Multimodal Interfaces for Human-Robot Interaction

by

Shokoofeh Pourmehr

M.Sc., K. N. Toosi University of Technology, 2011
B.Sc., Amirkabir University of Technology, 2008

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

in the
School of Computing Science
Faculty of Applied Science

SIMON FRASER UNIVERSITY
Fall 2016

Approval

Name:	**Shokoofeh Pourmehr**
Degree:	**Doctor of Philosophy**
Title:	***Multimodal Interfaces for Human-Robot Interaction***
Examining Committee:	**Chair:** Dr. Nick Sumner Assistant Professor

Dr. Richard Vaughan
Senior Supervisor
Associate Professor _____

Dr. Greg Mori
Co-Supervisor
Professor _____

Dr. Brian Funt
Internal Examiner
Professor _____

Dr. Joelle Pineau
External Examiner
Associate Professor
Department of Computer Science
McGill University _____

Date Defended: 23 December 2016

Abstract

Robots are becoming more popular in domestic human environments, from service applications to entertainment and education, where they share the workspace and interact directly with the general public in their everyday life. One long-term goal of human-robot interaction (HRI) research is to have robots work with and around people, taking instructions via simple, intuitive interfaces. For a successful, natural interaction robots are expected to be observant of the human present, recognize what they are doing and act appropriately to their attention-drawing behaviors such as gaze, body posture or gestures. We call such a system by which a robot can take notice of someone or something and consider it as interesting or relevant *attention system*. These systems enable robots to shift their focus of attention to a particular part of the information that is relevant and meaningful in a given situation based on the motivational and behavioral state of the robot. This awareness comes from interpreting the exchanged information between humans and robots. The exchange of information through a combination of different modalities is anticipated to be of most benefit. *Multimodal interfaces* can be used to take advantage of the existing strengths of each composite modality and overcome individual weaknesses. Also, it has been argued [1] that multimodal interfaces facilitate a more natural communication as by employing integrated systems users will be less concerned about how to communicate the intended commands or which modality to use, and therefore be free to focus on the task and goals at hand. This PhD thesis presents our contributions made in designing and implementing multimodal, sensor-mediated attention systems that enable users to interact directly with physically collocated robots using natural and intuitive communication methods. We focus on scenarios when there are multiple people or multiple robots in the environment. First, we introduce two multimodal human multi-robot interaction systems for selecting and commanding an individual or a group of robots from a population. In this context, we study how spatial configuration of user and robots may affect the efficiency of these interfaces in real-world settings. Next, we present a probabilistic approach for identifying attention-drawing signals from an interested party and controlling a mobile robot's attention toward the most promising interaction partner among a group of people. Finally, we report on a user study designed to assess the performance and usability of this proposed system for finding HRI partners in a crowd when used by the general public and compare it to manual control.

Keywords: Human-Robot Interaction, Sensor Fusion, Sensor-Mediated Interface, Interface Designs, Attention Systems

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

One long-term aim of autonomous robot research is to have robots work with and around people in their everyday environments, taking instructions via simple, intuitive human-robot interfaces. All else being equal, we would prefer that these interfaces require no special instrumentation of the humans and little or no training. We believe that in everyday use, the convenience of using an interface is an important aspect of its performance. In this manner, robots are expected to be observant of human presence, recognize what they are doing, decide why they are doing it and act appropriately [2]. This awareness comes from interpreting the information exchanged between humans and robots.

In general, this exchange of information happens through one or a combination of four primary classes of perceptual modalities: vision-based, audio-based, touch-based and range-based [3]. Different sensory channels exhibit different strengths and weaknesses regarding details, resolution, and accuracy. The exchange of information through a combination of different modalities is anticipated to be of most benefit. *Multimodal interfaces* can be used to take advantage of the strengths of each composite modality and overcome individual weaknesses. Moreover, according to Perzanowski et al. [1], multimodal interfaces facilitate a more *natural* communication. They suggest by employing integrated systems users will be less concerned about how to communicate the intended commands or which modality should be used, and are therefore free to focus on the task and goals at hand.

Another aspect of information exchange is the *flow direction* [4]. As stated by Wang et al. [5], the flow direction of information happens through one of the four distinct types, displayed in Figure 1.1. Case (a), or *physically collocated*, shows the situation where the human and the robot are physically in the same workspace; hence both can directly interact with each other and the world. In case (b), or *remote control*, the robot and the human are in different workspaces while all the changes happen in the robot's workspace. The human perceives the world through the robot's sensors and affects the world via the robot's interaction with the world. Case (c), or *virtually collocated*, describes the situations where the human and the robot can directly affect the world without any direct interaction. Case (d), or *assisted control*, is similar to case (b) but in this case, the robot affects the world only via the human's interaction with the workplace.

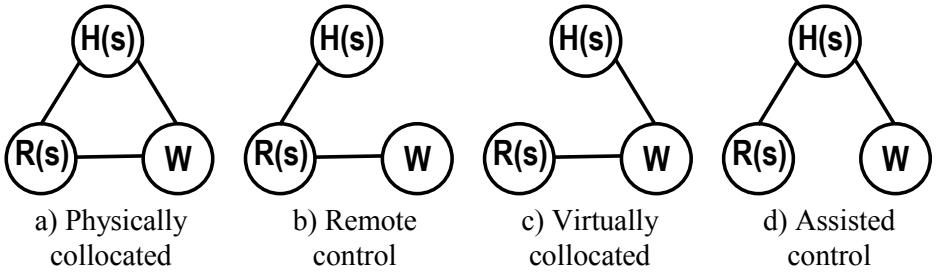


Figure 1.1: Information flow directions between human(s), robot(s) and the world [5]

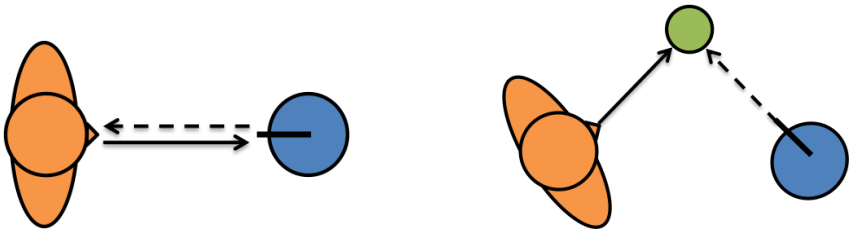
In this thesis, we focus on HRI systems that regulate a direct interaction between humans and robots via *sensor-mediated interfaces* (i.e. case (a) in Figure 1.1). In sensor-mediated interfaces, perceiving the human presence, intention and activity (i.e. human-oriented perception) are hardly relies on the robot's onboard sensors as opposed to remote controllers, desktop or GUI-based interfaces. Humans are essentially not instrumented with interfacing devices to be serviced by the robot [6]. They can interact with robots using human-like communication signals, such as speech, gesture, and gaze. Therefore, these interfaces facilitate a *natural and intuitive* interaction and require minimal or no special training [7].

It is beneficial for robots, intended to interact naturally with humans, to be able to detect and recognize human engagement behavior. Engagement is defined as the process whereby individuals in an interaction initiate, maintain and terminate their perceived connection to one another [8]. HRI research is mostly focused on the robot' and human's behaviors during the course of the interaction (*maintaining the interaction*), but not the *initiation* of it [9]. We emphasize the fact that HRI systems need to explicitly mark the beginning (or establishment of) the communication. This has not been sufficiently addressed in the field. We call such systems by which a robot can take notice of someone or something and consider it as interesting or relevant *attention systems*.

Attention systems enable robots to shift their focus of attention to a particular part of the information that is relevant and meaningful in a given situation based on the motivational and behavioral state of the robot [10, 11]. According to Kopp [12], "An attentive system must be able to identify relevant objects in the scene; select one of the identified objects; direct its sensors towards the selected object, and maintain its focus on it." Attentive robots can assess human intentions and properly react to their attention-drawing cues such as gaze, head pose, body posture, gestures, referential words or a combination of these.

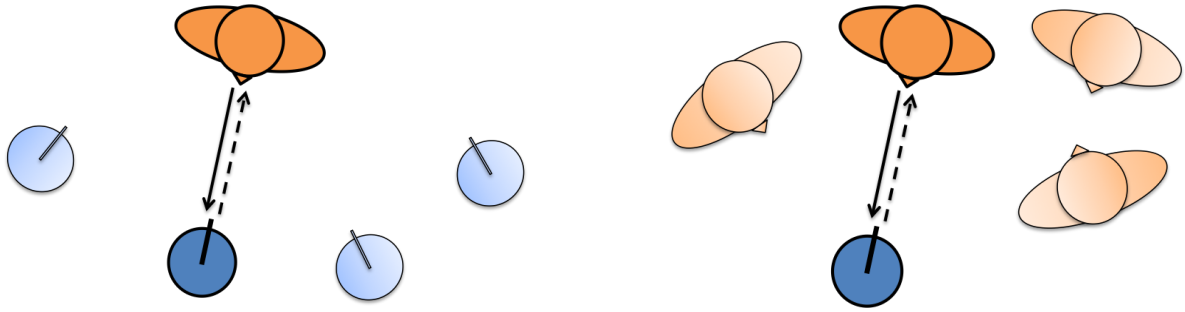
The robot's attention can be captured by a human to himself or an object or place in their environment. Based on the addressee location, attention is classified to *triadic* and *dyadic* [11]. In dyadic attention, humans and robots engage in a conversation-like interaction (*mutual attention*) (Figure 1.2a). One initiates the interaction by sending a signal, such as a waving gesture or spoken command, and the other acknowledges. In triadic attention, humans and robots intentionally focus on a shared object or space in the environment as referenced by one of the parties (*joint attention*) (Figure 1.2b).

In this thesis, we focus on designing attention systems for identifying human's behaviors in establishing *mutual attention* when there are multiple people or multiple robots in the environment. The presence of many humans or multiple robots notably poses additional challenges for recognizing interaction initiation signals.



(a) Mutual Attention (b) Joint Attention

Figure 1.2: The solid line shows the engagement initiating signal and the dashed line shows the acknowledgment signal.



(a) Single-Human Multi-Robot Scenario (b) Multi-Human Single-Robot Scenario

Figure 1.3: Different scenarios in establishing mutual attention where there are more than one human or one robot in the environment

In scenarios with multiple robots, the user must first designate a particular robot or robots of interest as the selected robots he intends to address in further interaction (Figure 1.3a). For sensor-mediated interfaces, the challenge is to disambiguate which robot(s) the user is attending to, as the communication channel (such as gaze or gesture) might be shared between several robots at the same time. One solution can be assigning names to each robot to be used to specify which one the user is attending. Selecting a robot by verbally calling its name cues all robots that the user wishes to interact with the mentioned robot. However, this can be significantly difficult as the number of robots grows.

In scenarios with multiple people, the robot must decide which human (if any) to interact with (Figure 1.3b). We want the robot to be able to automatically recognize potentially interested humans present in its workspace and then evaluate the posture, gesture or other salient features of each person to determine their intent to interact, and selectively engage in a one-to-one communication with the most promising interaction partner among all detected people.

The objectives of this thesis are:

- Proposing simple, easy-to-use and sensor-mediated (multi-)human (multi-)robot interaction systems

- Examining the effects of human-robot(s) spatial configuration and mechanism of composing teams of robot on the efficacy and efficiency of sensor-mediated interfaces
- Exploring whether a straightforward and robust sensor-mediated interface is an improvement over an alternative teleoperation

In what is to follow, we report our survey on sensor-mediated attention systems in Chapter 2. We have concluded that many of the proposed methods simplified the task of human interest detection with the stationary robots or static sensors assumptions (e.g. overhead cameras) or with the user(s) pose constraints. We suggest not only humans and robots should be unconstrained in position, but also be able to move freely around the workspace. Also, it is preferred for robots to be self-contained and rely only on their onboard sensing rather than external sensors or perception from some absolute point of reference.

1.1 Multi-Robot Scenarios: Recognizing Selection Commands

In Chapters 3, 4 and 5 we present our proposed multimodal interaction systems which facilitate the selection of an individual or a group of robots from a population for subsequent one-to-one interaction. One challenge of such systems designed for direct interaction is to identify which robot(s) among others are being selected since several robots might perceive the selection or commanding signals at the same time. Chapter 3 presents a complete and robust HRI system with several sensing modes, multi-phase robot behavior, and rich audio feedback. We show that using a pointing-based gesture combined with distributed election, an uninstrumented human can pick one robot while both robots and humans are moving freely around the workspace. In Chapters 4 and 5 we describe an interaction system designed for creating, modifying and commanding groups of robots from a population. We show that human-robot face engagement[1] can be used to determine the subject or subjects of verbal commands using indirect speech. This system is extended to the *Touch-To-Name* interaction system, whereby a user can identify an individual or a group of robots using haptic stimuli, and name them using a voice command. In Chapter 6 we study how the mechanics of composing teams for concurrent control and the relative positioning of humans and robots will affect the interface performance. We extend the concept of *Fan-out* introduced by Olsen et al. [14] to explicitly consider spatial constraints in real-world settings.

[1] Throughout this dissertation, we will use the term *face engagement* as coined by Goffman [13] to describe the process in which people use eye contact, gaze, and facial gestures to interact with or engage each other.

1.2 Multi-Human Scenarios: Finding a Potential Interaction Partner

In Chapter 7 we describe our proposed interaction system for controlling a mobile robot's attention in a medium-range multi-human scenario in semi-real world settings. We present a simple but effective method for integrating the outputs of multimodal human detectors.The performance and usability of this system are evaluated in a detailed user study with the general public in a semi-controlled setting. We studied how the robot's level of autonomy and response impact the user's interaction experience and workload. The results are reported in Chapter 8.

1.3 List of Publications

The following list of publications contain the preliminary reports describing methods and findings of this thesis:

1. Shokoofeh Pourmehr, Jack Thomas, and Richard Vaughan. Finding an interaction partner in a crowd, a robust interaction system and user study. *Manuscript in preparation*, 2016

2. Shokoofeh Pourmehr, Jack Thomas, and Richard T. Vaughan. What untrained people do when asked "make the robot come to you". In *Proceedings of the ACM/IEEE International Conference on Human-robot Interaction (HRI'16) (Late-Breaking Abstract)*, March 2016

3. Shokoofeh Pourmehr, Jake Bruce, Jens Wawerla, and Richard T. Vaughan. A sensor fusion framework for finding an HRI partner in crowd. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA'17), (Submitted)*

4. Shokoofeh Pourmehr, Jens Wawerla, Richard T. Vaughan, and Greg Mori. On the scalability of spatially embedded human multi-robot interfaces. In *Proceedings of the 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI'15) Workshop on Human-Robot Teaming*, Portland, USA, March 2015

5. Valiallah Mani Monajjemi, Shokoofeh Pourmehr, Seyed Abbas Sadat, Fei Zhan, Jens Wawerla, Greg Mori, and Richard T. Vaughan. Integrating multimodal interfaces to command UAVs. In *Proceedings of the 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI'14)*, pages 106–106, March 2014

6. Shokoofeh Pourmehr, Valiallah Mani Monajjemi, Seyed Abbas Sadat, Fei Zhan, Jens Wawerla, Greg Mori, and Richard T. Vaughan. You are Green: a Touch-to-Name interaction in an integrated multimodal multi-robot HRI system. In *Proceedings of*

the 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI'14), pages 266–267, March 2014

7. Shokoofeh Pourmehr, Valiallah Mani Monajjemi, Richard T. Vaughan, and Greg Mori. "You two! Take off!": Creating, modifying and commanding groups of robots using face engagement and indirect speech in voice commands. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'13)*, pages 137–142, November 2013

8. Shokoofeh Pourmehr, Valiallah Monajjemi, Jens Wawerla, Richard T. Vaughan, and Greg Mori. A robust integrated system for selecting and commanding multiple mobile robots. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA'13)*, pages 2874–2879, May 2013

Chapter 2

Approaches in Designing Attention Systems

This chapter gives an overview of the proposed methods for recognizing human intention in initiatings interaction. We focus on sensor-mediated interfaces that regulate a direct interaction between humans and robots sharing a physical workspace, as opposed to remote or manual control. The approaches are classified into methods for detecting: (a) signals for developing mutual attention and (b) cues for establishing joint attention.

