitive, efficient, and satisfactory in assisting them with the task. The preference for the SA scenario underscores the importance of designing systems that are intuitive to use and supportive of user needs.

6.5 Subjective Assessment

The subjective assessment of teaming scenarios revealed valuable insights into the strengths and areas for improvement in both HA and SA scenarios. Participants shared detailed experiences and provided constructive feedback that can inform the refinement of systems for various human-robot collaborative tasks, particularly in scenarios involving emergency response and reconnaissance.

In the HA mode, participants demonstrated a preference for intuitive decision-making (ex - moving forward, looking around, size relates to danger), leveraging factors such as the likelihood of finding objects and the immediacy of danger to and around humans. However, challenges such as slow turning and limited peripheral vision were noted, highlighting the importance of improving physical interfaces and enhancing situational awareness. This can be also seen in the case of the Likert scale response, where participants generally had neutral or positive feedback about the non-verbal interface intuitiveness but there were mixed responses regarding the effectiveness of their assistance in guiding the robot accurately. Suggestions for improvement included implementing graphical interfaces for prioritizing items and enhancing navigation capabilities through features like independent perception control with depth feedback, joystick control, and sound cues. Participants also suggested to be provided with real-time feedback.

Conversely, in the SA mode, participants acknowledged the potential of automation in streamlining tasks and providing immediate feedback, particularly through features like foveation and object detection. However, concerns regarding the system's inability to highlight critical elements consistently and challenges related to foveation-induced loss of information were raised. Participants emphasized the importance of refining algorithms for scene perception and enhancing camera feeds to improve overall system performance.

We notice that although participants feel they can help the robot effectively, they don't fully trust the system's understanding of the environment. This suggests they're confident in their ability to assist practically but are unsure about how well the system comprehends the surroundings. This highlights the importance of aligning participants' perceptions of the robot's intent with its actual capabilities to foster trust and collaboration.

Furthermore, discussions surrounding the handover of control between human operators and the robot highlighted the necessity of clear protocols and established cooperation practices. While participants expressed willingness to delegate control under certain conditions, such as when the operator possesses superior situational awareness or familiarity with the task, concerns regarding potential conflicts and the need for a hierarchical command structure were evident.

7 Conclusion and Future Direction

In this study, we set out to investigate the effectiveness and efficiency of mixed-reality-based systems in assisting operators during human-robot collaboration scenarios. Experiments with participants in a virtual search and rescue environment helped us explore this using multimodal interaction techniques. Participants demonstrated improved task performance, reduced workload, and higher usability ratings when utilizing SA methods compared to HA ones. In both cases, the results emphasize the complex interplay between human intuition and automated assistance in collaboration scenarios. Subjective assessments highlighted the importance of intuitive interfaces, real-time feedback, and clear protocols for effective collaboration between human operators and robotic systems. By addressing the identified challenges and incorporating user feedback, future developments in human-robot teaming can enhance operational effectiveness and promote seamless collaboration between human operators and robotic systems, ultimately advancing capabilities in domains such as emergency response and reconnaissance.

Acknowledgements

We would like to thank Jacek Malec and Björn Olofsson for their support during the work. This work was supported by ELLIIT - the Excellence Center at Linköping University and Lund University for Information Technology. The 3D environment employed in this research is provided by the WASP Research Arena for Public Safety (WARA-PS). The Wallenberg Artificial Intelligence, Autonomous Systems, and Software Program (WASP) funded this initiative through the Knut and Alice Wallenberg Foundation.

References

[AE17] Emre Akbas and Miguel P Eckstein. "Object detection through search with a foveated visual system". In: *PLoS computational biology* 13.10 (2017), e1005743.

[BKM08] Aaron Bangor, Philip T Kortum, and James T Miller. “An empirical evaluation of the system usability scale”. In: *Intl. Journal of Human–Computer Interaction* 24.6 (2008), pp. 574–594.

[Bro96] John Brooke. “Sus: a “quick and dirty’usability”. In: *Usability evaluation in industry* 189.3 (1996).

[HS88] Sandra G Hart and Lowell E Staveland. “Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research”. In: *Advances in psychology*. Vol. 52. Elsevier, 1988.

[JT23] Ayesha Jena and Elin Anna Topp. “Chaos to Control: Human Assisted Scene Inspection”. In: *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. 2023.

[LU23] LU. *Ethical review — staff.lu.se*. `https://www.staff.lu.se/research-and-education/research-support/research-ethics-and-animal-testing-ethics/ethical-review`. [Accessed 10-02-2024]. 2023.

[LHZ97] Thomas Laengle, Thomas Hoeniger, and Lanjuan Zhu. “Cooperation in human-robot-teams”. In: *ISIE’97 Proceeding of the IEEE international symposium on industrial electronics*. IEEE. 1997, pp. 1297–1301.

[Nat+23] Manisha Natarajan et al. “Human-robot teaming: grand challenges”. In: *Current Robotics Reports* 4.3 (2023), pp. 81–100.

[OR21] Linda Onnasch and Eileen Roesler. “A taxonomy to structure and analyze human-robot interaction”. In: *International Journal of Social Robotics* 13.4 (2021).

[Plo+22] Alexander Plopski et al. “The eye in extended reality: A survey on gaze interaction and eye tracking in head-worn extended reality”. In: *ACM Computing Surveys (CSUR)* 55.3 (2022).

[SG19] Ludwig Sidenmark and Hans Gellersen. “Eye&head: Synergetic eye and head movement for gaze pointing and selection”. In: *Proceedings of the 32nd annual ACM symposium on user interface software and technology*. 2019.

[TRW18] Nhan Tran, Josh Rands, and Tom Williams. “A hands-free virtual-reality teleoperation interface for wizard-of-oz control”. In: *Proceedings of the 1st International Workshop on Virtual, Augmented, and Mixed Reality for HRI (VAM-HRI)*. 2018.

[Tri+20] Eleftherios Triantafyllidis et al. “Study of multimodal interfaces and the improvements on teleoperation”. In: *IEEE Access* 8 (2020), pp. 78213–78227.

[VSU97] Boris Velichkovsky, Andreas Sprenger, and Pieter Unema. "Towards gaze-mediated interaction: Collecting solutions of the "Midas touch problem"". In: *Human-Computer Interaction INTERACT'97: IFIP TC13 International Conference on Human-Computer Interaction, 14th–18th July 1997, Sydney, Australia*. Springer. 1997.

[WP24] WARA-PS. *GitHub - wara-ps/cesium-unreal — github.com*. `https://github.com/wara-ps/cesium-unreal`. [Accessed 10-02-2024]. 2024.

[Wal+23] Michael Walker et al. "Virtual, augmented, and mixed reality for human-robot interaction: A survey and virtual design element taxonomy". In: *ACM Transactions on Human-Robot Interaction* 12.4 (2023), pp. 1–39.

[Zha+19b] Wei Zhang et al. "A gesture-based teleoperation system for compliant robot motion". In: *Applied Sciences* 9.24 (2019), p. 5290.

IMPACT OF GAZE-BASED INTERACTION AND AUGMENTATION ON HUMAN-ROBOT COLLABORATION IN CRITICAL TASKS

1 Abstract

We present a user study with 18 participants, analyzing head-gaze-based robot control and foveated visual augmentation in a simulated search-and-rescue task. Results show that foveated augmentation significantly improves task performance, reduces cognitive load by 38%, and shortens task time by over 60%. Head-gaze patterns analysed over both the entire task duration and shorter time segments show that near and far attention capture is essential to better understand user intention in critical scenarios. Our findings highlight the potential of foveation as an augmentation technique and the need to further study gaze measures to leverage them during critical tasks.

Ayesha Jena, Stefan Reitmann, and Elin Anna Topp. "Impact of Gaze-Based Interaction and Augmentation on Human-Robot Collaboration in Critical Tasks." In *2025 International Conference on Social Robotics + AI. Springer Nature. Naples, Italy*.
Accepted and to be published.

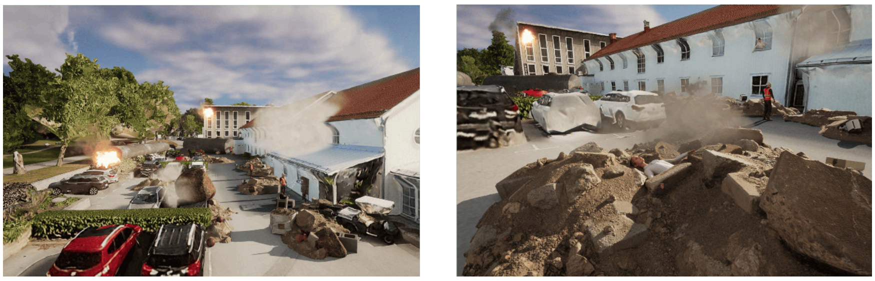


Figure 1: Left: Bird's eye view of the test environment; Right: A region of the environment showing a trapped human.

2 Introduction

Advancements in the field of robotics have led to a need for seamless collaboration and effective communication between humans and robots in different scenarios [Hen+19]. While communication methods such as speech and gestures have been extensively studied, they require explicit commands which limits their effectiveness in dynamic real-world interaction scenarios. In such cases, methods like gaze-based interaction offer an intuitive way of communication [LN+24]. In addition to being a method of interaction, gaze also shows operator's intentions regarding where to focus during task execution [Bel+24]. This is crucial in high-stakes scenarios, where fast recognition of user intent through gaze could enhance performance and improve collaboration.

In human-robot collaboration, gaze tracking has proved effective in combination with augmentation techniques to improve collaboration efficiency, situational awareness, and productivity [Sch+25]. While gaze is used for controlling the robot, augmentation techniques are used for visually enhancing the information provided by the system. Their effect in critical domains have been largely unexplored [She16], which leads to an adoption gap in understanding operator intentions and supporting collaboration in high-stake scenarios.

We explore this gap through a user study using our collaborative interface [JT23, JT24].

This work is novel in its design and integration of head-gaze based interaction and real-time foveation-based visual augmentation for collaboration in a critical search-and-rescue scenario, as shown in Fig. 1. We will explain how we interpret the terms augmentation and foveation, in the context of our study in sections 3 and 4.

Our study explored two different interaction designs within the interface: manual head-gaze based control interface - human assisted (HA), and a dynamic foveation based interface - system assisted (SA). We found that foveation based augmentation enhances task performance, reduces cognitive load, and improves collaboration in comparison with direct gaze-based control. While head-gaze directly indicates user attention, it performed suboptimally as a primary control modality in critical scenarios. Our analysis, however, shows the usability of foveated augmentation to guide attention also in such scenarios. We further found instances of extrafoveal attention capture which would be accounted for in future studies with an adaptive system that incorporates complex gaze behavior.

3 Related Work

When human and robots collaborate in teams for critical tasks and missions, the cognitive capabilities of operators tend to decline over time. This is because of the inherent nature of such scenarios where rapid actions need to be taken while receiving, processing and combining information from multiple sources at the same time [MF19]. This often leads to errors which in turn can provoke catastrophic outcomes [Liu+24]. Both explicit and implicit modes of communication have been studied to minimize the operator's effort while maintaining effective information exchange during task execution. Early research focuses on explicit communication modalities such as speech, gestures, and haptic interfaces [US23]. While these methods are effective in structured environments, they often lead to additional cognitive demands on operators, particularly in dynamic, high-stakes scenarios [ULS20]. This has led to interest in implicit communication cues, such as gaze, which enable more natural and intuitive human-robot coordination [LN+24].

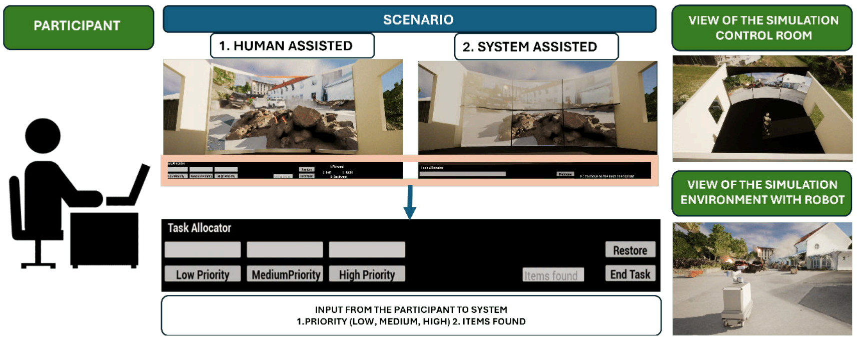

PARTICIPANT
SCENARIO
1. HUMAN ASSISTED
2. SYSTEM ASSISTED
VIEW OF THE SIMULATION CONTROL ROOM
VIEW OF THE SIMULATION ENVIRONMENT WITH ROBOT
Task Allocator
Low Priority
MediumPriority
High Priority
Items found
Restore
End Task
INPUT FROM THE PARTICIPANT TO SYSTEM
1.PRIORITY (LOW, MEDIUM, HIGH) 2. ITEMS FOUND

Figure 2: The setup of the user study. Left to right: Screen view of the interface shown to participants during the experiment, View of the simulation control room which receives visual input from the robot's camera and provides inputs from the participants to the system, View of the simulation environment showing the robot in the search-and-rescue test-bed.

Gaze-only interfaces encounter challenges like the "Midas touch problem", where unintentional gaze inputs trigger undesired actions [VSU97]. There is also the need for special hardware and software for eye tracking, and difficulties in real-world scenarios due to illumination effects, head rotations, or occlusions [JB16].

A similar, yet effective, approach is to use head pose instead. Studies have shown a high correlation between gaze direction and head pose in real-world scenarios, proving its effectiveness [JB16].

Building on methods to improve collaboration through the use of a visual interface, we explored different augmentation techniques that would enhance task performance while reducing cognitive overload [Su+23]. Techniques like highlighting important objects seemed ineffective in our case due to clutter and background distractions. Instead, foveation seemed to be a promising approach in this regard where focused regions are rendered in high resolution and outlying areas are blurred [Su+23]. Traditionally, this is used as a graphics-performance optimization technique, so this study is the first to apply foveation as a visual augmentation method in a search-and-rescue setting. Considering the complexities of real-world situations, this study investigates head-gaze based control and foveation based augmentation in simulation and analyzes the results using task performance metrics, subjective measurements, and gaze-based heatmaps.

4 Methodology

The study was conducted in a simulation environment where participants performed a search task with two interaction designs: HA and SA. Participants' head-gaze behavior was recorded and then analyzed to assess efficiency, cognitive workload, and task performance. Head-gaze tracking was performed using a camera-based tracking system that detected participants' head orientation and head-gaze direction. An overview and breakdown of the experimental conditions are shown in Fig. 2.

4.1 Simulation Environment

A search-and-rescue (SAR) test environment (28m x 83m) was developed from the 3D template of the real-world location provided by the WASP Research Arena for Public Safety (WARA-PS) [WP24] to simulate a post-disaster scenario using Unreal Engine 5.1 [Epi22]. The primary motivation behind creating this virtual SAR test-bed was to replicate challenging and high-stake conditions in a controlled, safe environment. The added wreckage, obstacles, and people were strategically placed throughout the environment. It also had multiple other areas of interest, such as fire outbreaks and electrical hazards. Additionally, it included simulated SAR personnel to replicate realistic operational constraints and potential coordination efforts in real-time rescue missions. This was done on a collaborative interface developed earlier [JT23, JT24]. The interface sends gaze data to Unreal Engine, where it's mapped with 3D vectors to highlight screen regions in orange. A key press turns the region green, forming a dual-confirmation input. It also includes text fields and buttons for selecting priorities, and overlays foveation cues on the camera feed from the robot (Fig. 2). Some distance away from the test area, a control room, as shown in Fig. 2, was designed to replicate real-world SAR operation centers where users could interact with the disaster site by navigating a mobile robot fitted with cameras and sensors. The robot used for this setup is a combination of mobile robot platform MiR200, a Universal Robots collaborative arm UR5e, attached with an Intel Realsense D435 RGBD camera on the plate of Schunk WSG-50 parallel gripper [JT23].

4.2 Experiment Procedure

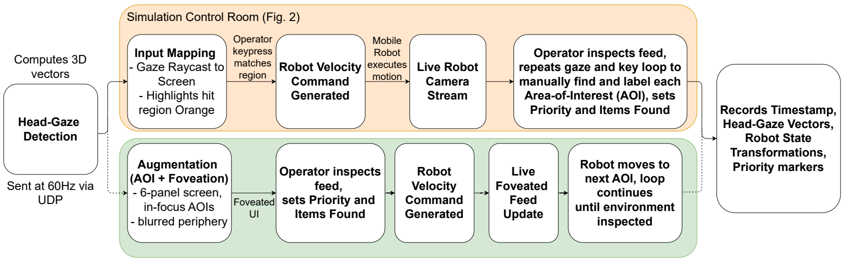

Simulation Control Room (Fig. 2)
Computes 3D vectors
Head-Gaze Detection
Sent at 60Hz via UDP
Input Mapping
- Gaze Raycast to Screen
- Highlights hit region Orange
Operator keypress matches region
Robot Velocity Command Generated
Mobile Robot executes motion
Live Robot Camera Stream
Operator inspects feed, repeats gaze and key loop to manually find and label each Area-of-Interest (AOI), sets Priority and Items Found
Augmentation (AOI + Foveation)
- 6-panel screen, in-focus AOIs
- blurred periphery
Foveated UI
Operator inspects feed, sets Priority and Items Found
Robot Velocity Command Generated
Live Foveated Feed Update
Robot moves to next AOI, loop continues until environment inspected
Records Timestamp, Head-Gaze Vectors, Robot State Transformations, Priority markers

Figure 3: The experiment procedure for both scenarios. Orange shows the HA scenario. Green shows the SA scenario.

Each session lasted about 60 minutes and began with a system check, consent and demographic forms, and a brief training. Participants then completed two counterbalanced scenarios - HA and SA, each involving teleoperating the robot through the interface (Fig. 3). In each scenario, they navigated within the test environment to find areas that needed inspection also referred to as areas of interests (AOIs), counted and classified objects (victims, debris, hazards), and assigned priority levels (Low, Medium, High) through the interface as can be seen in Fig. 2. Before the experiment it was made sure that the ethical protocols were followed per university guidelines [LU23]. During each session, we recorded head-gaze, dual-confirmation key presses, robot state transformations, NASA-TLX [HS88], System Usability Scale (SUS) [Bro96], post-study interview responses, and quantitative metrics (task times, AOI accuracy), all anonymized under data protection rules. A total of 18 participants (13 males, 4 females, 1 unspecified) took part in the study. 4 participants had experience in providing disaster relief. The mean age of participants was 31.29 years (SD = 9.78) excluding 1 participant who declined to report their age.

Human-Assisted (HA) Scenario

In this interaction design scenario, participants controlled the robot using head-gaze for directions and keypresses for confirmations. The experimental procedure can be seen in Fig. 3.

System-Assisted (SA) Scenario

In this scenario, the 6 screens shown to the participants were foveated based on the AOIs present and the density of hazard. Screens with AOIs were in focus whereas others were blurred. Although the system guided participants to prioritized regions, participants remained the final decision-makers, confirming or adjusting the importance levels of each item in AOIs (Low, Medium, High) through the interface. The six-section design was optimized to align with human visual working memory limits (5–9 chunks [Mil56]), ensuring rapid parsing without overwhelming users during critical scenarios.

5 Results and Discussions

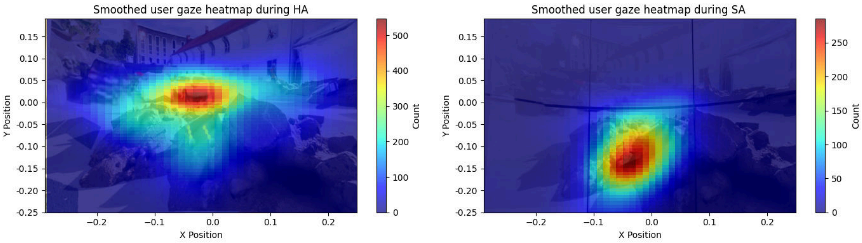


Figure 4: Distribution of head-gaze heatmap over the 2D screen area presented to the participants.

Table 1: Objective and Subjective Results for HA and SA scenarios. $*$ indicates significant result ($p < 0.001$)

Scenario	Total Time Taken (in s)	Total Humans Saved	Avg Humans Saved	NASA TLX Score	SUS
HA	678.88 ± 233.98	20 (Out of 54)	1.11 ± 0.75	53.85 ± 18.06	58.61 ± 14.80
SA	$*$**274.41 ± 52.95**	$*$**47** (Out of 54)	$*$**2.61 ± 0.69**	$*$**33.4 ± 15.24**	$*$**80.13 ± 16.30**

In both the scenarios, participants were instructed to identify critical AOIs, mark them, and assign points according to their priority. The results are discussed below.

5.1 Performance Metrics

The performance of the participants was measured using task completion time and the number of humans successfully located within AOIs. There were 3 trapped humans in each scenario within AOIs, resulting in a total of 54 instances of trapped humans across all participants. From the results shown in Table 1, we can see that participants performed better with foveation. The experience of the 4 participants with disaster relief did not affect the results in any manner.

5.2 Subjective Measurements

Mental Workload

The participants mental workload results (Table 1) from the NASA TLX questionnaire showed a 38% lower perceived workload in SA scenario ($M = 33.41$, $SD = 15.24$) compared to the HA scenario ($M = 53.85$, $SD = 18.07$).

System Usability

Participants answered 10 questions from the system usability questionnaire using a Likert scale (1 = strongly disagree to 5 = strongly agree). Results from the Table 1 show that participants perceived foveation to have higher usability during the search task.

Subjective Questionnaire

The subjective questionnaires contained long-form answers and five-point Likert scale questions. Participants provided neutral-to-positive ratings for the interface's natural intuitiveness, noting that its reliance on manual gaze-driven navigation mirrored "natural human exploration". However, this familiarity with gaze-based control came with challenges such as "slow turning" and "limited peripheral vision". On the other hand in SA, augmentation helped them streamline focus and execute the task faster, while reducing their cognitive load. However, participants also expected explainability regarding the augmentation decisions made by the system. In addition, participants wanted to be able to dynamically look at other sections of the screen to make sure the system made the correct decisions regarding foveation.

5.3 Head-gaze Analysis

In the HA scenario, participants' head-gaze was evenly spread across the screen, looking equally to the left and right, but with a stronger focus on the upper part. This scenario required them to use their head-gaze to drive the robot, causing them to scan the screen more and focus on the upper part for forward navigation based on the interface design. In the SA scenario, participants did not need to orient their head-gaze, so they adopted a more relaxed posture and looked mostly at the lower part of the screen as can be seen in Fig. 4. Additionally, analyzing shorter time segments in this scenario during task execution highlight that gazed region and the foveated region aligned 67% of the time. In the remaining 33%, participants' gaze slightly deviated around the foveated area. Upon further analysis of view counts, view percentages, and average gaze position across the screen area, it was found that this occurred due to high-priority items and people located outside the immediate field of view. This observation can be explained by extrafoveal attention capture which is when objects or individuals in more distant areas draw the user's attention over a longer distance [Nut+19]. Similar instances of extrafoveal attention captures were also observed in the HA scenario where users verbally inquired about denoting distant AOIs in comparison to focusing on immediate AOIs.

6 Conclusion

In summary, we conducted a user study to investigate the impact of gaze-based control and foveated augmentation for human-robot teleoperation during critical task execution. The study was performed with a system designed in simulation and the effects were studied through two scenarios. Based on the results of the user study, foveation based augmentation scenario outperformed gaze-based control leading to effective human-robot collaboration. Additionally, analysis of head-gaze behavior revealed alignment between gaze and foveated regions in a majority of cases, indicating that foveation can effectively direct attention. However, deviations due to extrafoveal attention capture highlight the complexity of user attention patterns and suggest the need for systems to accommodate such behavior.

Overall, this study demonstrates that head-gaze based foveated augmentation improves performance and user experience. While head-gaze reliably conveys where the user's attention is directed, our results suggest it is not optimal as a direct control modality in critical tasks. Further analysis is required to leverage gaze measures during critical tasks. The future work includes improving the interface by combining automated scene analysis with flexible, gaze dependent foveation that operators can invoke as an explicit confirmation cue.

Acknowledgements

We would like to thank Jacek Malec and Björn Olofsson for their support during the work. This work was funded by ELLIIT - the Excellence Center at Linköping University and Lund University for Information Technology, and partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. We would also like to thank the WASP Research Arena for Public Safety (WARA-PS) for providing the 3D template.

Disclosure of Interests

The authors have no competing interests to declare that are relevant to the content of this article.

References

[Bel+24] Valerio Belcamino et al. “Gaze-based intention recognition for human-robot collaboration”. In: *Proceedings of the 2024 International Conference on Advanced Visual Interfaces*. 2024, pp. 1–5.

[Bro96] John Brooke. “Sus: a “quick and dirty’usability”. In: *Usability evaluation in industry* 189.3 (1996).

[Epi22] Epic Games. *Unreal Engine*. Version 5.1. Nov. 15, 2022.

[HS88] Sandra G Hart and Lowell E Staveland. “Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research”. In: *Advances in psychology*. Vol. 52. Elsevier, 1988.

[Hen+19] Abdelfetah Hentout et al. “Human–robot interaction in industrial collaborative robotics: a literature review of the decade 2008–2017”. In: *Advanced Robotics* 33.15-16 (2019).

[JT24] Ayesha Jena and Elin A Topp. “Towards Understanding the Role of Humans in Collaborative Tasks”. In: *7th International Workshop on Virtual, Augmented, and Mixed-Reality for Human-Robot Interactions*. 2024.

[JT23] Ayesha Jena and Elin Anna Topp. “Chaos to Control: Human Assisted Scene Inspection”. In: *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. 2023.

[JB16] Sumit Jha and Carlos Busso. “Analyzing the relationship between head pose and gaze to model driver visual attention”. In: *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE. 2016.

[LU23] LU. *Ethical review — staff.lu.se*. `https://www.staff.lu.se/research-and-education/research-support/research-ethics-and-animal-testing-ethics/ethical-review`. [Accessed 10-02-2024]. 2023.

[LN+24] Matteo Lavit Nicora et al. “Gaze detection as a social cue to initiate natural human-robot collaboration in an assembly task”. In: *Frontiers in Robotics and AI* 11 (2024).

[Liu+24] Yuan Liu et al. “What affects human decision making in human–robot collaboration?: a scoping review”. In: *Robotics* 13.2 (2024).

[Mil56] George A Miller. "The magical number seven, plus or minus two: Some limits on our capacity for processing information." In: *Psychological review* 63.2 (1956).

[MF19] Milad Mirbabaie and Jennifer Fromm. "Reducing the cognitive load of decision-makers in emergency management through augmented reality". In: (2019).

[Nut+19] Antje Nuthmann et al. "Extrafoveal attentional capture by object semantics". In: *PLoS One* 14.5 (2019).

[Sch+25] Fabian Schirmer et al. "Utilizing Eye Gaze for Human-Robot Collaborative Assembly". In: *Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction*. 2025.

[She16] Thomas B Sheridan. "Human–robot interaction: status and challenges". In: *Human factors* 58.4 (2016).

[Su+23] Yun-Peng Su et al. "Integrating virtual, mixed, and augmented reality into remote robotic applications: a brief review of extended reality-enhanced robotic systems for intuitive telemanipulation and telemanufacturing tasks in hazardous conditions". In: *Applied Sciences* 13.22 (2023).

[ULS20] Vaibhav V Unhelkar, Shen Li, and Julie A Shah. "Decision-making for bidirectional communication in sequential human-robot collaborative tasks". In: *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. 2020.

[US23] Jacqueline Urakami and Katie Seaborn. "Nonverbal cues in human–robot interaction: A communication studies perspective". In: *ACM Transactions on Human-Robot Interaction* 12.2 (2023).

[VSU97] Boris Velichkovsky, Andreas Sprenger, and Pieter Unema. "Towards gaze-mediated interaction: Collecting solutions of the "Midas touch problem"". In: *Human-Computer Interaction INTERACT'97: IFIP TC13 International Conference on Human-Computer Interaction, 14th–18th July 1997, Sydney, Australia*. Springer. 1997.

[WP24] WARA-PS. *GitHub - wara-ps/cesium-unreal — github.com*. `https://github.com/wara-ps/cesium-unreal`. [Accessed 10-02-2024]. 2024.

BRIDGING MINDS AND MACHINES: A COMPREHENSIVE REVIEW OF INTENTION RECOGNITION IN HUMAN-ROBOT INTERACTION

Abstract

Advancements in the field of robotics, specifically human-robot interaction (HRI), have led to a need for real-time systems being able to infer and respond to human intentions. With intention being a multifaceted concept of varying interpretations in each domain based on their needs and frameworks, we have categorized it into low-level and high-level processes. This review provides an in-depth analysis using this categorization by exploring the psychological theories and computational models of the state-of-the-art techniques for intention recognition. By systematically exploring and reviewing the applications of intention recognition across domains such as healthcare, manufacturing, and search-and-rescue, among others,

Ayesha Jena, and Elin Anna Topp. "Bridging Minds and Machines: A Comprehensive Review of Intention Recognition in Human-Robot Interaction". Submitted to the ACM Transactions on Human-Robot Interaction.

this paper highlights key advancements, from probabilistic models to deep learning approaches, and identifies critical gaps in current research. Additionally, this paper provides insights into challenges related to ambiguity, generalization, scalability, and real-time inference, proposing directions aimed at more intuitive robotic systems.

1 Introduction

In the age of rapid technological advancement, the integration of artificial agents into various domains of our social life as social companions, industrial assistants, and service providers is reshaping how machines and humans interact [Bro17]. A crucial development within this landscape is **intention recognition**, a pivotal aspect that enables robots to understand and predict human intentions or to negotiate the terms between several different agents. The impact on real-time applications would be wide and varied. In sectors like manufacturing, understanding the intentions of human workers on the assembly line could help to better synchronize the actions of the robots for efficient goal fulfilment. Similarly, in healthcare, intention-aware robots could assist therapists by comprehending patient movements during rehabilitation exercises and suggesting interventions for personalized recovery. This extends to disaster response as well, where robots can work in cooperation with human responders, predicting actions and helping in search and rescue operations. In all these varied applications, robots with intention recognition capabilities would transform the way we perform collaborative tasks. However, this rapidly evolving field makes it essential to have a structured review to identify the methods, progress, challenges, and future opportunities.

This review aims to address critical gaps in the existing literature of intention recognition in robotics by providing a comprehensive analysis of the methodological and technological approaches that have shaped the current state-of-the-art. Differing from existing surveys which often focus on particular aspects of intention recognition, this review emphasizes on categorization of intention recognition into high-level and low-level processes, theoretical frameworks, and computational methodologies.

Intention recognition leverages a range of sensors, data analysis techniques, and machine learning algorithms to interpret human behaviour through gestures, facial expressions, speech, and movements. This ability to decode human intentions is transforming how robots assist in manufacturing, search-and-rescue, healthcare, education, and social interaction, driving more efficient, safe, and collaborative human-robot interactions.

Intention recognition is an inherent aspect of human-human interaction. Consider the scenario of being at a dinner table with friends or family, where one individual subtly gestures towards a salt shaker near you. Without any verbal communication, you perceive their glance and slight gesture as a request to pass the

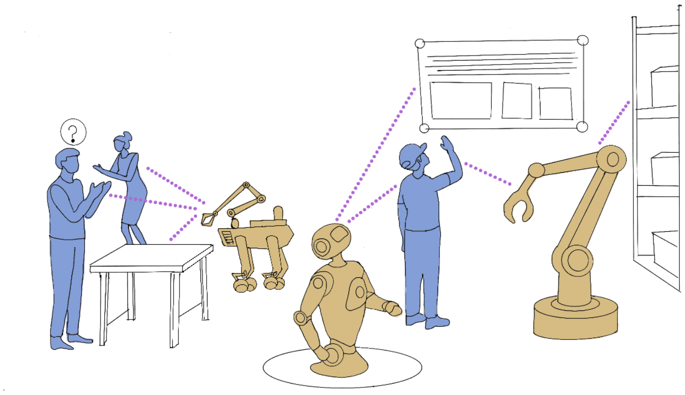


Figure 1: Intention recognition allows robots to look at humans and their surroundings to figure out the intentions behind their actions and assist them accordingly. One robot can attend to a group of people as can be seen in the left section of the image or a group of robots can assist a single person based on their capabilities as can be seen in the right section of the image.

salt, understanding their intention to avoid interrupting the conversation. Similarly, in a football match, a teammate may indicate their intention to pass the ball through body orientation and eye direction, rather than shooting at the goal themselves. Here, the interpretation of non-verbal cues—such as posture, gaze, and ball positioning—enables us to anticipate their actions and prepare to receive the ball. These examples highlight how humans can decode intentions through body language, especially in dynamic and fast-paced situations, facilitating coordinated actions without explicit verbal communication. This capability involves synthesizing multiple sources of information to infer others' thoughts or goals, illustrating a key aspect of human social cognition. Reflecting about whether the goals and beliefs of oneself and others are object or situation directed, means to ascribe *intentionality* [MK97, Per22] to these inferred goals or thoughts. While the ability to do so has been extensively studied, particularly in the context of robotics, there remains no universally accepted definition of intention recognition despite its significance and frequent application. The following few examples support the above statement and show a glimpse of the varying meanings of intention recognition in the field of robotics.

A search for the word 'intention' in Crossref with 'intention', 'recognition', and 'robotics' as keywords gives us a list of papers on intention recognition across various categories of robotics. For instance, in rehabilitation robotics, [Suz+07] used intention recognition to estimate the weight shift of paraplegics from the ground reaction force to control their Robot suite Hybrid Assistive Limb (HAL) to assist them in walking. Here, intention is identifying when the user wants to walk and it represents an action which affects the user by themselves. [Wan+13] recognise the intent of players in the direction they want to hit the ball in a game of table tennis using a modified Gaussian process dynamics model. In this case, the intention to be recognised is the direction of the return so that the robot can decide before the ball is hit by the player. The intention here affects an object which is manipulated by the user as opposed to the previous case where the effect was directly on the user. In the field of autonomous driving, [Bai+15] used intention recognition to predict the intention of pedestrians to safely select actions for autonomous driving amidst pedestrians. Here, intention is from the point of view of the pedestrians and what is the next action they are going to take. On a similar note, [Li+16] used a combination of the Hidden Markov model (HMM) and Bayesian Filtering to predict lane changing behaviour of drivers using inputs of steering angle, lateral acceleration, and yaw rate to assist in driving and preventing accidents. In this case, the intention is from the point of view of the driver and when they want to change lanes. Others have also looked at intention recognition in autonomous driving [Pet+19, MÅ18, TFA10, Xin+20] and pedestrians [Gol+19, Var+18, Li+17, KM17, Völ+15, Völ+16, PL16, DSS15, BED08, Zha+20]. These examples show how intention recognition is used to predict an action which may affect the user themselves or an object in their surrounding. However, intention in the above-mentioned cases is understood on a low level of actions and their immediate outcomes. [VTZ16] explain the role of intention in cognitive robotics and how, on a higher level, it is important to see the world from other people's perspectives by forming a theory of mind.

It is challenging to categorise intentions because of the varied definitions people have across different domains [Per+13, AA07, CG93, Hei04, Tah06]. We also struggled with the decision to select one categorization over the other, but, based on a few other works, felt that the distinction of *high-level intention recognition* and *low-level intention recognition* [KS08, Saf+15b, Saf+15a, GCR21] aligns best with our motivations. This can be seen in Figure 3.

The structure of this paper is as follows: **Section 2** covers the methodology used to scope the literature for relevant papers to be included in this review, **Section 3** introduces the categorization of intention recognition into high-level and low-level processes, **Section 4** explores the psychological theories contributing to human intention recognition, **Section 5** delves into the computational models and highlights the various state-of-the-art techniques and methodologies used for intention recognition, **Section 6** expands upon the categorization introduced in Section 3 and discusses broadly the various sub-categories, and **Section 7** identifies challenges, limitations, and emerging trends in intention recognition, offering recommendations for future research.