2.1 Establishing Mutual attention

Here, we review proposed methods for recognizing human intention in creating *mutual attention* where an interested human can call a robot's attention to herself or himself through a communication initiation cue such as gaze, gesture or stepping into a zone. We divided this section into two parts: (1) attention systems for a robot to find its potential interaction partner, (2) attention systems for human multi-robot interaction.

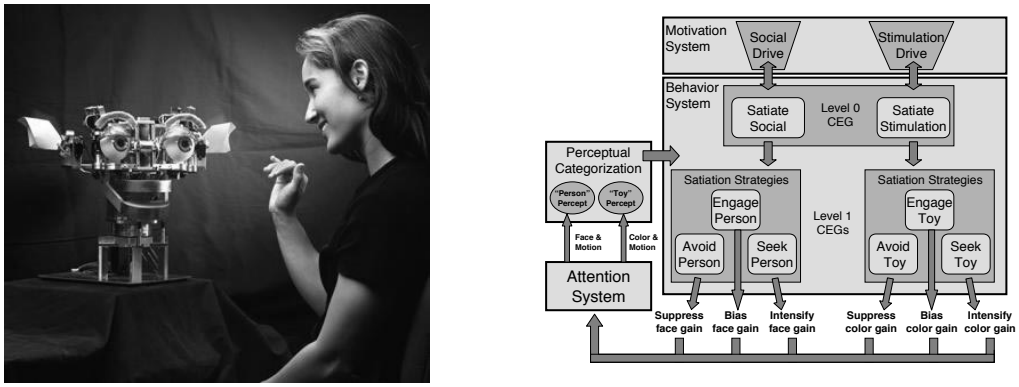
2.1.1 Finding an Interaction Partner

In the first phase of interaction, the robot has to find its potential interaction partner and (a) determine if a detected person is interested in engaging in a communication and (b) decide which person among all detected people in more interested in having a one-to-one interaction.

Interaction with a Single Person

The attention system for the robot Kismet (Figure 2.1a), proposed by Breazeal and Scassellati, is one of the first works on enabling a robot to bias its attention to facilitate interaction with a human [10]. They integrated visual perception and influences from the behavioral state of the robot to create a context-dependent attention system. The behavior states of the robot include *seeking people*, *avoiding people*, *seeking toys* and *avoiding toys*. The vision perception stimuli are divided into social stimuli (motion and face "*pop-outs*"), and non-social stimuli (motion and color saliency). The bias face gain is influenced by the social behaviors (i.e. seeking people and avoiding people), and the bias color gain is affected by the non-social behaviors (i.e. seeking toys and avoiding toys) (Figure 2.1b). The robot shows its shift of attention by eye movement. In a similar approach, Kozima et al. [23] proposed a vision-based attention coupling mechanism for the robot Infanoid [24]. They consider eye contact as a signal for interaction request. The robot actively searches for human's frontal faces using a skin-color filter and average-face templates. If any face is detected, the robot turns its gaze or face toward the human to establish eye contact.

Methods for combining audio and visual information were reported in [25] and [9]. Yonezawa et al. [25] used gaze and utterance for recognizing the user's intention in interaction. If the proportion of the user's gaze at the robot during her/his utterance is more than



(a) (b)

Figure 2.1: (a) Kismet, (b): Stimuli that drives the robot's attention [10]

50 percent, the robot assumes that it is the target of the user's speech. They employed single-camera-based gaze-tracking [26] to detect the user's attention direction, which consists of facial feature tracking and 3D eyeball model estimations. Another example is the work by Chen and Fitzgerald [9]. They proposed an audiovisual system, incorporating speech recognition, audio source localization, and face detection, to detect the human's intention. They assume the interested person will approach the robot while calling its name.

These attention systems [9,10,23,26] require humans to be located close to the robot due to the use of face engagement or gaze direction. However, the long-distance perception of human head pose and gaze is hardly possible with the current computer vision methodss. [27] and [28] introduced attention systems by which a human user can establish mutual attention with a distant robot. Monajjemi et al. [27] presented the first demonstration of establishing mutual attention with a distant (>25) outdoor flying robot (Figure 2.2). An uninstrumented human can attract the robot's attention, which actively searches for people, by waving at it. Once detected, the robot signals its attention to the user by performing a "wing-wobble" behavior.

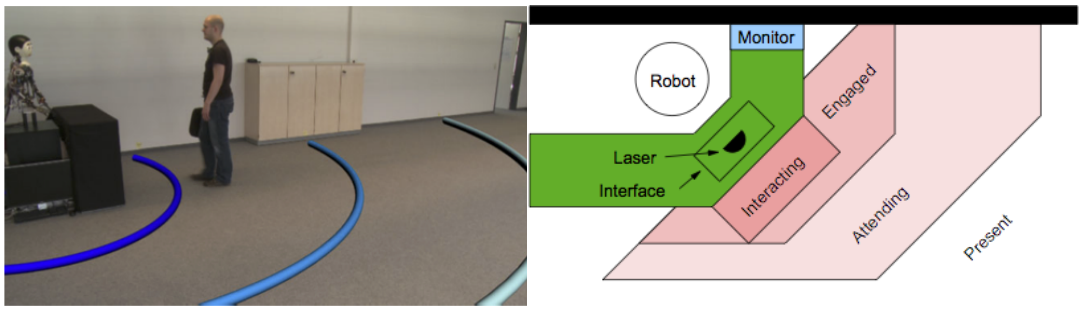
Another example is the system implemented by Bruce et al. [28] on a mobile ground-based robot, equipped with a consumer video camera, to detect people acting waving gestures from long distance (up to 35 meters). They extract the motion-containing regions of an image stream. If there is a region that has a strong periodic signal in the frequency range of human waving gestures, the human target is perceived. A key limitation of this method is that the periodic gesture detection module requires the robot to be stationary.

Distance and other spatial relationships have also been used as a basis for evaluating engagement, and as cues for interaction openings. The assumption is that people far away from a robot are probably less interested in a close interaction (or one-to-one interaction) with the robot compared to people coming towards it [29–34]. Nabe et al. [29] analyzed the behavior of 238 subjects in interacting with ROBOVIE-M, a humanoid robot, in a

Figure 2.2: A human waving at a UAV flying >25 meters above the ground [27]

science museum in relation to information from sound level and distance. They identified three zones in which a human shows *observing*, *talking* or *physical* interaction behavior depending on their distances to the robot. They concluded that people who stepped into the talking or verbal zone were willing to start an interaction.

Holthaus et al. [30, 31] designed an attention model that allows a receptionist robot to use the distance to a human as an input that triggers different behavioral outputs. Following Hall [35], four zones are defined based on the relative spatial positioning of a human with respect to the robot: intimate (<=45 cm), personal (<=120 cm), social (<=360 cm), or public (>=360 cm). The user, entering the personal zone, while facing the robot, will get its attention. Michalowski et al. [32, 33, 36] also designed systems for modeling the social engagement of a robotic receptionist, based on spatial information from a laser range scanner and head pose information from a video camera. They suggest the direction, and the speed of motion are more appropriate measures of engagement than location alone.



(a) (b)

Figure 2.3: Spatial attention: (a) A human in social distance to the robot [30, 31]. (b) Spatial regions used as initial estimate of engagement [33].

Some researchers used a person's movement trajectory information to determine who is willing to interact and who is not. Finke et al. [34] assumed that only people who approach the robot closer than one meter are interested in starting an interaction. The distance of one

meter is chosen regarding Hall's "social distance" [35]. Kato et al. [37] collected trajectories of pedestrians around a robot in a shopping mall and trained a classifier for estimating the people's intention in interaction. They chose four features including the distance to the robot, the stability of walking, the time being stopped at, and the size of a fan-shaped area in front of the robot that intersects with the human trajectory (Figure 2.4). Another data-driven approach for classifying a human's intentions and activities based on people trajectories is proposed by [38]. They trained Hidden Markov Models for intent recognition based on robot's observation of the distance and angle between its heading and the direction of a person. For example, these two features decrease as humans approach the robot.

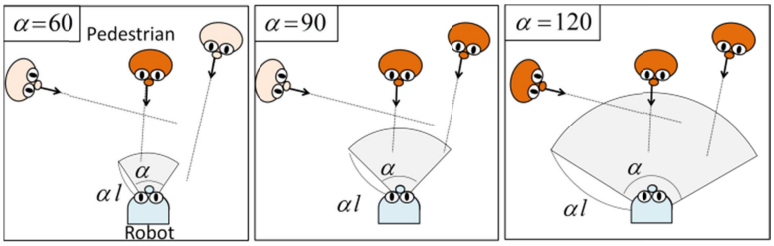


Figure 2.4: The fan shaped area in front of the robot that intersects with human trajectories is one of the features of the *"intention to interact"* class [37].

In a system proposed by Bruce et al. [39], the robot, instead of waiting motionless for people to approach, takes action to determine whether a person is interested in interaction. Detecting an intercept with a human, the robot changes its trajectory to remove the intercept. If the human corrects her or his trajectory to recreate an intercept, it will be considered as a signal that the human intends to initiate an interaction (Figure 2.5).

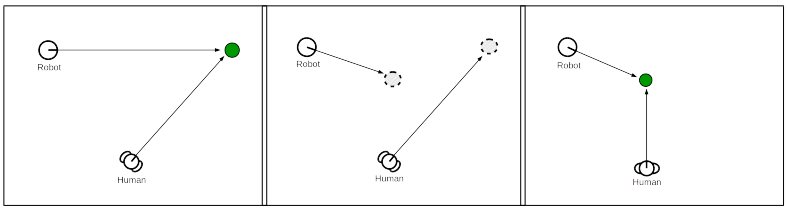


Figure 2.5: Probing intention to intercept. Left: Robot detects an intercept. Middle: Robot changes its trajectory to remove the intercept. Right: Interested human corrects its trajectory to recreate an intercept [39].

Interaction with Multiple People

In scenarios with multiple people, the robot may have to discern the most likely interaction partner among a group of individuals. The robot should be able to evaluate the posture, gesture or other salient features of each person to determine their intent to interact, and

Table 2.1: Comparison of surveyed studies on attention systems for establishing mutual attention in the single-human single-robot scenario.

Study	Platform	Cue of Attention	Awareness behavior	Evaluation[a]
Breazeal et al. [10]	active head	face engagement	gaze	SP
Kozima et al. [23]	upper torso	gaze	gaze and body orientation	US
Atienza et al. [40]	active head	face engagement/waving gesture	head orientation	POC
Yonezawa et al. [25]	toy robot	gaze and speech	speech and gaze	US
Chen et al. [9]	humanoid robot	approach and calling	head orientation	SP
Monajjemi et al. [27]	flying robot	waving gesture	wobble	SP
Bruce et al. [28]	mobile robot	waving gesture	approach	SP
Holthaus et al. [30,31]	humanoid robot	proxemics	gaze and body orientation	US
Michalowski et al. [32,33]	mobile robot	proxemics and head pose	speech and head orientation	US
Nabe et al. [29]	humanoid robot	proxemics	speech and body orientation	US
Finke et al. [34]	mobile robot	approach	head orientation	SP
Kato et al. [37]	humanoid robot	trajectory	speech and approach	SP
Kelley et al. [38]	mobile robot	trajectory	-	SP
Bruce et al. [39]	mobile robot	co-operative trajectory	approach	POC

[a] POC-proof of concept, SP-system performance, US-usability study

selectively engage in a one-to-one interaction with the person with the highest interest among all detected people.

Various strategies are proposed to decide which person gets the robot's attention. Some authors used visual information such as hand gestures [41–43], head pose [44] and body posture [45], some employed data provided by multiple modalities such as distance and visual information [46, 47], audio-visual information [7, 48–55] and a combination of three modalities [6, 56–58].

Stiefelhagen et al. [44] suggested that head pose can be used to help a robot to determine whether it was addressed by a person. They presented a system to model the focus of attention in a meeting scenario. People's faces are detected and tracked in a panoramic image (Figure 2.6). Then probability distributions of looking at other people are estimated from head poses using an unsupervised learning approach. Then these distributions are used to predict the focus of attention given a head pose.

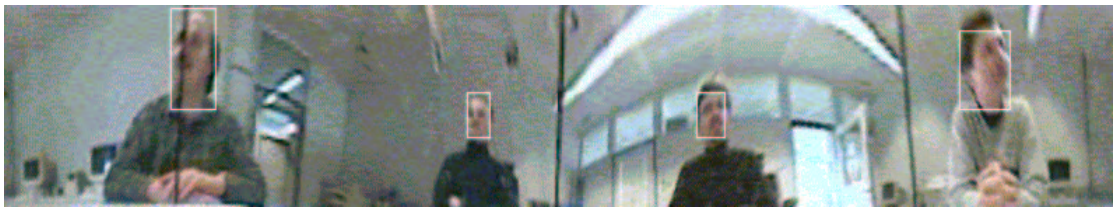


Figure 2.6: Panoramic view of the scene around the table [44]

Aguirre et al. [47] presented a vision-based approach for estimating the interest degree of people in surroundings based on their positions. It is assumed that a person near the

robot, standing centered with respect to it, has more interest than when he is far or at the either sides of the robot. The degree of attention is detected by analyzing the head pose.

Another example of vision-based attention systems is the work by Kobayashi et al. [43]. The authors designed a care robot to serve tea in a multi-party setting. People can make a request by raising their hand. The robot can accept multiple requests. If it detects another person raised his hand, asking for its service, while serving tea to someone else, it will turn its gaze to acknowledge that order is received. *JAMES* [45] is another robot capable of serving drinks. It can recognize people's intentions in ordering drinks at a bar. A depth camera and an RGB camera are used for detecting the body posture and head pose as non-verbal social cues for requesting attention. The spatial arrangement of people in groups is also incorporated as the third feature to recognize the customer's request for attention.

McKeague et al. [42, 59] also proposed to use depth information to detect human hands and bodies for multiple people detection (Figure 2.7). They employed a Bayesian framework for hand-body association. Geodesic distances are used to highlight important regions of a hand in crowded environments. The robot detects and associates multiple bodies and hands simultaneously and directs its attention towards the detected person who waves at it.

Figure 2.7: The hand and body associated method. Tracked bodies are displayed in red. Tracked hands are displayed in blue. Associations are shown with a connected white line [42].

The attention system proposed by Bohme et al. [7] utilized audio-visual information. They fused the information from vision-based movement detection and sound localization to detect people. Among detected people, the robot turns toward the one who approaches the robot. In a similar approach, Ruesch et al. [48] aggregated visual (color, intensity, and motion) with acoustic saliency maps to direct the robot's gaze toward the most salient location. Several authors consider the person that is currently speaking as the target person. For example, Okuno et al. [49,50] developed a real-time auditory and visual multiple-speaker

tracking system to control the focus of attention of an upper-torso humanoid robot called SIG [60]. The multiple human tracking is done by integrating the visual events (detected faces) and audio events (sound source directions). Matsusaka [51, 52] followed a similar strategy.

Spatial positioning has also been employed as a cue for willingness in interaction. Bennewitz [53, 54] assumed the person that tries to get the robot's attention would come closer to the robot. Tasaki et al. [55] set the priority of the interaction partner from multiple people based on the proxemics. The robot can locate people using face and sound source localization, and the distance is estimated using the size of the bounding box of the detected face. Poschmann et al. [46] developed a system for an interactive museum tour-guide robot to turn its head and look at the person of interest. The robot favors nearby people, with little head movement, and not being attended for a while, or at all.

Lang et al. [56] proposed an attention system for the mobile robot, BIRON [58], to estimate the position of the partner of interest and maintain attention during an interaction. Their method is based on *multimodal anchoring* person detection and tracking [61]. Anchoring means establishing connections (*anchor*) between processes that work on the symbolic level, or the level of abstract representations of objects in the world, and processes that are responsible for the physical observation of these objects (sensory level). The goal of multimodal anchoring is to link the description of an object to different types of perceptions, originating from different sensory modalities. The integration of all component's anchors will be done in a *composite anchoring process*. In this system, the object is the person and the components are face, speech and legs. A pan-tilt camera is used for face detection (and recognition if the identity of the person is known), stereo microphones are used to locate the direction of sound sources and a laser range finder is used to detect legs. Each person in the robot's vicinity will be tracked as soon as his or her legs *or* face is recognized by the system. If a detected person starts talking she will be recognized as the person of interest and the robot shifts its attention by turning its camera toward her (Figure 2.9). If she stops talking for more than 2 seconds it will lose the robot's attention to another talking person. This system covers distances up to 2 meters and angle up to 80 degrees. To prevent the robot not to turn towards the direction of not-valid voices (e.g. TV or radio), they extended the person attention system by voice validation and short-term memory [62].

Lin et al. [6] defined the interested user as the first person appearing at the direction of the detected sound, or the direction the robot is instructed to go. They developed a multimodal attention system that utilizes an array of eight digital microphones for speaker direction estimation, a laser scanner to detect legs and a camera for face detection. If all these human features are observed in the same direction, he or she will be considered as a potential interaction partner (Figure 2.10).

In Table 2.2 methods presented in this section are compared regarding the robot's platform and awareness behavior, the human's attention-drawing signals, and evaluation strate-

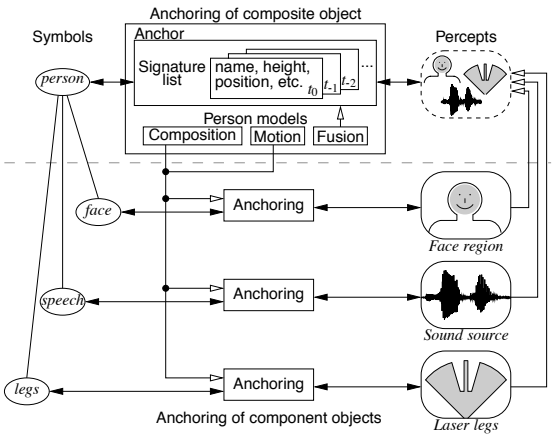


Figure 2.8: Multimodal anchoring [61]

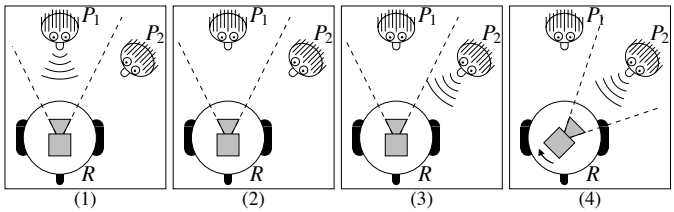


Figure 2.9: Sample behavior with two persons P_1 and P_2 standing near the robot R: In (1) P_1 is considered as communication partner, thus the robot directs its attention towards P_1. Then P_1 stops speaking but remains the person of interest (2). In (3) P_2 begins to speak. Therefore the robot's attention shifts to P_2 by turning its camera (4). Since P_2 is facing the robot, P_2 is considered as new communication partner [56].

gies. In contrast to the single human interacting with a single robot scenario, interface designs for multi-human settings are mostly multimodal with robots mostly acknowledge their attention by turning the body or gaze toward the interaction partner or speech.

2.1.2 Recognizing The Selection Command

In this section we review the proposed methods for establishing mutual attention between a human and an individual or a group of robots among a population. The human "multi"-robot interaction imposes additional challenges compared to regular dyadic human-robot interaction. Before starting an interaction, the user has to first somehow designate a particular robot, several of them or the whole group for further interaction. Payton [63, 64] proposed one of the first examples of such systems. He made use of a remote control to select the robot of interest using a narrow directional IR LED or to communicate with all robots via means of an omnidirectional IR LED. However, this interface is not sensor-mediated by our definition, as the interaction does not occur directly between human and robots. The user has to utilize an external device (i.e. remote control) for selecting robots.

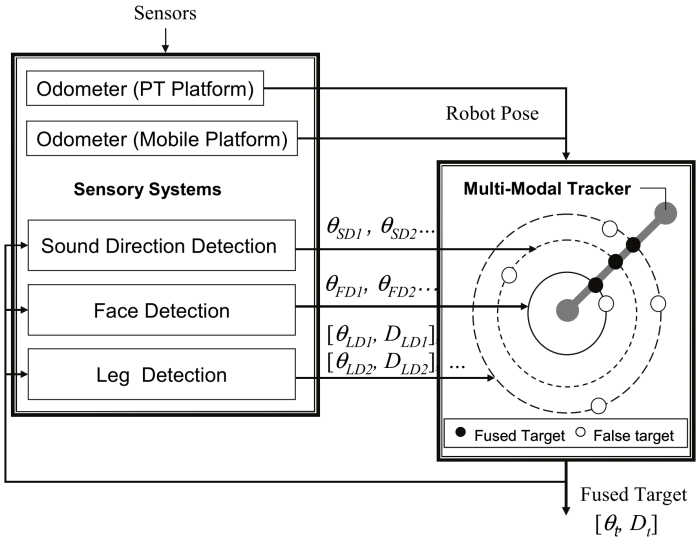


Figure 2.10: The gray bar shows that more than two trackers point toward the same direction, so the multimodal tracker will determine it as a target [6].

For sensor-mediated interfaces, the challenge is to disambiguate which robot(s) the user is trying to attend, as the communication channel (such as gaze or gesture) might be shared between several robots at the same time. One solution is assigning names to each robot that can be used to specify which one the human is attending to [65, 66]. Selecting a robot by verbally calling its name cues all robots to the fact that the user wishes to interact with the mentioned robot. However, all potential users have to learn the names of all robots.

Approaches to robot designation in multi-robot systems can be classified into two categories based on sensors or sources of perception. In the first category, the interaction mechanisms utilize shared sensors among all the participating robots for human-oriented perception. Thus, all robots perceive the human equally. A central unit interprets the human's selection commands and broadcast the data to the robots. In [67], for example, the user selects robots by looking at them. The user's gaze is traced by a head-mounted eye tracker. The tracker outputs 2D coordinates of the location where the human is looking and transfers this data to all robots in the user's surroundings.

Podevijn et al. [68] presented a gesture-based interface that allows an operator to select a particular group of robots from a swarm (subswarms) and command them as a single entity (Figure 2.11). The gestures are recognized by interpreting the 3D data from a Kinect sensor. When the operator sends the selection command, by opening his right arm, one subswarm gets chosen at random. Thus, the operator has to keep re-issuing the command until the intended subswarm gets selected.

For a robot to know if it is being attended and, therefore, belongs to a selected subswarm, it compares its subswarm ID with the subswarm-selected variable. They used a distributed algorithm to make sure that at any given moment, all robots across all different subswarms

Table 2.2: Comparison of surveyed studies on attention systems for establishing mutual attention in the multi-human single-robot scenario.

Study	platform	Cue of Attention	Awareness behavior	Evaluation[a]
Kumar et al. [41]	humanoid robot	waving gestures	shift of gaze	POC
McKeague et al. [42]	mobile robot	waving gestures	-	SP
Kobayashi et al. [43]	humanoid robot	hand raising gestures	shift of gaze	WOZ US
Stiefelhagen et al. [44]	simulated robot	head pose	-	SP
Gaschler et al. [45]	humanoid robot	head pose, body posture and spatial arrangement	-	SP
Aguirre et al. [47]	mobile robot	head pose and distance	-	SP
Poschmann et al. [46]	mobile robot	head movement and distance	head orientation	SP
Bohme et al. [7]	mobile robot	approach and sound	speech and body orientation	POC
Ruesch et al. [48]	upper torso	motion and sound	gaze	POC
Okuno et al. [49, 50]	upper torso	sound and face engagement	speech and body orientation	POC
Matsusaka et al. [51, 52]	humanoid robot	speech and gaze	speech and gaze	POC
Bennewitz et al. [53, 54]	humanoid robot	speech and face engagement and distance	gaze	US
Tasaki et al. [55]	upper torso	sound and face engagement and proxemics	speech and body orientation	POC
Lang et al. [56, 57]	mobile robot	face engagement and speech	body orientation	US, SP
Lin et al. [6]	mobile robot	face engagement and speech	head orientation	SP

[a] POC-proof of concept, SP-system performance, US-usability study

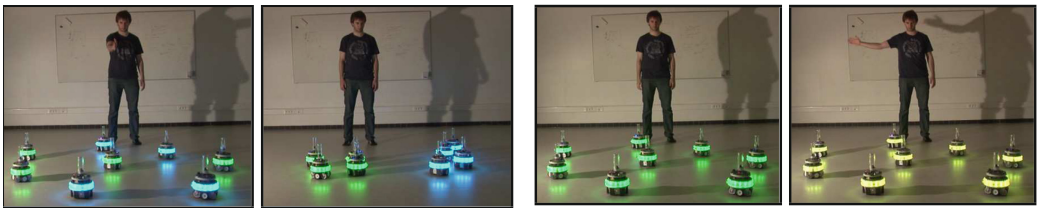


Figure 2.11: A user interacts with a swarm of robots using hand gestures. From left: (1) The user faces a swarm of 8 robots, (2) The user selects all robots by raising his right arm, (3) The user splits selected robots to two subswarms, (4) The user has two subswarms to interact with independently [68].

have the subswarm-selected variable set to the same ID. When the operator issues the selection command, the subswarm-selected variable will get updated based on the context. If none of the subswarms are selected, the variable will set to the smallest subswarm ID. After this, if another selection command is sent, the subswarm-selected value gets updated with the lowest ID in the list greater than the previous value. If there are IDs higher than the subswarm-select, no subswarm is selected anymore.

Another similar gesture-based interface is proposed by Alonso-Mora et al. [69] for human-swarm interaction. The user can select or deselect an individual robot by pointing at it, and a group of robots, by drawing a closed shape that encompasses them. When the user points to a position for more than one second, the closest robot to that position gets selected. A closed shape, convex or non-convex is recognized if two pointed positions are close to each other in space, but not in time. Figure 2.12 shows the setup of the system. A Kinect sensor is used for gesture recognition and an overhead camera for localization.

Similarly, Ozgur et al. [70] used pointing gestures to select and command individual modular robots (Roombots) on their grid. Their system, called Natural Roombots User

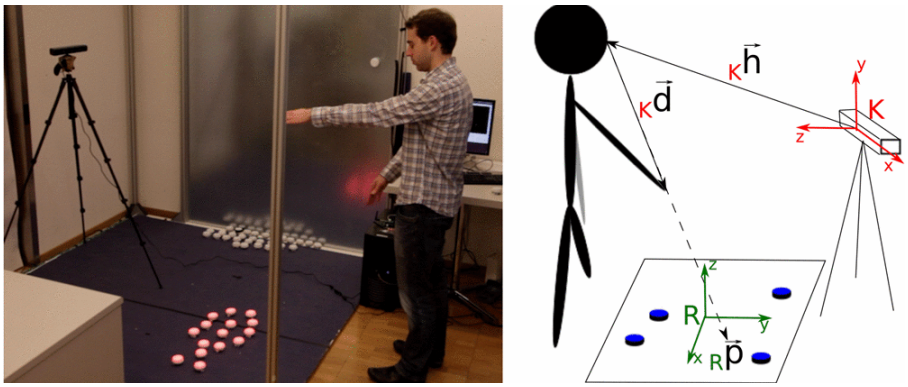
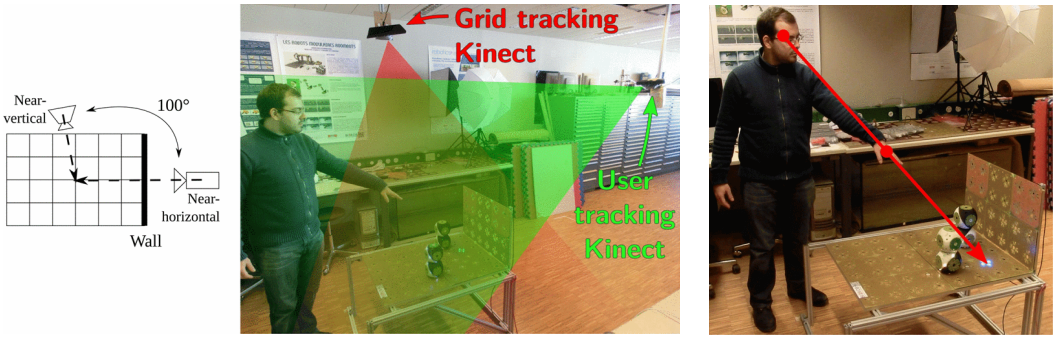


Figure 2.12: A user interacts with swarm of robots using hand gestures. Left: System setup, Right: Robot (R) and sensor (K) coordinate systems [69].

Interface, utilized a dual depth sensor setup to track the user's body and grid state (Figure 2.13). To move a robot, the user first selects the robot and then the target grid. For selecting a robot or a grid, the user points at them for 2 seconds. They will be deselected by being pointed at for 2.5 seconds.



(a) (b)

Figure 2.13: A human moves Roombots using pointing gesture: (a) Dual Kinect setup, (b) The head-hand pointing gesture [70].

Pointing gestures have also been employed for selecting and shuffling flying robots. Lichtenstern et al. [71] describe a prototype system in which multiple UAVs can be selected. Pointing gestures are recognized using a Kinect sensor, carried by a UAV hovering in front of an operator (Figure 2.14). The user can select each robot by pointing at it with his right arm and touching his right arm with his left arm to confirm it.

In all these proposed systems [68–71] a shared (RGB-D) sensor is used to determine which robot is attended by the user. The attention is detected by a central processing system and broadcast to all robots. The interaction environment is instrumented with an external device (fixed sensor), and the user has to stand in front of it to designate the robot(s) of interest.

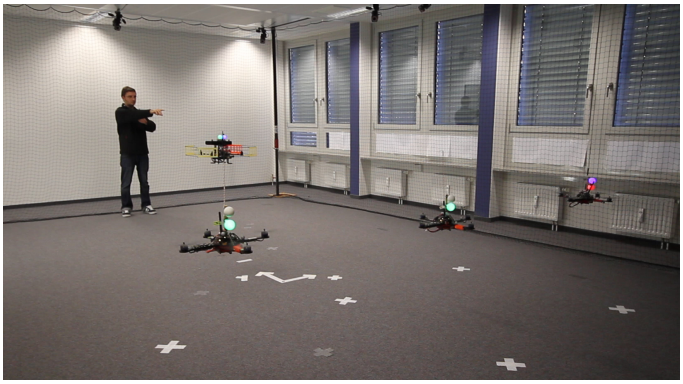
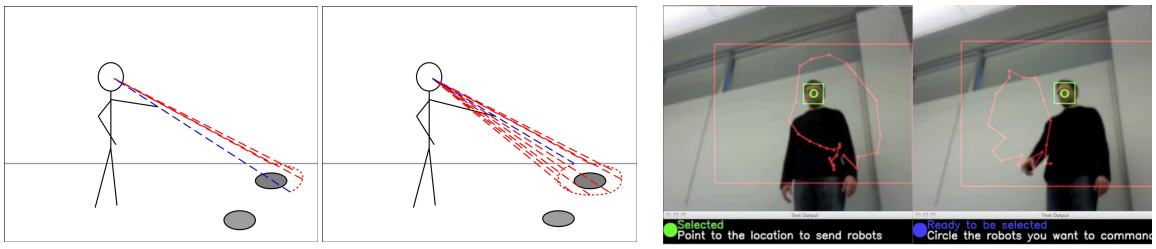


Figure 2.14: System setup for shuffling UAVs. The operator is pointing at the robot in his left to select it [71].