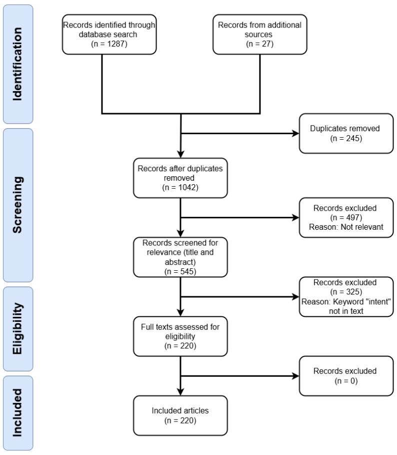

Identification
Screening
Eligibility
Included
Records identified through database search (n = 1287)
Records from additional sources (n = 27)
Duplicates removed (n = 245)
Records after duplicates removed (n = 1042)
Records excluded (n = 497) Reason: Not relevant
Records screened for relevance (title and abstract) (n = 545)
Records excluded (n = 325) Reason: Keyword "intent" not in text
Full texts assessed for eligibility (n = 220)
Records excluded (n = 0)
Included articles (n = 220)

Figure 2: A systematic review process using PRISMA

2 Methodology

In order to conduct a systematic review of intention recognition, we performed a scoping literature review following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) [Pag+21] methodology described as follows:

The initial search consisted of using the keywords and phrases: “intent”, “intention”, “human-intent”, “robot-intent”, “intention recognition”, “intention prediction”, “intention estimation”, “intention modelling”, “methods”, “intention recognition in robotics”, and “human-robot interaction”. These keywords and phrases were used in combination with operators such as AND, and OR to refine the results. The detailed step-by-step review process using PRISMA is shown in Figure 2.

Inclusion Criteria: Our study focused on including the papers that mentioned the keyword “intent” or “intention” in their text. All the publications included were in English. In terms of time-frame, studies and research done over the years are considered taking into account the contributions to the field of work that have been accumulated over a long period.

3 Categorisation

The examples in the previous section show how complex the topic of intention recognition is. They also highlighted the importance of creating systems that can understand and predict human behaviour. In a dyadic robotics context, intention recognition can be viewed in three different ways based on involved agents.

- In the first scenario, a human agent can try to recognize the intention of the robotic agent.

- In the second scenario, a robotic agent can try to recognize the intentions of the human agent.

- And in the third scenario, a robotic agent can try to recognize the intention of another robotic agent.

An intuitive understanding of robotics would allow one to conclude that the third scenario is solvable if the robots involved can communicate with each other. Unlike the complex nature of verbal and non-verbal communication persistent between humans, robots' intentions can be conveyed to one another through the exchange of goals and action plans via various communication channels such as wireless connection, signal lights, etc.

In the past, people have made attempts to categorize intentions to aid in reviewing intention recognition.

The notion of distal (D-intentions), proximal (P-intentions), and motor (M-intentions) intentions was introduced by [Pac08, Per22] and talks about how distal intentions broadly relate to [Bra87]'s notion of future intention and are formed during the initial stages of planning and decision-making. P-intentions are closer to the actual execution of actions. They emerge from D-intentions and are more specific, involving the decision to start acting in the present moment. M-intentions are generated from P-intentions and involve specific motor programs that execute the actions. These intentions are responsible for the fine-grained control of bodily movements. They operate at a very detailed and fast time scale, often outside conscious awareness, ensuring the smooth execution of movements.

[Omo+08] talk about active and passive intentions where passive is the intention which is inferred from observing other agents whereas active is the one which influences the other agent based on the actions of the self.

We will explore the definition of intention in more detail in the discussion section, building on the review of theories and methods in intention recognition throughout the paper.

4 Psychological Theories on Intention Recognition

Understanding the intentions of others is a natural ability of humans and is studied extensively in the field of psychology. [Hei13] suggests that people have a "folk psychology" which they use to infer the meaning behind the actions and ideas of other people for a given situation. Following this is the prominent work done by [PW78] to coin the term Theory of Mind (TOM), which describes it as the ability to recognize the mental states of other people. This work was further confirmed

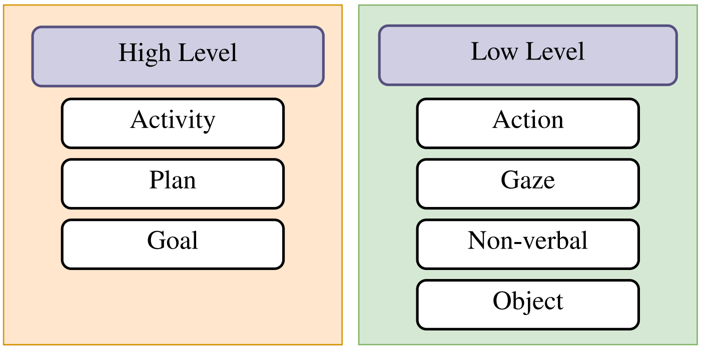


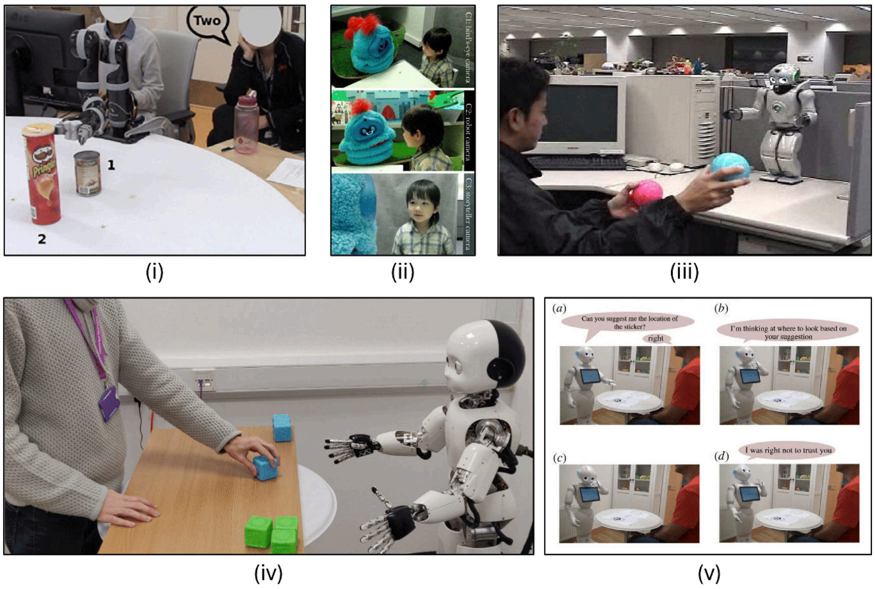
Figure 3: An image of the categorization of intentions is presented in this paper. In the figure, we have categorized intention into high-level and low-level. High-level indicates broader goals and long-term plans behind a person's actions and includes activity, plan, and goal in this category. Low-level indicates observable short-term actions and comprises action, gaze, non-verbal, and object-based subcategories. These categories focus on recognizing human intentions. In addition to human intentions, robots can also exhibit intentions, which are either explicitly programmed or inferred through mental states attributed to them by humans. We refer to this category as robot-based intentions, which are discussed in detail in Section 3.3.

by [WP83] as they studied and established the ability of children to associate relationships between two or more mental states by the ages of 4 to 6 years. [PBCR95] showed the relation between mental state association and autism in children, where children with autism fail to associate mental states with others. [BC+95] further went on to associate the inability of autistic children to assign a mental state to others with the failure to read intention through eye gaze.

A rather interesting outlook on intention is from the Wittgensteinian perspective as argued by [Kal19]. Here, Kalis explores the concept of intention from a philosophical viewpoint, challenging the traditional cognitive science approach that often tries to identify neural correlates of intentions. Traditionally, intentions are abstract, discrete mental states that cause actions. Kalis proposes that intentions are not mental states located in the brain but are better understood as patterns of behaviour extended over time and context. If intentions are not discrete mental states but rather patterns of behaviour, the difficulty in finding neural correlates for intentions makes sense. The paper suggests that neuroscience may be looking for something that does not exist in the way it is traditionally conceived. If this is true, then low-level intention recognition performed by artificial agents by observing the motion of other agents with respect to contextual understanding would be the correct direction of research for intention recognition in social robotics.

Another significant theory in the study of intention recognition is the mirror neuron mechanism theory [RC04] which states that humans show evidence of having a mirror neuron system which fires neurons when a particular action is observed as well as performed by them. This system of neurons is involved in action observation, imitation, and possibly in the understanding of language.

5 Computational Methods


Two
1
2
C1: bird's-eye camera
C2: robot camera
C3: storyteller camera
(a)
Can you suggest me the location of the sticker?
right
(b)
I'm thinking at where to look based on your suggestion
(c)
(d)
I was right not to trust you

(i) (ii) (iii)

(iv) (v)

Figure 4: Images showing psychological intention recognition from relevant references. (i) demonstrates a robot executing a trajectory to reach an object and the human predicts the intention during this motion [Per+11]. (ii) shows a storyteller and listener type experiment where the children tell stories to the robot which conveys attentiveness in a Bayesian theory of mind approach [LSB19]. (iii) shows a robot mirroring the user's hand motion to learn movement patterns [TFA10]. (iv) demonstrates a block-building game where the robot predicts the intentions of the user and collaborates to complete the task [VTZ16]. (v) shows an experiment where the robot either trusts or does not trust the human based on its belief derived from the theory of mind and verifies this via actions [VGC19].

The study of intention recognition has not been limited to theoretical psychology; it has also inspired the development of computational models in robotics and artificial intelligence. Humans tend to anthropomorphise non-living objects [EWC07, Cro+19] to better understand and explain what is not readily explainable. Hence, it is safe to assume that social robots with cognitive capabilities resembling that of humans would be an interesting research topic.

The credit for applying the theory of mind on a humanoid robot for the first time can be given to [Sca02] who relied on the models of the theory of mind proposed and discussed by [Les94] and [CA98]. He proposed that using the theory of mind would not only benefit interaction by allowing effective communication but also allow the artificial agent to learn from these interactions. To implement theory of mind in robots, Scassellati systematically used modules representing the various aspects of the traditional psychological models of theory of mind such as eye direction detection, intentionality detection, shared attention mechanism, as well as recognizing human actions and taking perspectives [Sca02, NAH02, Sch96, JD05, FMJ02, Bre+05, Gra+05, Tra+06].

In recent years, the architectures inspired by the theory of mind have gotten increasingly complex. [GBB07] and [SKK08] used the Leonardo robot to infer the beliefs, desires, and intentions of the human partner in a collaboration task using real-time behaviour. The task involves two other participants. All the participants and the robot observe chips and cookies being hidden in their respective boxes. One participant goes out and the other participant swaps the contents of the boxes in the presence of the robot and leaves the room. The other participant returns and goes to the box which had chips earlier and now has cookies. The robot's task is to best assist the participant in achieving her goal. In this case, the robot needs to reason about the belief of the participant that there are still chips in the box. With this belief and the truth it knows from observing the other participant make the swap, it needs to assist the participant by giving them the chips or cookies from the additional set it has.

In their review, [BO19a] talk about seven different implementations of complex architectures in robots that provide advantages in terms of false beliefs, active perception, learning preferences, and proactive behaviour in interactions [BO19a]. [BO19b] also talk about improving social robotics architectures by incorporating the teleological theory which is used to infer intentions behind others' actions based on the outcomes of these actions and the simulation theory which is used to simulate the mental states of others internally to understand them.

Various other computational models of the theory of mind have also been developed to understand the mental states and cognitive processes [POWW21]. One such model is the Bayesian theory of mind (BToM) model by [BST09] which formulates the problem of understanding actions as a Bayesian inference problem. The method is to use rational probabilistic planning in Markov decision problems to model the causal relationship between goals, actions, and beliefs. This can then be inverted to figure out the goals and beliefs of the agent.

They in turn also test the effectiveness of the Bayesian theory of mind model in mentalizing in two experiments [Bak+17]. In the first experiment, they tested participants' ability to jointly attribute beliefs and desires to an agent based on their observed actions. Participants viewed animated scenarios where an agent (e.g., a student) navigated through a grid-like environment with obstacles and food trucks. They were asked to infer the agent's desires (preferences for different food trucks) and beliefs about the unseen parts of the environment (e.g., whether the agent thought a specific food truck was behind a building). The BToM model was used to predict these inferences, and its predictions were compared to human judgments. The results showed that BToM could quantitatively predict participants' judgments, supporting the model's validity.

In the second experiment, they tested whether participants could use their theory of mind to infer both the agent's beliefs and percepts (what the agent could see) and aspects of the environment that only the agent could perceive. Here, participants observed the agent searching for a preferred food cart in a more complex environment where the locations and availability of the carts were hidden from the participants but observable by the agent. Participants had to infer the location of all carts based on the agent's actions. The BToM model's predictions were again compared to human judgments, with results showing that BToM successfully captured the complexity of human reasoning about the agent's percepts and the hidden state of the world.

Probabilistic and bayesian methods have also been used in other works for intention recognition [Sin+20, Per+11, Tah06, SH05, KH10, JA18, Dun+15, Tam+12]. These methods are beneficial for modeling uncertainty and making inferences, which is often the case in real-world scenarios. Furthermore, works by [Top17] and [CT16] provide additional proof-of-concept studies in this area, but did not appear in our primary literature search based on the keywords used.

Another work in BToM is done by [LSB19] in nonverbal communication where they propose a dual computational approach to model the interactions between a storyteller and a listener. The storyteller's role is modelled as a Partially Observable Markov Decision Process (POMDP), where they use nonverbal cues to influence and infer the listener's attentive state. Whereas, the listener's role is modelled using a Dynamic Bayesian Network (DBN) [AZN98, For+95, PW13] with a myopic policy, focusing on conveying attentiveness and influencing the storyteller's perception. They showed that their storyteller model outperformed state-of-the-art attention recognition methods whereas their listener model communicated attentiveness to the audience better than traditional signalling methods. Dynamic Bayesian Network is derived from Bayesian Network which is often used in intention recognition models as it helps solve uncertainty based problems [Pea14].

[VGC19, Vin+19] proposed and formed a cognitive system for artificial agents based on developmental robotics that incorporated the theory of mind, trust, and episodic memory. They address the less-explored scenario where a robot, rather than a human, acts as the trustor in a joint task, assessing the trustworthiness of its human partners. The model is tested through experiments that simulate a developmental psychology task. The proposed model achieves performance similar to children 5 years of age and older. The experiment involved identifying helpers from tricksters based on the cues provided.

With the popularity of neural networks in the past couple of years, researchers have explored the development of cognitive architectures using these techniques. This is what was proposed by [Rab+18] in the form of ToMnet that makes use of a meta-learning approach to build models to infer an agent's mental state. The model originally learns a strong prior for the future behaviour of the agents and can then produce richer predictions about them from a small number of observations of the agents' behaviour. Experiments show the effectiveness of this model as it passes standard theory of mind tasks involving recognising holders of false beliefs.

An interesting use case of the computational theory of mind model is to infer the mental states of multiple agents involved in urban search and rescue tasks which was researched by [Li+22a]. The model they proposed uses Deep Neural Networks to represent and update beliefs, predict actions, and generate inferences based on team behaviours. The study validated the ToM model by comparing its performance to human observers and found that the model outperformed the average human inferences on several tests.

In robotics, [TIS04] use recurrent neural networks with parametric biases (RNNPB) to implement mirror neuron-like systems for robots. Their findings suggest that the RNNPB model could provide insights into how memory consolidation occurs in biological systems and how complex behaviours can emerge from simpler learned patterns. [RPF13] also worked towards a similar implementation with a multi-layer connectionist model for an iCub robot. The model emphasizes bidirectional interactions between visual and motor areas, implementing a learning algorithm inspired by the biologically plausible GeneRec algorithm. The model was tested in a simulated environment where the iCub robot learned to perform grasping tasks and the experiments showed that the model successfully linked visual and motor representations, allowing the robot to recognize and understand actions from various perspectives. [HK10] combine mirror neuron and simulation theory for intention reading of humans in human-robot interaction scenarios.

Work done by [Pre+19] explore the concept of mental time travel (MTT) which refers to the ability to mentally project oneself into the past or the future. They propose that the underlying brain systems for this concept can be modelled computationally and describe the implementation of a multimodal memory system in the iCub robot based on Gaussian process latent variable models. Through experiments designed around face recognition, speaker recognition, emotion recognition, touch interaction, and action recognition, they demonstrate the viability of their system.

Overall, the advancement in computational methods, starting from implementations of theory of mind in robots to complex neural network architectures, have contributed significantly to the advancements in the field. These methods have enabled artificial agents to infer human intentions, enhancing human-robot interaction and collaboration across various tasks. However, despite significant progress, developing systems that provide a comprehensive prediction of human intentions remains a challenge. In the following sections, we will talk about the high-level and low-level categorization of intention recognition followed by the challenges and limitations in greater detail.

6 Intention Recognition

Intention recognition in its entirety is a complex topic to analyze. To facilitate a systematic and comprehensive understanding, we categorize it into high-level and low-level processes. This distinction has been discussed below.

6.1 High-Level

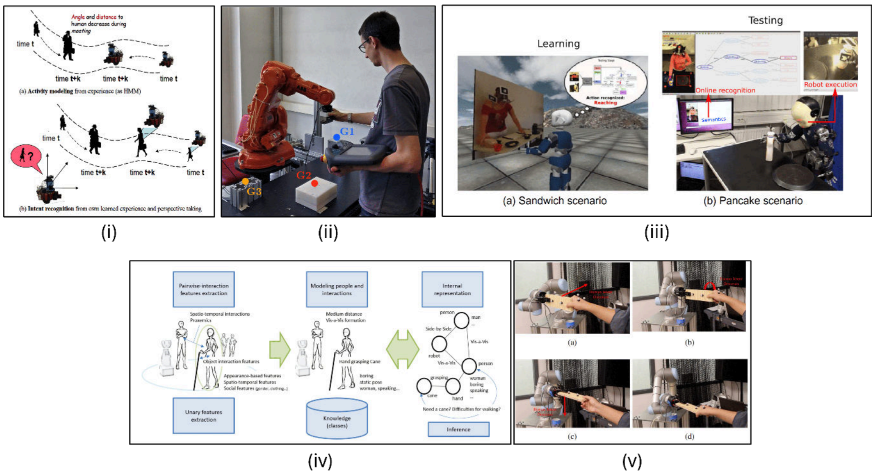
High-level intention recognition focuses on comprehending the broader goals and plans behind a person's actions, enabling systems to grasp complex and abstract intentions that often span longer timeframes [KS08, Saf+15b, Saf+15a, GCR21]. In this category, we have included papers from psychological intention recognition which describe conceptual works related to *intention recognition*, *activity recognition* which looks at a series of actions to achieve a certain goal, *plan recognition* which focuses on inferring the steps an agent is taking or supposed to take to achieve its intentions, and *goal recognition* which involves determining the end goal in terms of an achievement or an object that marks the end of the intention of the agent.

Activity

Before delving into activity recognition, there is a need to clarify the difference between *action* and *activity recognition* and that quite often in research these are used interchangeably. There is also a need to understand the hierarchy of steps that move from simple to complex to finally arrive at intent recognition.

The first step is the conceptually simplest form which is gesture/motion recognition. This can be defined as the human intent to move any single body part to achieve a goal. The next step, action recognition, involves identifying the intent behind performing a specific action. This consists of one or more gestures in a sequence. Activity is a series of actions performed by the human to achieve a goal and activity recognition comes as the third step in the intent recognition process. Once the activity has been recognised, based on the context of the activity, the goal of the human can be predicted in turn leading to recognising the underlying intent which comes as the final step of this process.

Activity recognition refers to the process of identifying and classifying specific activities that an individual or agent is performing. For decades, researchers have focused on recognising activity (walking, cooking, typing, etc.) as a first step to further determine why the person is doing those activities and infer their underlying goals, plans, or intentions. [Kel+10] base their work on the theory


time t
time t+k
(a) Activity modelling from experience (as HMM)
(b) Intent recognition from own learned experience and perspective taking
G1
G2
G3
Learning
Testing
Online recognition
Robot execution
Semantics
(a) Sandwich scenario
(b) Pancake scenario
Pairwise-interaction features extraction
Modeling people and interactions
Internal representation
Unary features extraction
Knowledge (classes)
Inference
(a)
(b)
(c)
(d)

(i) (ii) (iii)

(iv) (v)

Figure 5: Activity-based intention recognition examples from relevant references. (i) demonstrates a robot modeling the activity using HMMs that it needs to recognize later [Kel+08]. (ii) shows manual guidance of the robot to indicate possible goals for recognizing user's intention to follow certain trajectories [NZR18]. (iii) shows two cases - (a) is the learning stage where the robot automatically segments and recognizes human activities, and (b) which compares different scenarios and finds the semantics of observed activity to remain the same as (a) [Rak+19]. (iv) shows the major modules of a system used for modelling and inferring intentions [TC17]. (v) demonstrates an experiment where the human and the robot were holding a piece of wood together and the robot tried to recognize the intentions of the human to move and performed similar motion [Wan+19]

of mind and introduce a system which uses vision-based capabilities and Hidden Markov Models (HMMs) to model and recognize human activities. An interesting feature of their system is the ability to disambiguate similar actions by using contextual information. This allows for identifying different underlying intentions for similar-looking activities. HMM-based intention recognition is also seen in many other notable works [Cra+18, Kel+08, Tav+07, Pet+19, SPB10, AK06, BED08, ZSS08, ZCS08].

[RABC15] proposed a framework to perform activity recognition using a humanoid robot. The framework enables the robot to use semantic reasoning to infer high-level behaviours from low-level sensor data. They implemented this framework on an iCub robot and tested it in three experiments - pancake making, sandwich making, and setting the table, and showed that the robot can correctly segment and recognize human activities even with a simple perception system.

Delving into the details of human activity recognition (HAR) requires its own review. The research in this field has provided an in-depth understanding of various methods and techniques used for HAR. We have limited the focus of our paper to works where it is used for intention recognition. [SRG17] provide an in-depth analysis of activity recognition in videos. They examine and talk about various aspects of activity recognition such as datasets, evaluation metrics, and algorithms, and also talk about challenges in the field. Existing datasets like Charades, ActivityNet, and THUMOS are reviewed by the authors in this work. The authors provide valuable insights regarding the ambiguity of temporal boundaries of activities (difficulty in telling when an activity actually starts and ends) as well as errors in state-of-the-art activity recognition algorithms due to similar activities. Similarly, [VNK15] also provided a comprehensive review of the different methodologies used for HAR in the fields of computer vision and robotics. They also emphasize the role of publicly available datasets in advancing HAR research, discussing the characteristics of an ideal HAR dataset that should include a variety of activities, contexts, and environmental conditions. In comparison, [JBD19] provides a review of state-of-the-art methods used in HAR with a focus on traditional machine learning techniques such as Decision Trees, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Hidden Markov Models (HMM), as well as neural network based techniques such as Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), and Recurrent Neural Networks (RNN). Some notable works using SVMs, Decision Trees, KNNs, and other rule based algorithms can be seen in the works [DSS15, Wan+17, Völ+15, KM17, MÅ18, NZR18, Zha+20, Wan+19] whereas those for neural networks such as RNNs, CNNs, and Extreme Large Machine Algorithms for intention recognition are present in [Zha+19a, LH19, Gol+19, NZR18, Var+18, Wan+18, Völ+16, PL16, Li+17, Rak+19, Yan+21]. In the field of social robotics, a review of the recent approaches and techniques is presented by [Tap+19], focusing on how robots can perceive and understand human activities and social interactions. They discuss various methodologies for HAR, both at the individual level and within groups and focus on the importance of context in accurately interpreting human behaviour, noting that the same action can have different meanings depending on the social or environmental context.

Overall, the advancements in the field of activity recognition, from traditional machine learning techniques like Decision Trees and KNNs to modern deep learning approaches such as ANNs and RNNs, have improved our ability to recognize intentions. However, accurately recognizing and interpreting human intentions still remains a complex challenge due to factors like ambiguous actions, lack of generalization, model complexity, and the need for real-time processing. In the next section, we will explore how plan is used for intention recognition followed by a detailed discussion of the potential challenges in the field.

Plan

Figure 6: The robot learns to predict the human's intention—specifically when they deliberately plan to obstruct the robot's path [Par+16].

Inferring the underlying plan behind an action or a series of actions [AA07, Sad11] allows robots to proactively support their human counterparts in the task at hand.

[OMM19] developed a proactive intention recognition system to support operators in joint human-robot search and rescue missions. The system focuses on the movements of the human responder and concentrates search around those areas to increase search efficiency. The authors propose an intention recognition paradigm based on Monte Carlo planning techniques and POMDP environments that supports the robot's exploration strategy by providing an entropy reduction bonus to the reward function. Testing the system in various simulated environments with a drone shows improved efficiency of search and rescue compared to other baseline techniques.

Similarly, [Zha+23] address multi-agent intention recognition by introducing landmarks in the behavioural model to identify common intentions among agents in a multi-agent system. They modeled the environment using MDP and defined intention models through behaviour trees and landmark-based intention models which are further refined with a robust clustering mechanism for grouping the intentions of multiple agents and recognising them successfully. Experiments performed on two separate systems show improved performance of agents in tasks requiring collaboration.

[Den+19] presented a method to enhance navigation assistance for wheelchair robots in complex environments. This was done using clothoidal (Euler spiral) paths for narrow doorways and dynamic obstacles. Earlier work done by the authors used a circular local path template to suggest possible paths for robots but was limited in its ability to plan around complex environments. By introducing a Local State Lattice which generates a set of clothoidal paths, they provided smoother transition curves compared to circular paths which are also better at finding paths through narrow passageways. The advantage of using clothoidal paths is that it opens up new path estimations in complex environments which can then be used to infer the intentions of the wheelchair user and support them.

In the context of social robotics, [Par+16] proposed an innovative navigation approach by predicting human intentions and adjusting their navigation strategy to avoid potential conflict such as a human blocking the robot's path intentionally. The authors train a classifier offline capable of predicting if a human plans to interact with or block the robot. They use Gaussian process models to classify intent as well as regression to predict future trajectories which are then used in the robot's path planning for successful navigation.

In the case of multi-agent systems, [AA14] explored the concept of intent recognition and argued that intent recognition is more than just recognising the plan of the agents. They use a plan library and use it to compare the observed actions of the agents and test three types of agents (no recognition, plan-recognition, and intent-recognition) using the Repast Simphony platform on a collective box-pushing task. Results showed that intent-recognition agents were able to assist in more flexible ways, with actions they had not observed before. We will further look into the challenges and limitations concerning intention recognition in the later section.

Overall, the advancements in the field of plan recognition, from Monte Carlo planning techniques to using plan libraries, have improved our ability to recognize intentions. However, accurately recognizing and interpreting human intentions still remains a complex challenge due to factors like computational overhead, high false-positive rates, and desired adaptability to dynamic environments. In the next section, we will explore how goal is used for intention recognition followed by a detailed discussion of the potential challenges in the field.

Goal

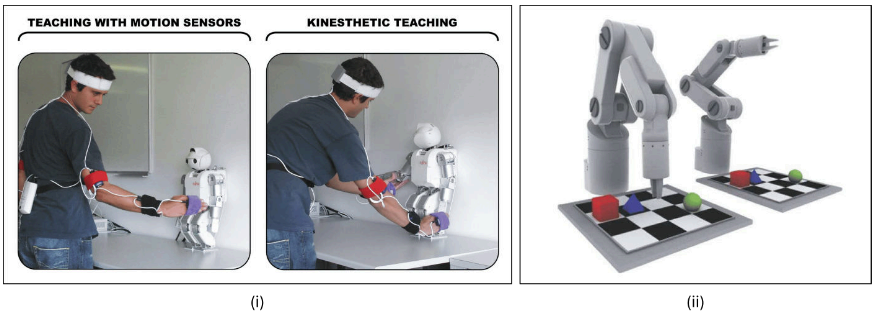


Figure 7: Image depicting goal recognition from relevant references. (i) shows two different types of teaching methods for replicating motions [CB07]. (ii) shows recognizing goal using a demonstrator and imitator combination [JB06a]

Goal recognition involves inferring the most likely goal from a sequence of observed actions. This is essential for robots to actively provide support in collaborative task settings. We will now discuss the various frameworks and models proposed in order to support intention recognition through goals. [Mur02] proposed a framework that uses DBNs to model and predict a person's goals by analyzing their movement patterns within a known environment. The environment is monitored using stationary laser range finders, which track the person's (x, y) position as they move through the space. The DBN model's purpose is to predict which predefined landmark (or goal) the person is heading toward next.

Building upon the idea that goals are crucial to understanding human actions, [BWG00] suggested that imitation is goal-directed and not simply a copy of the observed actions. Based on this rationale, [Bre+06] and [Erl+06] have shown learning to imitate actions relies on the ability of the imitating agent to infer the intentions of the demonstrator. However, not all motions are goal-directed, and intention recognition in such cases requires building a mental model of the demonstrator [JB06a, CDI06] as well as carefully analysing social cues as supportive knowledge for accurate intention prediction [SND06, Bre+06, CB07].

Extending upon the idea of imitation and goal inference, [JB06b] proposed a computational model to understand the underlying goals behind an action and imitate the demonstrator. The model proposed by the authors requires multiple interactions between the demonstrator and the imitator to learn the underlying goal and this is achieved by the imitator by considering the perspective of the demonstrator and maintaining and updating a model of the beliefs of the demonstrator. The model was tested in a simulated environment, showing that the robot could effectively learn and infer intentions over repeated imitation games. However, the model requires a deterministic environment and a large number of interactions which poses multiple challenges of scaling this approach to more complex, real-world scenarios.

While the previous models focus on imitation and mental modelling, another perspective involves assessing the rationality of actions. [BDK14] focus on computational models that simulate how humans detect if an action was intentional and predict the likely goal based on these observed actions. The authors assume that people follow the principle of rationality and propose a model that can detect the intentions of humans from observed sequences based on the rationality of actions. The two main components of their models are determining whether the actions are intentional by measuring how efficiently they move towards a goal and predicting the goal from an incomplete sequence of actions. The authors tested this model with 140 participants and found that the performance of the model closely matched that of the human participants in determining the intentionality of the actions and the predicted goals.

In another approach where the agent's behaviour is dynamic and unpredictable, [Zen+18] presented a novel work on goal recognition in dynamic network interdiction scenarios using Inverse Reinforcement Learning (IRL). The task involves inferring an agent's goals from observed actions. Traditional methods often rely on predefined libraries of plans or policies, but as mentioned, this approach is limited when the agent's behaviour is dynamic and unpredictable. The authors propose using IRL to model the opponent's behaviour based on observed trajectories, allowing for more accurate goal prediction. IRL is used to learn the reward function that best explains the observed behaviour of the opponent. This reward function is then used to predict future actions and goals. The proposed method was tested using human movement data on a simulated Chicago road network in which the IRL-based behaviour model outperformed other models in tracking accuracy and effectiveness in network interdiction.

Addressing goal recognition in smart environments, like home settings, [SBP22] present a framework for reasoning about intentions using probabilistic logic programming. The model proposed by the authors considers various observable actions and environmental properties to predict the intentions of the users, such as the intention to make coffee or prepare a meal in a smart home setting. When tested in this setting with full observability, the model achieved intention recognition accuracy of 75 per cent which decreased as the observability was reduced but the model still remained robust.

Additionally, when encountering uncertainty in collaborative tasks, [Bra+22] proposed a method to reduce uncertainty in robot's planning due to noise and a lack of knowledge of the human's goal. The method uses an unscented Kalman filter for state estimation, an HMM-based model for goal recognition, and a model predictive belief space controller based on Belief Iterative Linear Quadratic Gaussian (i-LQG) for control. The approach is evaluated in a simulated scenario where a mobile robot cooperates with a human in a two-dimensional space demonstrating that the proposed system can effectively reduce uncertainty about the human's goal while pursuing the cooperative task. [TFA10] also used a Kalman filter to avoid collisions between robots and humans in a cooperative environment based on the human's intentions.

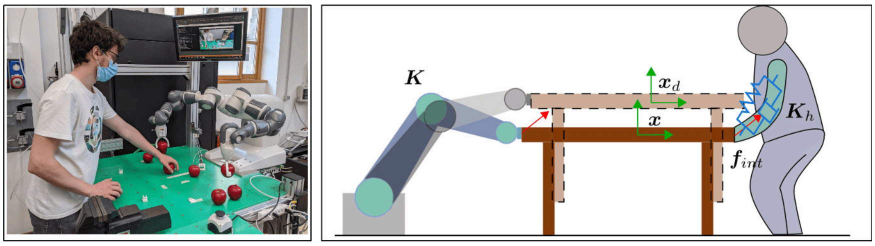
In an effort of addressing explainability in goal recognition, [AMV23] presented a model that focuses on providing answers to why a certain goal was recognized but not a certain other goal. The proposed model adapts the Weight of Evidence (WoE) framework from information theory to predict goals by weighing the strength of evidence of one goal hypothesis over the other based on the observed actions. The model is tested on eight goal recognition benchmark domains and a separate human study, showing that it can generate explanations efficiently without significantly increasing computational overhead and the explanations generated by the model improve participants' understanding of the model's decisions and enhance their trust in the system.

Furthermore, [ZKL23] explored goal recognition while focusing on the timing of the action taken by an agent to carry out certain actions. The proposed algorithm is evaluated using both synthetic data and real human data and the results suggest that incorporating timing information significantly improves goal recognition accuracy, particularly in scenarios where only a few actions have been observed.

Traditionally, goal recognition has relied on model-based methods, however, [Chi+23] present a deep-learning approach to goal recognition. The authors introduce GRNet, a system that uses an RNN architecture to process a sequence of observed actions and predict how likely each possible proposition (goal) is part of the agent's goal. To perform goal recognition within a domain, the network is trained once in the given domain. The model shows better accuracy and runtime compared to the state-of-the-art system - LGR, and combining the model with LGR further enhances performance, especially in incomplete or partial information scenarios. We will further look into the challenges and limitations concerning intention recognition in the later section.

As observed, intention recognition has been applied in numerous applications and still finds use in many more [Sad11] such as understanding stories [CG90], human-computer interaction [AA07, Hon01, Les98], monitoring traffic [PW13], assistive care [Gei02, Hai+03, Per+10, PH11, PA11, Roy+09, Tah06], and military activities [Hei04, MG04].

Overall, these works highlight significant advancements in goal recognition methodologies, ranging from probabilistic and other traditional model-based methods to deep learning approaches. By inferring human goals through goal recognition, these models enhance the ability of robots and AI systems to collaborate with humans across various domains and task settings. In the next section, we will look into low-level intention recognition followed by the challenges and limitations concerning intention recognition.



(i)

(ii)

Figure 8: Images showing action-based intention recognition from relevant references. (i) demonstrates a robot inferring the most likely future destination of the object picked by the human from recent change in position [For+21]. (ii) shows a robot moving a shared load using intention prediction in a number of ways - inferring direction and speed from dynamic action, following desired path from learned trajectory, and matching gain with the human for a neutral interaction [Li+22b].

6.2 Low Level

In contrast to high-level intention recognition, low-level intention recognition deals with more immediate, specific actions and behaviours, providing a granular understanding of the moment-to-moment intentions [KS08, Saf+15b, Saf+15a, GCR21]. In this category, we have included papers from *action recognition* which focuses on inferring specific actions of the agents, *gaze intention* which focuses primarily on using the gaze of the agents to infer their intentions, and *non-verbal intention* which focuses on other body parts used to infer intentions.