The other category of approaches to robot selection in multi-robot systems consists of methods where each robot has its individual point of view of the human. These systems use a minimal model of the environment and do not need to know the global location of each robot. There is no requirement for external sensors, overhead view or special properties in the environment to interpret the intention of the user. Examples include the work by Milligan [72] for selecting groups of robots without using any external instruments. Users can select robots by drawing a circle around the intended robots with their hand (Figure 2.15a). Each robot can determine whether it has been chosen by tracking the user's hand and face using an onboard RGB camera. The robot is selected if, from the robot's point of view, the user's face is within the polygon drawn by the hand motion (Figure 2.15b). The system, however, is not robust to incorrect selection. The user cannot deselect a designated robot or remove an individual from a group.

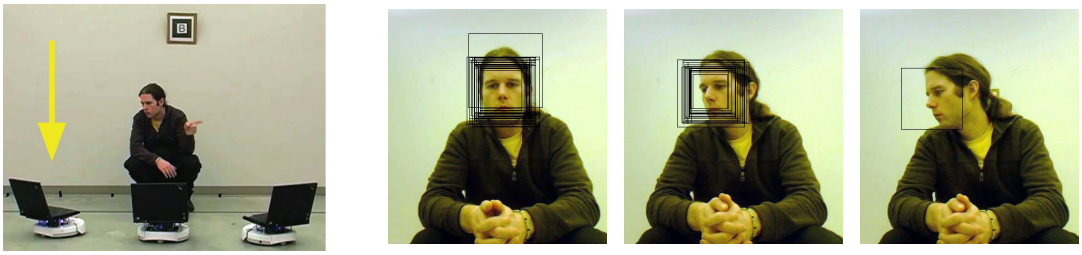


(a) (b)

Figure 2.15: (a) A user selects robots by drawing a circle around them, (b) An example of a positive selection (left) and a negative selection (right) [72].

Couture-Beil et al. [73] also employed a vision-based approach for the robot selection problem: a user selects an individual robot by focusing his gaze on it (i.e. making face engagement), and then commands the chosen robot with a motion-based gesture (Figure 2.16). The system uses face detection to estimate the gaze. Similar to [72], each robot is equipped with an RGB camera for face and gesture detection. The user's face might be simultane-

ously visible to multiple robots, so they should decide among themselves which one is being looked at. To solve this problem, the authors proposed to use the "face score": a measure of how directly the user is looking at the camera (i.e. the number of candidate detected faces). Robots use a wireless communication channel to collectively agree on the selection of a single robot at any time. Only the designated robot will watch for the gestural command.



(a) (b)

Figure 2.16: a) An uninstrumented human selects robots by simply looking at them. b) An example of three robots simultaneous camera views while arranged around a human user [73].

Inspired by Couture-Beil's work [73], Monajjemi et al. [74] proposed a selection method for flying robots. They used vision-based gestures (obtained from a passive monocular camera) to select and command a team of UAVs (Figure 2.17). The user focuses his attention on a robot by looking at it (i.e. making face engagement), then commands the selected robot using a one-handed waving gesture to add or remove from the team, or two-handed waving gestures to begin task execution. To handle the problem of rapid ego-motion of the flying robot's onboard camera, they employed a Kalman Filter to smooth face position estimates.

Figure 2.17: An uninstrumented user selects and commands flying robots using face engagement and waving gesture [74].

In similar work, Nagi et al. [75] presented a method for selecting individual and groups of robots from a networked swarm of UAVs using spatial gestures (i.e. gestures based on angular distances). A vocabulary of four two-handed spatial pointing gestures is defined as shown in Figure 2.18. The relative spatial configuration between the user and robots

is calculated based on face detection and the assessment of face poses. The user has to put on a pair of colored gloves and a jacket so that his hands are detected by color-based segmentation. Spatial gestures are recognized by combining individual opinions of predicted gestures using a distributed consensus protocol.

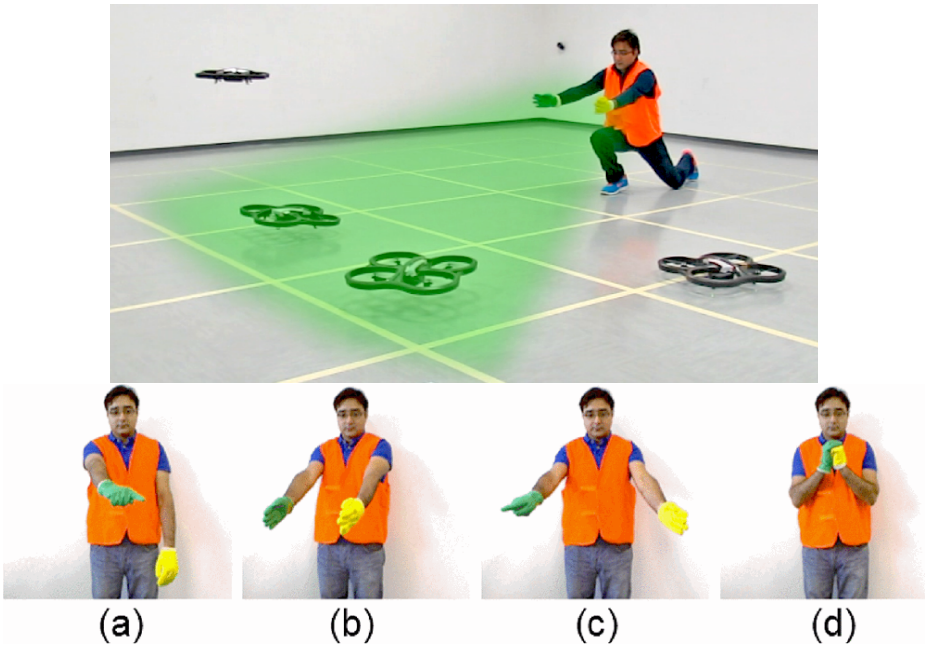


Figure 2.18: Top: An operator wearing tangible input device (i.e. known characteristics colors) selects two spatially located robots [76], Bottom: The two-handed gesture vocabulary for selecting: (a) individual robots, (b) group of robots, (c) individual and groups, (d) all robots [75].

In Table 2.2 the reviewed attention systems in this section are compared regarding the robot's platform and awareness behavior, the source of human-oriented perception, the cues of attention and evaluation strategies. As can be seen from the table, the robots exhibit their internal state mostly by changing their LED colors.

Table 2.3: Comparison of surveyed studies on attention systems for establishing mutual attention in the single-human multi-robot scenario.

Study	Platform	Perception	Cue of Attention	Awareness behavior	Evaluation[a]
Briggs et al. [66]	mobile robots	shared	calling their name	speech	POC
Xu et al. [67]	humanoid robots	shared	gaze	gaze	SP
Podevijn et al. [68]	mobile robots	shared	hand gestures	lighting LEDs	POC
Alonso et al. [69]	mobile robots	shared	pointing gesture	lighting LEDs	SP, US
Ozgur et al. [70]	mobile robots	shared	pointing gesture	lighting LEDs	POC
Lichtenstern et al. [71]	flying robots	shared	pointing gesture	lighting LEDs	POC
Milligan et al. [72]	mobile robots	individual	hand gestures	lighting LEDs	SP
Couture et al. [73]	mobile robots	individual	face engagement	lighting LEDs	SP
Monajjemi et al. [74]	flying robots	individual	face engagement	lighting LEDs	SP
Nahi et al. [75]	flying robots	individual	hand gestures	lighting LEDs	SP

[a] POC-proof of concept, SP-system performance, US-usability study

2.2 Establishing Joint Attention

The majority of studies on attention systems in HRI address the issue of how to detect a human's intention to establish *joint* attention by directing the focus of the robot's attention to a particular object or place in the environment, through social cues such as gaze, pointing gestures. Joint attention is defined as *spatio-temporal coordination of each other's attention* for social human-robot interaction [23]. Ferreira et al. [11] provided a survey of contributions to the modeling and implementation of the attentional mechanisms for socially interactive robots.

Several models have been proposed to use deictic gaze or referential looking for transferring knowledge about the position of intended objects. The user switches her focus of attention from the robot towards the intended object or location. In this case, the objects should be in the line of sight of the human. The work by Scasselati [77] was one of the first efforts for implementing the robot's gaze following for establishing the joint attention. Shon et al. [78, 79] proposed a system that utilizes a probabilistic model to follow the gaze of a human. They employed the gaze vectors (inferred from head pose) along with the color saliency maps of the visual scenes to estimate the position of the given objects attended by the user. The algorithm is implemented on an active stereo vision robotic head. Similar approaches [23, 80–82] have used head pose as an estimate for the human gaze direction toward intended referent object. In [81], the depth of the object along the direction of gaze is also inferred from the head orientation, which is used to estimate the center-of-mass of the object. Atienza and Zelinsky [40, 83] proposed an active vision system to determine the focus of attention of a user by searching along the human's gaze line. When the robot locates a human (detecting any moving object that has a color similar to human skin tone), it searches for an object that the user is staring at and picks it up and hands it over.

Oliveria et al. [84] proposed another gaze-tracing algorithm for attention systems. They combined the gaze direction with the short-term memory of the egocentric representation of the environment to enable the robot to recognize where the user is attending. In a different approach, Nagai et al. [85] employed a constructive model to learn the human gaze direction to a salient and interesting object by investigating the correlation in the sensorimotor coordination. However, this model only utilized static information (i.e. face direction) from the user. The constraint of this method was that the human had to tell the robot when to shift its focus of attention. To relax this constraint, in a follow-up study, Nagai [86] presented a learning model which used both motion and static information obtained from observing the user's gaze shift (Figure 2.19). Motion cues were also employed as the trigger to shift robot's gaze [87]. They reported that motion information accelerated the start-up time of learning the joint attention.

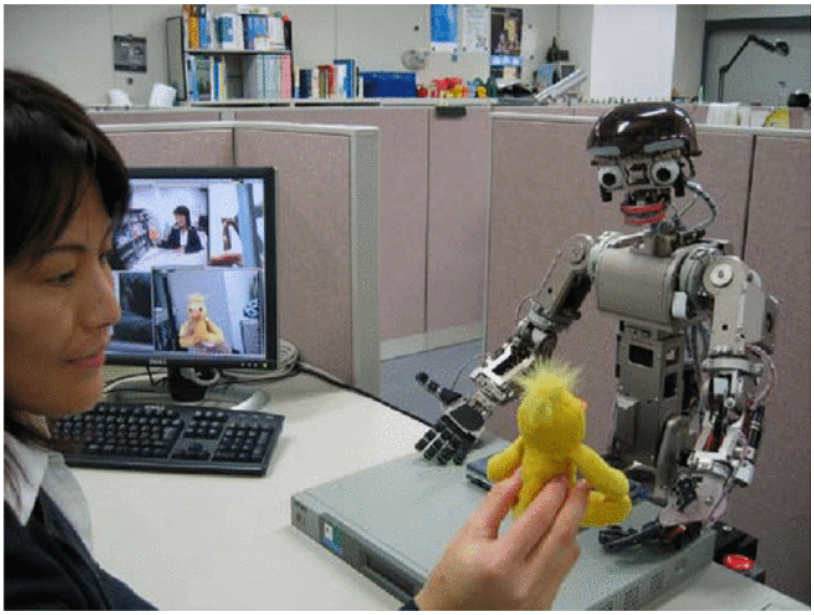


Figure 2.19: The robot looks at the toy in the user's hand by following her gaze [86]

After gaze direction, the most common attentional signal played for establishing joint attention is declarative pointing gestures [88]. In this case, the intended objects are pointed at by the human arm and hand. Examples include the work by Schauerte et al. [89]. The authors proposed a system that combines visual saliency with the directional information obtained from pointing gestures for estimating the position of pointed at objects. Pointing gestures are estimated by head-shoulder detector based on histograms of oriented gradients. The visual object-based saliency map is calculated based on spectral whitening of the image signal (Figure 2.20).

In the system proposed by Hofemann et al. [90], pointing gestures are recognized based on the trajectory of the hand that is detected by applying skin color segmentation. Then

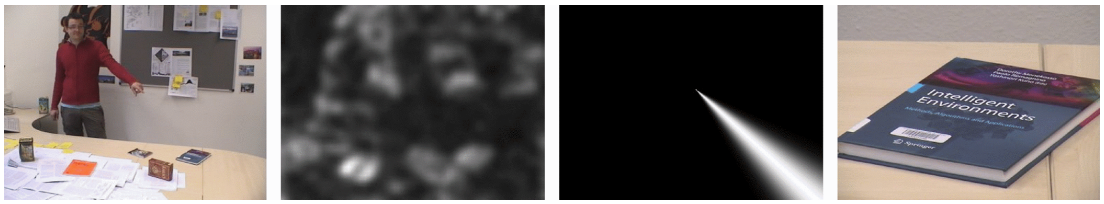
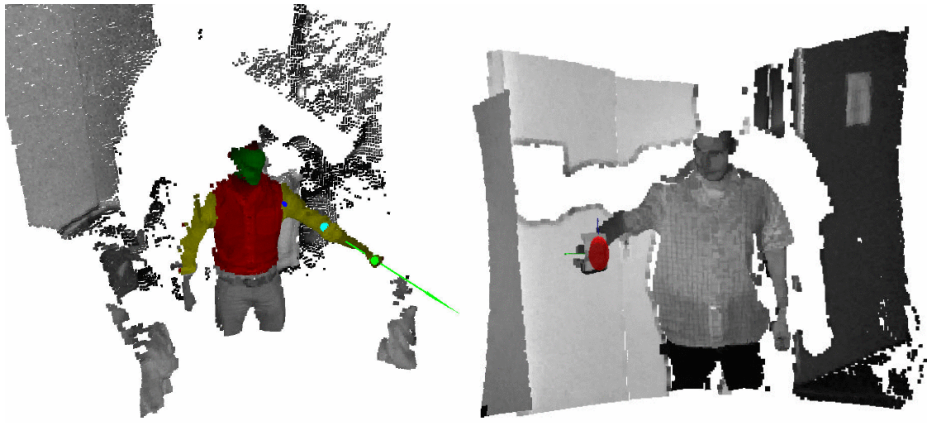


Figure 2.20: From left to right: the input image, object-based saliency map, pointing saliency map and attended object [89].

using the information from the direction of the hand motion, the target object is searched for in the expected area.

Similarly, Droeschel et al. [91, 92] used pointing and showing gestures to draw a robot's attention to a target object. They used a Time-of-Flight camera for perceiving gestures and estimating the pointing direction and a conventional color camera for object recognition (Figure 2.21). Pointing gestures are recognized by detecting and localizing the human's head, elbow, and hand. Two pointing vectors including eye-hand and elbow-hand are calculated. The robot picks the object that is closer to the pointing vector.



(a) Pointing gesture (b) Showing gesture

Figure 2.21: Recognizing pointing and sharing gestures [91]

These gestures can also be used to direct a robot's movements by pointing at the direction of the route [93]. Van den Bergh et al. [94] implemented a system for an interactive mobile robot by which a user can show the direction of the robot's navigation goal with a pointing gesture (Figure 2.22). The robot first looks for people to initiate an interaction. After detecting a human, it follows the direction of a detected pointing gesture and then searches for a new person. Human detection is achieved by integration of skin, face and leg detection. Then the robot asks the detected human to wave at the robot if he or she is willing to interact. Hand gestures are recognized using Kinect depth data and the 3D pointing direction is defined as the line that connects the wrist with the center of the hand. Other

examples of methods for pointing gesture recognition for joint attention include [95], [96] and [97].

Furthermore, objects can be drawn to the robot's attention when touched or manipulated by the human. Ito and Tani [98] used only cyclic movement patterns for generating joint visual attention. The robot learns the user's hand movements by tracking colored balls on the user's hands in an offline training phase and reproduces them when the user performs similar kinds of movement patterns.