Action

Low-level intention recognition focuses on identifying immediate, specific actions or gestures that provide insight into a person's intentions. This type of recognition is often subdivided into various approaches, each targeting different aspects of action analysis, such as recognizing simple gestures or sequences of movements. For example, action recognition examines short-duration actions, and various methods have been developed to predict the intent behind these actions, using technologies like computer vision, neural networks, and sensor data. The subdivision into techniques allows for more precise and specialized intention inference, enabling applications in fields like human-robot collaboration, assistive robotics, and teleoperation.

As described earlier, action refers to a short-duration sequence of gestures or a single gesture. Building on this understanding, [Zun+17] introduce a novel Intent from Motion paradigm which uses just the initial motion to predict the intention without any contextual information. This may sound impossible, but in their experiments involving reaching for a water bottle with 4 different intentions, the computer vision methods outperform humans in predicting the intention behind the observed action. This shows how subtle motion cues can reveal intentions and how technology can be used to predict these intentions in real-time.

Expanding on the use of motion cues, specifically for the human arm, [RD17] used actions to infer intentions and define intention as the 3D end goal of the actions performed by the user. The authors use a neural network to model the non-linear dynamics of the human arm motion and treat intention inference as a parameter estimation problem in a state-space model. They use an extended Kalman filter for expectation maximization to infer the intentions accurately and test the methodology in two studies, human-robot collaboration and assistive robotics, with each study having 3 experiments. The Extended Kalman filter has also been used by [RTD19] in their later works and by others such as [AH10] as well. Parameter based methods for intention recognition have also seen popularity over the years [KH10, Per+11, JA18, Dun+15, Tam+12].

In the scope of probabilistic modeling, [Tah06] proposed a system where machines can infer human intentions using a modified intention–action–state scenario modelled by DBNs. The paper presents an example of a human commanding a mobile robot remotely. The robot uses DBNs to recognize the human's intentions (e.g., moving towards or away from an object) and adjusts its actions to comply with these intentions. The interaction is tested in a simulation environment in which the robot demonstrates compliance with the recognized intentions, leading to smoother and more effective control.

Alternatively, [Omo+08] take a different approach to intention recognition by making the robot perform actions which are easy to interpret by the user, hoping to induce a known response from them. They term this approach as active intention leading (AIL) and argue in its favour, stating that, unlike traditional intention recognition which requires one to observe the other user to infer their intentions (passive intention recognition), this approach is computationally lighter since it works towards the self. Testing this approach on a hunter game with two different tasks shows the effectiveness of AIL in improving the performance of catching the prey earlier.

Further, to minimize uncertainties in human intention recognition, [JKC08] proposed a unique approach using ontology-based hierarchical user intentions. The authors use a RuleML-based intention recognition module which combines domain knowledge and sensor information such as temperature, humidity, vision, and auditory to infer the intentions of the humans from their actions. They also use conditional entropy to make the robot's behaviour proactive by selecting the appropriate actions based on the user's intentions. Testing the system on their simulator under three different scenarios - when multiple intentions are possible, when all candidate intentions are considered, or when only one intention is evident, results in successful recognition of user intentions, even in complex or ambiguous situations.

In a further attempt to understand the human intention even when the perception system is rudimentay, [RABC15] explored semantic perception in their work. The proposed system has two levels - a low-level feature extraction that uses colour to extract simple actions such as move, not move, and tool use, and a high-level activity recognition that uses a decision-tree approach to infer higher-level activity such as cut, pour, etc. The framework was implemented on an iCub robot and tested against human participants in three tasks - making pancakes, making sandwiches, and setting a table. The robot achieved 85 per cent accuracy in inferring human intentions and took about 0.12 seconds to make decisions and perform similar activities.

In an attempt to show feasibility in real-time application, [Der+17] in their work with the iCub robot presented a method for enabling the humanoid robot to predict human intentions during physical interaction using Probabilistic Movement Primitives (ProMPs). The primitives are learned in simulation where humans demonstrate multiple actions which can then be inferred in the very early phase of the action with the help of the ProMPs. Since ProMPs do not require explicit knowledge of the goal, the implemented method worked well on the real iCub robot when being tested on tasks involving reaching for objects and sorting.

In a similar attempt for teleoperation and shared control tasks, [TC17] presented a novel method for improving the performance of teleoperated robots in remote manipulation tasks. This is done particularly in challenging environments where communication delays, limited bandwidth, and environmental differences pose significant issues by using a task-parameterized generative model. The authors introduce a hidden semi-Markov model (HSMM) that learns from teleoperator demonstrations by segmenting the demonstrations into meaningful parts and encoding the transition patterns among these segments. This allows the robot to either assist the teleoperator through shared control or autonomously perform tasks which was tested on a Baxter robot in tasks such as reaching a movable target and opening a valve. Some other works in teleoperation and intention recognition are [JA18, SPB10, AK06].

Additionally, in order to address the ambiguity of indirect speect acts (ISAs) within the context of task based interactions, [BWS17] proposed a rule-based mechanism to understand directives from indirect speech acts and evaluate the framework in experiments involving simple tasks of knocking over coloured towers.

In warehouse environments, [PMP18] presented a framework for recognizing human intentions based on the theory of mind using Markov Decision Processes (MDP) to model the decision-making process of warehouse workers and a Hidden Markov Model (HMM) that estimates the worker's intentions based on observable actions. Testing the framework in simulated scenarios shows high accuracy in predicting the intentions of the workers even when they decide to change their goal midway.

Exploring shared control in physical human-robot interaction (pHRI), [Los+18] presented a comprehensive review, focusing on three critical components: intent detection, arbitration, and communication. The intent detection part of the review explores in detail how robots can detect human intentions, either binary or more complex and continuous ones, using force sensors, muscle activity sensors, and neural activity.

Based on neural network models, [LH19] presented a method to recognize human motion intention using a Radial Basis Function Neural Network (RBFNN) model. The model learns offline from the data collected using the adaptive impedance control method and is later used online to infer the desired velocity of the human limb during the collaborative task. Testing it against the adaptive impedance control method shows better performance in terms of improved synchronisation and reduced force applied by humans. Others have also been interested in using RBFNN for intention recognition [LG13, Jan+14, CNP06, Zha+19a, Gol+19, NZR18, Var+18, Wan+18, Völ+16, PL16, Li+17, Rak+19, Yan+21]. Control methods for intention recognition have also seen widespread interest in research [Hua+15b, Li+18, Len+12, Li+17, Wil+17, Wei+19, Buc94].

Further, with the approach of deep neural networks, [LZD20] proposed using a DNN to process RGB images and optical flow for intention recognition. The authors use two-stream architecture - a spatial network to look at skeleton joints for spatial information and a temporal network to look at optical flow for temporal dynamics and information. The output from both the networks is fused using average fusion and achieves intention recognition accuracy of 74 per cent on their dataset and 77 per cent on the Intention from Motion (IFM) dataset.

Focusing on enhancing human-robot interaction (HRI), [Li+21] presented an approach that enables robots to better understand human intentions through the integration of visual semantics and natural language processing (NLP). The authors use image segmentation to classify objects in the field of vision of the robot, use rule matching and Conditional Random Fields (CRF) to parse natural language instructions and combine this information to perform the task while looking at feedback from the users through their facial expressions.

Comparing neural network architectures for human intention prediction, [PF23] evaluated three different architectures - Long Short-Term Memory (LSTM), Transformer, and MLP-Mixer - specifically in predicting arm movement. A custom dataset was prepared using a VR environment where the participants were asked to perform hand movements towards boxes where their gazes were tracked and the gaze direction, head position, and controller position were recorded. After training the neural networks on the data, their performance based on accuracy, movement classification accuracy, and ahead-of-time movement prediction was compared. The transformer encoder model performed the best with 82.74 per cent accuracy for predicting movements and 80.06 per cent in correctly classifying the movements at least once. The next close performer was the MLP-mixer and had lower computational complexity than the transformer model. The LSTM model was the worst-performing model of the three.

In a similar setting, [God+22] presented a novel approach for recognizing hand gestures based on the lightmyography (LMG) signals and transformer based deep learning models. The authors use two transformer based deep learning models - Temporal Multi-Channel Transformer (TMC-T) and Temporal Multi-Channel Vision Transfer (TMC-ViT) - to classify hand gestures and compare these models against other machine learning models such as Convolutional Neural Networks (CNN), Bidirectional Long-Short-Term Memory (Bi-LSTM), Linear Discriminant Analysis (LDA), Support Vector Machines (SVM), and Random Forests (RF). The two transformer models outperform other models and achieve accuracies of 94.03 per cent and 93.69 per cent respectively in subject-specific gesture recognition of five hand gestures (power, pinch, extension, rest, and tripod).

Focusing on multiple actions simultaneously, [WVS23] presented a multi-modal visual and tactile sensors-based system to recognize human intention from observed actions. In the proposed system, a supervised machine learning algorithm is used to train the model with various interaction characteristics, including touch location, human pose, and gaze direction and is then later used to classify whether a human touch is intentional or not with an accuracy of 86 per cent which is much more accurate than similar systems with single modules for inferring intentions.

Taking a novel approach by focusing on the robot rather than the human operator, [Tsa+23] proposed a Machine Learning Operator Intent Inference (MLOII) model that uses an offline supervised learning method to learn from the spatial data of the robot and then is used to infer the navigational intent online. When tested against the Bayesian Operator Intent Recognition (BOIR) method, MLOII performed better in cases involving fewer obstacles and more direct routes whereas under-performed in cases of complex environments with more obstacles and large areas.

Emphasizing learning from demonstration, [Wan+18] proposed a teaching-learning-prediction (TLP) model that allows robots to learn from human demonstrations and predict hand-over intentions in real-time using multi-modal sensory data from inertial measurement unit (IMU) and EMG sensors. During the learning phase, natural language instructions are used along with the sensor data to learn from hand-over demonstrations. The model learns from about 5000 sets of hand-over demonstrations from six subjects and shows a prediction accuracy of 99.7 per cent in predicting hand-over intentions even from partial motion.

In a later attempt to interpret motion cues, [Dua+18] explored the ability to anticipate actions by interpreting cues such as body movements and eye gaze. The authors conducted human-human experiments to collect data and study how different cues such as eye gaze, head orientation, and arm movement influence the ability to anticipate actions. They also present a computational model based on recorded human actions that uses Gaussian Mixture Models (GMM) to simulate the arm trajectories and incorporate eye gaze patterns to predict action intentions. This model was incorporated in an iCub humanoid robot and tested in a second experiment where the robot was the actor and the human participants had to anticipate the robot's actions which they were successful in doing when both body movement and gaze were used by the robot. Although this study shows the importance of actions in intent recognition, it does not directly address any system for intention recognition from actions in humans which can be implemented on a robot.

[VGC19] used the theory of mind-based architecture to perform intent recognition from actions. They performed low-level processing of human skeletal data to form clusters representing different postures which the robot observes during a training phase and encodes actions as sequences of transitions through these clusters. They used a high-level module to infer the human partner's goals based on the observed sequences of cluster transitions and a Hidden Semi-Markov Model (HSMM). This architecture was implemented on an iCub robot and tested in an experiment where the robot had to collaborate with a human partner to build structures using toy blocks. Experiments showed that the robot was successfully able to predict the intentions of the human partners and assist in building the structures with an average time of 4.49 seconds to make predictions.

In the application area of agricultural robotics, [GB20] explored the design of a robot behavioural controller that learns human intention from body pose detected using OpenPose and then classified into behaviours. The robot's belief-desire-intention (BDI) system [RG+95] decides the next course of action based on the classified behaviour of the human, such as delivering or exchanging crates or avoiding the human to prevent interference.

Similarly, in the field of rehabilitation robotics, [WC20] introduced an adaptive neural cooperative control strategy that integrates human motion intention into the control of a rehabilitation robot to improve the effectiveness of therapy especially for upper-limb patients. The authors obtain muscular forces using surface electromyography (sEMG) signals and process these signals through a Kalman filter. A Gaussian Radial Basis Function Network (RBFN) is used to estimate the motion intention from the above-filtered forces and make necessary adjustments to the forces of the rehabilitation robot to assist in therapy. The method was tested with 10 volunteers on an upper-limb rehabilitation robot which was able to accurately follow and adjust the assistance it provided to the participants when they were moving their limbs along a predefined trajectory.

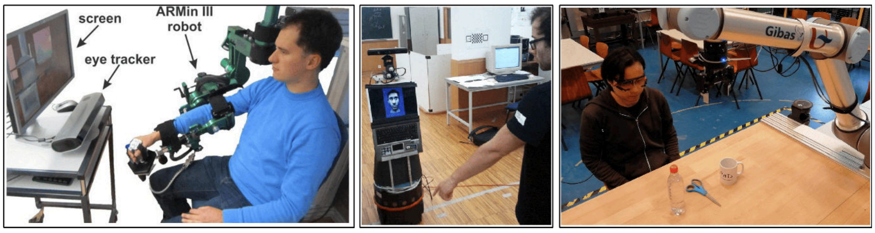
Moving further from just recognizing actions, [Liu+21a] introduced a multi-task model that integrates human action recognition and hand-held object identification to achieve more accurate and context-aware human intention recognition. The multi-task model has two sub-tasks, one that fuses Spatial Temporal Graph Convolutional Networks (ST-GCN) with Long Short-Term Memory (LSTM) networks (ST-GCN-LSTM) to effectively capture and recognize human actions from 3D skeleton data, and the other uses an improved YOLO v3 model to detect and identify objects that a human is holding. The authors evaluate the sub-tasks on datasets and the framework for human-robot interaction based on the multi-task model in a real-world setting to show that these models significantly improve action intention prediction accuracy.

Demonstrating a newer approach using deep learning algorithms, [For+21] proposed a method to use RGB-D camera and deep learning algorithms to track the 3D positions of objects in an environment and predict the likely location the human would place them. The predictions made by the model are based on the positions of key body joints (shoulders, elbows, wrists) over time, processed through a recurrent neural network (RNN) with a GRU-based encoder-decoder architecture. The proposed human intention prediction system when combined with a YOLOv3 model for object detection in a collaborative task with a dual-arm ABB Yumi robot reduced collisions by 38 per cent compared to a setup that only used human tracking and by 70 per cent compared to a system that only used object detection without any human intent prediction.

Based on the forces exerted during a physical human-robot interaction, [Lai+22] explored using physical human-robot interaction to estimate user intention. The authors proposed a novel method called Physical Human-Robot Interaction Primitives (pHRIP), which extends existing interaction primitives to capture the user's intent based on the forces exerted during the interaction. The method was implemented on a 7-dof robot arm and experiments with the model showed that the system was able to accurately predict user intentions while performing tasks such as target-directed reaching and obstacle avoidance. [Li+22b] went along a slightly different route and provided a comprehensive overview of the current state of interaction control in, what they call, contact robots that focuses on robots that physically interact with human users, particularly in tasks requiring close cooperation, such as rehabilitation, teleoperation, and collaborative manufacturing. The authors discuss the use of EMG, EEG, and tactile sensors to perform action intent recognition along with a combination of sensor inputs with machine learning models to classify intended actions.

Finally, [Ni+23] proposed a cross-view human intention recognition method that uses views of body and face from different angles and combines them to acquire meaningful semantics for intention recognition. The authors use a generative model to generate a different view from a given view, increasing the semantic information available as well as using an RNN to fuse spatial and temporal information for inferring the intentions of the user. The method is tested using a collaborative assembly task where the cross-view method significantly improves the fluency and efficiency of the robots and brings their performance closer to that of human participants. We will further look into the challenges and limitations concerning intention recognition in the later section.

Overall, these manuscripts highlight the advancements in action-based intention recognition, ranging from using Kalman filters to deep learning approaches such as transformers. By effectively inferring the intentions from observed human actions, these models enhance the ability of robots to work in social environments. In the next section, we will look into gaze-based intention recognition followed by the challenges and limitations concerning intention recognition.



(i) (ii) (iii)

Figure 9: Images showing gaze-based intention recognition from relevant references. (i) shows the human interacting with a robot which detects their intentions through an eye tracker [NR13]. (ii) demonstrates a robot understanding context and intention of the user from their gaze [Qui+12]. (iii) shows a robot understanding the intention of the human to pick up an object based on the gaze information acquired using an eye tracker [SCV19].

Gaze

Given the extensive research on gaze intention, it might appear as one of the most popular methods of intention recognition in the field of robotics. As in the case of human-human interaction, the eye and gaze direction plays a crucial role in interpreting attention and underlying intentions. Similarly, in the case of robots, this can be used to gain insights into the user's focus of attention, intention, and underlying mental state.

[Qui+12] talk about the importance of context awareness and intention understanding in robotic systems and how gaze estimation and gesture interpretation can be used as modalities to understand human intentions in different contexts. They specifically talk about the gaze estimation process that involves detecting the user's face, mapping it onto an ellipsoid model, and tracking the eye movements to estimate the direction of gaze. The gaze tracking system, implemented on a standard laptop with a webcam, is effective within a certain distance, providing a reasonable estimation of where the user is looking. With a predefined ontology, recognizing gaze as well as gesture helps the robot understand the context in which a particular action occurs, enabling a more accurate interpretation of user intentions.

The use of gaze in rehabilitaion systems is explored by [NR13]. The authors explored using gaze-based intention detection in a virtual environment (VE), allowing patients to interact more freely and naturally with the rehabilitation system. The system proposed by the authors utilizes the ARMin III upper extremity rehabilitation robot, which supports the patient's arm movements. The robot uses gaze data to infer the patient's intentions and provides the necessary support to complete the intended action.

Similarly, in a collaborative task setting, [Hua+15a] looked into using gaze patterns to predict a person's intention. The authors use support vector machines

(SVM) to classify and predict the customer's intended ingredient choice based on features derived from their gaze patterns achieving 76 per cent accuracy. The model was also able to anticipate the customer's request approximately 1.8 seconds before the verbal request was made.

Addressing assistive applications, [LZ17] proposed a novel framework to infer intentions from eye gaze to have a robot assist elderly or impaired individuals with activities of daily living (ADL). The system detects where a user is focusing their gaze, identifying "intentional gaze" (when the user looks at an object with the intent to manipulate it) versus "intention-free gaze" using a support vector machine (SVM) classifier based on features like gaze dwelling time, pupil size variation, and gaze speed. The framework was tested in a simulated homecare environment with a set of objects related to common daily tasks (e.g., making coffee, taking medicine) with intention inference achieving a correctness rate of up to 75 per cent.

Here, by combining gaze-based data with traditional model-based AI planning, [Sin+20] explored a novel approach to enhance the accuracy of human intention recognition. Gaze data is utilized to create a probability distribution over potential intentions. This model is combined with traditional model-based AI planning algorithms to predict intentions approximately 90 seconds earlier and with a 22 per cent increase in accuracy. The model combines both short-term (proximal) and long-term (distal) intentions into a single set to perform intention inference. Even during semi-rational or deceptive gaze behaviours, where individuals might try to mislead the system by looking at irrelevant areas, the combined model still performed better than using gaze or action data alone, showing a 9 per cent improvement in accuracy.

Similarly, [RKD18] introduced the Gaze-based Multiple Model Intention Estimator (G-MMIE) algorithm, which is designed to predict the goal intention of human reaching actions. Just like the previous work, this work fuses gaze information with motion data to improve the accuracy and timeliness of intention inference in human-robot collaboration scenarios with the difference being the models used for gaze and action prediction.

In an attempt to reduce errors due to saccadic eye movements, [SCV19] proposed a method that compares the similarity between hypothetical gazes on objects, generated from saliency maps, and actual gazes collected from eye-tracking devices while using the Earth Mover's Distance (EMD) to measure the similarity and employs a 1-Nearest Neighbor (1-NN) classifier to determine which object the human is focusing on. Results from the experiments show that the proposed method outperforms fixation-based methods in detecting human intention and achieves 92.2 per cent accuracy in predicting the object to be selected by humans in the interaction scenario.

Exploring first-person perspectives, [Kim+19] talked about using a first-person camera and an egocentric view to learn the user's intention through spatial and temporal information, allowing a soft wearable hand robot to assist users in grasping and releasing objects. The experiments conducted, including those involving a patient with Spinal cord injury, provide evidence of the system's effectiveness while comparison with EMG-based intention detection methods shows that the proposed method can predict user intentions faster and with high accuracy.

Circumventing the limitations of a fixation-based gaze system, [SCV20] proposed a system which uses a wearable eye tracker and a deep learning model to classify whether a human is looking at an object, based on the gaze data. The proposed model uses a Fully Convolutional Network (FCN) as the backbone, augmented with Convolutional Block Attention Modules (CBAM) and Residual Blocks and is designed to handle gaze data and object bounding boxes, aiming to predict the human's visual intention with high accuracy. With this system, the authors address issues related to fixation-based gaze systems, similar to how the previous authors did. The proposed model was able to achieve F1 scores of 0.971 for single-object scenarios and 0.962 for multiple-object scenarios.

Furthermore, in the context of shared autonomy, [FB21] investigated the use of gaze as a predictive cue for intention recognition in pick-and-place tasks. The study uses Gaussian Hidden Markov Models (GHMM) to model and predict user intention primarily based on scan paths derived from gaze data. The proposed models demonstrated strong generalizability across different users and task configurations, with accuracy rates significantly higher than random chance. The authors also explored a semantic model that abstracts the intention recognition process to generalize across different objects and tasks.

Similarly, focussing on proactive assistance, [GCR21] looked at gaze as the primary input to predict the user's intentions, allowing robots to anticipate actions and respond accordingly. The authors use a new algorithm which combines the cascade effect hypothesis which suggests that the more attention an object receives through gaze, the more likely it is to be selected by the user and an LSTM-based neural network trained to classify user intentions from gaze data with 75 per cent accuracy and up to 2 seconds before the user makes a selection.

On a similar effort to enhance collaborative processes, [Zha+21] utilized a novel interaction method that combines eye-tracking and gesture recognition to predict the intention of the operator and a finite state machine (FSM) to control the robot to make assistive manoeuvres. The study utilizes a neural network to classify nine different hand gestures based on joint angles. This classification helps in identifying the user's commands to the robot. The combination of eye gaze and hand pointing is used to select objects in the virtual environment. The Probabilistic Roadmap Planner (PRM) enables the robot to move objects without collisions efficiently. The method is tested in a virtual environment and shows improved efficiency in the assembly task when compared between two groups of participants.

In the scope of probabilistic modelling similar to previous recognition methods, [Bel+22] also presented a framework for predicting human intentions during teleoperated object manipulation tasks which combine human input with robotic assistance, by accurately and promptly estimating the user's intention based on their gaze and motion data. The authors propose a probabilistic model that combines gaze data, hand motion trajectories, and grasping triggers to predict which object a user intends to pick up or place. Intention estimation is performed using GHMMs trained on sequences of gaze and motion data and tested in a simulated environment with a cluttered scene, where objects are partially occluded, and users can manipulate them with either hand. The models perform predictions with an average duration of 0.5s remaining before the end of the actions performed by the users.

Lastly, with the use of modern eye-tracking technologies, [Yan+23] conducted a study focused on recognizing grasping intentions in human-robot interaction, particularly for individuals with disabilities that prevent them from performing physical tasks. The proposed system includes a binocular eye-tracker to capture gaze data, a scene camera for capturing the user's view, and algorithms for detecting object centroids and grasping positions. The setup is designed to identify where a user is looking and relate this information to their grasping intentions. In experiments involving performing viewing and grasping tasks, gaze data was analysed to understand the differences between gaze patterns associated with the two tasks. In grasping tasks, fixations were found to be more concentrated and closer to the grasping point, particularly the index finger, compared to viewing tasks where fixations were more dispersed. The authors used different machine learning techniques such as SVM, KNN, SGD, and decision trees to classify user intentions based on extracted features of gaze data and achieved an accuracy of 89 per cent on training objects and 85 per cent on new unseen objects with the best performing model. We will further look into the challenges and limitations concerning intention recognition in the later section.

Overall, these papers highlight the progress in gaze based intention recognition, ranging from using support vector machines to convolutional neural networks. By effectively inferring the intentions from gaze, these models enhance the ability of robots to understand context and operate in social environments. In the next section, we will look into non-verbal cues-based intention recognition followed by the challenges and limitations concerning intention recognition.

Non-verbal

By non-verbal intention, we refer to intention recognition through gestures which could be demonstrated via various body parts. These are done without the need for any verbal communication and thus, play a crucial role in human-robot interaction.

In order to facilitate intention recognition, [Neh+05] classify gestures into five broad categories manipulative, expressive side effects, symbolic, interactional, and referential gestures to facilitate intent recognition. The authors emphasize the importance of context in enabling robots to differentiate between similar gestures.

Building on the classification of gestures, [AH10] proposed using weighted probabilistic state machines to recognize intentions of humans in a human-robot

(i) (ii)

Figure 10: Images showing non-verbal cues based on intention recognition from relevant references. (i) demonstrates a robot adjusting its behaviour based on the emotional intention of the user [Che+17]. (ii) shows a robot identifies different intentions related to similar gestures [Neh+05].

collaboration scenario. The approach proposed by the authors is used to recognize explicit intentions (human actions explicity express the intentions) and implicit intentions (human actions need to be understood based on context to understand intention). In the proposed method, each intention is modelled as a separate state machine. The system updates the probabilities of each of the state machines based on the observed gesture of the human. The system was tested using an industrial robot arm and was able to successfully recognize explicit actions such as picking, placing, and passing objects, and implicit actions such as piling and un-piling objects.

Similarly, work done by [Qui+12] which was also mentioned in the gaze subsection, focused on periodic (repetitive) and deictic (pointing) gesture recognition to aid in intention recognition. The authors detect periodic gestures by analyzing the frequency and trajectory of hand movements and they use deictic gestures to divert the robot's attention to important places and objects in the environment.

Additionally, the classification of intentional driving behavior was explored by [WSH12]. They proposed a novel methodology for classifying driving behaviour using a hierarchical Perception-Action (P-A) model. The authors built on the idea that a cognitive agent's perceptual domain is developed based on the outcomes of its actions, rather than the traditional model where perception drives action, which simplifies visual processing by maintaining the complexity of the perceptual domain in relation to the agent's motor capabilities. They use the hierarchical P-A model to model the intentions at different levels of abstraction, from high-level protocols (like traffic rules) to low-level motor actions and evaluate different classification methods, including generative models (like Prolog-based first-order logic systems), discriminative models (like decision trees), and hybrid approaches. Experiments were conducted using data collected from an instrumented vehicle such as eye-tracking, control inputs, and environmental features. The results show that a hybrid approach combining generative and discriminative models yields the best performance in classifying driver intentions.

In the domain of brain-computer interfaces, [Mea+14] presented their work on a system that uses electroencephalography (EEG) signals to interpret the intention to perform gestures on a remote robotic hand. The proposed system is based on the Steady-State Visual Evoked Potential (SSVEP) approach which makes use of visual stimuli to elicit a brain response that can be detected and interpreted to understand the intention of the user.

In the case of emotion recognition, [Che+17] used facial expressions to interpret emotional intention and use it to direct the robot's actions in a human-robot interaction scenario. The authors use the Candide3 face model to recognize seven basic emotions (happiness, neutral, sad, surprise, fear, disgust, and anger) and further infer which drink the user would like to order. The behaviour of the robot is adapted to the intention of the user using an information-driven fuzzy friend-Q (IDFFQ) learning mechanism. While testing it on the bartender scenario, the system achieved 80.36 per cent accuracy in emotion recognition and 85.71 per cent accuracy in intention understanding.

Addressing the joint problem of intention recognition, attention, and tasks, [Wei+18] proposed using a hierarchical model called the Human-Attention-Object (HAO) graph. The authors used RGB-D videos to find out where the user was looking (attention), why they were looking there (intention), and what they were doing (task) using the proposed model. They use a beam search algorithm to infer the best matching task label, intention sequence, and attention points from the input and evaluate the model across three different tasks of *attention recognition*, *intention recognition*, and *task recognition*, significantly outperforming the considered baselines.

In shared autonomy settings, [JA19] presented a framework for human intent recognition where the robots can effectively infer the intentions of human operators to provide meaningful assistance during teleoperation tasks. The authors propose a recursive Bayesian filtering framework for intent recognition that models and combines multiple non-verbal observations to probabilistically reason about the user's intended goal. The model incorporates the user's actions as goal-directed behaviours with varying levels of rationality, allowing for a more personalized and accurate intent recognition and also has a feature called adjustable rationality to account for suboptimal or inconsistent behaviour from users.

In another work with emotional intent recognition, [Yan+22] proposed using facial expressions along with body actions from RGB videos. The authors propose a stacking model which fuses features from facial expressions and body actions to enhance the accuracy and robustness of intention recognition. The proposed model significantly outperforms single cue methods and achieves an accuracy of 94.57 per cent.

Further exploring emotion intention recognition, [Che+20] proposed a fuzzy deep neural network with a sparse autoencoder (FDNNSA). The model outperforms other state-of-the-art models in CK+, CASIA, and FABO datasets and shows better accuracy in emotion intention recognition in a bartender experimental study compared to softmax regression and deep neural networks. We will further look into the challenges and limitations concerning intention recognition in the later section. Works done by [YO07] and [Kur+19] are also notable in using a fuzzy learning approach for intention recognition.

Overall, these works highlight the progress in non-verbal cue based intention recognition, ranging from using probabilistic state machines to fuzzy deep neural networks with sparse autoencoders. By effectively inferring the intentions from non-verbal cues of humans, these models enhance the ability of robots to operate effectively in social environments. In the next section, we will look into object based intention recognition followed by the challenges and limitations concerning intention recognition.

Object

It is essential to understand human intention during interactions with objects. However, it is a significant challenge faced by researchers in the field of human-robot interaction. Various methods are proposed to address this and to enhance the adaptability of intelligent systems leading to better human-robot collaborations.

Using selected objects used for daily activities, [Liu+21b] discuss a method to infer the intention of object usage and interaction using Markov Random Fields (MRF). The model uses relevant features from the environment to infer the likely intentions behind object usage and outperforms methods like Recursive Bayesian Incremental Learning.

Further, based on the analysis of surrounding scenes and objects, the work done by [Dun14] presents a comprehensive approach to recognising and adapting to human intentions in real-time. The author proposes a novel probabilistic graphical model called Object-Action Intention Network (OAIN) that recognizes human intentions based on the objects in the scene and the potential actions associated with these objects. We will further look into the challenges and limitations concerning intention recognition in the later section.

In the next section, we will look into robot intent followed by the challenges and limitations concerning intention recognition.

6.3 Robot Intent

We previously categorised intention recognition into primarily two categories: *high-level* and *low-level*. However, a certain section of intention recognition work does not fit into either of these two categories. That is inferring the intentions of a robotic agent. Thus, we introduce this as a third category of intention recognition in the scope of this paper, which focuses on a human agent understanding a robotic agent's intentions either through passive inference or through active communication.

In case of humanoid robots, [Mut+09] explored the possibility of leaking gaze-like cues from two humanoid robots and observing whether participants successfully infer the intention of the robots in a guessing experiment. Results show that

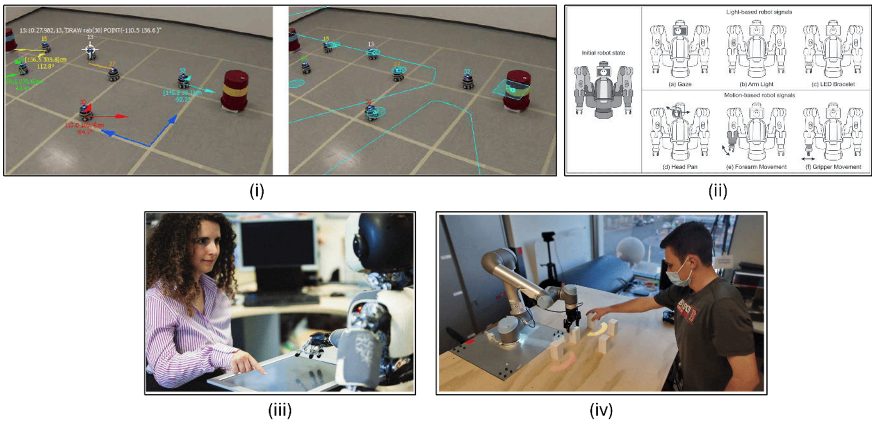


Figure 11: Images showing robotic agent's intent from relevant references. (i) demonstrates the robot data and intentions using augmented reality system [Ghi+14]. (ii) shows different methods used to express robot intent [Lem+21]. (iii) demonstrates robot intention recognition through gaze behaviour and movement patterns [Sci+15]. (iv) shows the projection of the robot intention [SA22].

participants performed better in the guessing game when the robots used gaze cues, indicating that they were able to detect and use these cues to make inferences about the robot's intentions. Although this work does not talk about predicting human intention from gaze, it reinforces the fact that humans use gaze to recognize intent and make decisions according to it. We have been looking at human intention recognition until now. However, for effective communication and collaboration between humans and robots, humans should also be able to infer the intentions of the robot.

Using light and motion based approaches of expressing robot intention, [Lem+21] tested three light-based and three motion-based methods while working with a human on a sorting task. Results showed that the light-based LED signal on the wrist of the robot was the most noticeable among all the six signals. The head pan was the most noticeable signal amongst the motion-based methods.

In an effort to see if humans could attribute intentions to robots, [Koa+13] used visual communication cues to explore if untrained humans could comprehend robot intentions correctly. They conducted a user study by operating a robot like a hearing dog and used a wizard-of-oz approach to guide the participants to one of the two sound sources as part of the experiment. Results from their experiments showed that participants were successful in identifying the intent of the robot and that gaze and head movements were important aspects for communicating the intention of the robot to the participants visually.

Expanding on the use of humanoid robots for intention recognition, [Sci+15] explored how humanoid robots can be used as tools to study the ability of humans to read and predict the intentions of others based on their movements. The authors highlight the issues with using video-based (passive observation compared to actual interaction) and VR environments (disconnected from the physical world and laws) for studying intention understanding and propose using humanoid robots as a better alternative offering modularity of control in robot actions, more natural interaction due to sharing the same physical space, and a second-person interaction scenario to preserve the natural reciprocity of the interaction between the subject and the robot. However, the authors also point out some issues with humanoid robots such as imperfect human-like motion and experimental validity of triggering the same cognitive process as other humans in subjects.

[HB18] explored the concept of robot understandability by addressing key questions, including its definition, significance, and the principles for designing robots that can be easily understood by humans. The authors present the definition for understanding and also introduce a term called "communicative actions" which are actions used to support other's understanding of the agent. They also go on to introduce a model of interaction understanding which describes how humans and robots can benefit from first-level and second-level theory of mind to facilitate understanding of each other. Based on the proposed model, the authors also suggest several guidelines for designing understandable robots such as determining what information should be communicated to the human, how the robot should infer the human's mental state, and how communicative actions should be generated and directed.

Additionally, an in-depth review of VR, AR, and XR interfaces used in Human-Robot Interaction (HRI) by [Wal+23] provided useful insights to the topic. A key issue highlighted in this review is the challenge humans face in understanding robot capabilities and intentions and that XR technologies offer a solution by helping communicate the robot's motion intentions and behaviours, which enhances collaboration. Incorporating robot intentions into shared workspaces has also been a focus, particularly through spatial AR projections aimed at optimizing collaboration [Mat06, LHS13, LKK11, PK09, CA15, Coo+14]. Similarly, [SA22] used various approaches such as visual and auditory modalities to convey intent and also used colour and light intensity changes to reflect the state of the robot. Another early implementation of the projection technique to control mobile robots via navigation areas marked using gestures was proposed by [Ish+09]. There have been other advances as well such as workspace projections, like the MAR-CPS System [Omi+15, Ghi+14], and the visualization of object handover points through AR headsets like HoloLens, which help improve understanding [New+22] and safety in shared environments [Ros+19].