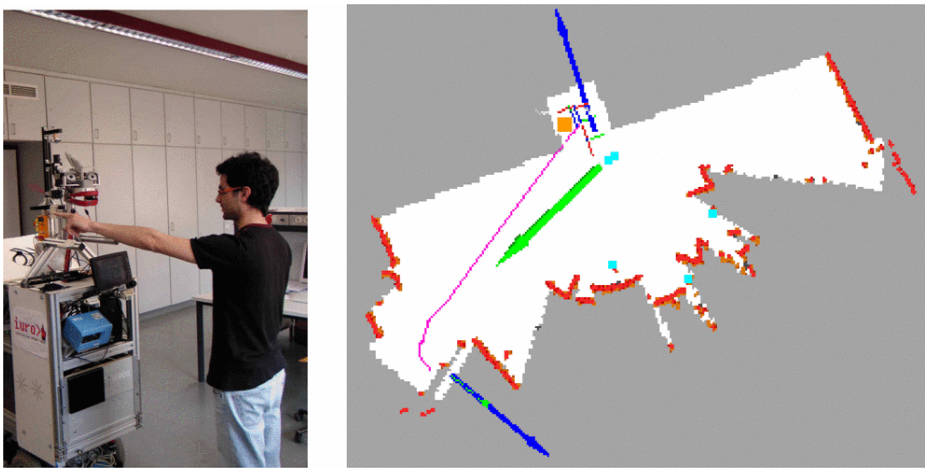


Figure 2.22: Person pointing in the direction of the next exploration goal [94]

Other research efforts include studying joint attention from a multisensory perspective. In these scenarios, the person points or looks at an object and says something about it such as "look at that" or "this is a book".

Haasch et al. [99] developed a multimodal object attention system (OAS) for identifying objects pointed to by the user with gestures, verbal instructions and salient object features (e.g. color, shape). The goal is to direct the robot's attention toward an unknown object and teach them what to focus on. When a lexical cue such as 'this' or 'that' is detected, the gesture recognition module gets activated. Then the area of the camera image that is pointed at by the user is searched for extracting the description of the object (e.g. type, color, owner).

Kuno et al. [100] presented a "Vision for Communication" system to enable a robot to understand human utterances and intentions. The robot can respond to simplified implicit utterances (e.g. "Get that for me") by recognizing human actions such as face and hand movement or pointing gestures. Similar approaches that utilized pointing gestures and referential words include [101] and [102, 103]. In a different audiovisual approach [104], speech is used to add information to the belief about detected target objects. The human points toward an object and utters a label for it (e.g. "this is a red button"). The visual features of the pointed-at object (e.g. color, state and location) are combined with speech information, allowing the human to refer to it by name.

Table 2.4: Comparison of surveyed studies on attention systems for establishing joint attention.

Study	platform	Cue of Attention	Awareness behavior	Evaluation[a]
[77], [80], [104]	humanoid robot	gazing at the object	gaze following	POC
[78], [81], [85]	humanoid robot	gazing at the object	gaze following	SP
[40, 83]	active head	gazing at the object	picking up the object	POC
[84]	?	gazing at the object	-	SP
[89]	?	pointing at the object	camera orientation	SP
[90]	mobile robot	pointing at the object	-	SP
[91, 92]	humanoid robot	pointing at the object	-	SP
[94]	humanoid robot	pointing to the direction of the goal	body movement	SP
[98]	humanoid robot	manipulating the object	body movement	SP
[99]	mobile robot	verbal referencing + pointing at the object	camera orientation	POC
[100]	humanoid robot	verbal referencing + gazing at the object	gaze following	SP
[101]	?	verbal referencing + pointing at the object	-	SP
[102, 103]	humanoid robot	verbal referencing + pointing at the object	arm/head movement	US

[a] POC-proof of concept, SP-system performance, US-usability study

In Table 2.4 methods presented in this section are compared regarding the robot's platforms and awareness behaviors, the human's attention-drawing signals, and the evaluation strategies. Designed interfaces for establishing joint attention have been mostly tested on humanoid robots. In some cases, the authors did not mention the platform (indicated by "?" in the table).

2.3 Evaluation Methods

There is no consensus on a standard set of metrics for evaluating HRI systems. Interface systems differ widely based on tasks, applications, user skills, platforms and sensors, and identifying a set of generic metrics that can provide a framework for assessing a broad range of interface designs is not straightforward. Murphy and Schrechenghost [105] conducted an analysis of 29 papers that proposed metrics for HRI, and identified 42 metrics (see Figure 2.23). They were categorized based on the object directly measured: the human (7), the robot (6), or the system (29). The **system** category was subdivided into *productivity*, *efficiency*, *reliability*, *safety* and *coactivity*. Some metrics have been implemented in different ways and thus may appear in more than one category.

Other evaluation metrics that can be added to the **system** category include *usability*, *social acceptance*, *user experience*, and *societal impact*, which are proposed from a human-centred HRI perspective [106]. Also, Clarkson and Arkin [107] developed seven usability heuristics for evaluating interface system designs: *visibility of system status*, *appropriate information presentation*, *use natural cues*, *synthesis of system and interface*, *help users recognize, diagnose, and recover from errors*, *the flexibility of interaction architecture and aesthetic* and *minimalistic design*.

The usability of interaction systems also depends on the necessity of being user adaptive or training and instrumenting the user (user possessing, wearing or carrying a specific device to communicate with the robot). Thus, infrequent and short training time for potential users are other evaluation factors [108].

In another survey of evaluation metrics, Steinfeld et al. [109] argued that criteria for *good performance* differ substantially depending on the robot or the task purpose of the interface design. For example, in evaluating the *social effectiveness*, metrics should be chosen based on the internal design of the robot, whether it has a basis in cognitive science or it is a functionally designed robot, and based on that determining which metrics (engineering, psychological, sociological) are most appropriate for evaluating social effectiveness. They proposed to generalize common metrics in terms of five task categories: *navigation*, *perception*, *management*, *manipulation* and *social*.

In the surveyed studies, several methods have been used for reporting the accuracy, robustness, responsiveness and effectiveness of the designed attention system. The most common objective criterion for measuring success in establishing mutual or joint attention is the false positive rate of detecting the target user or object among distractors in real robot trials. Chen et al. [9] evaluated the performance of their proposed multimodal human interest detection algorithm with over 100 hours of real life experiments. They used precision, recall and F-measure for estimating the accuracy and robustness of the interest detection algorithm, and the robot's reaction time for assessing the responsiveness.

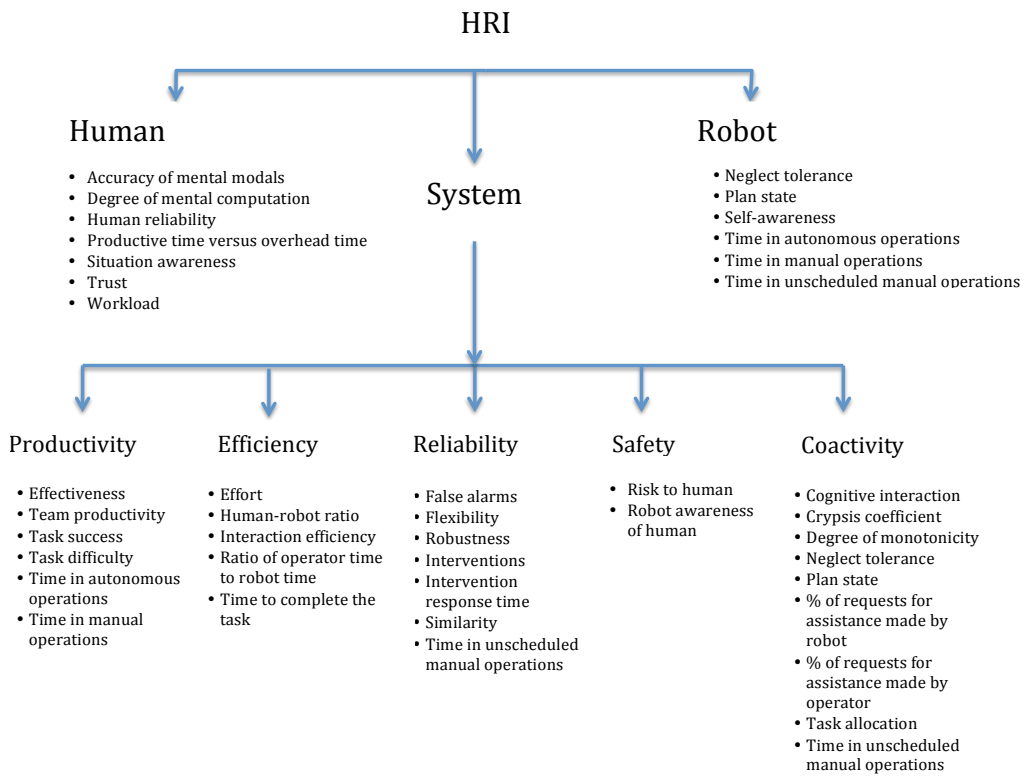


Figure 2.23: Survey on metrics for HRI [105]

Some approaches used multiple conditions such as different target objects and users [40], various integration schemes [46], or the accuracy of the sub-systems [6, 34]. In some cases, the spatial success rate [75], or temporal success rate is measured [10, 67]. Some validated their approach with the proof-of-concept demonstrations of an exemplary interaction cycle on real robots [39, 71]. Monajjemi et al. [27] demonstrated the practicality of their approach using a series of real-world trials.

Several proposed attention systems have been evaluated with the help of subjective questionnaires to study how the attention behavior change the people's perception of the robot and how it can improve the social aspects of human-robot interaction. Yonezawa et al. [25] used various attentive behaviors of the robot to show the importance and effectiveness of their proposed crossmodal awareness in creating positive impressions of the robot.

Kobayashi et al. [43] conducted a Wizard of Oz study to assess the effectiveness of the robot's capability to display acknowledgment by turning its gaze toward the person who wants to initiate interaction. The result verified that people feel more satisfied with the robot even if it cannot immediately deal with their request. Similarly, Holthaus et al. [30,31] showed that a robot that displays its attention and intentions is perceived as more interested in the human compared to having no attentional behavior.

2.4 Conclusion

We surveyed research that addressed the problem of how robots can detect a human intention in initiating interaction. Attention systems enable robots to assess and react to human attention-drawing behaviors such as gaze direction, body posture, gesturing or spatial positioning. The robot's shift of attention can be exhibited in several ways depending on the robot's platform such as turning its camera or body to the direction of the target or changing its LED colors.

The contributions made in designing such systems are studied and compared in the context of establishing joint or mutual attention. Joint attention refers to situations where humans and robots jointly shift their focus to a particular object or place in the environment. In the surveyed studies, pointing and gazing at the target are the most commonly used signals for coordinating joint attention. Mutual attention refers to a dyadic one-to-one interaction. Methods for recognizing mutual-attention development signals, such as waving gesture, head pose or proxemics, are studied in the contexts of single-human single-robot interaction, single-human multi-robot interaction, and multi-human single-robot interaction.

Research in attention systems in HRI often focused on establishing joint attention or mutual attention between a single human and a single robot and quite a few researchers have studied the problem of creating mutual attention when there are multiple people or multiple robots in the environment.

In scenarios with multiple robots, before engaging in an interaction, the user should somehow designate particular robot(s) for subsequent one-to-one interaction. This is a challenging problem since each robot must decide if the user is paying attention to it or its peers. They have to cooperate to combine their independent observations of the user and determine which robots were intended. Pointing gestures, waving gestures and face engagement are three proposed methods in the surveyed studies for selecting robots from a population. One extension to these attention systems can be integrating indirect speech (e.g. "*You Two*") similar to making joint attention, but with multiple robots. This will add the capability of selecting arbitrary numbers of robots by saying "*You N*", where N is the number of desired robots. Incorporating voice command can enable the user to select groups of robots from a population and create a direct addressing scheme to previously anonymous robots (e.g."*You three are Red Team*"), and command a team of arbitrary size with a single command (e.g. "*Red Team, take off!* "). We explore this in Chapter 5.

In scenarios with multiple people, the robot should be able to evaluate the posture, gesture or other salient features of each person to determine which of several people in their surroundings wants to initiate an interaction. In surveyed works, several passive and active stimuli have been used for designating a particular person among a population for subsequent one-to-one interactions such as spatial arrangement, body posture, hand-raising gestures, and speech. While these studies typically focus on close range human-robot

distances, the problem of controlling a mobile robot's attention in distant multi-human robot interaction is much less well explored. This is a challenging task. In addition to ordinary sensor noise, other people may be moving around the environment and occlude the target; people walking by or performing other tasks will change their appearance to the robot's sensors, and the robot's ego-motion changes the sensor readings at every sample.

Many of the proposed methods simplified the task of human interest detection with stationary robots or static sensors assumptions (e.g. overhead cameras) or with the user(s) pose constraints. We suggest that both humans and robots should be unconstrained in position and motion and be able to move freely around the workspace. It is preferred for robots to be self-contained and rely on their onboard sensing rather than external sensors or perception from some absolute point of reference. In our view, HRI systems do not require the user to possess, wear or carry any device or receive special training to interact with robots.

However, working with sensor-mediated interfaces is limited by the quality of sensing and the spatial arrangement of users and robots. While state-of-the-art techniques provide excellent face tracking, human pose estimation, speech recognition, and etc., in ideal conditions, we often have occlusion, motion-induced blur and false positives from background noise and clutter, which makes developing a robust HRI system rather challenging. To achieve robust operation despite persistent sensing problems, we propose to integrate multiple sources of sensory information, human detection, and tracking. This method differs from [61] in that different human percepts are detected and tracked independently, allowing for parallel detection and tracking, as oppose to [61] where the robot starts looking for faces as soon as it detect any leg, in their case. Also our proposed probabilistic framework, differs form [6] in that the spatial position of all multimodal percepts are described by Gaussian distributions instead of scalar values and integrated probabilistically.

Chapter 3

A Robust Integrated System for Selecting and Commanding Multiple Mobile Robots

In this chapter we describe a system whereby multiple humans and mobile robots interact robustly using a combination of sensing and signaling modalities. Extending the work by Couture Beil et al. [73] on selecting an individual robot from a population by face engagement, we show that reaching toward a robot - a specialization of pointing - can be used to designate a particular robot for subsequent one-on-one interaction. To achieve robust operation despite frequent sensing problems, the robot uses three phases of human detection and tracking, and emits audio cues to solicit interaction and guides the behavior of the human. A series of real-world trials demonstrates the practicality of our approach.

3.1 Introduction

Our goal is to develop methods for uninstrumented humans to give commands to individual and groups of robots using simple and natural interfaces. By "simple and natural" we mean that humans interact with robots in ways familiar from human-human or human-animal interactions, such as pointing gestures, gaze direction, and spoken commands. It has been argued that using these familiar interaction modes for HRI could mean that people require less training or have lower cognitive load compared to using a novel unique-to-robots interface [110]. Certainly, a common approach has two distinct benefits for teams of many humans working with one or more robots. First, humans can interact with robots and human teammates in the same way, which means that in both only one method must be learned and that a single execution of a command could be received by a mixed team of robots and humans. Second, uninstrumented and untrained non-teammate human observers can potentially understand the HRI they are watching.

For example, our research group has previously shown that an individual robot can be selected from a population for one-on-one interaction by a user simply looking directly at that robot [73]. This also works for humans and many other animals; we are acutely sensitive to a steady gaze. In this approach, each robot carries a camera and uses a standard face detection algorithm to evaluate how well it can see the user's face. The innovation of this work was using explicit wireless communication between robots to perform a distributed election algorithm to unambiguously decide which robot (if any) was being looked at directly, and was thus the subject of attention. Once elected, the single selected robot would watch the user for subsequent commands. Non-selected robots would not attend to this command, and indeed would not waste resources looking for it. Below, we show that this method can be generalized by replacing the face engagement with a pointing-based gesture.