Lastly, a significant application of AR systems is explored by [Bam+19] and [Yua+19], where AR is used for dynamic trajectory updates. These updates communicate a robot's intended movements in response to human presence, effectively conveying critical information such as grasp indicators and target positions.

We will further look into the challenges and limitations concerning intention recognition in the later section.

Overall, these works highlight the progress in robot intent expression and recognition, ranging from using light and sound-based cues to projection techniques. By effectively inferring the intentions of the robots, humans in shared spaces can collaborate effectively and safely with the robots. In the next section, we will look at recent developments and challenges concerning intention recognition.

7 Recent Developments and Challenges

Currently, technological advancements in machine learning and artificial intelligence have led to significant progress in the development of large language models (LLMs) which are capable of reasoning in certain contexts. The capability to understand complex concepts, generate human-like texts and assist in a wide range of tasks, has the potential to infer and execute human commands.

[Wan+24b] in their overview of large language models for robotics, talk about GIRAF [Lin+23] which uses LLMs to accurately understand the intention behind human gestures and execute tasks accordingly through a robot. Building upon the capabilities of LLMs, [VBK24] also explored the possibility of LLMs as proxies of human observers in HRI tasks. They specifically investigate if current LLMs possess Theory of Mind abilities and how this capability can be applied to HRI. This is done by conducting a study where LLMs look at a robot performing a task and are tasked with predicting how a human observer would interpret that behaviour. To validate the experimental setup, the authors test it on human participants and have them interpret the behaviour of the robot. Initial results suggest that LLMs might possess ToM abilities, as their responses often align with human interpretations. However, providing perturbations to prompts causes the LLMs to falter in their responses suggesting pattern recognition or retrieval-based reasoning rather than true theory of mind understanding.

There are cases where human operators explicitly state what they desire from the robot using natural language. However, since natural language is not directly understandable and usable by robots to plan an action, LLMs form a bridge to parse, understand, and translate the natural language instructions from the human operators to achievable goals in the form of an action plan for the robot. Extensive work has also been done by [Vem+23] in this regard, talking about a high-level library which is used by LLMs to bridge the gap between human instructions and robotic actions.

Further exploring the capabilities of LLMs in intention recognition, [AAW24] explored the use of LLMs in collaborative tasks using a combination of non-verbal cues (hand gestures, body poses, and facial expressions), verbal cues, and environmental states. The intention inferring is performed by the perceptive reasoning framework which then combines its output with the task reasoning framework to execute appropriate robot actions. The proposed model is implemented on a NICOL robot and tested with a collaborative task requiring categorizing various objects based on shape, colour, and purpose. Results show that amongst the compared LLMs, GPT 4 showed the best performance and LLMs from OpenAI performed better across the board.

Additionally, [Hua+24] discussed a novel approach to enhance human-robot collaboration by using LLMs and vlms for proactive intention tracking with a robot to assist a human user in cooking tasks. The authors propose a Language-driven Intention Tracking (LIT) system which uses a task graph to track the intention over the long term and predict future actions of the human. The setup is tested on a preliminary study which involves the task of salad making showing the robot able to smoothly assist the human in the task.

Moreover, [Wan+24a] presented their system which involves using LLMs for multi-modal HRI. The authors use three major modules in this system: *scene narrator*, *planner*, and *expresser* where the expresser uses atomic action animation clips to control the actuators and create facial expressions on the robot to express its intentions. We will further look into the challenges and limitations concerning intention recognition in the later section.

LLMs are still in their emergent stage and are being actively researched. Despite significant research done in the field, intention recognition done using LLMs faces unique challenges. The uniformity in plan generation often results in inadequate adaptability to complex or dynamic environments, limiting the system's effectiveness in real-world applications. The reliance on well-crafted prompts demands expertise, making the technology less accessible and hindering usability in unpredictable scenarios. Additionally, LLMs exhibit fragility and inconsistency, particularly in Theory of Mind tasks, where minor contextual changes can lead to drastically different outcomes, undermining reliability. Latency in processing and response generation further complicates real-time decision-making, while the need for a safety layer to verify generated actions adds another layer of complexity. Moreover, LLMs struggle to interpret and act upon non-verbal cues and more complex environments, which are crucial in nuanced intention recognition tasks. Balancing model expressiveness with responsiveness, managing multi-user scenarios, and optimizing function granularity also present significant hurdles, particularly when scaling the systems for broader applications. These challenges collectively highlight the need for further refinement and evaluation to enhance the practicality and reliability of LLMs in intention recognition.

7.1 Psychological

The Theory of Mind (ToM) in robotics faces several limitations that hinder its effectiveness in intention recognition, particularly in complex, real-world scenarios. Early implementations often suffered from oversimplification due to limited computational resources, leading to a reduction in the richness and depth of behaviours that can be modelled [Sca02]. Current approaches are primarily reactive rather than proactive, relying on passive perception where robots respond to stimuli rather than actively seeking information, which limits their adaptability and understanding in dynamic environments. Additionally, many models rely on large, contextually dependent datasets and simplified assumptions about human cognition, which may not fully capture the nuances of real-world interactions and may struggle outside of controlled laboratory conditions. The computational demands of these models also necessitate approximations, potentially reducing accuracy and generalizability. Moreover, current implementations often fail to differentiate between instrumental and epistemic goals, limiting the depth of understanding and prediction of human intentions. These challenges are further compounded by issues such as the uncanny valley effect, the requirement for more complex sensory processing, and difficulties in generalizing findings from controlled environments to unpredictable real-world contexts.

The implementation of the Mirror Neuron System (MNS) in robotics encounters several key limitations. Current models, such as those utilizing recurrent neural networks with parametric biases (RNNPB), struggle to hierarchically organize behaviours, making it difficult to handle the complexity of real-world actions. These models primarily replicate learned movement patterns rather than understanding and reproducing goal-directed behaviours, limiting their adaptability and effectiveness. Furthermore, the self-organization of neurons that can recognize actions across different perspectives or grasp types remains a significant challenge, hindering the development of a more robust and versatile MNS in robotic systems.

In the context of psychological intention recognition in robotics, it becomes clear that current methodologies face significant challenges. The limitations of the Theory of Mind (ToM) and the Mirror Neuron System (MNS) models underscore the complexity of accurately recognizing and predicting human intentions, especially in dynamic, real-world scenarios. The traditional approaches, often constrained by computational resources and overly simplistic models, fall short of capturing the nuanced, context-dependent nature of human behaviour. As [Kal19] suggests, if intentions are better understood as patterns of behaviour rather than discrete mental states, then the future of intention recognition in robotics may lie in developing systems that focus on observing and interpreting these behavioural patterns in context, rather than relying on traditional cognitive models. This shift could pave the way for more adaptive and accurate recognition systems, better equipped to handle the complexities of real-world interactions.

7.2 Activity

Despite promising research in the field of activity recognition, several key challenges persist. Static context models must be rebuilt when the robot moves, while traditional models like HMMs struggle with sensor noise, leading to false positives and negatives. Visual-based methods face difficulties with segmentation, especially when distinguishing between fine-grained activities that involve similar objects or actions. Human activity recognition systems are often hindered by a lack of generalization across different environments and conditions, requiring extensive data that is often unavailable or not diverse enough. The placement and number of sensors, as well as distinguishing activities in multi-person scenarios, complicate the process further. Moreover, balancing the trade-off between model complexity and real-time processing remains a critical challenge. Finally, the transition from third-person to first-person perspectives introduces additional noise and complexity, particularly in dynamic environments, making accurate recognition in real-world settings difficult.

7.3 Plan

Plan recognition also faces similar challenges as activity recognition in terms of computational overhead, high false-positive rates, and reliance on offline learning leading to compromise in accuracy and adaptability of systems in dynamic environments. Furthermore, in the case of tree-based approaches, the sensitivity of strategies to search tree depth can hinder efficiency, particularly in resource-constrained settings. The automation of landmark identification, critical for intention understanding, is still unresolved, limiting scalability. Finally, agents may make erroneous decisions due to incomplete observations, highlighting the need for robust feedback mechanisms.

7.4 Goal

The major challenges in goal recognition across various approaches include the need for extensive training data and interactions, which hampers real-time applicability and scalability, as seen in models like GRNet and the imitation learning model. The reliance on deterministic environments and manually constructed models, such as in the works by [JB06b] and [SBP22], limits applicability in dynamic and uncertain real-world scenarios. Additionally, the inability to handle noisy data and variability in human behaviour, highlighted in [AMV23]'s and [ZKL23]'s studies, further restricts the generalizability and accuracy of goal recognition systems. Finally, the difficulty in managing multiple concurrent intentions and the challenge of capturing the complexity of human planning underscores the limitations in current models' robustness and explanatory power. The field of goal recognition is still evolving, with each approach offering unique strengths and encountering distinct limitations. Traditional methods, such as those based on imitation learning or probabilistic models, have laid the groundwork for understanding the intricacies of goal inference but often fall short in real-world applications due to their reliance on structured environments and extensive prior knowledge. Recent developments in machine learning, particularly deep learning, have introduced more flexible and scalable solutions. These methods, exemplified by systems like GRNet, show promise in handling more complex and less predictable scenarios. However, they also introduce new challenges, such as the need for vast amounts of training data and computational resources. As research continues to push the boundaries of what is possible, the focus is shifting towards creating more robust, generalizable, and explainable models that can effectively operate in real-time and under the uncertainties of real-world conditions.

7.5 Action

Action recognition in intention prediction faces several significant challenges. One major issue is the reliance on kinematic cues from RGB video data, which may not capture subtle motion details as effectively as 3D kinematics, limiting the robustness of predictions in real-world scenarios. The lack of contextual information further hampers model performance, as human intention often depends heavily on environmental context. Outlier motions and unusual conditions are poorly represented in training data, leading to inaccuracies and reduced generalization across diverse tasks and scenarios. The computational intensity of real-time intention inference, coupled with the challenges of handling noisy sensory data, makes robust and scalable implementation difficult. Moreover, systems that depend on predefined ontologies or static models struggle with adaptability, particularly when encountering novel or ambiguous situations. These challenges are exacerbated by the variability in human actions and biosignals, which complicates the creation of models that can consistently and accurately predict intentions across different users and conditions. Additionally, the co-adaptation of humans and robots during interaction is often overlooked where the humans react to the robot's actions and change their behaviour accordingly, reducing the effectiveness of intention recognition systems in dynamic environments. Overall, these limitations highlight the need for more refined models that can integrate multiple sensory inputs, handle context effectively, and adapt to evolving human-robot interactions.

7.6 Gaze

The challenges in gaze-based intention recognition are multifaceted, arising from both technical limitations and user variability. The precision of gaze tracking is often compromised by factors such as the error margin in gaze estimation, user discomfort during prolonged use, and sensitivity to environmental conditions like occlusions and cluttered scenes. Moreover, systems frequently struggle with scalability, as they may require retraining when additional targets are introduced, and their performance can degrade in complex or dynamic environments. The 'Midas touch' problem, where unintentional gaze triggers unintended actions, remains a significant challenge, along with the difficulty of accurately interpreting gaze patterns in diverse, real-world scenarios. Additionally, the generalizability of these systems is limited due to the dependency on context-specific data, the variability in user behaviour, and the restricted diversity of participant samples in studies. These challenges highlight the need for improved gaze tracking accuracy, better handling of user variability, and enhanced scalability to make gaze-based intention recognition more reliable and broadly applicable.

7.7 Non-verbal

Non-verbal intention recognition in human-robot interaction spreads across multiple domains ranging from emotional intention to any kind of gesture made from one of the several body parts. This category of intention recognition faces several significant challenges. One major difficulty is the need to incorporate contextual understanding and interaction history, as gesture recognition cannot rely solely on kinematics, and cultural and individual variations further complicate this task. The complexity of modelling human intentions, which often involve unseen mental states or motivations, adds another layer of difficulty, especially when relying on action features alone. The limitations of current technologies, such as the low signal quality in EEG-based systems and non-convex optimization in probabilistic models, hinder real-time and accurate intent inference. Additionally, challenges arise from the need for extensive data preprocessing, the difficulty of adapting to changes in user intentions mid-task, and issues related to object segmentation in cluttered environments. The scalability of systems to handle interactions involving many humans and robots, and the accuracy of emotion recognition, are also critical issues that impact the overall reliability and effectiveness of intention recognition systems.

7.8 Object

The major challenges in object intention recognition include the static nature of models, which struggle to adapt to changing user preferences over time, even with adaptive algorithms like Q-learning. Accurately interpreting visual affordances is difficult due to objects affording multiple actions, complicating the prediction of user intent. Additionally, object segmentation in cluttered or occluded environments often leads to errors, impacting tasks like categorization and pose estimation. Moreover, the system's performance is inconsistently affected by learning rates, with scene-dependent factors playing a significant role.

7.9 Robot Intent

As discussed earlier, one aspect of intention recognition is inferring the intentions of the robot by human agents which has its own set of issues. The effectiveness of nonverbal cues, such as gaze direction, is influenced by the robot's design, with highly human-like robots potentially causing cognitive overload or discomfort. Signalling intent in shared workspaces is complicated by the limitations of current methods, such as light-based projecting signals requiring flat surfaces and motion-based signals being easily confused with routine actions. The environment and user traits further complicate interpretation, with cluttered spaces and specific personality traits affecting signal comprehension. Additionally, the use of humanoid robots in research is limited by their inability to perfectly replicate human movements, which may hinder accurate intention reading. The inherent complexity of designing robots that can effectively communicate intentions, compounded by asymmetries in human-robot sensing and acting capabilities, and the challenge of directing communicative actions in multi-human environments, highlights the multifaceted difficulties in advancing robot intention recognition. Robot intention recognition is also constrained in the domains of Virtual Reality (VR), Extended Reality (XR), and Mixed Reality (MR). The integration of object and robot intentions with action probabilities could pave the way for richer, more intuitive visualizations, tailoring projections to convey personalized information effectively. Future research should also delve into the comparative impacts of static versus continuous projections in conveying a robot's forthcoming goals, extending beyond mere motion intent to visualize shared objectives, requisite objects, and comprehensive states of the robot. An understanding of how to balance visual information to avoid overload, perhaps through real-world studies, is crucial. Additionally, investigating the influence of anthropomorphic versus non-anthropomorphic gestures on user perception could provide significant insights into optimizing user experiences in HRI contexts.

8 Discussion and Future Work

This review outlines the broad and varied nature of intention recognition in robotics. By categorizing intention recognition into *high-level*, *low-level*, and *robot intention* recognition, we provide a comprehensive understanding of the different theories, approaches, and methods used in robotics.

We understood that *high-level* intention recognition focuses on broader objectives and long-term plans behind human actions and generally encompasses intention recognized from activity, plans, and goals. In contrast, *low-level* intention recognition prioritizes immediate actions and cues such as gaze and other nonverbal gestures. Taking these into consideration, intention could be defined as an agent's goal with a dynamic action plan within a defined environment and intention recognition then becomes the key which allows us to understand the desired state and the sequence of actions necessary to achieve this goal. Therefore, understanding the goal, action plan, and environment enables us to predict the agent's intended outcomes and the steps they will take to reach them within the specific context.

The review emphasizes the evolving role of environmental factors in enhancing context based information to improve the ability to interpret human intentions in regard to its immediate surrounding. It identifies the challenges and limitations in the high-level and low-level intention categories, and addressing these identified challenges as discussed in the previous section can lead to more adaptable and resilient robotic systems. Thus, making them capable of operating effectively in dynamic and unpredictable environments.

Building on the review, we conclude that future research in the field should focus on the integration of multi-modal information sources. This is done to improve the accuracy and robustness of intention recognition systems. Exploring the synergy between high-level and low-level recognition processes along with relative scene information can lead to better prediction of human intentions.

Acknowledgements

We would like to thank Jacek Malec and Björn Olofsson for their support during the work. This work was supported by the Excellence Center at Linköping University, Lund University for Information Technology (ELLIIT), and the Wallenberg Artificial Intelligence, Autonomous Systems, and Software Program (WASP).

References

[AK06] Daniel Aarno and Danica Kragic. "Layered HMM for motion intention recognition". In: *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE. 2006, pp. 5130–5135.

[AA14] Najla Ahmad and Arvin Agah. "Plan and intent recognition in a multi-agent system for collective box pushing". In: *Journal of Intelligent Systems* 23.1 (2014), pp. 95–108.

[AZN98] David W Albrecht, Ingrid Zukerman, and An E Nicholson. "Bayesian models for keyhole plan recognition in an adventure game". In: *User modeling and user-adapted interaction* 8 (1998), pp. 5–47.

[AAW24] Hassan Ali, Philipp Allgeuer, and Stefan Wermter. "Comparing Apples to Oranges: LLM-powered Multimodal Intention Prediction in an Object Categorization Task". In: *arXiv preprint arXiv:2404.08424* (2024).

[AMV23] Abeer Alshehri, Tim Miller, and Mor Vered. “Explainable goal recognition: a framework based on weight of evidence”. In: *Proceedings of the International Conference on Automated Planning and Scheduling*. Vol. 33. 2023, pp. 7–16.

[AA07] Marcelo Gabriel Armentano and Analía Amandi. “Plan recognition for interface agents: state of the art”. In: *Artificial Intelligence Review* 28.2 (2007), pp. 131–162.

[AH10] Muhammad Awais and Dominik Henrich. “Human-robot collaboration by intention recognition using probabilistic state machines”. In: *19th International Workshop on Robotics in Alpe-Adria-Danube Region (RAAD 2010)*. IEEE. 2010, pp. 75–80.

[Bai+15] Haoyu Bai et al. “Intention-aware online POMDP planning for autonomous driving in a crowd”. In: *2015 ieee international conference on robotics and automation (icra)*. IEEE. 2015, pp. 454–460.

[BST09] Chris L Baker, Rebecca Saxe, and Joshua B Tenenbaum. “Action understanding as inverse planning”. In: *Cognition* 113.3 (2009), pp. 329–349.

[Bak+17] Chris L Baker et al. “Rational quantitative attribution of beliefs, desires and percepts in human mentalizing”. In: *Nature Human Behaviour* 1.4 (2017), p. 0064.

[Bam+19] Daniel Bambuŝek et al. “Combining interactive spatial augmented reality with head-mounted display for end-user collaborative robot programming”. In: *2019 28th IEEE international conference on robot and human interactive communication (RO-MAN)*. IEEE. 2019, pp. 1–8.

[BC+95] Simon Baron-Cohen et al. “Are children with autism blind to the mentalistic significance of the eyes?” In: *British Journal of Developmental Psychology* 13.4 (1995), pp. 379–398.

[BWG00] Harold Bekkering, Andreas WohlschlaÈger, and Merideth Gattis. “Imitation of gestures in children is goal-directed”. In: *The Quarterly Journal of Experimental Psychology Section A* 53.1 (2000), pp. 153–164.

[Bel+22] Anna Belardinelli et al. “Intention estimation from gaze and motion features for human-robot shared-control object manipulation”. In: *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2022, pp. 9806–9813.

[BED08] Holger Berndt, Jorg Emmert, and Klaus Dietmayer. “Continuous driver intention recognition with hidden markov models”. In: *2008 11th International IEEE Conference on Intelligent Transportation Systems*. IEEE. 2008, pp. 1189–1194.

[BO19a] Francesca Bianco and Dimitri Ognibene. “Functional advantages of an adaptive theory of mind for robotics: a review of current architectures”. In: *2019 11th Computer Science and Electronic Engineering (CEEC)* (2019), pp. 139–143.

[BO19b] Francesca Bianco and Dimitri Ognibene. “Transferring adaptive theory of mind to social robots: Insights from developmental psychology to robotics”. In: *Social Robotics: 11th International Conference, ICSR 2019, Madrid, Spain, November 26–29, 2019, Proceedings 11*. Springer. 2019, pp. 77–87.

[BDK14] Elisheva Bonchek-Dokow and Gal A Kaminka. “Towards computational models of intention detection and intention prediction”. In: *Cognitive Systems Research* 28 (2014), pp. 44–79.

[Bra87] Michael Bratman. “Intention, plans, and practical reason”. In: (1987).

[Bra+22] Christian Alexander Braun et al. “Belief Space Control with Intention Recognition for Human-Robot Cooperation”. In: *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE. 2022, pp. 1897–1903.

[Bre+05] Cynthia Breazeal et al. “Learning from and about others: Towards using imitation to bootstrap the social understanding of others by robots”. In: *Artificial life* 11.1-2 (2005), pp. 31–62.

[Bre+06] Cynthia Breazeal et al. “Using perspective taking to learn from ambiguous demonstrations”. In: *Robotics and autonomous systems* 54.5 (2006), pp. 385–393.

[BWS17] Gordon Briggs, Tom Williams, and Matthias Scheutz. “Enabling robots to understand indirect speech acts in task-based interactions”. In: *Journal of Human-Robot Interaction* 6.1 (2017), pp. 64–94.

[Bro17] Elizabeth Broadbent. “Interactions with robots: The truths we reveal about ourselves”. In: *Annual review of psychology* 68.1 (2017), pp. 627–652.

[Buc94] Michael Buckland. “On the nature of records management theory”. In: *The American Archivist* (1994), pp. 346–351.

[CB07] Sylvain Calinon and Aude Billard. “Incremental learning of gestures by imitation in a humanoid robot”. In: *Proceedings of the ACM/IEEE international conference on Human-robot interaction*. 2007, pp. 255–262.

[CT16] Felip Martí Carrillo and Elin Anna Topp. “Interaction and task patterns in symbiotic, mixed-initiative human-robot interaction”. In: *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence*. 2016.

[CG93] Eugene Charniak and Robert P Goldman. “A Bayesian model of plan recognition”. In: *Artificial Intelligence* 64.1 (1993), pp. 53–79.

[CG90] Eugene Charniak and Robert Goldman. “Plan recognition in stories and in life”. In: *Machine Intelligence and Pattern Recognition*. Vol. 10. Elsevier, 1990, pp. 343–351.

[CDI06] Antonio Chella, Haris Dindo, and Ignazio Infantino. “A cognitive framework for imitation learning”. In: *Robotics and Autonomous Systems* 54.5 (2006), pp. 403–408.

[Che+17] Luefeng Chen et al. “Information-driven multirobot behavior adaptation to emotional intention in human–robot interaction”. In: *IEEE Transactions on Cognitive and Developmental Systems* 10.3 (2017), pp. 647–658.

[Che+20] Luefeng Chen et al. “A fuzzy deep neural network with sparse autoencoder for emotional intention understanding in human–robot interaction”. In: *IEEE Transactions on Fuzzy systems* 28.7 (2020), pp. 1252–1264.

[Chi+23] Mattia Chiari et al. “Goal recognition as a deep learning task: The grnet approach”. In: *Proceedings of the International Conference on Automated Planning and Scheduling*. Vol. 33. 1. 2023, pp. 560–568.

[CA98] Leslie B Cohen and Geoffrey Amsel. “Precursors to infants’ perception of the causality of a simple event”. In: *Infant behavior and development* 21.4 (1998), pp. 713–731.

[Coo+14] Michael D Coovert et al. “Spatial augmented reality as a method for a mobile robot to communicate intended movement”. In: *Computers in Human Behavior* 34 (2014), pp. 241–248.

[CA15] Nuno Costa and Artur Arsenio. “Augmented reality behind the wheel-human interactive assistance by mobile robots”. In: *2015 6th International Conference on Automation, Robotics and Applications (ICARA)*. IEEE. 2015, pp. 63–69.

[Cra+18] Martijn Cramer et al. “Towards robust intention estimation based on object affordance enabling natural human-robot collaboration in assembly tasks”. In: *Procedia CIRP* 78 (2018), pp. 255–260.

[CNP06] GCHE de Croon, S Nolfi, and EO Postma. “Towards pro-active embodied agents: on the importance of neural mechanisms suitable to process time information”. In: *Complex Engineered Systems: Science Meets Technology* (2006), pp. 338–363.

[Cro+19] Charles R Crowell et al. “Anthropomorphism of robots: study of appearance and agency”. In: *JMIR human factors* 6.2 (2019), e12629.

[Den+19] Kevin Denis et al. “Clothoidal Local Path Template for Intention Estimation by Assistive Mobile Robots”. In: *Proceedings of the International Conference on Automated Planning and Scheduling*. Vol. 29. 2019, pp. 689–697.

[Der+17] Oriane Dermy et al. “Prediction of intention during interaction with iCub with probabilistic movement primitives”. In: *Frontiers in Robotics and AI* 4 (2017), p. 45.

[DSS15] Frederik Diederichs, Tobias Schüttke, and Dieter Spath. “Driver intention algorithm for pedestrian protection and automated emergency braking systems”. In: *2015 IEEE 18th International Conference on Intelligent Transportation Systems*. IEEE. 2015, pp. 1049–1054.

[Dua+18] Nuno Ferreira Duarte et al. “Action anticipation: Reading the intentions of humans and robots”. In: *IEEE Robotics and Automation Letters* 3.4 (2018), pp. 4132–4139.

[Dun14] Kester Duncan. “Scene-dependent human intention recognition for an assistive robotic system”. In: (2014).

[Dun+15] Kester Duncan et al. “Scene-dependent intention recognition for task communication with reduced human-robot interaction”. In: *Computer Vision-ECCV 2014 Workshops: Zurich, Switzerland, September 6-7 and 12, 2014, Proceedings, Part III 13*. Springer. 2015, pp. 730–745.

[EWC07] Nicholas Epley, Adam Waytz, and John T Cacioppo. “On seeing human: a three-factor theory of anthropomorphism.” In: *Psychological review* 114.4 (2007), p. 864.

[Erl+06] Wolfram Erlhagen et al. “Goal-directed imitation for robots: A bio-inspired approach to action understanding and skill learning”. In: *Robotics and autonomous systems* 54.5 (2006), pp. 353–360.

[FMJ02] Ajo Fod, Maja J Matarić, and Odest Chadwicke Jenkins. "Automated derivation of primitives for movement classification". In: *Autonomous robots* 12 (2002), pp. 39–54.

[For+95] Jeff Forbes et al. "The batmobile: Towards a bayesian automated taxi". In: *IJCAI*. Vol. 95. 1995, pp. 1878–1885.

[For+21] Federico Formica et al. "Neural networks based human intent prediction for collaborative robotics applications". In: *2021 20th International Conference on Advanced Robotics (ICAR)*. IEEE. 2021, pp. 1018–1023.

[FB21] Stefan Fuchs and Anna Belardinelli. "Gaze-based intention estimation for shared autonomy in pick-and-place tasks". In: *Frontiers in Neurorobotics* 15 (2021), p. 647930.

[GB20] Alexander Gabriel and Paul Baxter. "Towards Intention Recognition for Human-Interacting Agricultural Robots". In: *Proceedings of The 3rd UK-RAS Conference*. 2020.

[Gei02] Christopher W Geib. "Problems with intent recognition for elder care". In: *Proceedings of the AAAI-02 Workshop "Automation as Caregiver*. 2002, pp. 13–17.

[Ghi+14] Fabrizio Ghiringhelli et al. "Interactive augmented reality for understanding and analyzing multi-robot systems". In: *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE. 2014, pp. 1195–1201.

[God+22] Ricardo V Godoy et al. "Lightmyography based decoding of human intention using temporal multi-channel transformers". In: *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2022, pp. 6087–6094.

[Gol+19] Michael Goldhammer et al. "Intentions of vulnerable road users—Detection and forecasting by means of machine learning". In: *IEEE transactions on intelligent transportation systems* 21.7 (2019), pp. 3035–3045.

[GCR21] Carlos Gomez Cubero and Matthias Rehm. "Intention recognition in human robot interaction based on eye tracking". In: *Human-Computer Interaction–INTERACT 2021: 18th IFIP TC 13 International Conference, Bari, Italy, August 30–September 3, 2021, Proceedings, Part III 18*. Springer. 2021, pp. 428–437.

[GBB07] Jesse Gray, Matt Berlin, and Cynthia Breazeal. "Intention recognition with divergent beliefs for collaborative robots". In: *Society for the study of artificial intelligence and simulation of behaviour (AISB-07)* (2007).

[Gra+05] Jesse Gray et al. “Action parsing and goal inference using self as simulator”. In: *ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication, 2005.* IEEE. 2005, pp. 202–209.

[Hai+03] Karen Zita Haigh et al. “Independent LifeStyle Assistant™(ILSA)”. In: *Honeywell Laboratories, Minneapolis, MN, Tech. Rep. ACSPO3023* (2003).

[HK10] Ji-Hyeong Han and Jong-Hwan Kim. “Human-robot interaction by reading human intention based on mirror-neuron system”. In: *2010 IEEE International Conference on Robotics and Biomimetics*. IEEE. 2010, pp. 561–566.

[Hei13] Fritz Heider. *The psychology of interpersonal relations*. Psychology Press, 2013.

[Hei04] Clint Heinze. “Modelling intention recognition for intelligent agent systems”. In: (2004).

[HB18] Thomas Hellström and Suna Bensch. “Understandable robots-what, why, and how”. In: *Paladyn, Journal of Behavioral Robotics* 9.1 (2018), pp. 110–123.

[Hon01] Jun Hong. “Goal recognition through goal graph analysis”. In: *Journal of Artificial Intelligence Research* 15 (2001), pp. 1–30.

[Hua+15a] Chien-Ming Huang et al. “Using gaze patterns to predict task intent in collaboration”. In: *Frontiers in psychology* 6 (2015), p. 1049.

[Hua+15b] Jian Huang et al. “Control of upper-limb power-assist exoskeleton using a human-robot interface based on motion intention recognition”. In: *IEEE transactions on automation science and engineering* 12.4 (2015), pp. 1257–1270.

[Hua+24] Zhe Huang et al. “LIT: Large Language Model Driven Intention Tracking for Proactive Human-Robot Collaboration–A Robot Sous-Chef Application”. In: *arXiv preprint arXiv:2406.13787* (2024).

[Ish+09] Kentaro Ishii et al. “Designing laser gesture interface for robot control”. In: *Human-Computer Interaction–INTERACT 2009: 12th IFIP TC 13 International Conference, Uppsala, Sweden, August 24-28, 2009, Proceedings, Part II 12*. Springer. 2009, pp. 479–492.

[JA18] Siddarth Jain and Brenna Argall. “Recursive Bayesian human intent recognition in shared-control robotics”. In: *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2018, pp. 3905–3912.

[JA19] Siddarth Jain and Brenna Argall. “Probabilistic human intent recognition for shared autonomy in assistive robotics”. In: *ACM Transactions on Human-Robot Interaction (THRI)* 9.1 (2019), pp. 1–23.

[Jan+14] Young-Min Jang et al. “Human intention recognition based on eyeball movement pattern and pupil size variation”. In: *Neurocomputing* 128 (2014), pp. 421–432.

[JB06a] Bart Jansen and Tony Belpaeme. “A computational model of intention reading in imitation”. In: *Robotics and Autonomous Systems* 54.5 (2006), pp. 394–402.

[JB06b] Bart Jansen and Tony Belpaeme. “A model for inferring the intention in imitation tasks”. In: *ROMAN 2006-The 15th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE. 2006, pp. 238–243.

[JKC08] Hochul Jeon, Taehwan Kim, and Joongmin Choi. “Ontology-based user intention recognition for proactive planning of intelligent robot behavior”. In: *2008 International Conference on Multimedia and Ubiquitous Engineering (mue 2008)*. IEEE. 2008, pp. 244–248.

[JBD19] Charmi Jobanputra, Jatna Bavishi, and Nishant Doshi. “Human activity recognition: A survey”. In: *Procedia Computer Science* 155 (2019), pp. 698–703.

[JD05] Matthew Johnson and Yiannis Demiris. “Perceptual perspective taking and action recognition”. In: *International Journal of Advanced Robotic Systems* 2.4 (2005), p. 32.

[Kal19] Annemarie Kalis. “No intentions in the brain: A Wittgensteinian perspective on the science of intention”. In: *Frontiers in psychology* 10 (2019), p. 946.

[Kel+08] Richard Kelley et al. “Understanding human intentions via hidden markov models in autonomous mobile robots”. In: *Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction*. 2008, pp. 367–374.

[Kel+10] Richard Kelley et al. “Understanding activities and intentions for human-robot interaction”. In: *Human-robot interaction*. IntechOpen. 2010, pp. 288–305.

[KS08] Peter Kiefer and Klaus Stein. “A Framework for Mobile Intention Recognition in Spatially Structured Environments.” In: *BMI*. 2008, pp. 28–41.

[Kim+19] Daekyum Kim et al. “Eyes are faster than hands: A soft wearable robot learns user intention from the egocentric view”. In: *Science Robotics* 4.26 (2019), eaav2949.

[Koa+13] Kheng Lee Koay et al. “Hey! There is someone at your door. A hearing robot using visual communication signals of hearing dogs to communicate intent”. In: *2013 IEEE symposium on artificial life (ALife)*. IEEE. 2013, pp. 90–97.

[KM17] Santiago Gerling Konrad and Favio R Masson. “Pedestrian intention estimation from egocentric data”. In: *2017 XVII Workshop on Information Processing and Control (RPIC)*. IEEE. 2017, pp. 1–5.

[KH10] Peter Krauthausen and Uwe D Hanebeck. “A model-predictive switching approach to efficient intention recognition”. In: *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE. 2010, pp. 4908–4913.

[Kur+19] Rifky Kurniawan et al. “Electric bionic legs used gyroscope and accelerometer with fuzzy method”. In: *2019 International Symposium on Electronics and Smart Devices (ISESD)*. IEEE. 2019, pp. 1–5.

[Lai+22] Yujun Lai et al. “User intent estimation during robot learning using physical human robot interaction primitives”. In: *Autonomous Robots* 46.2 (2022), pp. 421–436.

[LSB19] Jin Joo Lee, Fei Sha, and Cynthia Breazeal. “A Bayesian theory of mind approach to nonverbal communication”. In: *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE. 2019, pp. 487–496.

[LKK11] Joo-Haeng Lee, Junho Kim, and Hyun Kim. “A note on hybrid control of robotic spatial augmented reality”. In: *2011 8th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI)*. IEEE. 2011, pp. 621–626.

[Lem+21] Gregory Lemasurier et al. “Methods for expressing robot intent for human–robot collaboration in shared workspaces”. In: *ACM Transactions on Human-Robot Interaction (THRI)* 10.4 (2021), pp. 1–27.

[Len+12] Tommaso Lenzi et al. “Intention-based EMG control for powered exoskeletons”. In: *IEEE transactions on biomedical engineering* 59.8 (2012), pp. 2180–2190.

[Les98] Neal Brian Lesh. *Scalable and adaptive goal recognition*. University of Washington, 1998.

[Les94] Alan M Leslie. “ToMM, ToBy, and Agency: Core architecture and domain specificity”. In: *Mapping the mind: Domain specificity in cognition and culture* 29 (1994), pp. 119–48.

[LHS13] Florian Leutert, Christian Herrmann, and Klaus Schilling. “A spatial augmented reality system for intuitive display of robotic data”. In: *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE. 2013, pp. 179–180.

[Li+22a] Huao Li et al. “Theory of mind modeling in search and rescue teams”. In: *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE. 2022, pp. 483–489.

[Li+16] Keqiang Li et al. “Lane changing intention recognition based on speech recognition models”. In: *Transportation research part C: emerging technologies* 69 (2016), pp. 497–514.

[Li+18] Mantian Li et al. “Towards online estimation of human joint muscular torque with a lower limb exoskeleton robot”. In: *Applied Sciences* 8.9 (2018), p. 1610.

[Li+17] Renjie Li et al. “Driver-automation indirect shared control of highly automated vehicles with intention-aware authority transition”. In: *2017 IEEE intelligent vehicles symposium (IV)*. IEEE. 2017, pp. 26–32.