However, working with gesture-based interfaces is limited by the quality of sensing. While the state-of-the-art techniques provide excellent face tracking, human skeletal pose estimation, and etc., in ideal conditions, we often have occlusion, motion-induced blur and false positives from background clutter. This means that building a robust HRI system is challenging. The method described below employs multiple phases of human-detection

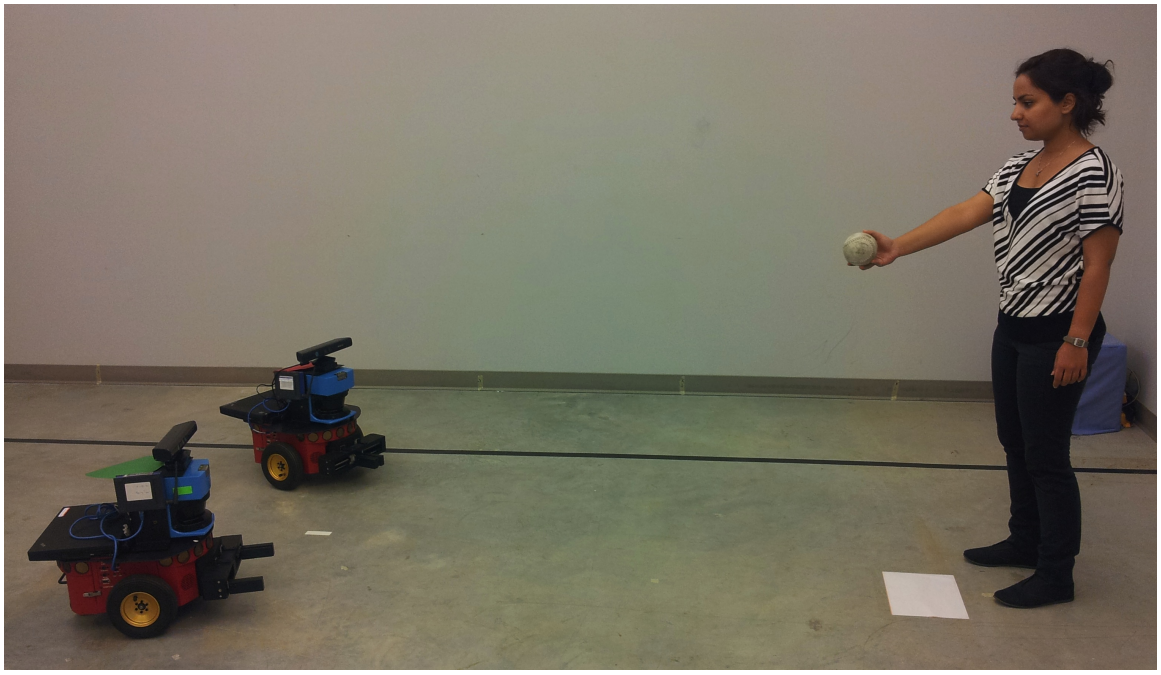


Figure 3.1: An uninstrumented person selecting one robot out of a group by offering it a ball - a modified pointing gesture.

with timeouts, retries, and fallback behavior to contribute to robustness. However, even with a significant engineering effort, we find that the robot still encounters a tough sensing condition every few minutes or roughly 10% of interactions. Our approach to this "last 10%" problem is to provide rich feedback to the user about the robot's state, so that the user can choose to make the problem easier; perhaps by adjusting their pose so the robot can see previously occluded limbs or joints. In practice we informally observe that this rich feedback makes interaction feel more responsive even when no problem occurs since no-touch interfaces have no built-in feedback as observed by Adams:

> The machine was rather difficult to operate. For years radios had been operated by means of pressing buttons and turning dials; then as the technology became more sophisticated [...] all you had to do was wave your hand in the general direction of the components and hope [111].

By providing carefully-designed audio feedback about all its interaction state changes, our robots quickly reassure users that their waving is working.

The contributions of this work are (i) the demonstration of using a pointing-based gesture combined with distributed election to guarantee that at most one robot is selected; (ii) the first demonstration of an uninstrumented human selecting a robot from a population while both robots and humans are moving freely around the workspace; and (iii) a case study of a complete and robust HRI system using several sensing modes, multi-phase robot

behavior and rich audio feedback to guide the user to resolve the "last 10%" of tricky sensing situations.

3.2 Background

3.2.1 Multimodal Interaction Systems

Existing researches on integrated interaction systems use a combination of different modalities as input. The work by Steifelhagen et al. [112] is a classic example of an integrated system, which includes speech recognition and vision for color-based hand and face tracking to estimate pointing direction. Wang et al. [113] describe a fusion approach using scanning LIDAR data for leg detection combined with vision-based human detection. Droeschel et al. [92] also use LIDAR for leg and torso detection, and subsequent vision-based detection. Further, they provide a study of pointing for human-robot interaction, defining a Gaussian process regression model for estimating pointing direction from depth data. In contrast, we consider a multi-robot scenario using pointing gesture for selection, and developed a verification approach for human localization that solicits human interaction to aid the robot in deciding if a candidate detection is valid or not.

3.2.2 Interaction with Multi Robot Systems

The works on human-robot interfaces for multi-robot systems can be broken up into two general cases:

Traditional Human-Computer Interfaces

Rather than interacting directly with robots, a traditional human-computer interface is used to represent the spatial configuration of the robots and allow the user to interact with the robots remotely. Examples include McLurkin et al. [114] that uses an overhead view of the swarm in a traditional point-and-click GUI named "SwarmCraft", and work by Kato that displays an overhead live video feed of the system on an interactive multi-touch computer table, which users can control the robots' paths by drawing a vector field over top of the world [115].

Embodied, World-Embedded Interactions

Embodied, world-embedded interactions occur directly between humans and robots, through mechanical or sensor-mediated interfaces. A useful property of this type of interaction is that since robots observe humans directly using their onboard sensing, they may not need to localize themselves in a shared coordinate frame in contrast to the GUI-based interfaces. Also, human users can walk and work among the robots, and are not tied to an operator station. Examples include work by Payton that uses an omnidirectional IR LED to broadcast

messages to all robots, and a narrow, directional IR LED to select and command individual robots [116]. Naghsh et al. present a similar system designed for firefighters, but do not discuss selecting individual robots [117]. Zhao et al. propose the user interacts with the environment by leaving fiducial-based "notes" (for example, "vacuum the floor" or "mop the floor") for the robots at work site locations [118]. Xue et al. introduces a clever fiducial design for imperfect visibility conditions and combines this with user-centric gestures in an underwater scenario [119].

Audio cues are also often used for human detection, including the recent work of Deleforge and Horaud [120] in which a "cocktail party robot" localizes a sound source with an active robot head with binaural sensors.

Our research group, previously developed face engagement [73] and circling-gesture [121] techniques for single- and multiple-robot selection. However, these systems had no strategy for human detection other than faces, and the vision system for interpreting circling gestures lacked robustness. In this work, we provide a novel, robust integrated system that includes human detection strategies, pointing estimation from a depth sensor, and solicits interaction to guide the human's behavior.

3.2.3 Gesture-Based Robot Interaction

There is a vast computer vision literature on gesture recognition: Mitra and Acharya [122] provide a survey. Several gesture-based robot interfaces exist; we do not attempt to provide an exhaustive survey, but rather mention some interesting examples. See Chapter 2. Systems may use static gestures where the user holds a certain pose or configuration, or dynamic gestures where the user performs a combination of actions.

Waldheer et al. [123] use both static and motion-based gestures to control a trash-collecting robot. Loper et al. [124] demonstrate an indoor/outdoor person-following robot that uses an active depth sensing camera to recognize static gestures. Earlier work by Kortenkamp et al. [125] presents a mobile robot that uses an active vision system to recognize static gestures by building a skeleton model of the human operator; a vector of the human's arm is used to direct the robot to a particular point. Perzanowski et al. [1] present a multimodal speech and gesture-based interface; an active vision system is used to interpret pointing gestures as directional vectors, and to measure the distance between the user's two hands.

All gesture-based systems discussed so far are designed to work with a single robot, with exception of [1]; however, the work by Couture Beil et al. [73] from our research group was among the firsts to allow for gesture-based interfaces designed for multi-robot systems which rely solely on non-verbal communication. In this chapter, we present a novel variant of this system: a user can select and command an individual robot using a pointing-like gesture.

3.3 Method

For the work presented in this chapter we use two Pioneer DX3 robots shown in Figure 3.1. Both robots are equipped with the well-known Sick LMS200 scanning LIDAR and the popular Microsoft Kinect active RGB-D sensor. Also, each robot has a 2 DOF gripper mounted in front. The mobile robot base, laser, and the gripper are controlled by the built-in computer running ROS. The Kinect sensor is connected to a laptop mounted on top of the robot. This computer provides the computational power needed for skeleton tracking based on the Kinect data, using the ROS Kinect stack. The robots are controlled by an off-board computer. Note this is done for convenience not for lack of onboard computational resources which are very modest compared to the skeleton extraction process.

The Kinect sensor is designed as a novel human interface for computer games. In normal use the sensor is stationary and human players move around nearby in front of a television. By mounting the sensor on a mobile platform we create two challenges, (i) the range and field of view are smaller than that of sensor traditionally used in robotics such as LIDAR and passive RGB cameras; and (ii) the sensor has difficulties acquiring a skeleton if the sensor itself is in motion. In the following, we describe how we address both problems.

3.3.1 Coarse Human Detection

The robot's first task is to find a human for interaction. The Kinect sensor's field of view covers only a small part of our 8×10m arena, so we use a 2D laser range finder with an 180 degree × 8m field of view. To find legs, we look for discontinuities with certain properties in the laser range data that corresponds to two human legs close to each other.

While this simple method is fast and effective, it is subject to false positives since many objects in the world, such as furniture or a pair of trash cans may appear similar to a pair of legs. Once a candidate leg-pair is detected the robot narrows its laser field of view to the section in the scan where the legs were last detected. This filters out subsequent leg detections which the robot would otherwise have to reject to stay on target, and provides an almost cost-free method of focusing the robot's attention on a single candidate detection. Next the robot servos towards the detected legs to either confirm or reject the presence of a human.

3.3.2 Fine Human Detection

Once the robot is close enough ($< 3m$) to the location of the detected legs to reliably use the Kinect sensor, the robot stops briefly. It now triggers the Kinect's built-in user detection algorithm based on the 3D measurements from the sensor. Figure 3.2 shows a successful user detection, marked by the color blobs. If a match is found, a human is successfully detected. The hypothesis of a human being present is rejected if no user is detected in the Kinect data, or the location of the legs does not match that of any detected user. The latter

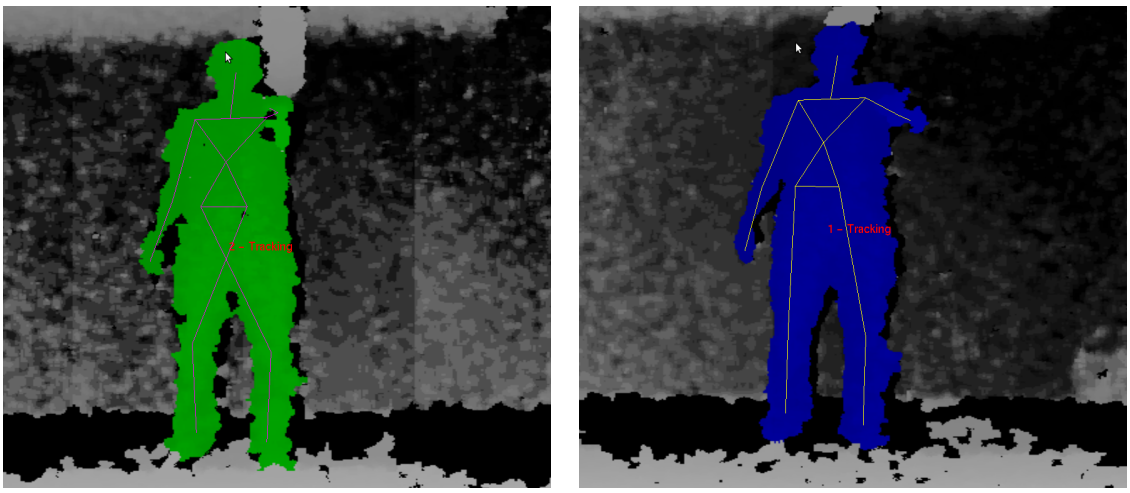


Figure 3.2: View of a human performing a *reaching gesture* from the intended robot (left) and the unintended robot (right) in a setup similar to that shown in Figure 3.1. The color blobs (blue and green) indicate that the user is successfully detected.

is important because the Kinect may also report false positives. In case of rejection, the robot returns to the laser-based leg detection turns away from the false positive detection and wanders around the world.

3.3.3 Gesture Recognition

Gestures are recognized by interpreting the skeleton data from the Kinect sensor. After a human is successfully detected the robot triggers the Kinect's skeleton recognition algorithm. Depending on the pose of the human, skeleton matching can fail, e.g. if joints are occluded. In this case, the robot plays a "sad" sound as a hint to the user to adjust her pose. If a skeleton cannot be detected after a threshold time, the robot gives up and returns to the leg detection behavior. A successful skeleton match triggers the playback of a "happy" sound. We found (informally) that this basic feedback greatly improves the usability of the system. It makes it easier for the user to assist the robot in situations that would otherwise be difficult for the robot to resolve by itself. For example changing the viewing angle usually does not resolve joint occlusions caused by an unnatural human pose or baggy clothing.

The robots are programmed to distinguish three simple gestures: pointing left, pointing right and a reaching gesture. Detecting no gesture indicates to the robot that the human is currently not interested in interacting; after some time watching the human with no gesture detected, the robot gives up and turns away to look for a different companion.

The reaching gesture indicates that the user wants the robot to either fetch or deliver an object, e.g. a ball. Whether the object is to be fetched or delivered is decided based on whether the robot currently holds the object or not. This can be directly measured by

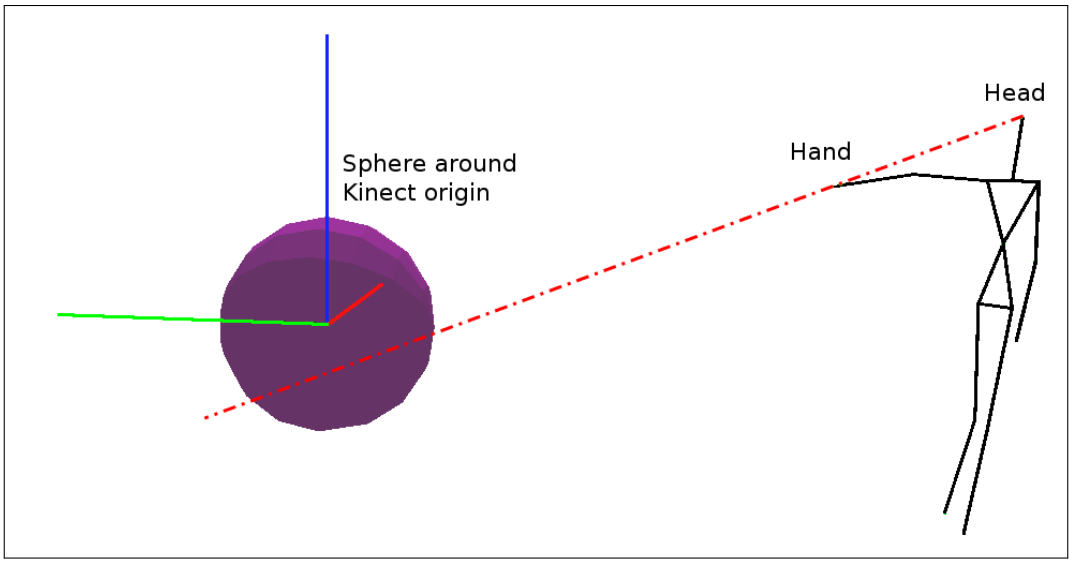


Figure 3.3: A reaching gesture is recognized if there is an intersection of a sphere around the origin of the Kinect sensor with the line between head and hand joints.

reading the touch sensors in the gripper paddles since our system knows there is exactly one ball in its world. Detecting a reaching gesture requires obtaining the position of the user's head and hand from the Kinect skeleton data. An example of skeleton data for the reaching gesture is shown in Figure 3.2. A gesture is classified as reaching if a line drawn through the head and hand joints intersects with a sphere centred at the origin of the Kinect sensor. This is shown in Figure 3.3. The required precision of the gesture can be adjusted via the radius of the sphere. This gesture works with either the left or right arm. Conceptually we consider reaching a modified pointing gesture - the only difference is the shape (and possibly content) of the hand.

By pointing either right or left the user can instruct the robot to turn in the respective direction. This allows a user uninterested in giving or receiving a ball to send the robot in the direction of someone who is. This gesture is detected by drawing a line through the points of the hand and elbow joints. If the orientation of this line in the Kinect frame of reference is within a given range (chosen experimentally), the gesture is classified as pointing. The angle α between the line and the x-axis has to be $-40° < \alpha < 40°$ for the left arm and $-40° < |\alpha| - 180° < 40°$ for the right arm. At the same time, the angle β to the y-axis and γ to the z-axis both have to be between 50° and 130°. Figure 3.4 illustrates the concept.

3.3.4 Sounds

The robots emit sounds to provide feedback to users at the moments described in the next section. All sounds for this demonstration are from the *Willow Garage Robot Sounds*

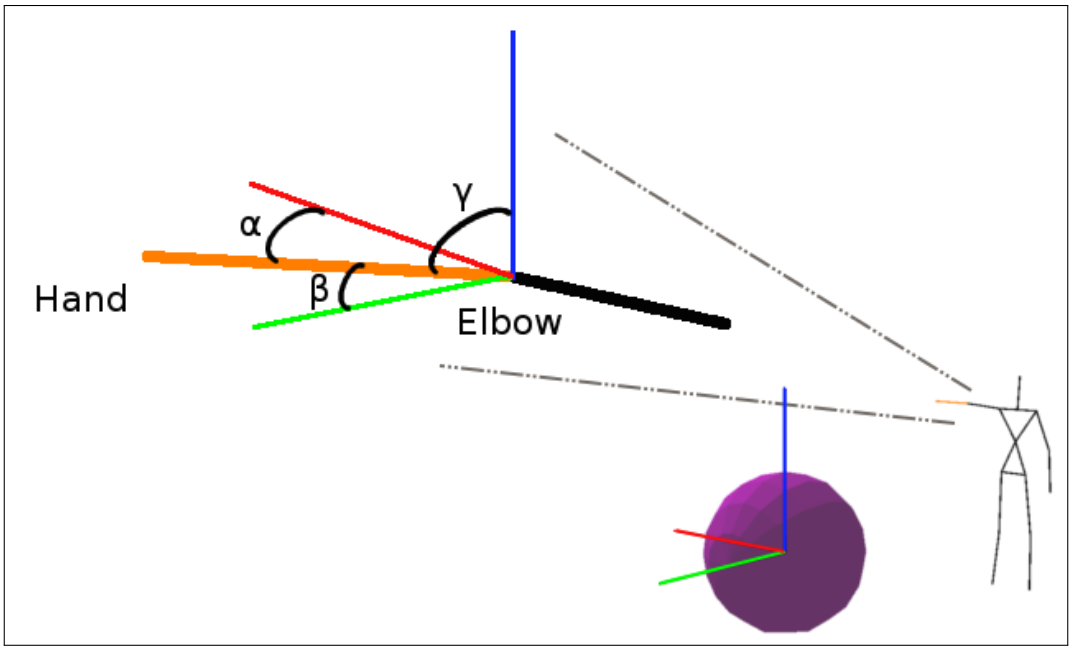


Figure 3.4: A pointing gesture is recognized by analyzing the orientation of the line between hand and elbow joints.

Library[1]. We believe the sounds make a significant contribution to the system robustness by informing the user about the robots' internal state. However, we do not provide evidence for this here: this is left for future work, along with the interesting topic of how to design effective sounds.

3.3.5 Multi-Robot System

The method described above works well (see experiments below) in a single robot setting with one or more humans present. It also works well if two robots approach a human from very different directions, that is the angle between the robot's trajectory is 45° or larger. If the robots approach more or less in parallel, as in Figure 3.1, the gesture classification algorithm has difficulties determining if the human intended to reach for robot A or B, see Figure 3.2. The problem is that the reaching-gesture-detection spheres can overlap in these situations and a reaching line (the line between head and hand joints) can intersect both spheres and hence both robots will positively identify the reaching gesture.

We address this problem with an election based method developed earlier by our group [73]. In the original version robots compared the quality of their human frontal face detections to determine which robot (if any) the user was looking at. Generalizing this idea, we seek to obtain a scalar value that varies from different robot viewpoints, broadcast that value and elect the robot with the best score. To disambiguate the reaching gesture we use the

[1] `http://hri.willowgarage.com/sounds/`

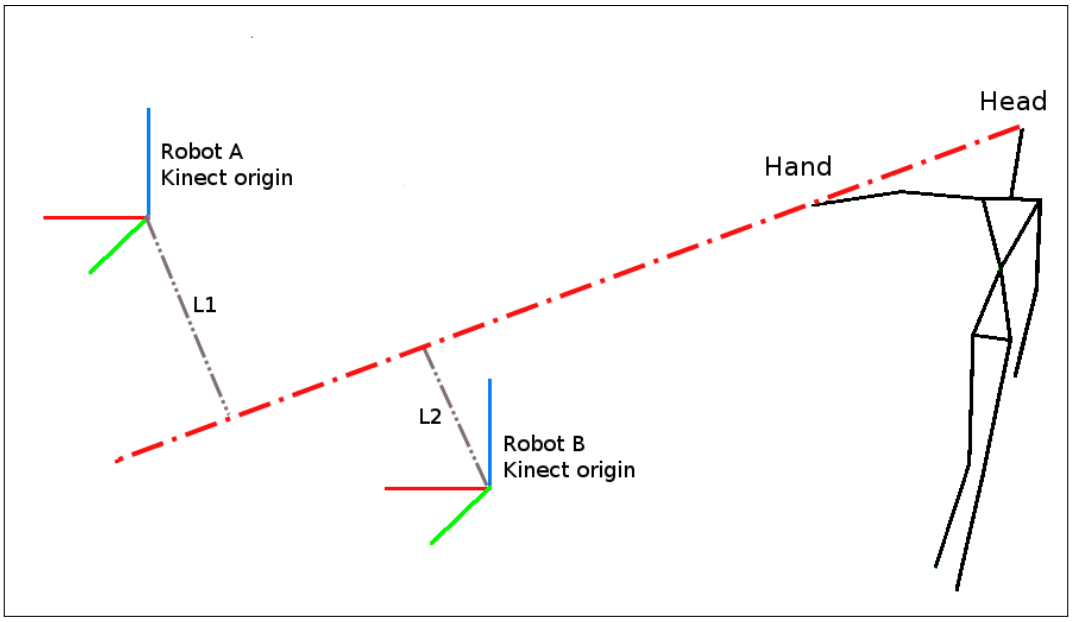


Figure 3.5: Disambiguating the reaching gesture for multiple robots (details in the text).

length of a normal to the reaching line through the origin of the Kinect sensor (shown in Figure 3.5). The robot with the smallest length value is the one the human intends to engage. Note this method does not require the robots to be localized or share a common reference frame, since each robot uses only the local appearance to score a gesture.

3.4 Demonstration and Discussion

We performed two different robot navigation scenarios as shown in the supplementary video: `https://youtu.be/WLwb7gJOl8w`. In the first scenario, one robot, and two human operators, one of whom starts holding a ball, are located in a $8 \times 10m$ room clear of obstacles. Each robot:

- first finds the users in the room, who are located at arbitrary locations
- attends to one user and approaches her, emitting a happy sound indicating readiness to interact
- receives a command, and executes it. The commands are either fetching or delivering the ball, or turning right or left to search for other users in the room.

After getting or giving the ball, robot starts searching for other users in the room. In our trials, the two users are instructed to execute the following interaction script:

1. user_1 sends the robot to the right by *pointing* to that direction.

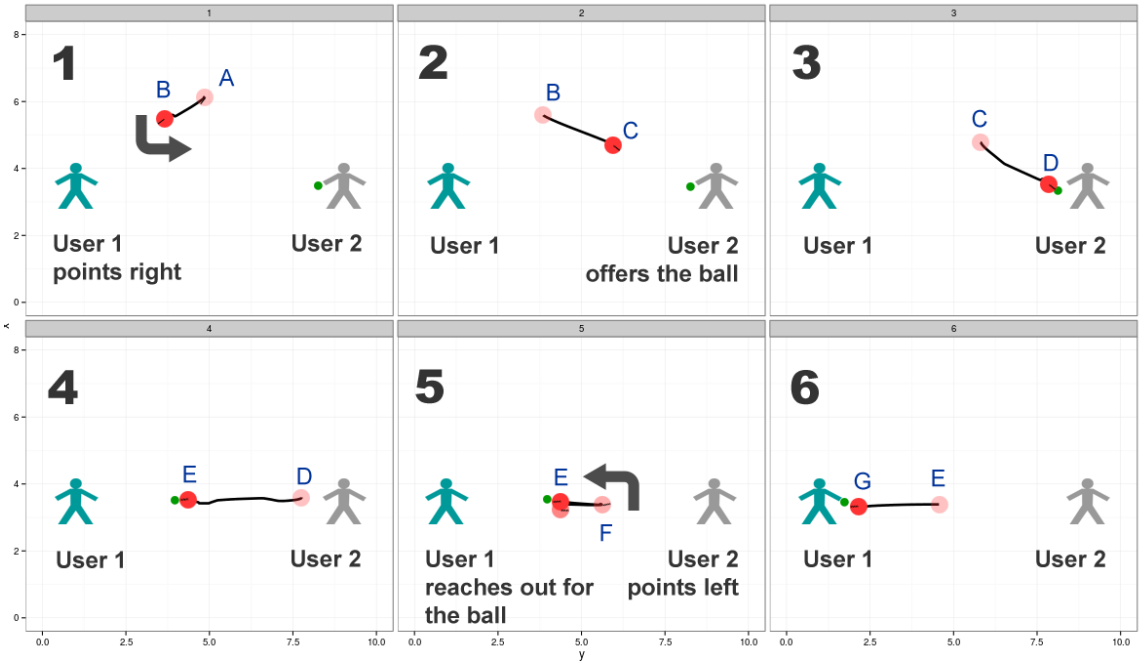


Figure 3.6: Robot and human behaviors during scenario 1, Trial #1, showing the interaction script performed perfectly: 1. Robot finds and approaches user$_1$. User$_1$ makes point-to-right gesture. 2. Robot turns right, finds and approaches user$_2$. User$_2$ makes reaching gesture. 3. Robot drives close to user$_2$ to receive the ball. 4. Robot finds and approaches user$_1$. 5. User$_1$ makes no gesture. Robot turns to find other users. Robot finds and approaches user$_2$. User$_2$ makes point-to-left gesture. Robot turns left, finds and approaches user$_1$. 6. User$_1$ makes reaching gesture. Robot goes closer to user$_1$ to deliver the ball.

2. user$_2$ has the ball, and gives it to the robot by offering it with the *reaching gesture*.

3. user$_1$ issues no commands (*no gesture*).

4. user$_2$ sends the robot left by *pointing* to the left.

5. user$_1$ requests the ball with the *reaching gesture*.

The users are instructed to attempt to recover from any failures, which happened on two occasions as described below.

An example trial where the script is executed perfectly is shown in Figure 3.6, based on data recorded from Trial #1. The robot trajectory is recorded using an overhead vision system not used in the robot control loop.

The results of 10 trials are presented in Table 3.1. The robot correctly detected users' legs on 48/50 opportunities (96% success rate). In Trial #5 the robot did not immediately detect user$_1$'s legs, so it targeted user$_2$ and approached him. Inspecting the video of the trial we see that user$_1$ was standing with legs tight together, causing the leg detector to fail. User$_2$ pointed to the right to help the robot find user$_1$, and the system subsequently

Table 3.1: Result of experiments with one robot

Trial No.	Leg Detection	User Detection	Point-to-Right Gesture Detection	Point-to-Left Gesture Detection	Reaching Gesture Detection	No Gesture
1, 2, 4, 7, 8, 10	30/30	30/30	Correct	Correct	12/12	Correct
3	5/5	5/5	Correct	Correct	1/2	Correct
5	4/5	5/5	Correct	Correct	2/2	Correct
6	5/5	5/5	Correct	Correct	1/2	Correct
9	4/5	5/5	Correct	Correct	2/2	Correct
Success Rate	96%	100%	100%	100%	90%	100%

executed the script without errors. RGB-D user detection worked in all 50 cases and point right/left gestures were detected in all 10 cases. The success rate of reaching gestures (offering and requesting the ball) was 18/20 (90%). All of the failures in gesture detection were occurred when skeleton could not be detected within the threshold time of 15 seconds, despite encouraging the user with sounds, so the robot began searching for other users.

In the second scenario, two robots interact with one user who starts holding a ball. The robots:

1. first find the user in the room, who is located at arbitrary location.
2. approach the user, emitting a happy sound indicating readiness to interact.
3. wait to be selected by the user. Once selected, a robot drives forward to collect the ball.

When both robots arrive at the user and indicate their attention by sound, the user chooses one robot by offering it the ball. An example (Trial #1) is shown in Figure 3.7, showing the robot trajectories and user behavior.

Ten trials, labeled 1 to 10, are performed with the robots initially located $3m$ apart. When they arrive at the user, waiting for a command, they are roughly $2m$ apart. Thus their selection spheres (see Section 3.3.5) barely overlap. The results of the trials are presented in Table 3.2, showing the success of the interaction step. Leg and body detections are omitted, since they worked in every case. In every trial except #7 the human skeleton was correctly observed and the correct robot was selected to collect the ball. In no case did both robots detect a reaching gesture intersecting their ego-sphere.

To test the ambiguity resolution mechanism, ten further trials, labeled 11^* to 20^*, are performed with robots initially placed as close together as possible. Their ego-spheres overlapped by around 50%. In nine of these trials, both robots observed the reaching gesture vector intersecting with their sphere, so the election algorithm was used to determine which robot was selected; in each case, the robot intended by the human was selected. In Trial $\#20^*$, robot R_1 could not acquire a skeleton track, so the other robot could not complete the election. In such cases, the system could decide to elect any robot that observes the gesture instead of failing, the option we chose.

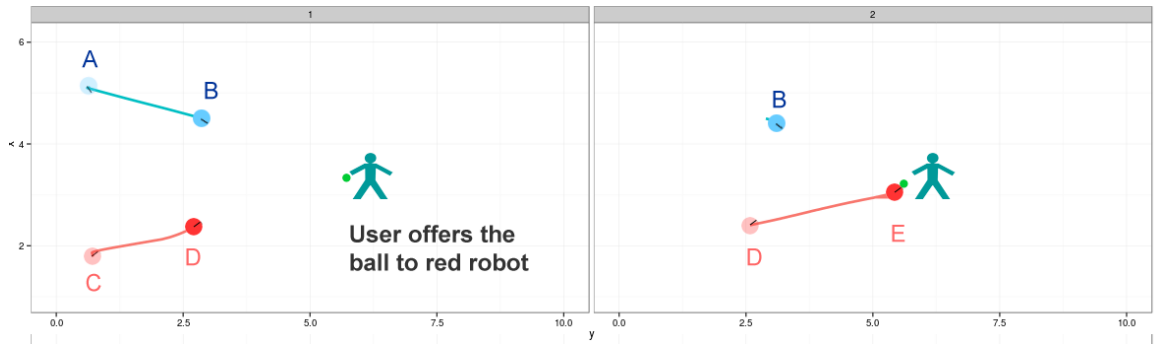


Figure 3.7: Robot and human behaviors during scenario 2, Trial #1: 1) Two robots find and approach the user. 2) The user selects red (lower) robot to receive the ball. Red robot goes closer to fetch the ball.