[LZD20] Shengchao Li, Lin Zhang, and Xiumin Diao. “Deep-learning-based human intention prediction using RGB images and optical flow”. In: *Journal of Intelligent & Robotic Systems* 97 (2020), pp. 95–107.

[LZ17] Songpo Li and Xiaoli Zhang. “Implicit intention communication in human–robot interaction through visual behavior studies”. In: *IEEE Transactions on Human-Machine Systems* 47.4 (2017), pp. 437–448.

[LG13] Yanan Li and Shuzhi Sam Ge. “Human–robot collaboration based on motion intention estimation”. In: *IEEE/ASME Transactions on Mechatronics* 19.3 (2013), pp. 1007–1014.

[Li+22b] Yanan Li et al. “A review on interaction control for contact robots through intent detection”. In: *Progress in Biomedical Engineering* 4.3 (2022), p. 032004.

[Li+21] Zhihao Li et al. “Intention understanding in human–robot interaction based on visual-NLP semantics”. In: *Frontiers in Neurorobotics* 14 (2021), p. 610139.

[Lin+23] Li-Heng Lin et al. “Gesture-informed robot assistance via foundation models”. In: *7th Annual Conference on Robot Learning*. 2023.

[Liu+21a] Chunfang Liu et al. “Robot recognizing humans intention and interacting with humans based on a multi-task model combining ST-GCN-LSTM model and YOLO model”. In: *Neurocomputing* 430 (2021), pp. 174–184.

[Liu+21b] Yan Liu et al. “An MRF-Based Intention Recognition Framework for WMRA with Selected Objects as Contextual Clues”. In: *Intelligent Robotics and Applications: 14th International Conference, ICIRA 2021, Yantai, China, October 22–25, 2021, Proceedings, Part III 14*. Springer. 2021, pp. 345–356.

[LH19] Zhiguang Liu and Jianhong Hao. “Intention Recognition in Physical Human-Robot Interaction Based on Radial Basis Function Neural Network”. In: *Journal of Robotics* 2019.1 (2019), p. 4141269.

[Los+18] Dylan P Losey et al. “A review of intent detection, arbitration, and communication aspects of shared control for physical human–robot interaction”. In: *Applied Mechanics Reviews* 70.1 (2018), p. 010804.

[MK97] Bertram F Malle and Joshua Knobe. “The folk concept of intentionality”. In: *Journal of experimental social psychology* 33.2 (1997), pp. 101–121.

[MG04] Wenji Mao and Jonathan Gratch. “A utility-based approach to intention recognition”. In: *AAMAS 2004 Workshop on Agent Tracking: Modeling Other Agents from Observations*. Vol. 46. 2004, p. 59.

[Mat06] Takafumi Matsumaru. “Mobile robot with preliminary-announcement and display function of forthcoming motion using projection equipment”. In: *ROMAN 2006-The 15th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE. 2006, pp. 443–450.

[Mea+14] Roberto Meattini et al. “Gestural art: A Steady State Visual Evoked Potential (SSVEP) based Brain Computer Interface to express intentions through a robotic hand”. In: *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*. IEEE. 2014, pp. 211–216.

[MÅ18] Naveed Muhammad and Björn Åstrand. “Intention estimation using set of reference trajectories as behaviour model”. In: *Sensors* 18.12 (2018), p. 4423.

[Mur02] Kevin Patrick Murphy. *Dynamic bayesian networks: representation, inference and learning*. University of California, Berkeley, 2002.

[Mut+09] Bilge Mutlu et al. “Nonverbal leakage in robots: communication of intentions through seemingly unintentional behavior”. In: *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*. 2009, pp. 69–76.

[NAH02] Yukie Nagai, Minoru Asada, and Koh Hosoda. “Developmental learning model for joint attention”. In: *IEEE/RSJ International Conference on Intelligent Robots and Systems*. Vol. 1. IEEE. 2002, pp. 932–937.

[Neh+05] Chrystopher L Nehaniv et al. “A methodological approach relating the classification of gesture to identification of human intent in the context of human-robot interaction”. In: *ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication, 2005*. IEEE. 2005, pp. 371–377.

[New+22] Rhys Newbury et al. “Visualizing robot intent for object handovers with augmented reality”. In: *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE. 2022, pp. 1264–1270.

[Ni+23] Shouxiang Ni et al. “Cross-View Human Intention Recognition for Human-Robot Collaboration”. In: *IEEE Wireless Communications* 30.3 (2023), pp. 189–195.

[NZR18] Davide Nicolis, Andrea Maria Zanchettin, and Paolo Rocco. “Human intention estimation based on neural networks for enhanced collaboration with robots”. In: *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2018, pp. 1326–1333.

[NR13] Domen Novak and Robert Riener. “Enhancing patient freedom in rehabilitation robotics using gaze-based intention detection”. In: *2013 IEEE 13th international conference on rehabilitation robotics (ICORR)*. IEEE. 2013, pp. 1–6.

[OMM19] Dimitri Ognibene, Lorenzo Mirante, and Letizia Marchegiani. “Proactive intention recognition for joint human-robot search and rescue missions through monte-carlo planning in pomdp environments”. In: *Social Robotics: 11th International Conference, ICSR 2019, Madrid, Spain, November 26–29, 2019, Proceedings 11*. Springer. 2019, pp. 332–343.

[Omi+15] Shayegan Omidshafiei et al. “Mar-cps: Measurable augmented reality for prototyping cyber-physical systems”. In: *AIAA Infotech@ Aerospace*. 2015, p. 0643.

[Omo+08] Takashi Omori et al. “Computational modeling of human-robot interaction based on active intention estimation”. In: *Neural Information Processing: 14th International Conference, ICONIP 2007, Kitakyushu, Japan, November 13-16, 2007, Revised Selected Papers, Part II 14*. Springer. 2008, pp. 185–192.

[Pac08] Elisabeth Pacherie. “The phenomenology of action: A conceptual framework”. In: *Cognition* 107.1 (2008), pp. 179–217.

[Pag+21] Matthew J Page et al. “The PRISMA 2020 statement: an updated guideline for reporting systematic reviews”. In: *bmj* 372 (2021).

[Par+16] Chonhyon Park et al. “Hi robot: Human intention-aware robot planning for safe and efficient navigation in crowds”. In: *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2016, pp. 3320–3326.

[PK09] Jongkyeong Park and Gerard Jounghyun Kim. “Robots with projectors: an alternative to anthropomorphic HRI”. In: *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*. 2009, pp. 221–222.

[PL16] Ki-Hee Park and Seong-Whan Lee. “Movement intention decoding based on deep learning for multiuser myoelectric interfaces”. In: *2016 4th international winter conference on brain-computer Interface (BCI)*. IEEE. 2016, pp. 1–2.

[Pea14] Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier, 2014.

[PA11] Luís Moniz Pereira and Han The Anh. “Intention recognition with evolution prospection and causal bayes networks”. In: *Computational intelligence for engineering systems: Emergent applications* (2011), pp. 1–33.

[PH11] Luís Moniz Pereira and The Anh Han. “Elder care via intention recognition and evolution prospection”. In: *Applications of Declarative Programming and Knowledge Management: 18th International Conference, INAP 2009, Évora, Portugal, November 3-5, 2009, Revised Selected Papers 18*. Springer. 2011, pp. 170–187.

[Per+10] Luís Moniz Pereira et al. “Proactive intention recognition for home ambient intelligence”. In: *Workshop Proceedings of the 6th International Conference on Intelligent Environments*. IOS Press. 2010, pp. 91–100.

[Per+11] Luís Moniz Pereira et al. “Context-dependent incremental intention recognition through Bayesian network model construction”. In: *CEUR Workshop Proceedings*. Vol. 818. CEUR Workshop Proceedings. 2011, pp. 50–58.

[Per+13] Luís Moniz Pereira et al. “State-of-the-art of intention recognition and its use in decision making”. In: *Ai Communications* 26.2 (2013), pp. 237–246.

[POWW21] Jairo Perez-Osorio, Eva Wiese, and Agnieszka Wykowska. “Theory of mind and joint attention”. In: *The Handbook on Socially Interactive Agents: 20 years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 1: Methods, Behavior, Cognition*. 2021, pp. 311–348.

[Per22] Michele Persiani. “Expressing and recognizing intentions”. PhD thesis. Umeå University, 2022.

[PMP18] Tomislav Petković, Ivan Marković, and Ivan Petrović. “Human intention recognition in flexible robotized warehouses based on markov decision processes”. In: *ROBOT 2017: Third Iberian Robotics Conference: Volume 2*. Springer. 2018, pp. 629–640.

[Pet+19] Tomislav Petković et al. “Human intention estimation based on hidden Markov model motion validation for safe flexible robotized warehouses”. In: *Robotics and Computer-Integrated Manufacturing* 57 (2019), pp. 182–196.

[PF23] Julius Pettersson and Petter Falkman. “Comparison of LSTM, Transformers, and MLP-mixer neural networks for gaze based human intention prediction”. In: *Frontiers in Neurorobotics* 17 (2023), p. 1157957.

[PBCR95] Wendy Phillips, Simon Baron-Cohen, and Michael Rutter. “To what extent can children with autism understand desire?” In: *Development and Psychopathology* 7.1 (1995), pp. 151–169.

[PW78] David Premack and Guy Woodruff. “Does the chimpanzee have a theory of mind?” In: *Behavioral and brain sciences* 1.4 (1978), pp. 515–526.

[Pre+19] Tony J Prescott et al. “Memory and mental time travel in humans and social robots”. In: *Philosophical Transactions of the Royal Society B* 374.1771 (2019), p. 20180025.

[PW13] David V Pynadath and Michael P Wellman. “Accounting for context in plan recognition, with application to traffic monitoring”. In: *arXiv preprint arXiv:1302.4980* (2013).

[Qui+12] Joao Quintas et al. “Context-based understanding of interaction intentions”. In: *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*. IEEE. 2012, pp. 515–520.

[Rab+18] Neil Rabinowitz et al. “Machine theory of mind”. In: *International conference on machine learning*. PMLR. 2018, pp. 4218–4227.

[Rak+19] Sanzhar Rakhimkul et al. “Autonomous object detection and grasping using deep learning for design of an intelligent assistive robot manipulation system”. In: *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*. IEEE. 2019, pp. 3962–3968.

[RABC15] Karinne Ramirez-Amaro, Michael Beetz, and Gordon Cheng. “Understanding the intention of human activities through semantic perception: observation, understanding and execution on a humanoid robot”. In: *Advanced Robotics* 29.5 (2015), pp. 345–362.

[RG+95] Anand S Rao, Michael P Georgeff, et al. “BDI agents: from theory to practice.” In: *Icmas*. Vol. 95. 1995, pp. 312–319.

[RD17] HC Ravichandar and A Dani. “Intention inference for human-robot collaboration in assistive robotics”. In: *Human Modelling for Bio-Inspired Robotics*. Elsevier, 2017, pp. 217–249.

[RKD18] Harish Chaandar Ravichandar, Avnish Kumar, and Ashwin Dani. “Gaze and motion information fusion for human intention inference”. In: *International Journal of Intelligent Robotics and Applications* 2 (2018), pp. 136–148.

[RTD19] Harish Chaandar Ravichandar, Daniel Trombetta, and Ashwin P Dani. “Human intention-driven learning control for trajectory synchronization in human-robot collaborative tasks”. In: *IFAC-PapersOnLine* 51.34 (2019), pp. 1–7.

[RPF13] Kristína Rebrová, Matej Pecháč, and Igor Farkaš. “Towards a robotic model of the mirror neuron system”. In: *2013 IEEE Third Joint International Conference on Development and Learning and Epigenetic Robotics (ICDL)*. IEEE. 2013, pp. 1–6.

[RC04] Giacomo Rizzolatti and Laila Craighero. “The mirror-neuron system”. In: *Annu. Rev. Neurosci.* 27.1 (2004), pp. 169–192.

[Ros+19] Eric Rosen et al. “Communicating and controlling robot arm motion intent through mixed-reality head-mounted displays”. In: *The International Journal of Robotics Research* 38.12-13 (2019), pp. 1513–1526.

[Roy+09] Patrice Roy et al. “A hybrid plan recognition model for Alzheimer’s patients: interleaved-erroneous dilemma”. In: *Web Intelligence and Agent Systems: An International Journal* 7.4 (2009), pp. 375–397.

[Sad11] Fariba Sadri. “Logic-based approaches to intention recognition”. In: *Handbook of research on ambient intelligence and smart environments: Trends and perspectives*. IGI Global, 2011, pp. 346–375.

[Saf+15a] Mohammad Taghi Saffar et al. “Intent recognition in a simulated maritime multi-agent domain”. In: *Machine Learning, Optimization, and Big Data: First International Workshop, MOD 2015, Taormina, Sicily, Italy, July 21-23, 2015, Revised Selected Papers 1*. Springer. 2015, pp. 158–170.

[Saf+15b] Mohammad Taghi Saffar et al. “Intent understanding using an activation spreading architecture”. In: *Robotics* 4.3 (2015), pp. 284–317.

[SND06] Joe Saunders, Chrystopher L Nehaniv, and Kerstin Dautenhahn. “Teaching robots by moulding behavior and scaffolding the environment”. In: *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*. 2006, pp. 118–125.

[Sca02] Brian Scassellati. “Theory of mind for a humanoid robot”. In: *Autonomous Robots* 12 (2002), pp. 13–24.

[Sch96] Stefan Schaal. “Learning from demonstration”. In: *Advances in neural information processing systems* 9 (1996).

[SH05] Oliver C Schrempf and Uwe D Hanebeck. “A generic model for estimating user intentions in human-robot cooperation”. In: *International Conference on Informatics in Control, Automation and Robotics*. Vol. 3. SCITEPRESS. 2005, pp. 251–256.

[Sci+15] Alessandra Sciutti et al. “Investigating the ability to read others’ intentions using humanoid robots”. In: *Frontiers in psychology* 6 (2015), p. 1362.

[SCV19] Lei Shi, Cosmin Copot, and Steve Vanlanduit. “What are you looking at? detecting human intention in gaze based human-robot interaction”. In: *arXiv preprint arXiv:1909.07953* (2019).

[SCV20] Lei Shi, Cosmin Copot, and Steve Vanlanduit. “Visual Intention Classification by Deep Learning for Gaze-based Human-Robot Interaction”. In: *IFAC-PapersOnLine* 53.5 (2020), pp. 750–755.

[SKK08] Bruno Siciliano, Oussama Khatib, and Torsten Kröger. *Springer handbook of robotics*. Vol. 200. Springer, 2008.

[SRG17] Gunnar A Sigurdsson, Olga Russakovsky, and Abhinav Gupta. “What actions are needed for understanding human actions in videos?” In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2137–2146.

[Sin+20] Ronal Singh et al. “Combining gaze and AI planning for online human intention recognition”. In: *Artificial Intelligence* 284 (2020), p. 103275.

[SBP22] Gary B Smith, Vaishak Belle, and Ronald PA Petrick. “Intention recognition with ProbLog”. In: *Frontiers in Artificial Intelligence* 5 (2022), p. 806262.

[SA22] Shubham Sonawani and Heni Amor. “When And Where Are You Going? A Mixed-Reality Framework for Human Robot Collaboration”. In: *5th International Workshop on Virtual, Augmented, and Mixed Reality for HRI*. 2022.

[SPB10] Nikolay Stefanov, Angelika Peer, and Martin Buss. “Online intention recognition for computer-assisted teleoperation”. In: *2010 IEEE International Conference on Robotics and Automation*. IEEE. 2010, pp. 5334–5339.

[Suz+07] Kenta Suzuki et al. “Intention-based walking support for paraplegia patients with Robot Suit HAL”. In: *Advanced Robotics* 21.12 (2007), pp. 1441–1469.

[Tah06] Karim A Tahboub. “Intelligent human-machine interaction based on dynamic bayesian networks probabilistic intention recognition”. In: *Journal of Intelligent and Robotic Systems* 45 (2006), pp. 31–52.

[TFA10] Yusuke Tamura, Tomohiro Fukuzawa, and Hajime Asama. “Smooth collision avoidance in human-robot coexisting environment”. In: *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE. 2010, pp. 3887–3892.

[Tam+12] Yusuke Tamura et al. “Development of pedestrian behavior model taking account of intention”. In: *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE. 2012, pp. 382–387.

[TIS04] Jun Tani, Masato Ito, and Yuuya Sugita. “Self-organization of distributedly represented multiple behavior schemata in a mirror system: reviews of robot experiments using RNNPB”. In: *Neural Networks* 17.8-9 (2004), pp. 1273–1289.

[TC17] Ajay Kumar Tanwani and Sylvain Calinon. “A generative model for intention recognition and manipulation assistance in teleoperation”. In: *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2017, pp. 43–50.

[Tap+19] Adriana Tapus et al. “Perceiving the person and their interactions with the others for social robotics–a review”. In: *Pattern Recognition Letters* 118 (2019), pp. 3–13.

[Tav+07] Alireza Tavakkoli et al. "A vision-based architecture for intent recognition". In: *Advances in Visual Computing: Third International Symposium, ISVC 2007, Lake Tahoe, NV, USA, November 26-28, 2007, Proceedings, Part II 3*. Springer. 2007, pp. 173–182.

[Top17] Elin Anna Topp. "Interaction patterns in human augmented mapping". In: *Advanced Robotics* 31.5 (2017), pp. 258–267.

[Tra+06] J Gregory Trafton et al. "Communicating and collaborating with robotic agents". In: *Cognition and multi-agent interaction: From cognitive modeling to social simulation* (2006), pp. 252–278.

[Tsa+23] Evangelos Tsagkournis et al. "A Supervised Machine Learning Approach to Operator Intent Recognition for Teleoperated Mobile Robot Navigation". In: *IFAC-PapersOnLine* 56.2 (2023), pp. 8333–8338.

[Var+18] Dimitrios Varytimidis et al. "Action and intention recognition of pedestrians in urban traffic". In: *2018 14th International conference on signal-image technology & internet-based systems (SITIS)*. IEEE. 2018, pp. 676–682.

[Vem+23] Sai Vemprala et al. "Chatgpt for robotics: Design principles and model abilities. 2023". In: *Published by Microsoft* (2023).

[VBK24] Mudit Verma, Siddhant Bhambri, and Subbarao Kambhampati. "Theory of Mind abilities of Large Language Models in Human-Robot Interaction: An Illusion?" In: *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*. 2024, pp. 36–45.

[VTZ16] David Vernon, Serge Thill, and Tom Ziemke. "The role of intention in cognitive robotics". In: *Toward Robotic Socially Believable Behaving Systems-Volume I: Modeling Emotions* (2016), pp. 15–27.

[VGC19] Samuele Vinanzi, Christian Goerick, and Angelo Cangelosi. "Mindreading for robots: Predicting intentions via dynamical clustering of human postures". In: *2019 Joint IEEE 9th international conference on development and learning and epigenetic robotics (ICDL-EpiRob)*. IEEE. 2019, pp. 272–277.

[Vin+19] Samuele Vinanzi et al. "Would a robot trust you? Developmental robotics model of trust and theory of mind". In: *Philosophical Transactions of the Royal Society B* 374.1771 (2019), p. 20180032.

[Völ+15] Benjamin Völz et al. "Feature relevance estimation for learning pedestrian behavior at crosswalks". In: *2015 IEEE 18th International Conference on Intelligent Transportation Systems*. IEEE. 2015, pp. 854–860.

[Völ+16] Benjamin Völz et al. “A data-driven approach for pedestrian intention estimation”. In: *2016 ieee 19th international conference on intelligent transportation systems (itsc)*. IEEE. 2016, pp. 2607–2612.

[VNK15] Michalis Vrigkas, Christophoros Nikou, and Ioannis A Kakadiaris. “A review of human activity recognition methods”. In: *Frontiers in Robotics and AI* 2 (2015), p. 28.

[Wal+23] Michael Walker et al. “Virtual, augmented, and mixed reality for human-robot interaction: A survey and virtual design element taxonomy”. In: *ACM Transactions on Human-Robot Interaction* 12.4 (2023), pp. 1–39.

[Wan+24a] Chao Wang et al. “LaMI: Large Language Models for Multi-Modal Human-Robot Interaction”. In: *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 2024, pp. 1–10.

[Wan+24b] Jiaqi Wang et al. “Large language models for robotics: Opportunities, challenges, and perspectives”. In: *arXiv preprint arXiv:2401.04334* (2024).

[Wan+18] Weitian Wang et al. “Human intention prediction in human-robot collaborative tasks”. In: *Companion of the 2018 ACM/IEEE international conference on human-robot interaction*. 2018, pp. 279–280.

[Wan+19] Wendong Wang et al. “Bionic control of exoskeleton robot based on motion intention for rehabilitation training”. In: *Advanced Robotics* 33.12 (2019), pp. 590–601.

[Wan+17] Yiwei Wang et al. “Human intention estimation with tactile sensors in human-robot collaboration”. In: *Dynamic Systems and Control Conference*. Vol. 58288. American Society of Mechanical Engineers. 2017, V002T04A007.

[Wan+13] Zhikun Wang et al. “Probabilistic movement modeling for intention inference in human–robot interaction”. In: *The International Journal of Robotics Research* 32.7 (2013), pp. 841–858.

[Wei+18] Ping Wei et al. “Where and why are they looking? jointly inferring human attention and intentions in complex tasks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 6801–6809.

[Wei+19] Shouyang Wei et al. “An integrated longitudinal and lateral vehicle following control system with radar and vehicle-to-vehicle communication”. In: *IEEE Transactions on Vehicular Technology* 68.2 (2019), pp. 1116–1127.

[Wil+17] Francis R Willett et al. “A comparison of intention estimation methods for decoder calibration in intracortical brain–computer interfaces”. In: *IEEE Transactions on Biomedical Engineering* 65.9 (2017), pp. 2066–2078.

[WP83] Heinz Wimmer and Josef Perner. “Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children’s understanding of deception”. In: *Cognition* 13.1 (1983), pp. 103–128.

[WSH12] David Windridge, Affan Shaukat, and Erik Hollnagel. “Characterizing driver intention via hierarchical perception–action modeling”. In: *IEEE Transactions on Human-Machine Systems* 43.1 (2012), pp. 17–31.

[WVS23] Christopher Yee Wong, Lucas Vergez, and Wael Suleiman. “Vision-and Tactile-Based Continuous Multimodal Intention and Attention Recognition for Safer Physical Human–Robot Interaction”. In: *IEEE Transactions on Automation Science and Engineering* (2023).

[WC20] Qingcong Wu and Ying Chen. “Development of an intention-based adaptive neural cooperative control strategy for upper-limb robotic rehabilitation”. In: *IEEE Robotics and Automation Letters* 6.2 (2020), pp. 335–342.

[Xin+20] Yang Xing et al. “An ensemble deep learning approach for driver lane change intention inference”. In: *Transportation Research Part C: Emerging Technologies* 115 (2020), p. 102615.

[Yan+23] Bo Yang et al. “Natural grasp intention recognition based on gaze in human–robot interaction”. In: *IEEE Journal of Biomedical and Health Informatics* 27.4 (2023), pp. 2059–2070.

[Yan+22] Shengtian Yang et al. “Interaction Intention Recognition via Human Emotion for Human-Robot Natural Interaction”. In: *2022 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)*. IEEE. 2022, pp. 380–385.

[Yan+21] Ziyi Yang et al. “An intention-based online bilateral training system for upper limb motor rehabilitation”. In: *Microsystem Technologies* 27 (2021), pp. 211–222.

[YO07] So-Jeong Youn and Kyung-Whan Oh. “Intention recognition using a graph representation”. In: *World Academy of Science, Engineering and Technology* 25 (2007).

[Yua+19] Liangzhe Yuan et al. “Human gaze-driven spatial tasking of an autonomous MAV”. In: *IEEE Robotics and Automation Letters* 4.2 (2019), pp. 1343–1350.

[Zen+18] Yunxiu Zeng et al. “Inverse reinforcement learning based human behavior modeling for goal recognition in dynamic local network interdiction”. In: *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence*. 2018.

[ZKL23] Chenyuan Zhang, Charles Kemp, and Nir Lipovetzky. “Goal recognition with timing information”. In: *Proceedings of the international conference on automated planning and scheduling*. Vol. 33. 1. 2023, pp. 443–451.

[Zha+19a] Dalin Zhang et al. “Making sense of spatio-temporal preserving representations for EEG-based human intention recognition”. In: *IEEE transactions on cybernetics* 50.7 (2019), pp. 3033–3044.

[Zha+20] Hongjia Zhang et al. “Research on a pedestrian crossing intention recognition model based on natural observation data”. In: *Sensors* 20.6 (2020), p. 1776.

[Zha+23] Zhang Zhang et al. “Intention recognition for multiple agents”. In: *Information Sciences* 628 (2023), pp. 360–376.

[Zha+21] Xue Zhao et al. “Human–robot collaborative assembly based on eye-hand and a finite state machine in a virtual environment”. In: *Applied Sciences* 11.12 (2021), p. 5754.

[ZCS08] Chun Zhu, Qi Cheng, and Weihua Sheng. “Human intention recognition in smart assisted living systems using a hierarchical hidden markov model”. In: *2008 IEEE International Conference on Automation Science and Engineering*. IEEE. 2008, pp. 253–258.

[ZSS08] Chun Zhu, Wei Sun, and Weihua Sheng. “Wearable sensors based human intention recognition in smart assisted living systems”. In: *2008 International Conference on Information and Automation*. IEEE. 2008, pp. 954–959.

[Zun+17] Andrea Zunino et al. “What will i do next? The intention from motion experiment”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 2017, pp. 1–8.

Appendix

Table 1: Summary of Psychological Theories on Intention Recognition

References	Overview	Findings/Contributions
[Hei13]	Suggested people have "folk psychology"	Used to infer the meaning behind the actions of others for a given situation.
[PW78]	Coined the term "Theory of Mind (ToM)"	Formalized ToM concept as the ability to recognize others' mental states.
[WP83]	Investigated children's understanding of multiple mental states and false beliefs	Ability of children to associate relationships between multiple mental states by the ages of 4 to 6 years.
[PBCR95]	Explored the relation between mental state association and autism in children	Children with autism fail to associate mental states with others.
[BC+95]	Explored difficulties reading intentions via eye gaze in children with autism	Associated the inability of children with autism to assign mental state to others with the failure to read intention through eye gaze.
[Kal19]	Viewed intention from Wittgensteinian perspective	Argueed intentions are patterns of behaviour extended over time and context.
[RC04]	Described mirror neuron mechanism involved in action observation and imitation	Humans show evidence of having a mirror neuron system which fires neurons when a particular action is observed as well as performed by them.

Table 2: Summary of Computational Methods

References	**Overview**	**Findings/Contributions**
Foundational ‘Theory of Mind’ (ToM) in Robots [Sca02, Les94, CA98, NAH02, Sch96, JD05, FMJ02, Bre+05, Gra+05, Tra+06]	First application of psychological ToM using different modules in a humanoid robot.	- Used modules such as eye direction detection, intentionality detection, shared attention mechanism, as well as recognizing human actions and taking perspectives. - Proposed implementing ToM-based modules improves communication and helped to learn from interactions.
[GBB07, SKK08]	Real-time ToM inference for beliefs, desires, intentions in multi-participant scenarios.	- Robot observed false-belief task of chips and cookies swap. - Assisted humans based on inferred beliefs in real time.
Teleological and Simulation Theories [BO19a, BO19b, POWW21]	Teleological (outcome-based) and simulation (internal modeling) theories.	- Talked about seven different implementations of complex architectures in robots. - Teleological theory is used to infer intentions behind others’ actions based on the outcomes of these actions. - Simulation theory is used to simulate the mental states of others internally to understand them. - Reviewed other computational ToM models.
Bayesian Theory of Mind (BToM) [BST09, Bak+17]	Formulated the problem of understanding actions as a Bayesian inference problem.	- Used rational probabilistic planning in Markov decision problems to model the causal relationship between goals, actions, and beliefs. - Validated via human participants on belief and desire attribution. - Achieved strong predictive accuracy, and capturing hidden states and percepts.
[Sin+20, Per+11, Tah06, SH05, KH10, JA18, Dun+15, Tam+12, Top17, CT16]	Probablistic and Bayesian Methods for intention recognition.	- These methods are beneficial for modeling uncertainty and making inferences in real-world scenarios. - Proof-of-concept studies in this area.

Dual Computational Approach [LSB19, AZN98, For+95, PW13, Pea14]	Storyteller–listener model using POMDP for the storyteller and DBN for the listener's attentiveness.	- Storyteller uses nonverbal cues to infer listener's attentive state. - Listener uses a myopic DBN policy. - Storyteller model outperformed state-of-the-art attention recognition methods. - Listener model communicated attentiveness better than traditional signalling methods.
[VGC19, Vin+19]	Integrated ToM, trust, and episodic memory in a cognitive system for artificial agents.	- Robot acted as trustor and discerned helpers versus tricksters. - Matched performance of 5-year-old children in trust tasks. - Incorporated psychological tasks to evaluate reliability of human partners.
ToMNet [Rab+18]	Meta-learning approach to infer agent mental states with minimal data.	- Learned strong priors for agent behavior. - Successfully handled false-belief tasks after few observations. - Illustrated how deep learning can achieve flexible ToM reasoning.
Multi-Agent ToM [Li+22a]	Used RNNP and multi-layer connectionist model to mimic mirror neuron systems.	- Used recurrent neural networks with parametric biases (RNNPB) to implement mirror neuron-like systems for robots. - RNNPB model showed how complex behaviours can emerge from simpler learned patterns. - Biologically inspired GeneRec algorithm for iCub's motor-visual integration. Combined mirror neuron and simulation theory for intention reading.

[Pre+19]	Mental Time Travel concept.	- Ability to mentally project oneself into the past or the future. - Based on Gaussian process latent variable models. - Experiments designed around memory-based tasks (face recognition, speaker recognition, emotion recognition, touch interaction, and action recognition).

Table 3: Summary of Activity-based Intention Recognition

References	**Overview**	**Findings/Contributions**
HMM-Based Activity -[Kel+10], -[Cra+18, Kel+08, Tav+07, Pet+19, SPB10, AK06, BED08, ZSS08, ZCS08]	- Introduced a Vision-based system and HMMs for human activity recognition, leveraging ToM concepts - Other HMM-based intention recognition works	- Disambiguated similar actions via contextual information. - Allowed for detecting different intentions behind visually similar activities.
[RABC15]	Proposed a framework to perform activity recognition using a humanoid robot.	- Enabled the iCub robot to use semantic reasoning to infer high-level behaviours from low-level sensor data.
[SRG17, VNK15, JBD19, Tap+19]	Provided overviews of state-of-the-art HAR methods, datasets, metrics, challenges	- Insight on ambiguity of temporal boundaries of activities. - Discussed traditional ML (SVM, KNN, Decision Trees) and modern NN approaches (ANN, CNN, RNN)
Traditional ML -[DSS15, Wan+17, Völ+15, KM17, MÅ18, NZR18, Zha+20, Wan+19]	SVM, Decision Trees, KNN, rule-based algorithms for intention recognition	-

Neural Net Approaches -[Zha+19a, LH19, Gol+19, NZR18, Var+18, Wan+18, Völ+16, PL16, Li+17, Rak+19, Yan+21]	RNNs, CNNs, and Extreme Large Machine Algorithms for intention recognition	-

Table 4: Summary of Plan-based Intention Recognition

References	**Overview**	**Findings/Contributions**
[AA07, Sad11]	Discussed frameworks for plan-based recognition	- Reviewed plan libraries, interface agents, and underlying theories. - Discussed the use of logic-based formalisms, deduction and abduction.
Monte Carlo and POMDP [OMM19]	Developed proactive intention recognition in search and rescue robotics	Supported the robot's exploration strategy by providing an entropy reduction bonus to the reward function.
Multi-Agent System [Zha+23], [AA14]	Recognized and clustered intentions in multi-agent settings.	Used MDP, behaviour trees, landmark-based models, and plan libraries.
Navigation Assistance [Den+19], [Par+16]	Inferred human plan for navigation (wheelchairs, robot motion)	- Used clothoidal (Euler) paths for complex environments. - Used Gaussian process models to classify intent to interact. - Smoother, conflict-free movement outcomes.

Table 5: Summary of Goal-based Intention Recognition

References	**Overview**	**Findings/Contributions**
DBN/Probabilistic [Mur02, SBP22, Bra+22, TFA10]	Predicted the next goal in terms of predefined landmarks, tasks, and user intentions.	- The DBN model predicted which predefined landmark (or goal) the person is heading toward next. - Achieved 75 per cent accuracy using probabilistic logic programming to reason about intentions (make coffee or prepare a meal) in smart home settings. - Used Kalman filters and HMM to reduce uncertainty in collaborative tasks.
Imitation and Goal-Inference [BWG00, Bre+06, Erl+06, JB06a, CDI06, JB06b, SND06, Bre+06, CB07]	- Suggested that imitation is goal-directed and not simply a copy of the observed actions. - For motions that were not goal-directed, a mental model of the demonstrator is needed.	- Learned to imitate actions relied on the ability of the imitating agent to infer the intentions of the demonstrator. - Used mental modelling and analyzed social cues for better intention detection.
Rationality of Actions [BDK14]	Focussed on computational models that simulated how humans detect if an action was intentional and thus predicted the likely goal.	- Assumed that people followed the principle of rationality. - Studies found the model's performance closely matched that of the human participants in determining the intentionality of the actions and the predicted goals.
Inverse Reinforcement Learning [Zen+18]	Inverse Reinforcement Learning for goal recognition in dynamic network interdiction scenarios.	- Learned a reward function from observed trajectories. - Outperformed other models in tracking accuracy and effectiveness in network interdiction.

Explainability and Timing - [AMV23] - [ZKL23]	- Provided explanations for recognized goals. - Incorporated timing information of the action taken by an agent to carry out certain actions.	- Adapted the Weight of Evidence framework to clarify why one goal hypothesis should be chosen over the other. - Timing information improved goal recognition accuracy in scenarios with few actions.
Deep Learning [Chi+23]	Introduced GRNet (uses RNN-based architecture) for goal recognition.	- Outperformed state-of-the-art systems - LGR in accuracy and runtime. - Combined GRNet with LGR yields better performance in partial information scenarios.
Broader Applications [Sad11, CG90, AA07, Hon01, Les98, PW13, Gei02, Hai+03, Per+10, PH11, PA11, Roy+09, Tah06, Hei04, MG04]	Intention recognition across stories, human-computer interaction, monitoring traffic, assistive care, military activities.	-

Table 6: Summary of Action-based Intention Recognition

References	Overview	Findings/Contributions
[Zun+17]	Predict intentions from the initial motion without contextual information	- Computer vision methods outperformed humans in predicting the intention behind the observed action. - Demonstrates subtle motion cues as intention indicators
Motion Cues [RD17, RTD19, AH10, KH10, Per+11, JA18, Dun+15, Tam+12]	Used a neural network to model the non-linear dynamics of the human arm motion.	- Defined intentions as the 3D end goal of the actions. - Used an extended Kalman filter for expectation maximization to infer the intentions. - Parameter-based methods have also seen popularity over the years.