Table 3.2: Result of experiments with two robots

Trial No.	Robot	Skeleton Detection	Line-Sphere Intersection	Gesture Detection	Selection
1-6, 8-10	R_1	Success	Success	Success	Correct
	R_2	Success	Failure	-	
7	R_1	Success	Failure	-	-
	R_2	Failure	-	-	
$11^* - 19^*$	R_1	Success	Success	Success	Correct
	R_2	Success	Success	Success	
20^*	R_1	Failure	-	-	-
	R_2	Success	Success	Success	

3.5 Conclusion

We described a system whereby multiple humans and mobile robots interact robustly using a combination of sensing and signaling modalities. Extending one of our research group previous works on selecting an individual robot from a population by face engagement, we showed that reaching toward a robot - a specialization of pointing - can be used to designate a particular robot for subsequent one-on-one interaction. To achieve robust operation despite frequent sensing problems, the robots use three phases of human detection and tracking, and emit audio cues to solicit interaction and guide the behavior of the human. A series of real-world trials demonstrates the practicality of our approach.

A proper user-study with naive participants would be required to justify a formal claim that this system is "intuitive" or better than any other method. We do not make this claim but note informally that selecting and commanding a robot to take a ball by simply holding out the ball to the chosen robot feels fun and right, as does holding out your empty hand to the robot with the ball and having the robot come and drop it at your feet. The first few times you try it, you have to smile.

We used a very small set of discrete gestures. The gesture set could be extended to allow a user to point to *any* arbitrary place in the environment. This has been done for a single robot system (e.g. [125, 126]); however, an interesting extension would be to exploit multiple robots to jointly estimate the vector given the system's ability to capture images of the user from multiple angles simultaneously.

Chapter 4

"You two! Take off!": Creating, Modifying and Commanding Groups of Robots Using Face Engagement and Indirect Speech in Voice Commands

Figure 4.1: An uninstrumented person selects and commands multiple robots out of a group by looking at them and saying the desired number of robots.

We present a multimodal system for creating, modifying and commanding groups of robots. Extending the work by Couture Beil et al. [73] on selecting an individual robot from a population by face engagement, we show that we can dynamically create groups of a desired number of robots by speaking the number we desire, e.g. "*You three*", and looking at the robots we intend to form the teams. We evaluate two different methods of detecting which robots are intended by the user, and show that an iterated election performs well in our setting. We also show that teams can be modified by adding and removing individual robots: "*And you. Not you*". The success of the system is examined for different spatial configurations of robots with respect to the user and each other to find the proper workspace of selection methods.

4.1 Introduction

We have been working on methods for uninstrumented humans to select and command individual and groups of robots using simple and intuitive interaction systems. Inspired by the ways humans interact with each other or with animals, we have utilized face engagements, pointing gestures and spoken commands to interact with teams of robots.

Some researchers incorporate multiple modalities as communication inputs. Perzanowski et al. [65] present a multimodal speech and gesture-based interface to work with teams of cooperative robots; they employ the knowledge of spatial relations obtained by speech input along with gesture information from an active vision system to build context predicates. The work by Briggs et al. [66] combine spoken inputs and vision components to update the belief model of autonomous agents. Prasov [127] describes the role of shared gaze between a human and an individual robot during remote spoken collaboration.

In this chapter we propose and demonstrate a new interaction mode for multi-robot HRI: standing in front of a population of robots, a user can designate a subgroup of determined size by looking at them and saying "*You two!*" (or three, or N). The robots cooperate to combine their independent observations of the user's face and determine which robots were intended. Membership of the group can then be modified by adding a robot with

"*And you*" or removing one with "*Not you*". The team can then be commanded as a unit with e.g. "*Take off!*" (Figure 4.1). The user wears a Bluetooth earpiece microphone but is otherwise uninstrumented. In a series of real-world experiments, we show that the method works reliably for a broad range of relative poses of user and three robots.

The contributions of this work are (i) to propose a new interaction modality using indirect speech ("*You two!*"); (ii) to show that human-robot face engagement can be used to determine the subject or subjects of verbal commands using indirect speech. To do this we introduce two methods for selecting groups by face engagement. We also provide the first analysis of the reliability of selection by face engagement as the spatial arrangement of user and robots varies.

4.2 Method

We assume that before assigning a task to a team of robots, the human operator must select some robots from the population to form the team. We seek to design HRI systems that make this easy. We believe that a good approach is for uninstrumented humans to interact with teams of autonomous robots as they would with teams of humans since this is familiar. This motivates our choice of face engagement and spoken commands.

In our system, each robot runs a face detector and communicates with a centralized voice recognition subsystem and database to coordinate their activity. Interfacing and message passing is enabled by ROS [128]. The system layout is summarized in Figure 4.2.

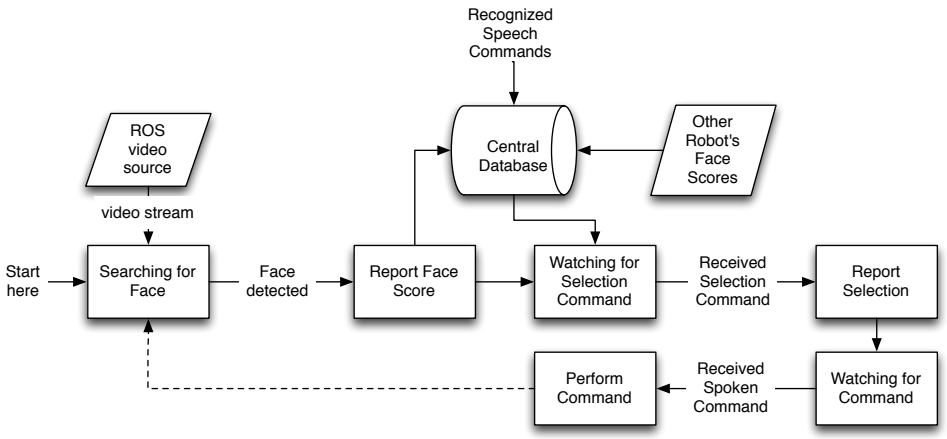


Figure 4.2: System diagram: the system runs on each robot.

4.2.1 Face Detection and Tracking

The first step of robot selection is to detect and track the human face. Each robot is equipped with a video camera, and faces are detected in each frame. Face detection is done

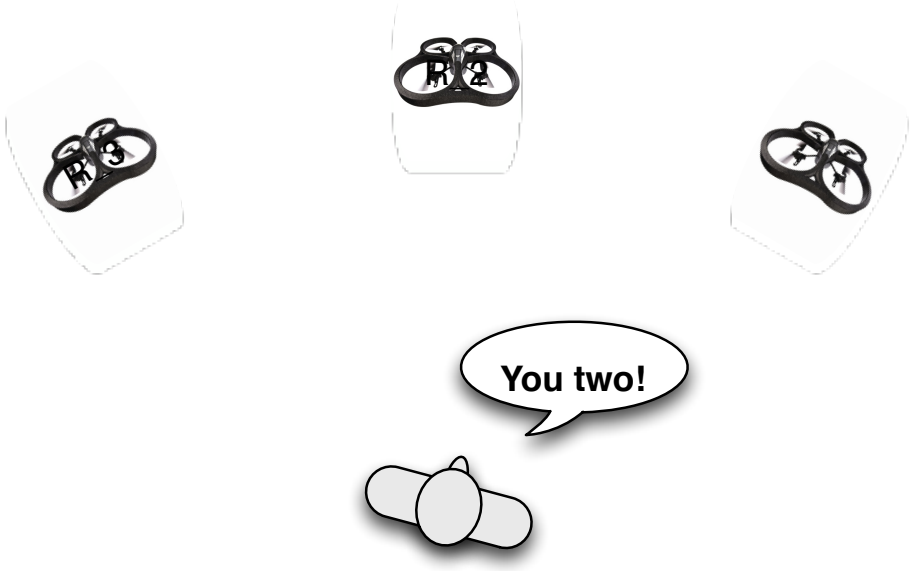


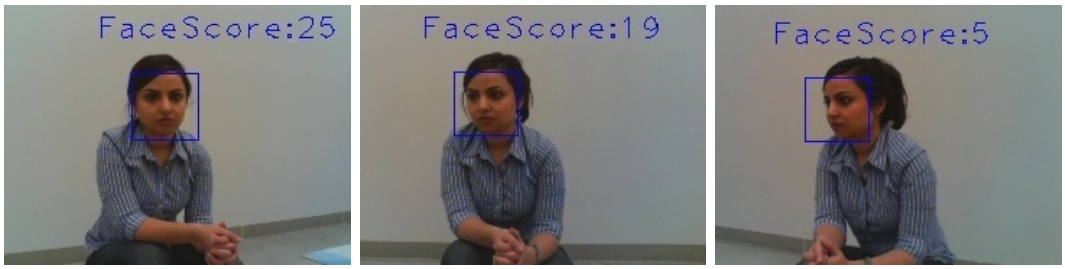
Figure 4.3: A human operator creates a team of robots by looking at them and uttering the desired number of robots.

using the OpenCV [129] implementation of the Viola-Jones face detector [130]. Once a face is detected, a Kalman Filter is used to track the detected face to achieve robustness to occasional false negative and false positive detections.

The robots use the face detector to understand if they are currently being looked at by the human. When the human face is visible to multiple robots at the same time we use a mechanism developed and successfully used earlier by our group [73] to determine which robot is currently being looked at by the user. The face detector, a cascade Haar classifier, finds a group of adjacent sub-windows around each candidate face. Since, the classifier is trained on the frontal faces only, the number of such sub-windows increases when the human is directly looking at the camera (Figure 4.4). We use this number as a score to assess the gaze direction of the currently tracked face. In the next section, we will describe how our system uses this so-called "face score" to determine which robot is currently being engaged by the user.

4.2.2 Voice Recognition

We employed the PocketSphinx library [131] to do speech recognition. PocketSphinx is an open source speech recognition system which matches voice commands with a predefined vocabulary. The vocabulary we used is defined by the words and phrases necessary for our system. It is a very small vocabulary which makes speech recognition very accurate in practice but requires the human operator to learn the set of allowed words and phrases.



(a) Right (b) Middle (c) Left

Figure 4.4: An example of three robots' simultaneous camera views while arranged around a human operator. The user intends to engage the right-hand robot (view a) and it has the highest face score.

4.2.3 Robot Selection

Our interaction design calls for the user to announce the desired number of robots (e.g. "*You two*") and look at them (Figure 4.3). When the keyword "*You*" is detected, all robots currently tracking the user's face announce their ID and face score to the central database (and display their state by changing the LED colors to the user). In the experimental section below we describe two variations of our basic election method where we define a team of size N as the either the N robots with the highest simultaneous face scores, or the best single face score, iterated N times, with the winner of each round not participating in subsequent rounds. In either case, the intention is that the N robots that are most attended to by the user form the group.

4.2.4 Combining Group and Individual Engagement

We add the keywords "*And you*" and "*Not you*" to add and remove individual robots, currently face-engaged, to and from the team. This allows us to create teams of adjacent robots and add individual distant robots afterward. It also makes it possible to recover from incorrect group allocations: if the wrong robot was added to the team we could remove it ("*Not you*") then add our preferred robot ("*And you*"). Once a group is correctly assembled, we can control the team as a unit. In the video demonstration[1] we show a team of UAVs being created, modified, and commanded to "*Take off*".

4.3 Experimental Results

To demonstrate and validate the system, we performed several experiments with different spatial configurations of robots with respect to the user and other robots (Figure 4.5). Since we are using face engagement to attract the robot's attention, the spatial arrangement of the workspace is important. Experiments are designed to find the spatial arrangements

[1] `https://youtu.be/I8sJud-OApw`

of the user and robots that work for our system. For convenience in these experiments we used laptops with integrated webcams to stand in for the robots; however, our video demonstration[2] shows the system working with three low-cost UAV robots.

The human operator uses a small lexicon for announcing the desired number of robots (e.g. "*You three*"), modifying the group (adding a new member (e.g. "*And you*") or removing (e.g. "*Not you*")), getting robot's attention or re-grouping (e.g. "*again*" or "*robots*") and commanding the selected group (e.g. "*Take off*").

As shown in Figure 4.5, for each experiment robots are located l meters from the user with θ degrees of separation and the user at the centre. l ranges from **1** to **2.5**m with **0.5**m steps and θ ranges from **15** to **90** degrees with **15** degree steps. In each configuration, the user attempts to select a single or multiple robots as required by the trial. Each experiment is repeated five times, so that overall 198 experiments were performed.

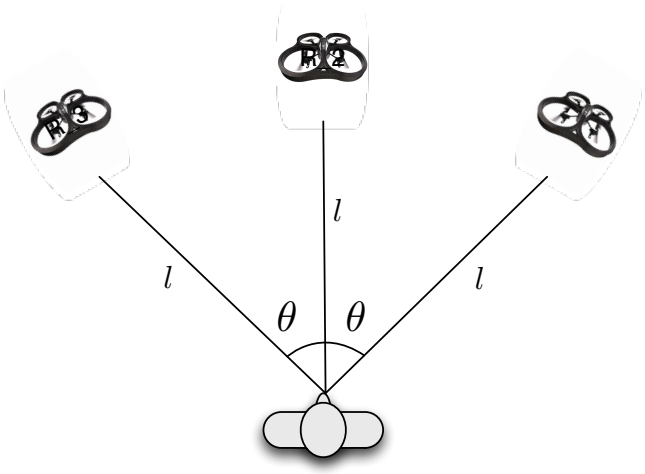


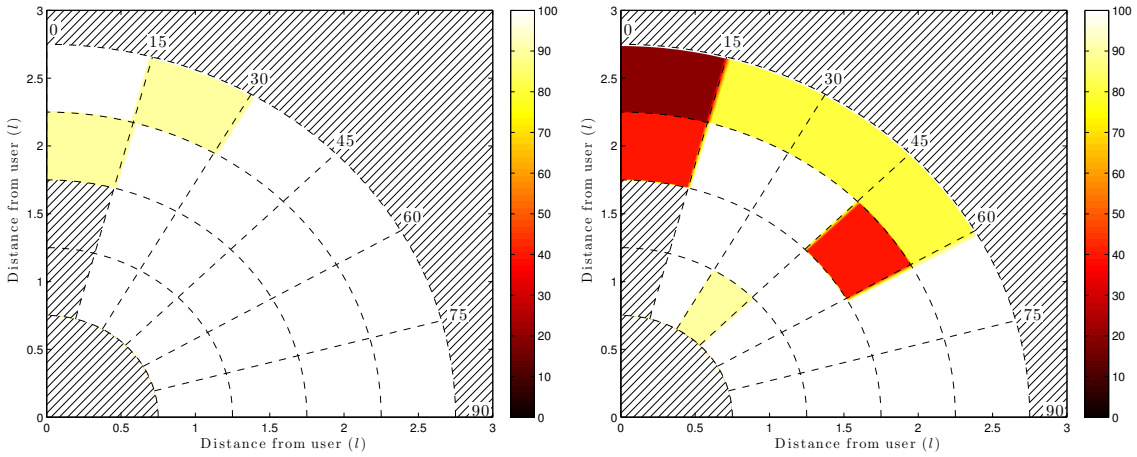
Figure 4.5: The configuration of robots with respect to the user and each other.

The results are shown in Figures 4.6 - 4.8. The graphs show the success rate of selecting the desired robots located at l meters from the user and θ degrees from other robots. Due to the symmetry between the left and right-hand robot cases, right and left results are combined. The results are presented as a heat map, where a white color indicates 100% success rate and black color 0% success rate. No experiments were performed in the hatched area, as it was either too close or too far for the face detection to work, or there was not room to fit three robots in that space.