Probabilistic Modeling [Tah06]	Used a modified intention–action–state scenario modelled by DBNs.	- DBNs to recognize the human's intentions (e.g., moving towards or away from an object). - Tested in a simulation environment. - Robot demonstrated compliance with the recognized intentions.
Active Intention Leading [Omo+08]	Proposed AIL where the robot acts so user can easily interpret its intention.	- Less computationally heavy than passive recognition. - Tested this approach in a "hunter game" scenario. - Proposed approach outperformed passive intention recognition in two different tasks.
Ontology-based hierarchical user intentions [JKC08]	Hierarchical user intentions with sensor data (temperature, humidity, vision, auditory).	- Used RuleML and domain knowledge. - Used conditional entropy to make the robot's behaviour proactive.
[RABC15]	Two-level approach (low-level color features and high-level decision trees)	- Implemented on iCub. - Achieved about 85% accuracy in 0.12s decision time.
Probabilistic Movement Primitives (ProMPs) [Der+17]	Enable iCub to predict human intention during physical interaction.	- ProMPs learned from demonstration. - Explicit goal information not required. - Successfully tested in real robot with tasks such as reaching and sorting.
Teleoperation and shared control task [TC17, JA18, SPB10, AK06]	Improved shared control in remote manipulation with HSMM learning from demonstrations.	- Used demonstrations by segmenting them into meaningful parts. - Allowed partial autonomy for the robot or full teleoperation. - Tested on a Baxter robot in tasks such as reaching a movable target and opening a valve.
Indirect speech acts (ISAs) [BWS17]	Understood directives from indirect speech acts.	- Focused on rule-based mechanism. - Evaluated in simple tasks such as knocking over colored towers.

[PMP18]	Combined Theory of Mind, MDP, and HMM for worker's intention recognition in warehouses.	- Achieved high accuracy in predicting worker's changing goals. - Supported dynamic goal switching mid-task.
Physical human-robot interaction (pHRI) [Los+18]	Reviewed intent detection, arbitration, and communication for physical human-robot interaction.	- Explored in detail how robots can detect human intentions, either binary or more complex and continuous ones. - Explored sensors (force, muscle activity, and neural activity).
Neural Network Models [LH19, LG13, Jan+14, CNP06, Zha+19a, Gol+19, NZR18, Var+18, Wan+18, Völ+16, PL16, Li+17, Rak+19, Yan+21, Hua+15b, Li+18, Len+12, Li+17, Wil+17, Wei+19, Buc94]	Used a range of neural or control-based approaches to infer motion intention.	- RBFNN improved synchronisation and force control. - Control methods for intention recognition have also seen widespread interest in research.
Deep Neural Networks [LZD20]	Used DNN to process RGB images and optical flow for intention recognition.	- Used two-stream architecture, a spatial network (skeleton joint information) and a temporal network (optical flow). - Achieved accuracy of 74 per cent on their dataset and 77 per cent on the Intention from Motion (IFM) dataset.
Visual semantics and natural language processing (NLP) [Li+21]	Integrated visual semantics with natural language to perform the tasks while looking at feedback from the users' facial expressions.	- Used image segmentation, CRF, user expressions, and rule matching.

Neural Network Architectures [PF23]	Compared LSTM, Transformer, MLP-Mixer for arm movement prediction in VR-based dataset.	- Transformer encoder model performed the best with 82.74 per cent accuracy for predicting movements. - MLP-mixer had lower computational complexity than the transformer model. - LSTM model was the worst performing model of the three.
[God+22]	Presented a novel approach for recognizing hand gestures based on the lightmyography (LMG) signals and transformer based deep learning models.	- Used two transformer based deep learning models: Temporal Multi-Channel Transformer (TMC-T) and Temporal Multi-Channel Vision Transfer (TMC-ViT) to classify gestures. - The two transformer models outperformed CNN, Bi-LSTM, LDA, SVM, and RF and achieved accuracies of 94.03 per cent and 93.69 per cent respectively.
[WVS23]	Recognized intention from human actions with visual and tactile sensors.	- A supervised machine learning algorithm is used to train the model with touch location, human pose, and gaze direction among others. - Achieved 86 per cent accuracy in classifying whether a human touch is intentional or not.
[Tsa+23]	Proposed an MLOII model that is used to infer the navigational intent online.	MLOII performed better with fewer obstacles and more direct routes as compared to complex environments (against BOIR model).
TLP Model [Wan+18]	Hand-over intentions from multi-modal sensors (IMU, EMG) combined with natural language instructions.	- Model learned from about 5000 sets of hand-over demonstrations. - Achieved prediction accuracy of 99.7 per cent even from partial motion.

Anticipating Actions with GMM [Dua+18]	Understood how eye gaze, head orientation, and arm movement helped predict actions.	- Applied GMM to simulate arm trajectories and incorporate gaze patterns for predictions. - Model was tested on an iCub robot.
ToM based Architecture [VGC19]	Processed low-level skeletal data and used a high-level module to infer the human partner's goal.	- Implemented architecture on an iCub robot to collaborate on a toy block task. - Took an average of 4.49s to predict intentions accurately.
[GB20]	Robot used belief-desire-intention with body poses being detected by the OpenPose library.	- Classified user behaviour to decide the next course of action. - Performed tasks such as delivering or exchanging crates or avoiding to humans to prevent interference.
[WC20]	Used sEMG to control a rehabilitation robot for upper-limb patients.	- Introduced an adaptive neural cooperative control strategy. - Used RBFN to estimate motion intention from filtered sEMG signals.
Context-aware Intention Recognition [Liu+21a]	Combined ST-GCN-LSTM for skeletal data and YOLO v3 for object data to infer context-aware intention.	Evaluated in real-world setting to show model prediction accuracy.
Deep learning algorithm [For+21]	Combined RGB-D camera and deep learning algorithms to predict the spatial location of object's final position.	- Made predictions based on positions of key body joints (shoulders, elbows, wrists) processed through an RNN with a GRU-based encoder-decoder architecture. - Reduced collisions by 38 per cent when combined with YOLO v3. - Reduced collision by 70 per cent when combined with human-tracking.

Physical Human-Robot Interaction Primitives [Lai+22]	Proposed Physical Human-Robot Interaction Primitives that capture user's force based intent.	- Tested on 7-dof robot arm for obstacle avoidance and target-directed reaching. - Predicts user's intention from forces exerted during the interaction.
Review of Contact Robots [Li+22b]	Surveyed interaction control in robots that physically interact with human users (rehabilitation, teleoperation, and collaborative manufacturing).	- Discussed the use of EMG, EEG, and tactile sensors. - Additionally, discussed the combination of sensor inputs with machine learning models to classify intended actions.
Cross-View Method [Ni+23]	Generated multi-angle body and face views to improve intention recognition in collaborative assembly tasks.	- Used generative model and RNN to fuse spatial and temporal information for inferring the intentions of the user. - Cross-view method improved the fluency and efficiency of the robots' performance to near human levels.

Table 7: Summary of Gaze-based Intention Recognition

References	**Overview**	**Findings/Contributions**
Context awareness and intention understanding [Qui+12]	Used gaze estimation and gesture interpretation to understand intentions in different contexts.	- Used face detection, ellipsoid mapping, and eye tracking to estimate the direction of gaze. - Predefined ontology helps in contextualizing user intentions.
Gaze in Rehabilitation System [NR13]	Gaze-based intention detection for ARMin III upper extremity rehabilitation robot.	- Provided more natural user-robot interaction in the virtual environment. - Allowed the robot to support patients with gaze data.

Collaborative Task [Hua+15a]	Used support vector machines (SVM) to classify and predict the customer's intended ingredient choice.	- Achieved 76% accuracy. - Anticipated user's request 1.8s before verbal cue.
Assistive Application [LZ17]	Intentional gaze vs intention-free gaze using support vector machine (SVM) classifier to assist elderly or impaired individuals with ADL.	- SVM classifier based on features like gaze dwelling time, pupil size variation, and gaze speed. - Approximately 75% correctness with intention inference in a simulated home care environment.
Gaze and Model-based AI [Sin+20]	Merged gaze data with traditional model-based AI planning to improve intention recognition.	- The model combined short-term (proximal) and long-term (distal) intentions. - 22% increase in accuracy over single-mode. - Robust to semi-rational or deceptive gaze behaviour.
Gaze information with Motion data [RKD18]	Fused gaze information with motion data using Gaze-based Multiple Model Intention Estimator (G-MMIE) algorithm.	- Provided more accurate intention inference.
Saccadic eye movements [SCV19]	Compared hypothetical gazes (from saliency maps) and actual gazes using Earth Mover's Distance (EMD) and 1-Nearest Neighbor (1-NN) classifier.	- 92.2% accuracy in predicting the object to be selected by humans. - Outperformed fixation-based methods.
Egocentric Perspectives [Kim+19]	First-person camera approach for a soft wearable hand robot assisting user in grasping and releasing tasks.	- Faster and higher accuracy than EMG-based intention detection methods. - Demonstrated success with a patient having spinal cord injury.

Deep Learning for Wearable Eye Tracker [SCV20]	FCN with CBAM and Residual Blocks to predict the human's visual intention with high accuracy.	- F1 score of 0.971 (single object), and 0.962 (multi-object). - Addressed issues related to fixation-based gaze systems.
Shared Autonomy [FB21]	GHMM to model and predict user intention primarily based on gaze scan paths for pick-and-place tasks.	- Demonstrated strong generalizability across users and tasks. - High accuracy rates.
Cascade Effect [GCR21]	Combined cascade effect hypothesis and LSTM-based neural network classification.	- 75% accuracy up to 2s before user's selection.
Eye-Tracking and Gesture [Zha+21]	Joint gaze and hand gestures for intention recognition in a virtual assembly environment.	- FSM to control the robot to make assistive manoeuvres. - Neural net recognized 9 hand gestures. - Probabilistic Roadmap Planner for collision-free robot assistance.
Probabilistic Modeling [Bel+22]	Probabilistic model in teleoperated object manipulation using gaze and motion data.	- GHMMs trained on sequences of gaze and motion data. - Models performed predictions 0.5s before an action ends. - Robust to cluttered scene with partial occlusions and bimanual tasks.
Modern Eye-tracking Technologies [Yan+23]	Recognized grasping intentions for disabled users by analyzing gaze and scene camera data.	- SVM, KNN, SGD, and Decision Trees to classify user intentions. - Achieved an accuracy of 89 per cent on training objects and 85 per cent on new unseen objects. - Distinguished “viewing” versus “grasping” fixations.

Table 8: Summary of Non-verbal Intention Recognition

References	Overview	Findings/Contributions
Gesture Classification [Neh+05]	Classified gestures into five broad categories: manipulative, expressive side effects, symbolic, interactional, and referential gestures to facilitate intent recognition.	- Emphasized context in distinguishing similar gestures.
Weighted Probabilistic State Machines [AH10]	Proposed approach is used to recognize implicit vs explicit intentions.	- Each intention is modelled as a separate state machine. - Successfully recognized picking, placing, passing objects (explicit) and piling, un-piling (implicit).
[Qui+12]	Focused on periodic (repetitive) and deictic (pointing) gesture recognition.	- Analyzed the frequency and trajectory of hand movements. - Used deictic gestures to divert the robot's attention to important places and objects.
Hierarchical Perception-Action Model [WSH12]	Classified intentional driving behaviours with generative, discriminative, hybrid approaches.	- Eye-tracking and environmental data collected from an instrumented vehicle. - Hybrid approach yields the best performance in classifying driver intentions.
[Mea+14]	Used EEG signals (SSVEP approach) to interpret the intention to perform gestures on a remote robotic hand.	Brain-computer interface for teleoperation tasks.
Emotional Expression [Che+17, Che+20, Yan+22, YO07, Kur+19]	Recognized emotional or affective intention from facial expressions.	- Used Candide3 face model to recognize seven basic emotions. - Used an information-driven fuzzy friend-Q (IDFFQ) learning mechanism to achieve 85.71 per cent accuracy in intention understanding. - Merged body actions with facial cues for 94.57% accuracy. - Fuzzy learning approaches.

Hierarchical Model [Wei+18]	Addressed the joint problem of intention recognition, attention, and tasks.	- Used RGB-D videos to infer attention, intention, and task using Human-Attention-Object (HAO) graph. - Outperforms baselines on RGB-D videos.
Recursive Bayesian Filtering Framework [JA19]	Accounted for multiple non-verbal signals to provide assistance to users during teleoperation tasks.	- Combined multiple non-verbal observations to probabilistically reason about the user's intended goal. - Models user's suboptimal or inconsistent behaviors with an adjustable rationality parameter. - Provided personalized assistance.

Table 9: Summary of Object-based Intention Recognition

References	**Overview**	**Findings/Contributions**
[Liu+21b]	Markov Random Fields for inferring user's likely object usage intentions	- Outperformed Recursive Bayesian Incremental Learning. - Targeted for objects used for daily activities.
[Dun14]	Proposed a novel probabilistic graphical model called Object-Action Intention Network (OAIN).	- Recognized human intentions based on the objects in the scene and the potential actions associated with these objects.

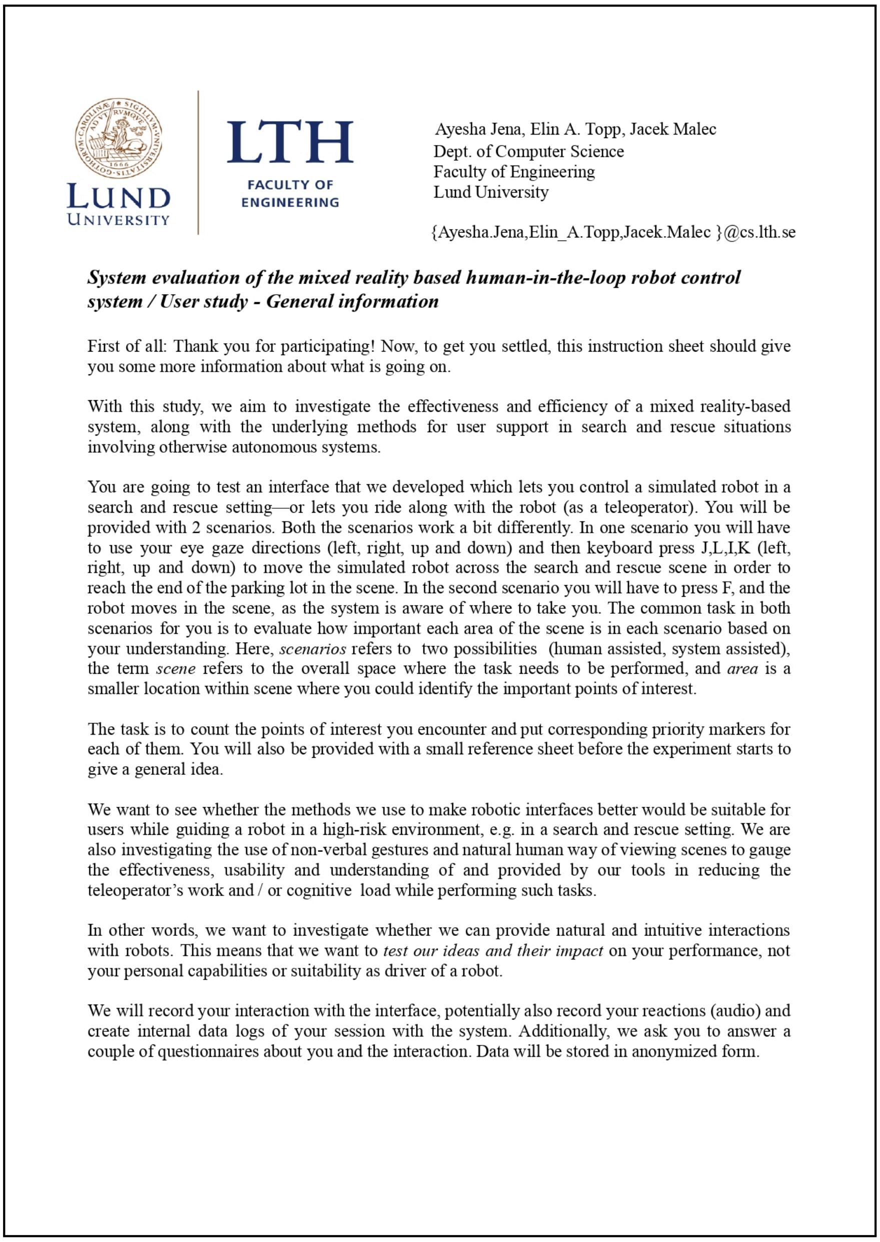
Table 10: Summary of Robot Intent Recognition

References	**Overview**	**Findings/Contributions**
Gaze-Like Cues in Humanoid Robots [Mut+09]	Tested whether humans could detect robot gaze cues to infer intentions.	- Participants recognized robot intent better with gaze cues. - Reinforced the importance of gaze in HRI.

Light and Motion-Based Approach [Lem+21]	Tested three light-based and three motion-based methods while working with a human.	- Light-based LED method on the wrist of the robot was the most noticeable among all six signals. - Head-pan was most noticeable among motion-based.
Visual Communication Cues [Koa+13]	Explored the possibility of untrained humans comprehending the robot's intentions correctly.	- Wizard-of-oz approach for guiding participants to specific sound sources. - Participants succeeded with gaze and head motion as important communication cues.
[Sci+15]	Used humanoid robots to study human ability to read movements as intention indicators.	- Robots offered control, natural reciprocity, and shared physical space.
Robot Understandability [HB18]	Defined robot understandability, significance, and design principles.	- Introduced a term called "communicative actions", and a model of interaction understanding based on the first-level and second-level theory of mind.
VR, AR, XR in HRI [Wal+23, Mat06, LHS13, LKK11, PK09, CA15, Coo+14, SA22, Ish+09, Omi+15, Ghi+14, New+22, Ros+19]	VR, AR, XR to help communicate robot's motion intentions and behaviours which enhances collaboration.	- Optimized collaboration. - Visual and auditory modalities to convey the intent and state of the robot. - MAR-CPS workspace projection system. - Navigation area markers using gestures.
[Bam+19, Yua+19]	AR to communicate dynamic trajectory updates.	Indicated grasp points and target positions in response to human presence.

9 Appendix

9.1 Appendix A - Documents for User Study



LUND UNIVERSITY

LTH FACULTY OF ENGINEERING

Ayesha Jena, Elin A. Topp, Jacek Malec
Dept. of Computer Science
Faculty of Engineering
Lund University

{Ayesha.Jena,Elin_A.Topp,Jacek.Malec }@cs.lth.se

System evaluation of the mixed reality based human-in-the-loop robot control system / User study - General information

First of all: Thank you for participating! Now, to get you settled, this instruction sheet should give you some more information about what is going on.

With this study, we aim to investigate the effectiveness and efficiency of a mixed reality-based system, along with the underlying methods for user support in search and rescue situations involving otherwise autonomous systems.

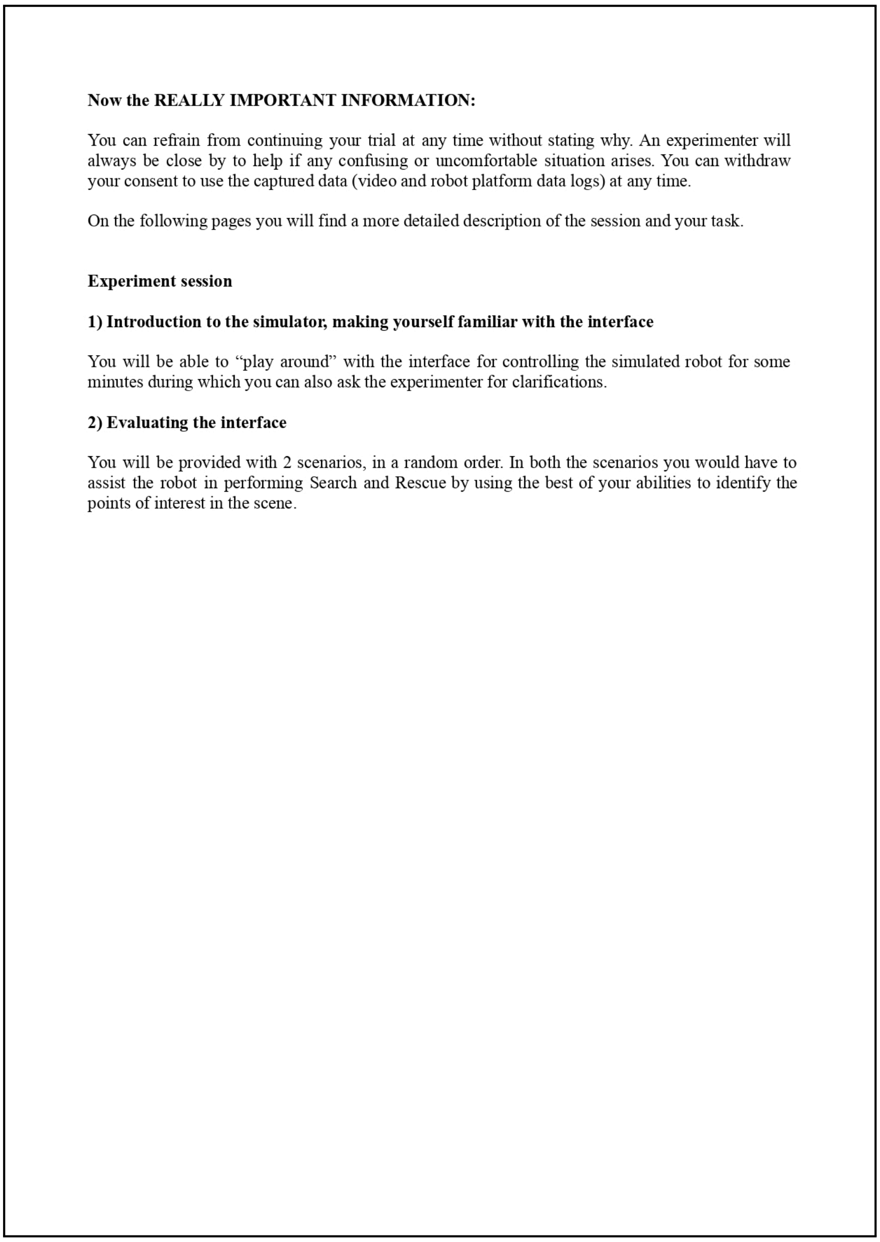
You are going to test an interface that we developed which lets you control a simulated robot in a search and rescue setting—or lets you ride along with the robot (as a teleoperator). You will be provided with 2 scenarios. Both the scenarios work a bit differently. In one scenario you will have to use your eye gaze directions (left, right, up and down) and then keyboard press J,L,I,K (left, right, up and down) to move the simulated robot across the search and rescue scene in order to reach the end of the parking lot in the scene. In the second scenario you will have to press F, and the robot moves in the scene, as the system is aware of where to take you. The common task in both scenarios for you is to evaluate how important each area of the scene is in each scenario based on your understanding. Here, *scenarios* refers to two possibilities (human assisted, system assisted), the term *scene* refers to the overall space where the task needs to be performed, and *area* is a smaller location within scene where you could identify the important points of interest.

The task is to count the points of interest you encounter and put corresponding priority markers for each of them. You will also be provided with a small reference sheet before the experiment starts to give a general idea.

We want to see whether the methods we use to make robotic interfaces better would be suitable for users while guiding a robot in a high-risk environment, e.g. in a search and rescue setting. We are also investigating the use of non-verbal gestures and natural human way of viewing scenes to gauge the effectiveness, usability and understanding of and provided by our tools in reducing the teleoperator's work and / or cognitive load while performing such tasks.

In other words, we want to investigate whether we can provide natural and intuitive interactions with robots. This means that we want to *test our ideas and their impact* on your performance, not your personal capabilities or suitability as driver of a robot.

We will record your interaction with the interface, potentially also record your reactions (audio) and create internal data logs of your session with the system. Additionally, we ask you to answer a couple of questionnaires about you and the interaction. Data will be stored in anonymized form.



Now the REALLY IMPORTANT INFORMATION:

You can refrain from continuing your trial at any time without stating why. An experimenter will always be close by to help if any confusing or uncomfortable situation arises. You can withdraw your consent to use the captured data (video and robot platform data logs) at any time.

On the following pages you will find a more detailed description of the session and your task.

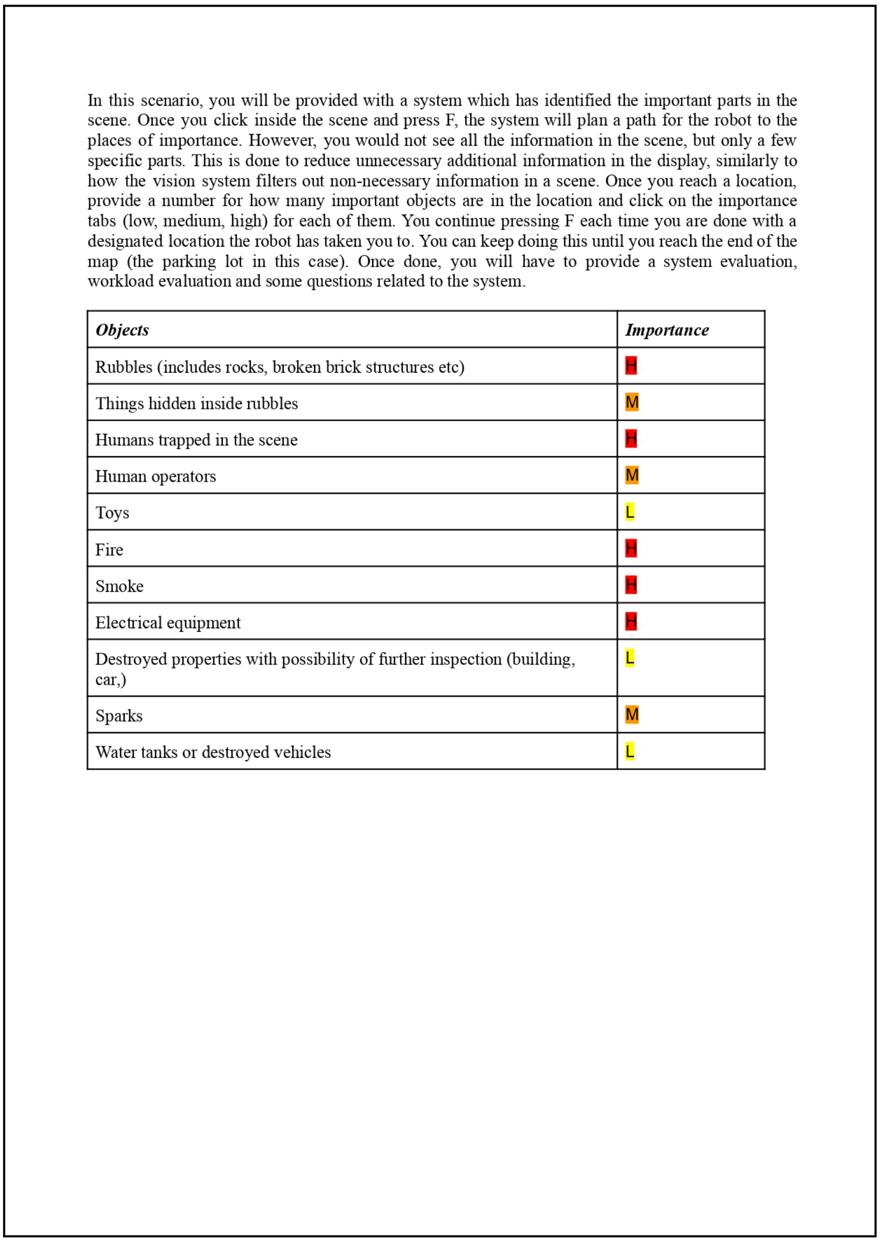
Experiment session

1) Introduction to the simulator, making yourself familiar with the interface

You will be able to “play around” with the interface for controlling the simulated robot for some minutes during which you can also ask the experimenter for clarifications.

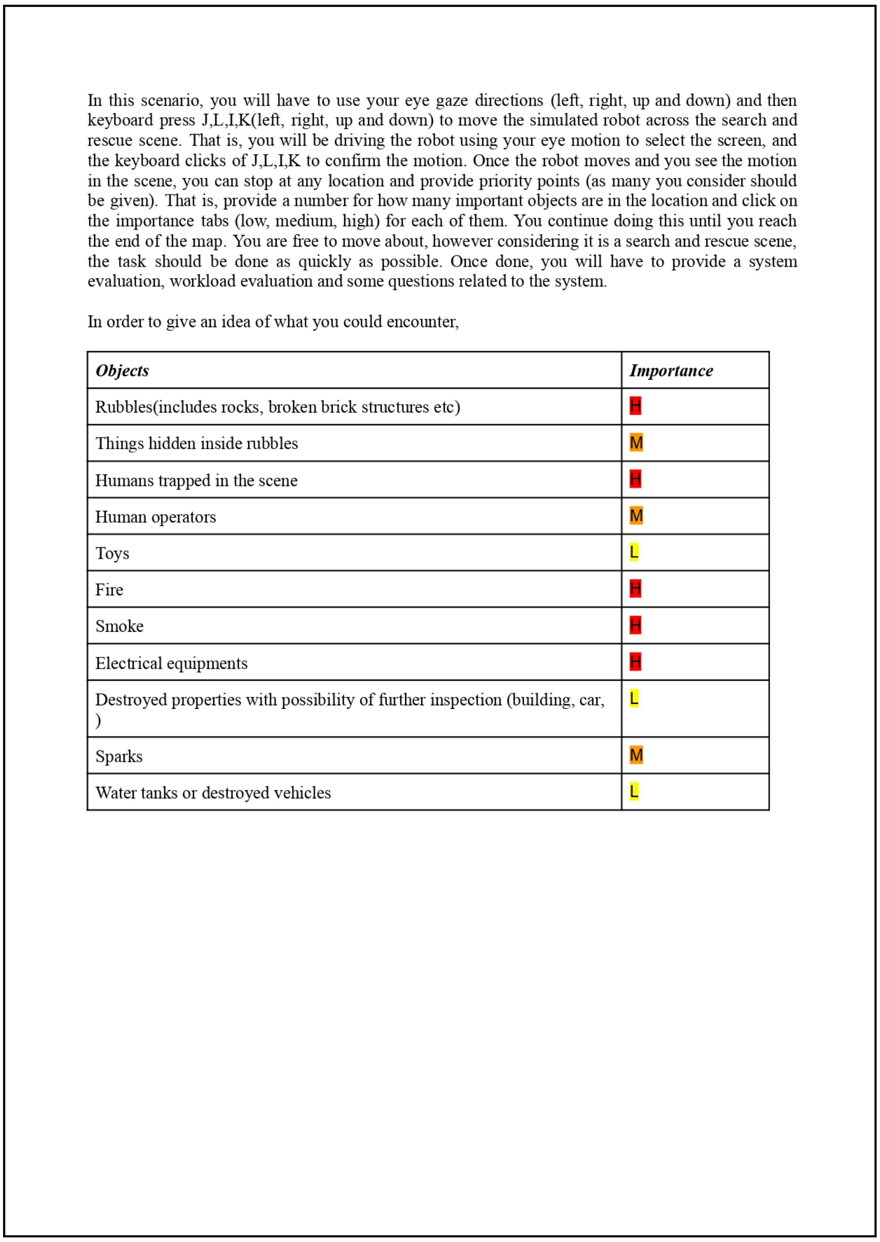
2) Evaluating the interface

You will be provided with 2 scenarios, in a random order. In both the scenarios you would have to assist the robot in performing Search and Rescue by using the best of your abilities to identify the points of interest in the scene.



In this scenario, you will be provided with a system which has identified the important parts in the scene. Once you click inside the scene and press F, the system will plan a path for the robot to the places of importance. However, you would not see all the information in the scene, but only a few specific parts. This is done to reduce unnecessary additional information in the display, similarly to how the vision system filters out non-necessary information in a scene. Once you reach a location, provide a number for how many important objects are in the location and click on the importance tabs (low, medium, high) for each of them. You continue pressing F each time you are done with a designated location the robot has taken you to. You can keep doing this until you reach the end of the map (the parking lot in this case). Once done, you will have to provide a system evaluation, workload evaluation and some questions related to the system.

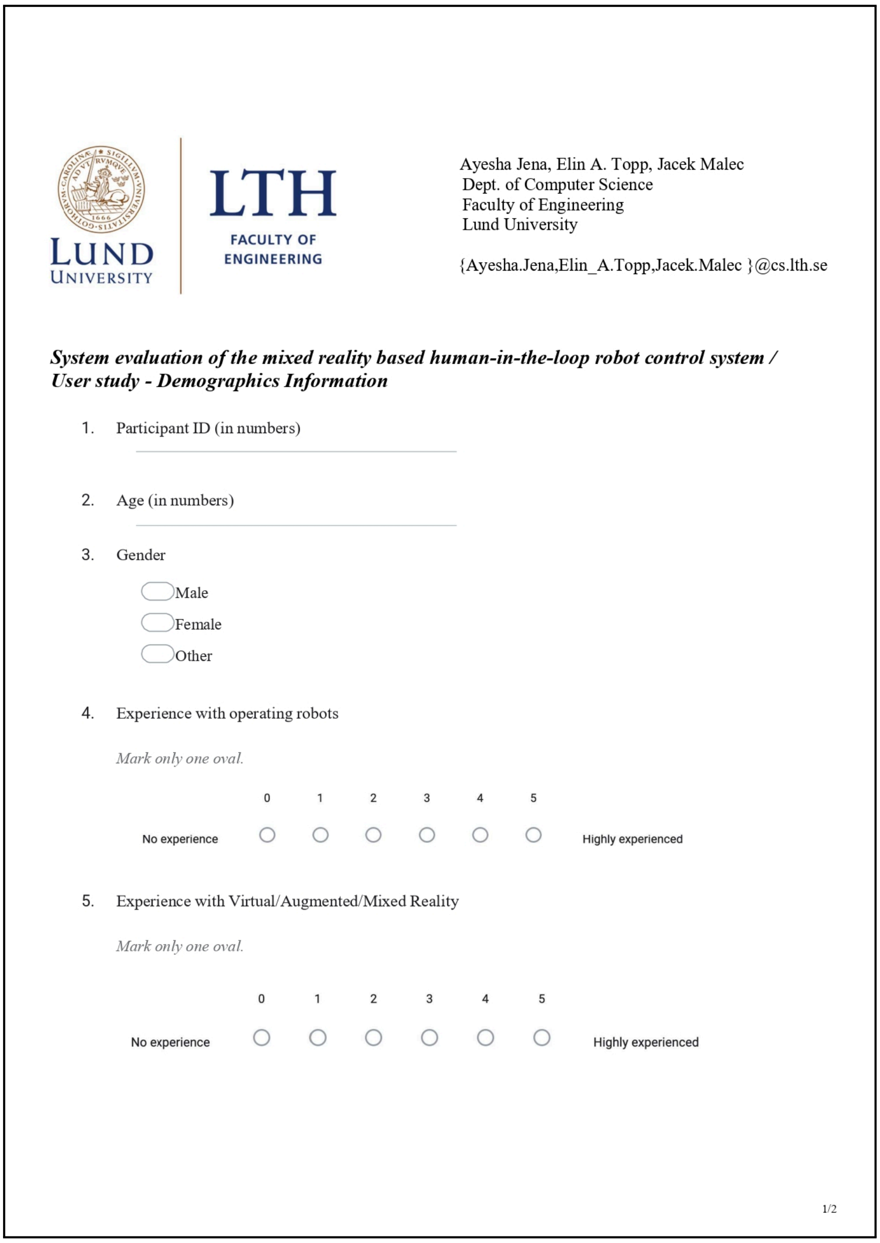
Objects	***Importance***
Rubbles (includes rocks, broken brick structures etc)	H
Things hidden inside rubbles	M
Humans trapped in the scene	H
Human operators	M
Toys	L
Fire	H
Smoke	H
Electrical equipment	H
Destroyed properties with possibility of further inspection (building, car,)	L
Sparks	M
Water tanks or destroyed vehicles	L



In this scenario, you will have to use your eye gaze directions (left, right, up and down) and then keyboard press J,L,I,K(left, right, up and down) to move the simulated robot across the search and rescue scene. That is, you will be driving the robot using your eye motion to select the screen, and the keyboard clicks of J,L,I,K to confirm the motion. Once the robot moves and you see the motion in the scene, you can stop at any location and provide priority points (as many you consider should be given). That is, provide a number for how many important objects are in the location and click on the importance tabs (low, medium, high) for each of them. You continue doing this until you reach the end of the map. You are free to move about, however considering it is a search and rescue scene, the task should be done as quickly as possible. Once done, you will have to provide a system evaluation, workload evaluation and some questions related to the system.

In order to give an idea of what you could encounter,

Objects	***Importance***
Rubbles(includes rocks, broken brick structures etc)	H
Things hidden inside rubbles	M
Humans trapped in the scene	H
Human operators	M
Toys	L
Fire	H
Smoke	H
Electrical equipments	H
Destroyed properties with possibility of further inspection (building, car,)	L
Sparks	M
Water tanks or destroyed vehicles	L



Ayesha Jena, Elin A. Topp, Jacek Malec
Dept. of Computer Science
Faculty of Engineering
Lund University

{Ayesha.Jena,Elin_A.Topp,Jacek.Malec }@cs.lth.se

System evaluation of the mixed reality based human-in-the-loop robot control system / User study - Demographics Information

1. Participant ID (in numbers)

   ____

2. Age (in numbers)

   ____

3. Gender

   ○ Male
   ○ Female
   ○ Other

4. Experience with operating robots

   Mark only one oval.

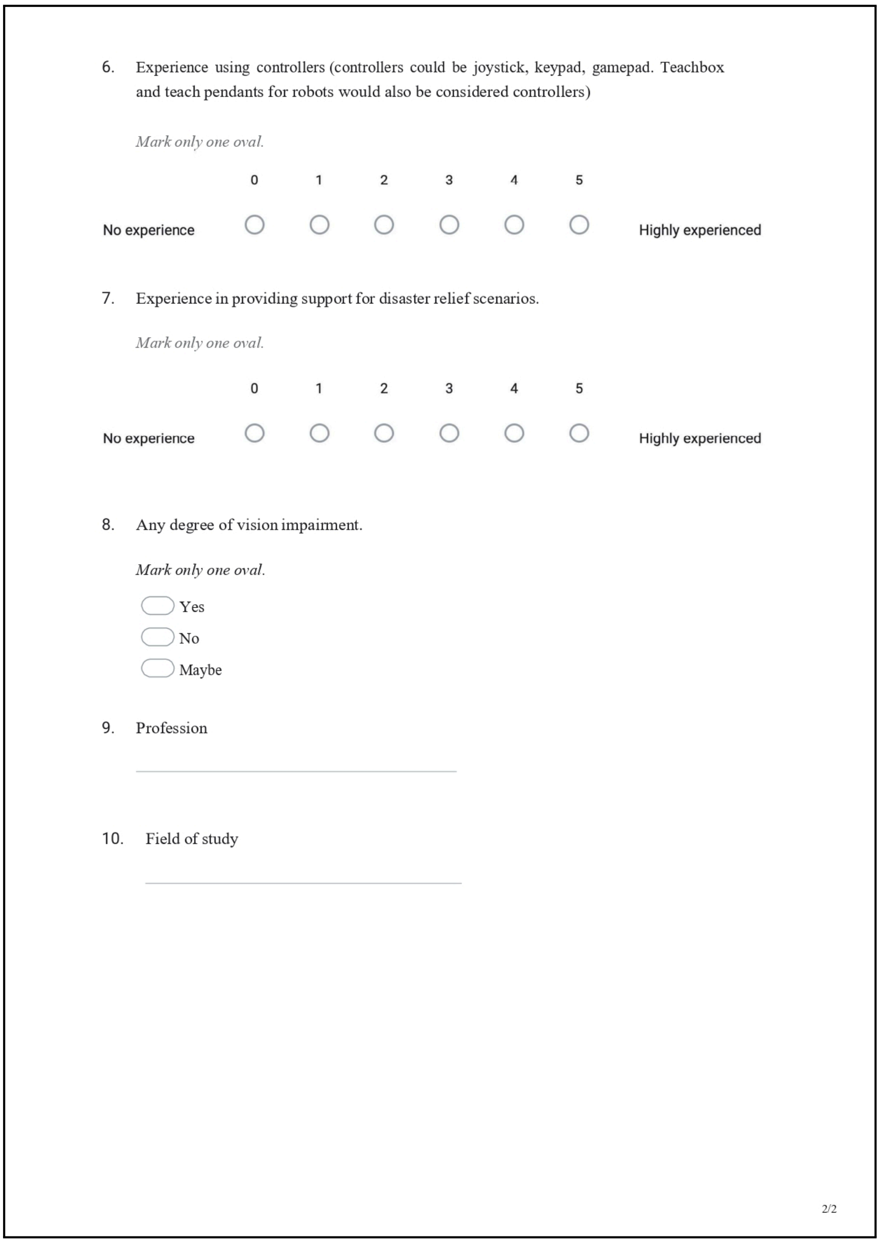
	0	1	2	3	4	5	
No experience	○	○	○	○	○	○	Highly experienced

5. Experience with Virtual/Augmented/Mixed Reality

   Mark only one oval.

	0	1	2	3	4	5	
No experience	○	○	○	○	○	○	Highly experienced

1/2



6. Experience using controllers (controllers could be joystick, keypad, gamepad. Teachbox and teach pendants for robots would also be considered controllers)

Mark only one oval.

	0	1	2	3	4	5	
No experience	○	○	○	○	○	○	Highly experienced

7. Experience in providing support for disaster relief scenarios.

Mark only one oval.

	0	1	2	3	4	5	
No experience	○	○	○	○	○	○	Highly experienced

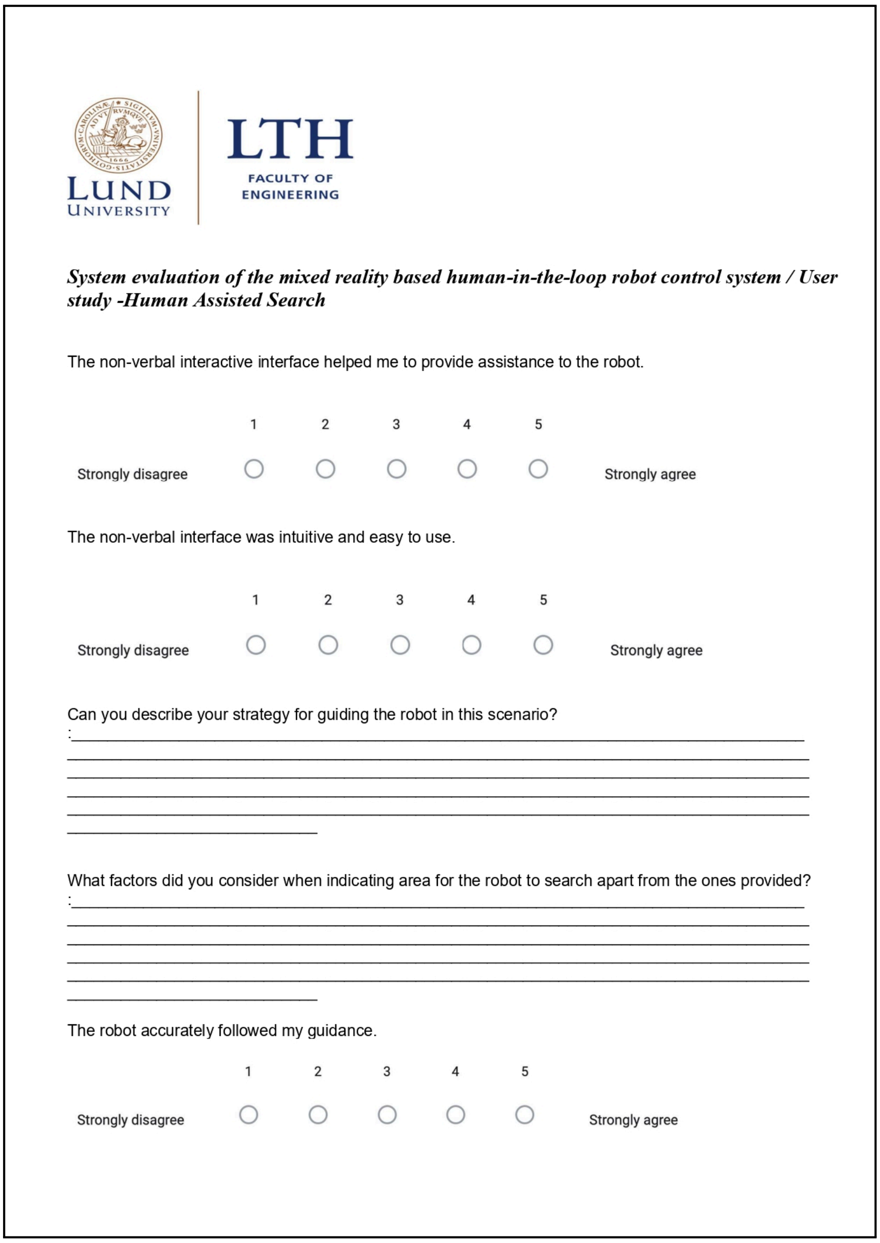
8. Any degree of vision impairment.

Mark only one oval.

○ Yes
○ No
○ Maybe

9. Profession

10. Field of study



LUND UNIVERSITY

LTH
FACULTY OF ENGINEERING

System evaluation of the mixed reality based human-in-the-loop robot control system / User study -Human Assisted Search

The non-verbal interactive interface helped me to provide assistance to the robot.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

The non-verbal interface was intuitive and easy to use.

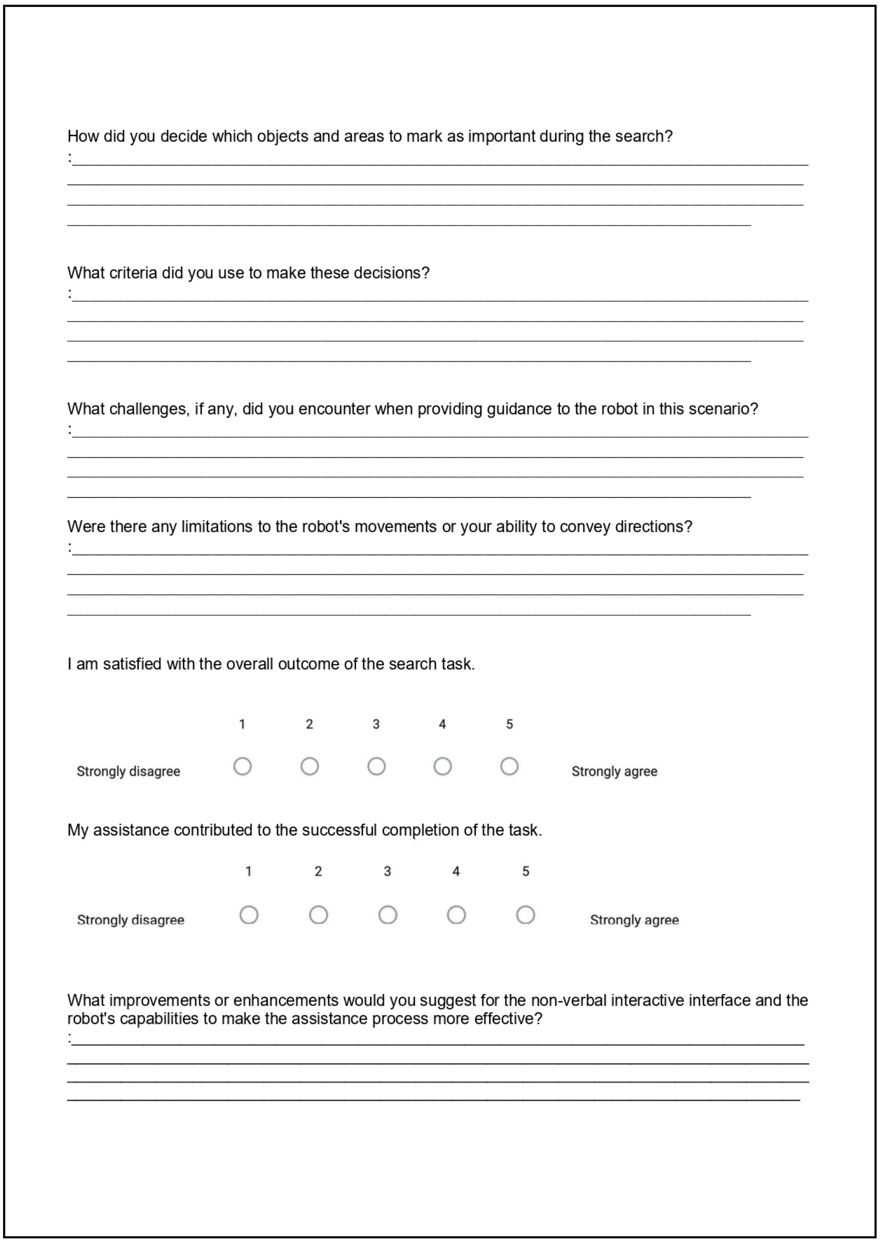
	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

Can you describe your strategy for guiding the robot in this scenario?
:____
____
____
____
____
____

What factors did you consider when indicating area for the robot to search apart from the ones provided?
:____
____
____
____
____
____

The robot accurately followed my guidance.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree



How did you decide which objects and areas to mark as important during the search?
:____
____
____
____

What criteria did you use to make these decisions?
:____
____
____
____

What challenges, if any, did you encounter when providing guidance to the robot in this scenario?
:____
____
____
____

Were there any limitations to the robot's movements or your ability to convey directions?
:____
____
____
____

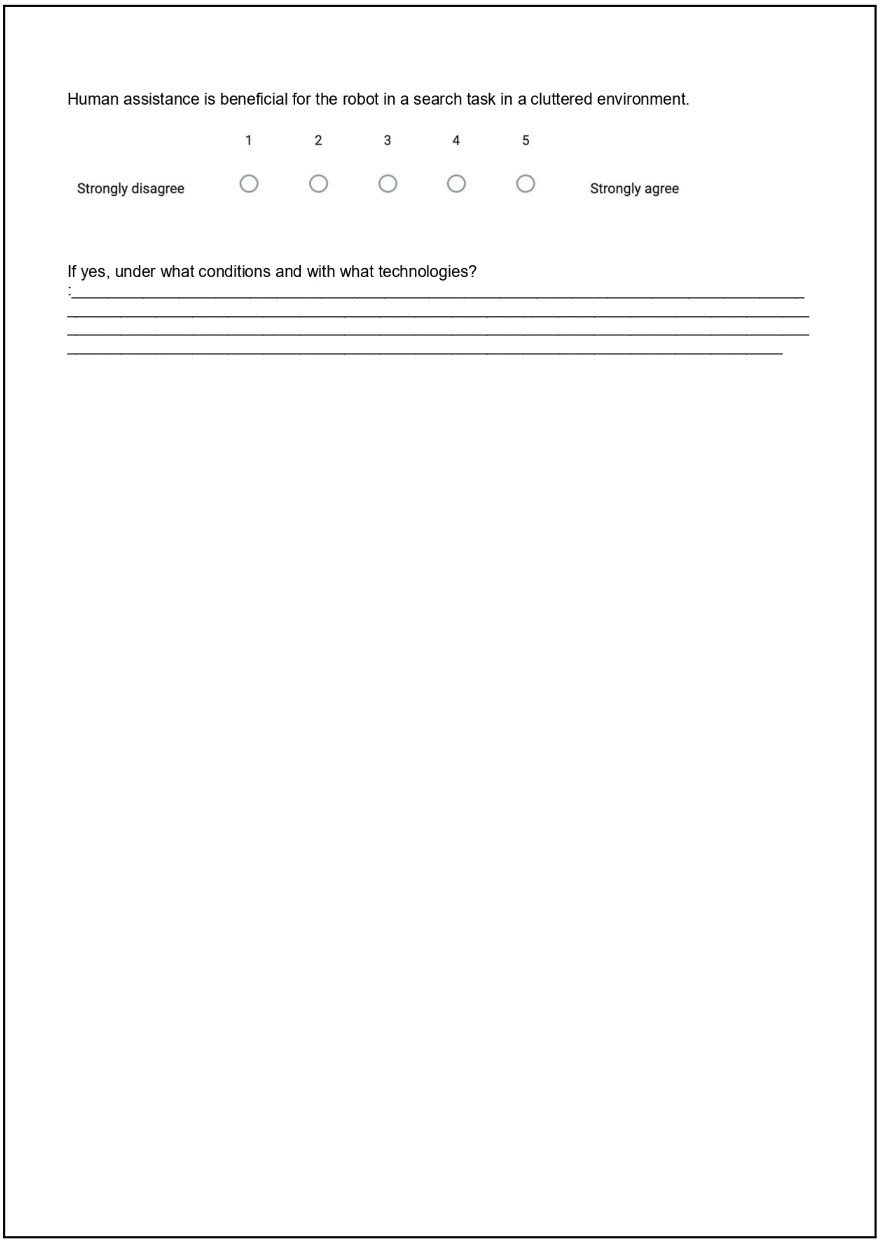
I am satisfied with the overall outcome of the search task.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

My assistance contributed to the successful completion of the task.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

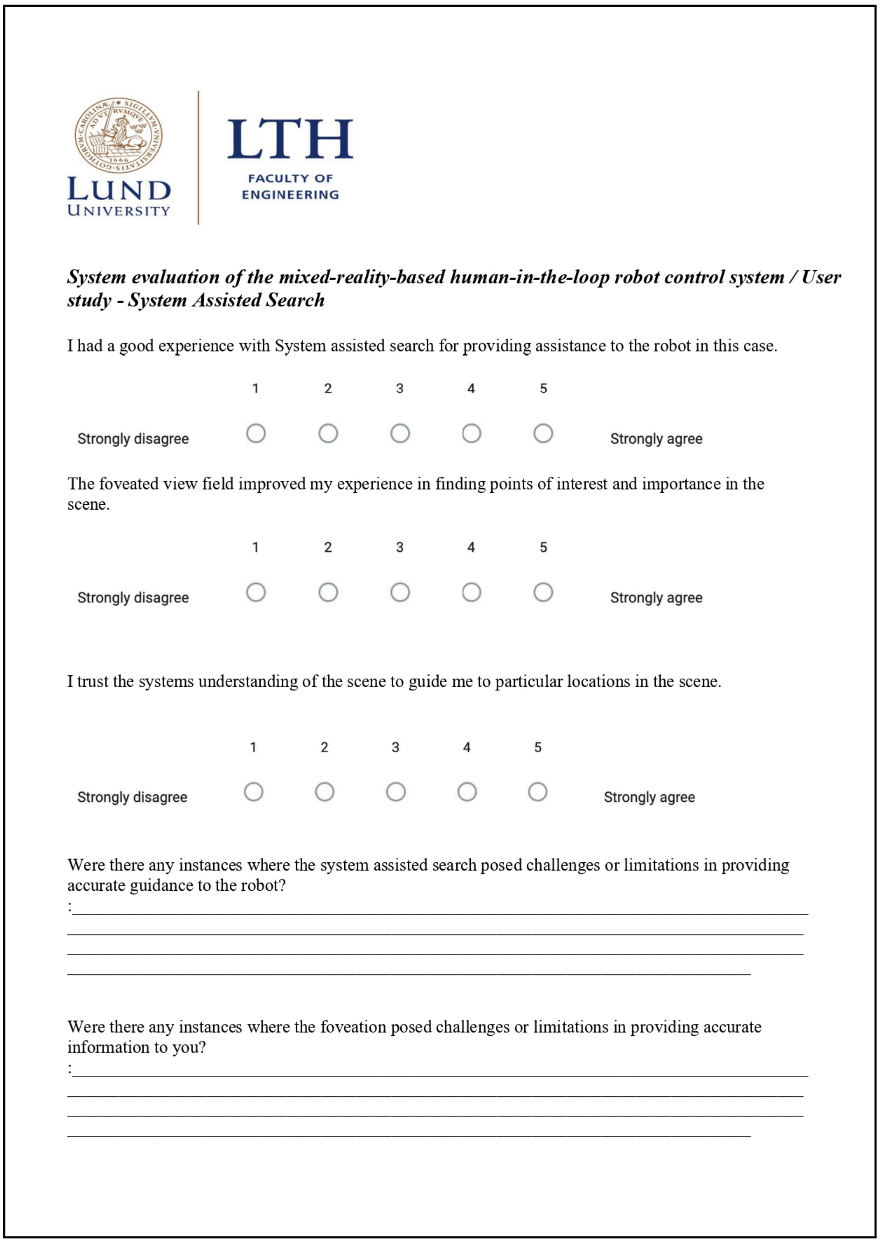
What improvements or enhancements would you suggest for the non-verbal interactive interface and the robot's capabilities to make the assistance process more effective?
:____
____
____
____



Human assistance is beneficial for the robot in a search task in a cluttered environment.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

If yes, under what conditions and with what technologies?
:____
____
____
____



LUND UNIVERSITY

LTH
FACULTY OF ENGINEERING

System evaluation of the mixed-reality-based human-in-the-loop robot control system / User study - System Assisted Search

I had a good experience with System assisted search for providing assistance to the robot in this case.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

The foveated view field improved my experience in finding points of interest and importance in the scene.

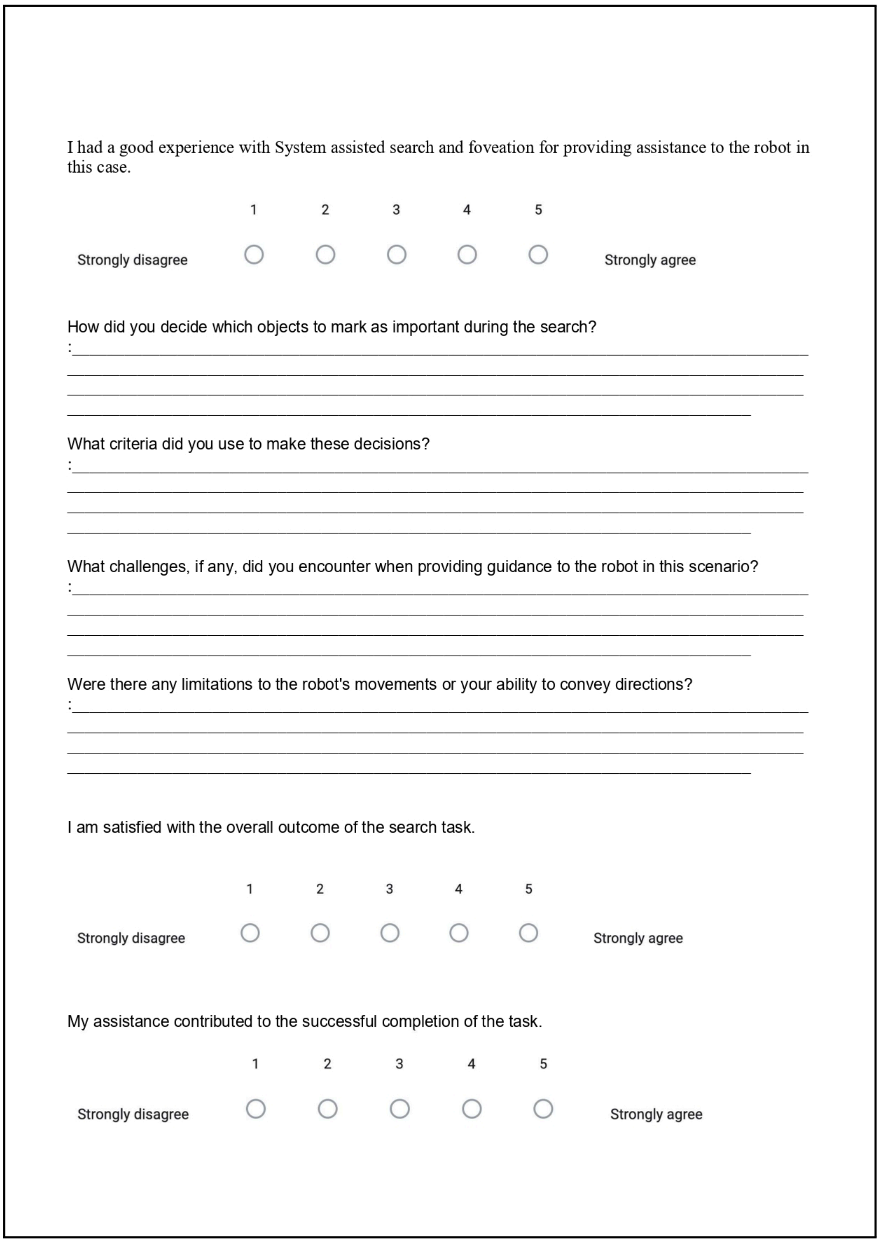
	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

I trust the systems understanding of the scene to guide me to particular locations in the scene.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

Were there any instances where the system assisted search posed challenges or limitations in providing accurate guidance to the robot?
:____
____
____
____

Were there any instances where the foveation posed challenges or limitations in providing accurate information to you?
:____
____
____
____



I had a good experience with System assisted search and foveation for providing assistance to the robot in this case.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

How did you decide which objects to mark as important during the search?

:____
____
____
____

What criteria did you use to make these decisions?

:____
____
____
____

What challenges, if any, did you encounter when providing guidance to the robot in this scenario?

:____
____
____
____

Were there any limitations to the robot's movements or your ability to convey directions?

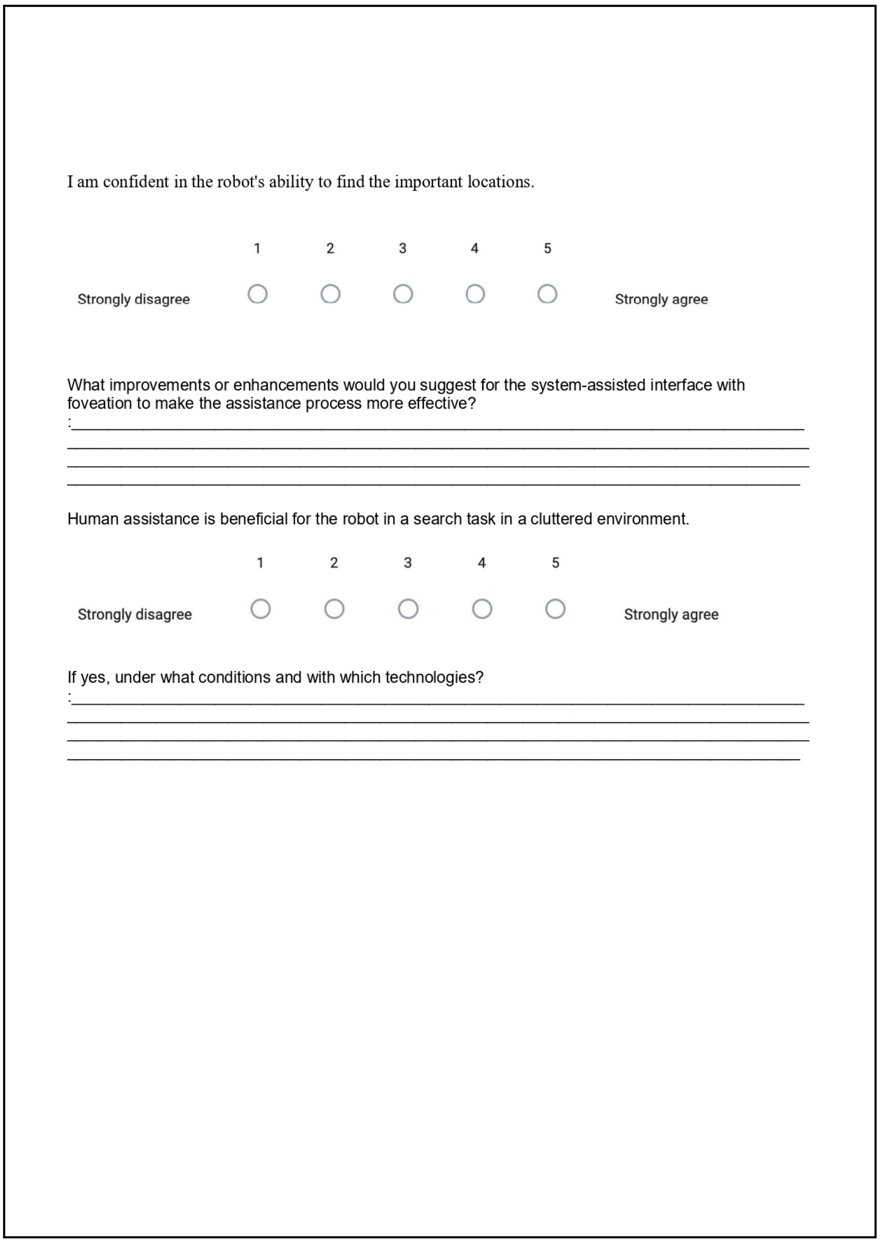
:____
____
____
____

I am satisfied with the overall outcome of the search task.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

My assistance contributed to the successful completion of the task.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree



I am confident in the robot's ability to find the important locations.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

What improvements or enhancements would you suggest for the system-assisted interface with foveation to make the assistance process more effective?

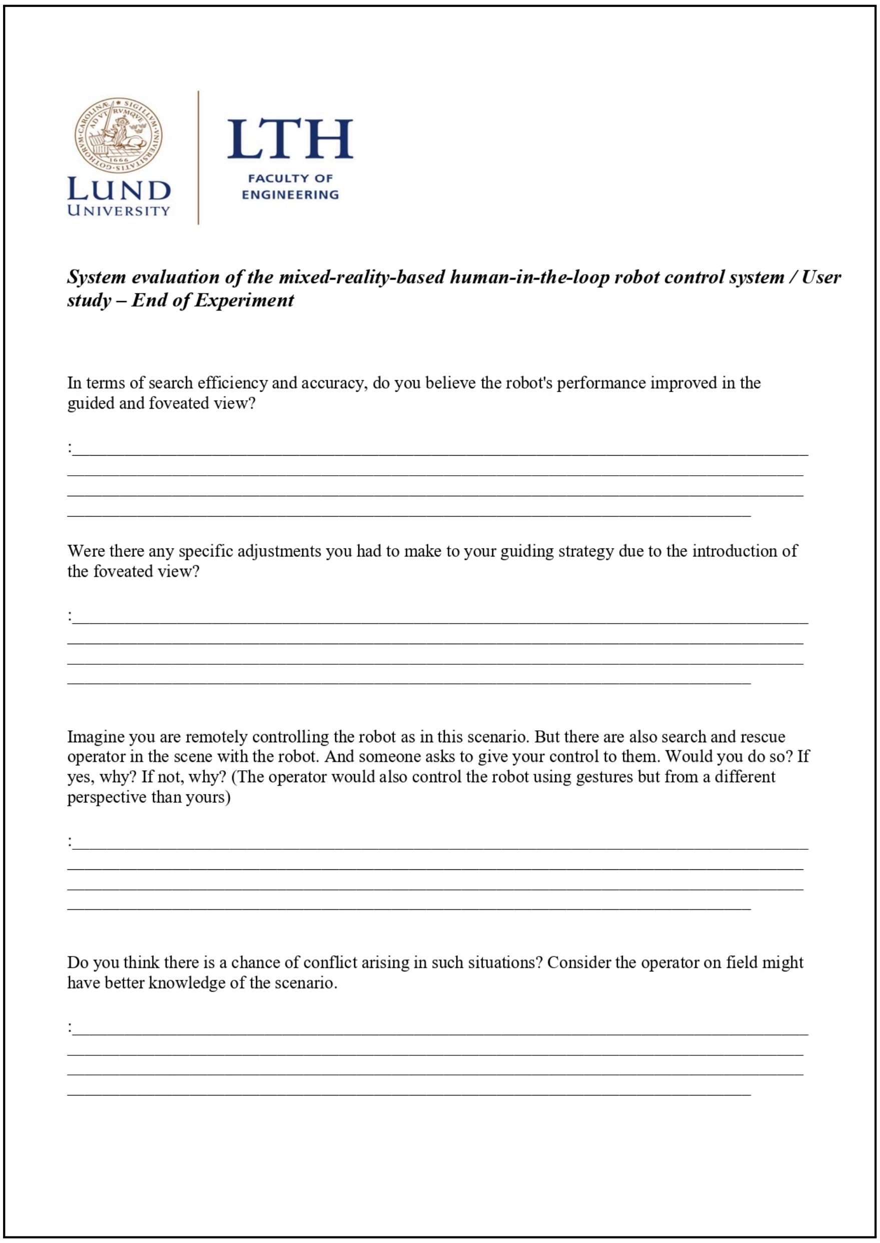
:____
____
____
____

Human assistance is beneficial for the robot in a search task in a cluttered environment.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

If yes, under what conditions and with which technologies?

:____
____
____
____



LUND
UNIVERSITY

LTH
FACULTY OF
ENGINEERING

System evaluation of the mixed-reality-based human-in-the-loop robot control system / User study – End of Experiment

In terms of search efficiency and accuracy, do you believe the robot's performance improved in the guided and foveated view?

:____
____
____
____

Were there any specific adjustments you had to make to your guiding strategy due to the introduction of the foveated view?

:____
____
____
____

Imagine you are remotely controlling the robot as in this scenario. But there are also search and rescue operator in the scene with the robot. And someone asks to give your control to them. Would you do so? If yes, why? If not, why? (The operator would also control the robot using gestures but from a different perspective than yours)

:____
____
____
____

Do you think there is a chance of conflict arising in such situations? Consider the operator on field might have better knowledge of the scenario.

:____
____
____
____

9.2 Appendix B - Posters

Poster I: Mixed-Initiative Interaction for Collaborative Robotics
Presented during the poster session at the COMPUTE Summer Retreat in Arild, 2022

Poster II: Mixed-Initiative Interaction for Collaborative Robotics
Presented during the poster session at the ELLIIT Annual Workshop in Linköping, 2022

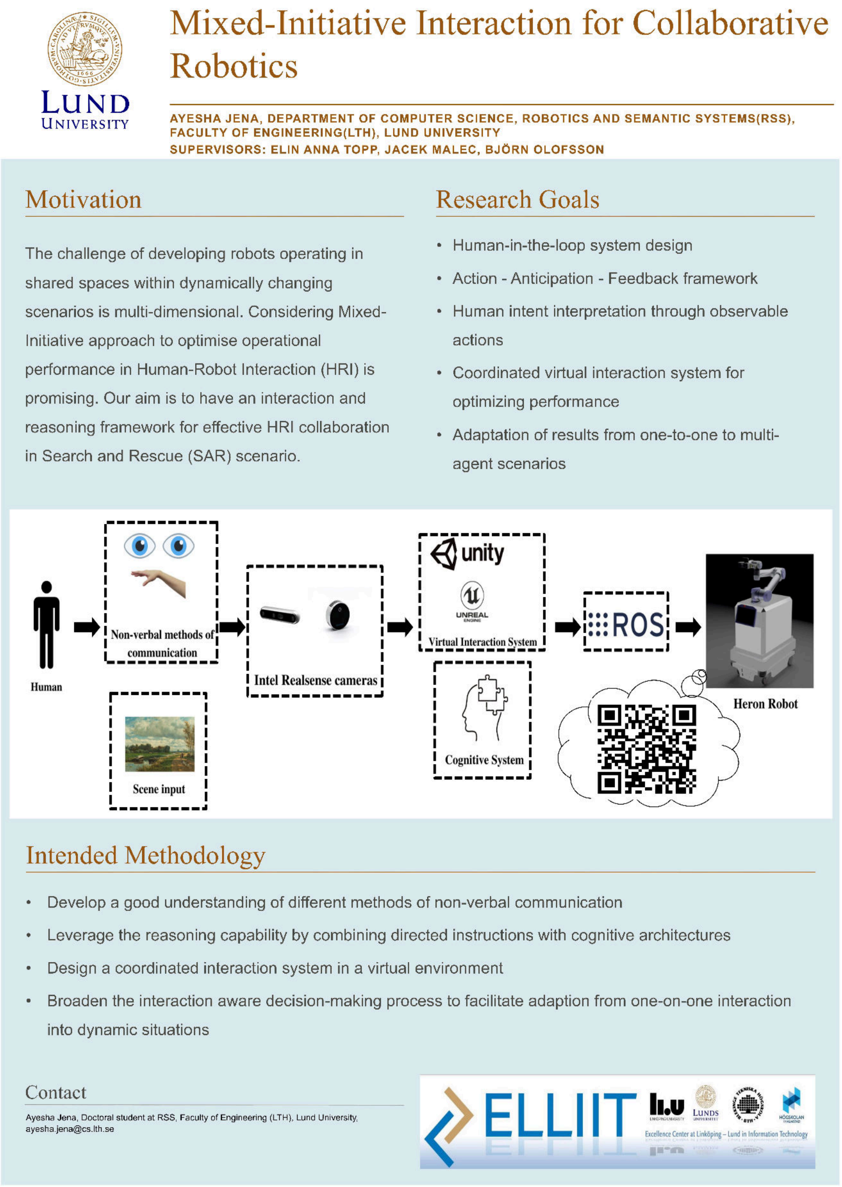
Poster III: Chaos to Control: Human Assisted Scene Inspection
Presented during the poster session at the HRI Conference in Stockholm, 2023

Poster IV: Mixed-Initiative Interaction for Collaborative Robot
Presented during the poster session at the WASP Winter Conference in Norrköping, 2024

Poster V: Towards Understanding the Role of Humans in Collaborative Tasks
Presented during the poster session at the ELLIIT Annual Workshop in Lund, 2024

Poster VI: Support for critical collaboration tasks through gaze guidance and visual augmentation
Presented during the poster session at the WASP Winter Conference in Norrköping, 2025

Poster VII: Impact of Gaze-Based Interaction and Augmentation on Human-Robot Collaboration in Critical Tasks
To be presented during the poster session at ROMAN Conference, 2025



Mixed-Initiative Interaction for Collaborative Robotics

AYESHA JENA, DEPARTMENT OF COMPUTER SCIENCE, ROBOTICS AND SEMANTIC SYSTEMS(RSS), FACULTY OF ENGINEERING(LTH), LUND UNIVERSITY
SUPERVISORS: ELIN ANNA TOPP, JACEK MALEC, BJÖRN OLOFSSON

Motivation

The challenge of developing robots operating in shared spaces within dynamically changing scenarios is multi-dimensional. Considering Mixed-Initiative approach to optimise operational performance in Human-Robot Interaction (HRI) is promising. Our aim is to have an interaction and reasoning framework for effective HRI collaboration in Search and Rescue (SAR) scenario.

Research Goals

- Human-in-the-loop system design
- Action - Anticipation - Feedback framework
- Human intent interpretation through observable actions
- Coordinated virtual interaction system for optimizing performance
- Adaptation of results from one-to-one to multi-agent scenarios

Human

Non-verbal methods of communication

Scene input

Intel Realsense cameras

unity

UNREAL ENGINE

Virtual Interaction System

Cognitive System

ROS

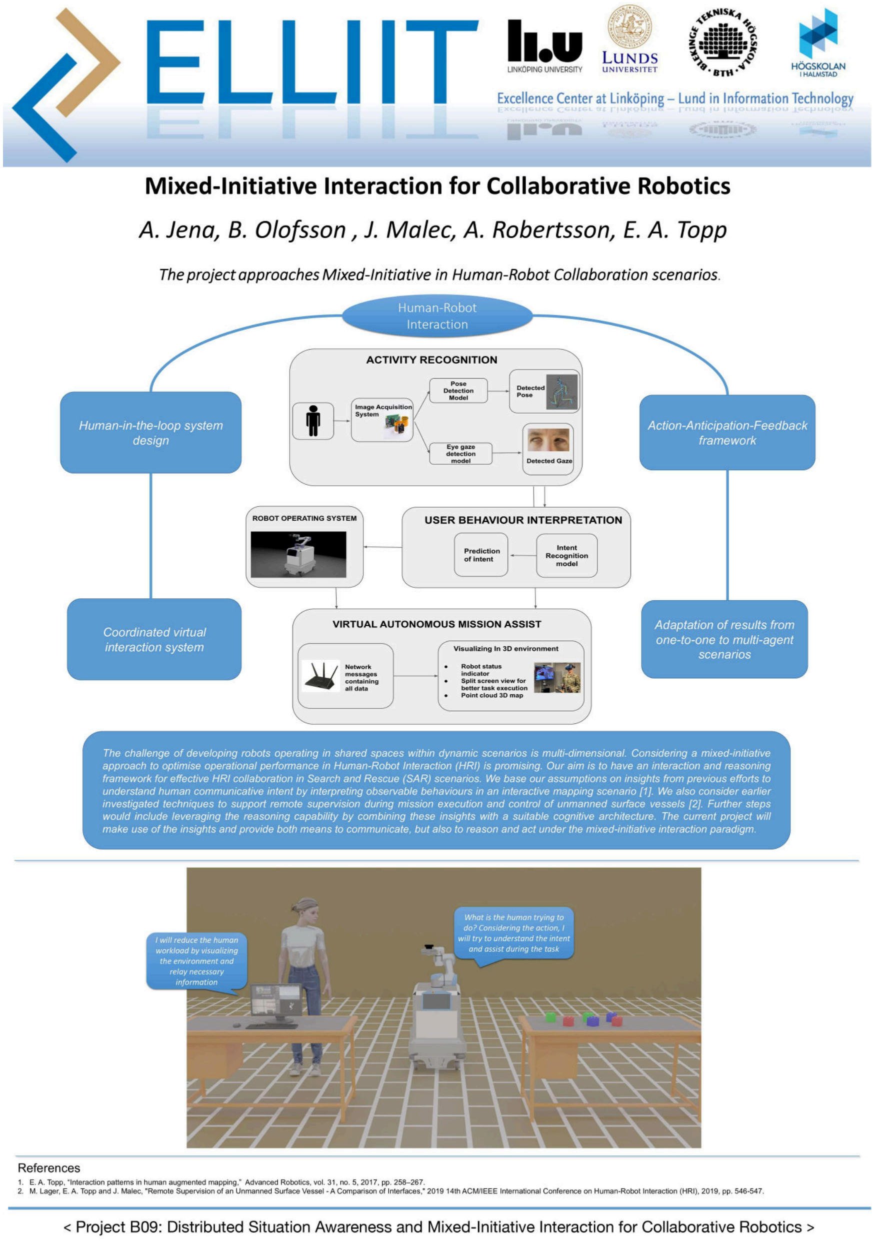
Heron Robot

Intended Methodology

- Develop a good understanding of different methods of non-verbal communication
- Leverage the reasoning capability by combining directed instructions with cognitive architectures
- Design a coordinated interaction system in a virtual environment
- Broaden the interaction aware decision-making process to facilitate adaption from one-on-one interaction into dynamic situations

Contact

Ayesha Jena, Doctoral student at RSS, Faculty of Engineering (LTH), Lund University,
ayesha.jena@cs.lth.se



ELLIIT
LINKÖPING UNIVERSITY
LUNDS UNIVERSITET
BLEKINGE TEKNISKA HÖGSKOLA · BTH ·
HÖGSKOLAN I HALMSTAD
Excellence Center at Linköping – Lund in Information Technology

Mixed-Initiative Interaction for Collaborative Robotics

A. Jena, B. Olofsson , J. Malec, A. Robertsson, E. A. Topp

The project approaches Mixed-Initiative in Human-Robot Collaboration scenarios.

Human-Robot Interaction

Human-in-the-loop system design

Action-Anticipation-Feedback framework

Coordinated virtual interaction system

Adaptation of results from one-to-one to multi-agent scenarios

ACTIVITY RECOGNITION

Image Acquisition System

Pose Detection Model

Detected Pose

Eye gaze detection model

Detected Gaze

ROBOT OPERATING SYSTEM

USER BEHAVIOUR INTERPRETATION

Prediction of intent

Intent Recognition model

VIRTUAL AUTONOMOUS MISSION ASSIST

Network messages containing all data

Visualizing In 3D environment

- Robot status indicator
- Split screen view for better task execution
- Point cloud 3D map

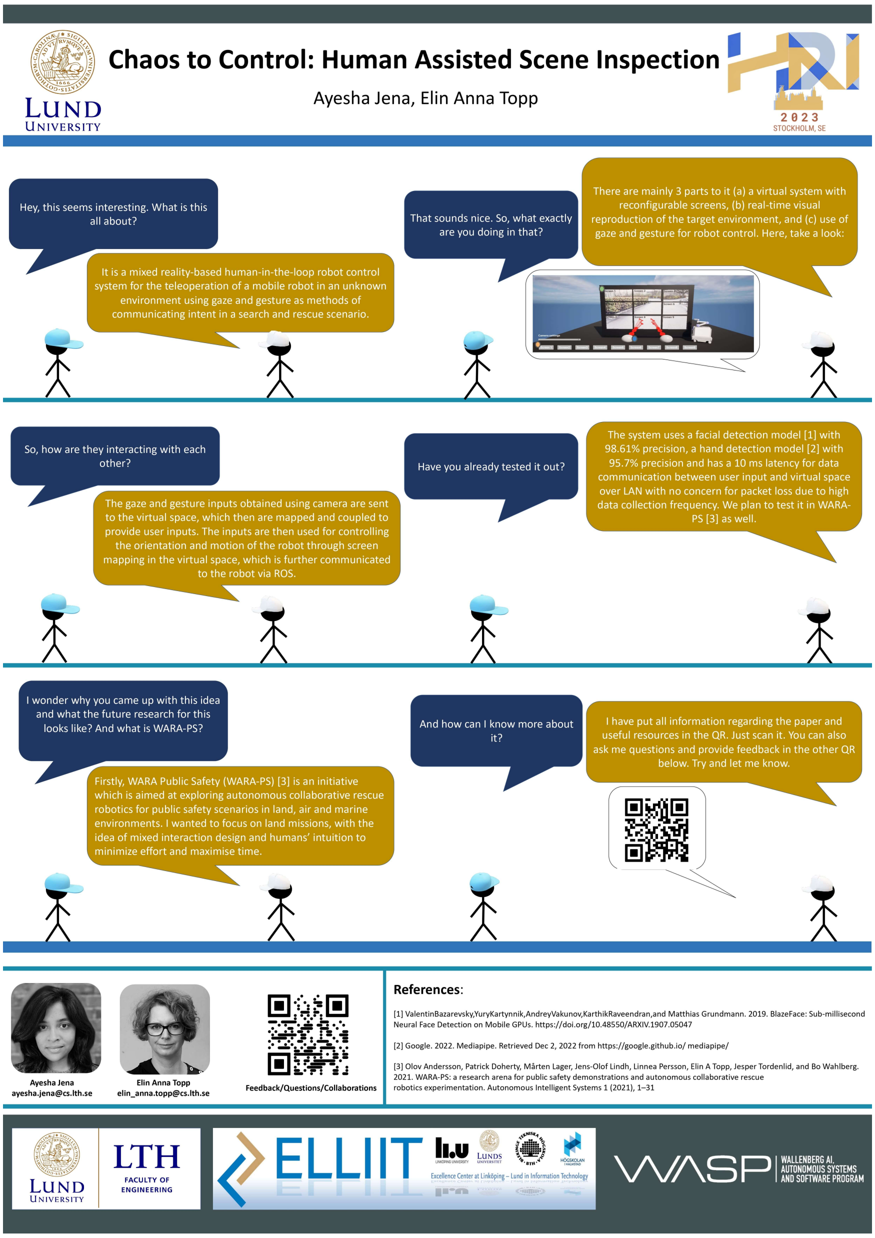
The challenge of developing robots operating in shared spaces within dynamic scenarios is multi-dimensional. Considering a mixed-initiative approach to optimise operational performance in Human-Robot Interaction (HRI) is promising. Our aim is to have an interaction and reasoning framework for effective HRI collaboration in Search and Rescue (SAR) scenarios. We base our assumptions on insights from previous efforts to understand human communicative intent by interpreting observable behaviours in an interactive mapping scenario [1]. We also consider earlier investigated techniques to support remote supervision during mission execution and control of unmanned surface vessels [2]. Further steps would include leveraging the reasoning capability by combining these insights with a suitable cognitive architecture. The current project will make use of the insights and provide both means to communicate, but also to reason and act under the mixed-initiative interaction paradigm.

I will reduce the human workload by visualizing the environment and relay necessary information
What is the human trying to do? Considering the action, I will try to understand the intent and assist during the task

References

1. E. A. Topp, "Interaction patterns in human augmented mapping," Advanced Robotics, vol. 31, no. 5, 2017, pp. 258–267.
2. M. Lager, E. A. Topp and J. Malec, "Remote Supervision of an Unmanned Surface Vessel - A Comparison of Interfaces," 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2019, pp. 546-547.

< Project B09: Distributed Situation Awareness and Mixed-Initiative Interaction for Collaborative Robotics >



Chaos to Control: Human Assisted Scene Inspection

Ayesha Jena, Elin Anna Topp

Hey, this seems interesting. What is this all about?

It is a mixed reality-based human-in-the-loop robot control system for the teleoperation of a mobile robot in an unknown environment using gaze and gesture as methods of communicating intent in a search and rescue scenario.

That sounds nice. So, what exactly are you doing in that?

There are mainly 3 parts to it (a) a virtual system with reconfigurable screens, (b) real-time visual reproduction of the target environment, and (c) use of gaze and gesture for robot control. Here, take a look:

So, how are they interacting with each other?

The gaze and gesture inputs obtained using camera are sent to the virtual space, which then are mapped and coupled to provide user inputs. The inputs are then used for controlling the orientation and motion of the robot through screen mapping in the virtual space, which is further communicated to the robot via ROS.

Have you already tested it out?

The system uses a facial detection model [1] with 98.61% precision, a hand detection model [2] with 95.7% precision and has a 10 ms latency for data communication between user input and virtual space over LAN with no concern for packet loss due to high data collection frequency. We plan to test it in WARA-PS [3] as well.

I wonder why you came up with this idea and what the future research for this looks like? And what is WARA-PS?

Firstly, WARA Public Safety (WARA-PS) [3] is an initiative which is aimed at exploring autonomous collaborative rescue robotics for public safety scenarios in land, air and marine environments. I wanted to focus on land missions, with the idea of mixed interaction design and humans' intuition to minimize effort and maximise time.

And how can I know more about it?

I have put all information regarding the paper and useful resources in the QR. Just scan it. You can also ask me questions and provide feedback in the other QR below. Try and let me know.

Ayesha Jena
ayesha.jena@cs.lth.se

Elin Anna Topp
elin_anna.topp@cs.lth.se

Feedback/Questions/Collaborations

References:

[1] ValentinBazarevsky,YuryKartynnik,AndreyVakunov,KarthikRaveendran,and Matthias Grundmann. 2019. BlazeFace: Sub-millisecond Neural Face Detection on Mobile GPUs. https://doi.org/10.48550/ARXIV.1907.05047

[2] Google. 2022. Mediapipe. Retrieved Dec 2, 2022 from https://google.github.io/ mediapipe/

[3] Olov Andersson, Patrick Doherty, Mårten Lager, Jens-Olof Lindh, Linnea Persson, Elin A Topp, Jesper Tordenlid, and Bo Wahlberg. 2021. WARA-PS: a research arena for public safety demonstrations and autonomous collaborative rescue robotics experimentation. Autonomous Intelligent Systems 1 (2021), 1–31

LUND UNIVERSITY
HRI 2023 STOCKHOLM, SE
LTH FACULTY OF ENGINEERING
ELLIIT
Excellence Center at Linköping – Lund in Information Technology
WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Mixed-Initiative Interaction for Collaborative Robots

Ayesha Jena[1], Jacek Malec[2], Björn Olofsson[3], Elin Anna Topp[4]
[1,2,4]Department of Computer Science, [3]Department of Automatic Control
Lund University

1 Introduction

The lack of mature algorithms and control schemes for autonomous systems makes it still difficult for them to operate safely in high-risk environments [1]. Our approach involves utilizing humans' intuition of the environment through intention recognition to reach the final desired state, i.e., goal, by having a defined action plan. Thus, to achieve this mixed-initiative interaction as a part of human-robot symbiosis, we need the means to describe both expectations of the humans and deviations from them to refine the knowledge and contextual understanding of the environment where this interaction takes place.

2 Methods

- We developed an interface system to assist operators to control a robot using gaze and hand signals [1].
- Currently, we expect to build an understanding of how humans operate in Search and Rescue (SAR) in terms of prioritizing various elements in a scene to maximize rescue efforts.
- Using the findings from the experiment and previous work in cognitive robotics [2],we plan to explore human intent interpretation through observable actions to signify a goal-oriented motive.

“Heron robot”, courtesy Robotlab LTH

3 Experimental Setup

- In first case, a human operator (fig. a) leverages their intuition to guide the robot in SAR using “gaze and hand signaling controls” while identifying areas of interest.
- In the other case, the system identifies potential parts of the scene (robot view is shown in fig. b and top view of the scene in fig. c) and relays this information selectively to the operator by a technique similar to foveated rendering.
- In both cases, the operator needs to indicate the level of priority in SAR considering factors such as the urgency of the situation, potential risks, and the likelihood of successful rescue.

4 Analysis

- The preliminary results indicate that models can accurately detect these signals from any camera, with low inference time, suggesting quick information processing [1].
- In the experiment, higher points are given to dangerous elements like fire, electrical equipment, and trapped humans, prioritizing rescue efforts where needed. Lower scores are assigned to less critical items with the scoring mechanism totaling to a hundred.
- The system uses these points to identify areas of interest, evaluating the operator's ability to identify important regions in one case and assisting the user in a concentrated and controlled search using points from wider reconnaissance in the second case.

Setup

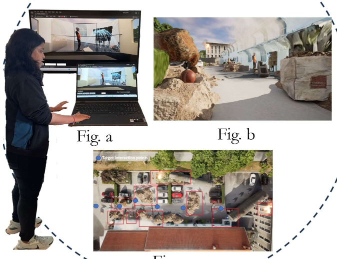


Fig. a

Fig. b

Fig. c

5 Main Takeaways

- Integration of the eye gaze and hand signaling with the navigation capabilities of a simulated robot.
- An interface system, its application in Search and Rescue scenarios, and how points are used to prioritize rescue efforts and guide the human operator or the system during different tasks.
- A method for improving the operation of autonomous systems in high-risk environments by incorporating human intuition and possible intention recognition.

Ayesha Jena
Faculty of Engineering
Lund University, Sweden
ayesha.jena@cs.lth.se

Jacek Malec
Faculty of Engineering
Lund University, Sweden
jacek.malec@cs.lth.se

Björn Olofsson
Faculty of Engineering
Lund University, Sweden
bjorn.olofsson@control.lth.se

Elin Anna Topp
Faculty of Engineering
Lund University, Sweden
elin_anna.topp@cs.lth.se

References:
[1] Jena, A. and Topp, E.A., 2023, March. Chaos to Control: Human Assisted Scene Inspection. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 491-494).
[2] Vinanzi, S., Goerick, C. and Cangelosi, A., 2019, August. Mindreading for robots: Predicting intentions via dynamical clustering of human postures. In *2019 Joint IEEE 9th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)* (pp. 272-277). IEEE.

ELLIIT Excellence Center at Linköping – Lund in Information Technology
WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

ELLIIT
Excellence Center at Linköping – Lund in Information Technology

Towards Understanding the Role of Humans in Collaborative Tasks

A. Jena, B. Olofsson , J. Malec, E. A. Topp

The project approaches Mixed-Initiative in Human-Robot Collaboration scenarios

This work investigates the dynamics of human-robot collaboration, focusing on human-assisted and system-assisted approaches within search and rescue operations. Utilizing virtual and mixed-reality interfaces, the study assessed task performance, workload, usability, and participant experiences.

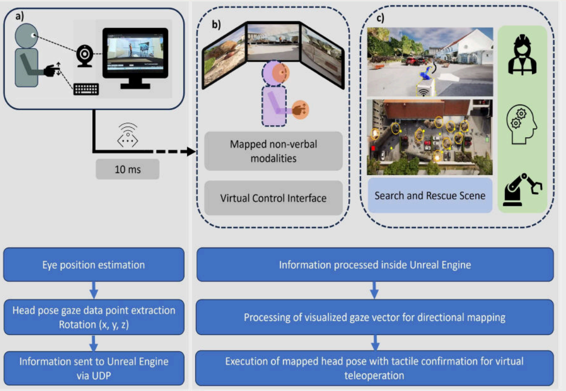

a)
b)
c)
10 ms
Mapped non-verbal modalities
Virtual Control Interface
Search and Rescue Scene
Eye position estimation
Head pose gaze data point extraction Rotation (x, y, z)
Information sent to Unreal Engine via UDP
Information processed inside Unreal Engine
Processing of visualized gaze vector for directional mapping
Execution of mapped head pose with tactile confirmation for virtual teleoperation

Figure 1: An image of the different aspects of mixed-reality based human-in-the-loop robot control system.

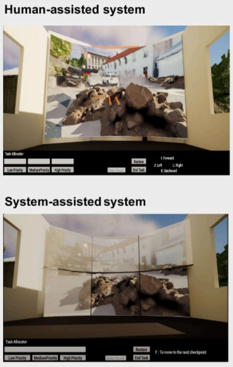
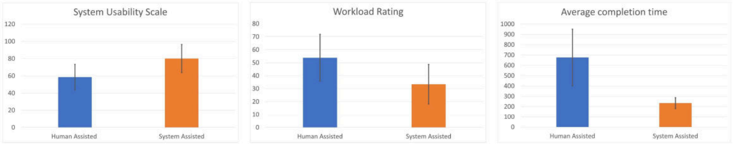


Figure 2: Field of view of the users in both scenarios

Results and Findings:



Subjective feedback provides insights for system improvements and protocol development, underscoring the value of integrating human collaboration to boost operational efficiency in complex, high-risk settings. The study also explores mixed-initiative interaction in Human-Robot Interaction (HRI), highlighting the importance of non-verbal cues in dynamic control sharing during missions.

Future Work:

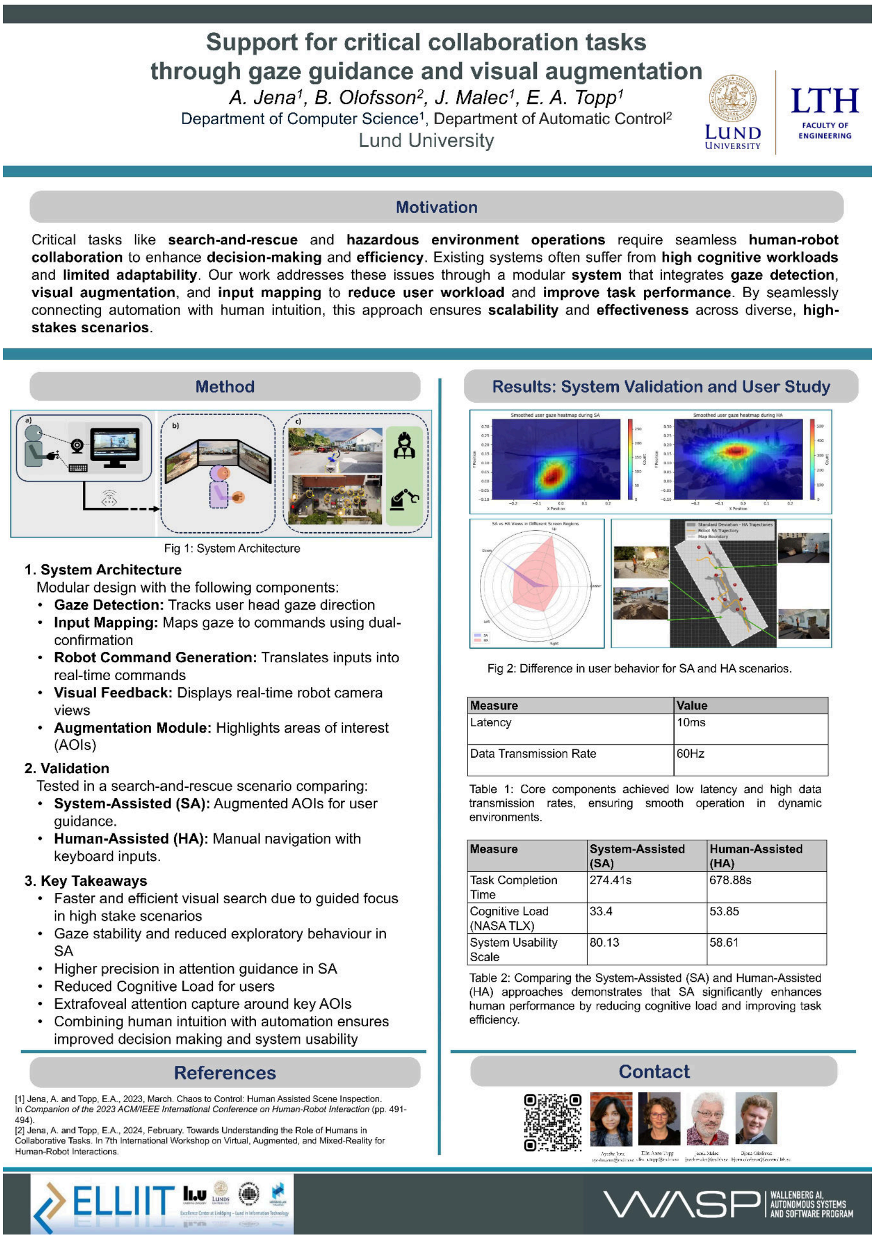
Preliminary findings on gaze and gesture control emphasize the potential of mixed-reality systems in understanding and categorizing non-verbal communication for effective human-robot teamwork, paving the way for advanced research in intention recognition and collaborative dynamics.

References:

[1] Jena, A. and Topp, E.A., 2023, March. Chaos to Control: Human Assisted Scene Inspection. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 491-494).

[2] Jena, A. and Topp, E.A., 2024, February. Towards Understanding the Role of Humans in Collaborative Tasks. In *7th International Workshop on Virtual, Augmented, and Mixed-Reality for Human-Robot Interactions.*

< Project B09: Distributed Situation Awareness and Mixed-Initiative Interaction for Collaborative Robotics >



Support for critical collaboration tasks through gaze guidance and visual augmentation

A. Jena[1], B. Olofsson[2], J. Malec[1], E. A. Topp[1]
Department of Computer Science[1], Department of Automatic Control[2]
Lund University

Motivation

Critical tasks like **search-and-rescue** and **hazardous environment operations** require seamless **human-robot collaboration** to enhance **decision-making** and **efficiency**. Existing systems often suffer from **high cognitive workloads** and **limited adaptability**. Our work addresses these issues through a modular **system** that integrates **gaze detection**, **visual augmentation**, and **input mapping** to **reduce user workload** and **improve task performance**. By seamlessly connecting automation with human intuition, this approach ensures **scalability** and **effectiveness** across diverse, **high-stakes scenarios**.

Method

Fig 1: System Architecture

1. System Architecture

Modular design with the following components:

- **Gaze Detection:** Tracks user head gaze direction
- **Input Mapping:** Maps gaze to commands using dual-confirmation
- **Robot Command Generation:** Translates inputs into real-time commands
- **Visual Feedback:** Displays real-time robot camera views
- **Augmentation Module:** Highlights areas of interest (AOIs)

2. Validation

Tested in a search-and-rescue scenario comparing:

- **System-Assisted (SA):** Augmented AOIs for user guidance.
- **Human-Assisted (HA):** Manual navigation with keyboard inputs.

3. Key Takeaways

- Faster and efficient visual search due to guided focus in high stake scenarios
- Gaze stability and reduced exploratory behaviour in SA
- Higher precision in attention guidance in SA
- Reduced Cognitive Load for users
- Extrafoveal attention capture around key AOIs
- Combining human intuition with automation ensures improved decision making and system usability

Results: System Validation and User Study

Smoothed user gaze heatmap during SA
Smoothed user gaze heatmap during HA
Y Position
X Position
Count
SA vs HA Views in Different Screen Regions
Up
Down
Center
Left
Right
SA
HA
Standard Deviation - HA Trajectories
Robot SA Trajectory
Map Boundary

Fig 2: Difference in user behavior for SA and HA scenarios.

Measure	Value
Latency	10ms
Data Transmission Rate	60Hz

Table 1: Core components achieved low latency and high data transmission rates, ensuring smooth operation in dynamic environments.

Measure	System-Assisted (SA)	Human-Assisted (HA)
Task Completion Time	274.41s	678.88s
Cognitive Load (NASA TLX)	33.4	53.85
System Usability Scale	80.13	58.61

Table 2: Comparing the System-Assisted (SA) and Human-Assisted (HA) approaches demonstrates that SA significantly enhances human performance by reducing cognitive load and improving task efficiency.

References

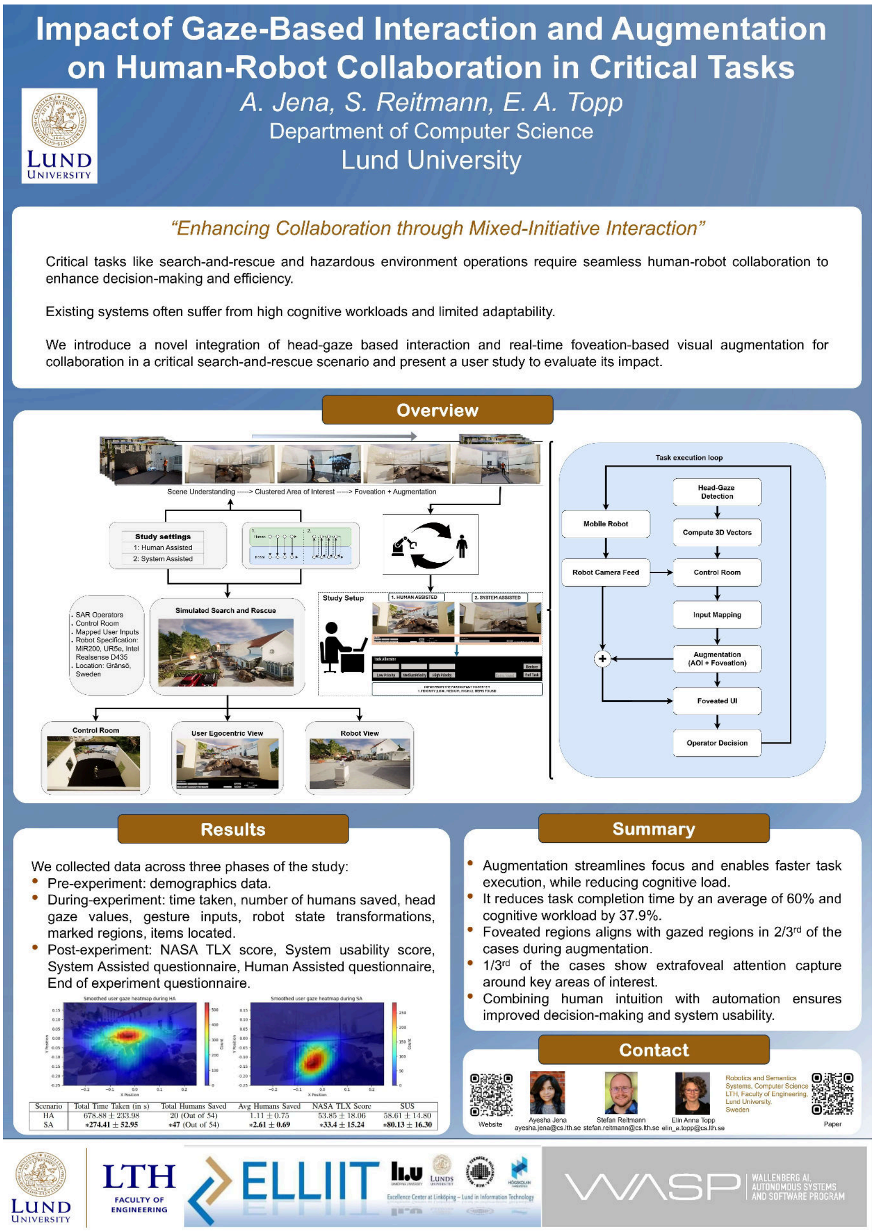
[1] Jena, A. and Topp, E.A., 2023, March. Chaos to Control: Human Assisted Scene Inspection. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 491-494).

[2] Jena, A. and Topp, E.A., 2024, February. Towards Understanding the Role of Humans in Collaborative Tasks. In 7th International Workshop on Virtual, Augmented, and Mixed-Reality for Human-Robot Interactions.

Contact

Ayesha Jena ayesha.jena@cs.lth.se
Elin Anna Topp elin_anna.topp@cs.lth.se
Jacek Malec jacek.malec@cs.lth.se
Björn Olofsson bjorn.olofsson@control.lth.se

ELLIIT Excellence Center at Linköping – Lund in Information Technology
WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM



Impactof Gaze-Based Interaction and Augmentation on Human-Robot Collaboration in Critical Tasks

A. Jena, S. Reitmann, E. A. Topp
Department of Computer Science
Lund University

"Enhancing Collaboration through Mixed-Initiative Interaction"

Critical tasks like search-and-rescue and hazardous environment operations require seamless human-robot collaboration to enhance decision-making and efficiency.

Existing systems often suffer from high cognitive workloads and limited adaptability.

We introduce a novel integration of head-gaze based interaction and real-time foveation-based visual augmentation for collaboration in a critical search-and-rescue scenario and present a user study to evaluate its impact.

Overview

Scene Understanding -----> Clustered Area of Interest -----> Foveation + Augmentation

Study settings
1: Human Assisted
2: System Assisted

- SAR Operators
- Control Room
- Mapped User Inputs
- Robot Specification: MiR200, UR5e, Intel Realsense D435
- Location: Gränsö, Sweden

Simulated Search and Rescue

Study Setup

Control Room

User Egocentric View

Robot View

Task execution loop

Mobile Robot

Robot Camera Feed

Head-Gaze Detection

Compute 3D Vectors

Control Room

Input Mapping

Augmentation (AOI + Foveation)

Foveated UI

Operator Decision

Results

We collected data across three phases of the study:

- Pre-experiment: demographics data.
- During-experiment: time taken, number of humans saved, head gaze values, gesture inputs, robot state transformations, marked regions, items located.
- Post-experiment: NASA TLX score, System usability score, System Assisted questionnaire, Human Assisted questionnaire, End of experiment questionnaire.

Smoothed user gaze heatmap during HA

Smoothed user gaze heatmap during SA

Scenario	Total Time Taken (in s)	Total Humans Saved	Avg Humans Saved	NASA TLX Score	SUS
HA	678.88 ± 233.98	20 (Out of 54)	1.11 ± 0.75	53.85 ± 18.06	58.61 ± 14.80
SA	***274.41 ± 52.95**	***47** (Out of 54)	***2.61 ± 0.69**	***33.4 ± 15.24**	***80.13 ± 16.30**

Summary

- Augmentation streamlines focus and enables faster task execution, while reducing cognitive load.
- It reduces task completion time by an average of 60% and cognitive workload by 37.9%.
- Foveated regions aligns with gazed regions in $2/3^{rd}$ of the cases during augmentation.
- $1/3^{rd}$ of the cases show extrafoveal attention capture around key areas of interest.
- Combining human intuition with automation ensures improved decision-making and system usability.

Contact

Website

Ayesha Jena
ayesha.jena@cs.lth.se

Stefan Reitmann
stefan.reitmann@cs.lth.se

Elin Anna Topp
elin_a.topp@cs.lth.se

Robotics and Semantics Systems, Computer Science
LTH, Faculty of Engineering,
Lund University,
Sweden

Paper

LUND UNIVERSITY
LTH FACULTY OF ENGINEERING
ELLIIT Excellence Center at Linköping – Lund in Information Technology
WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM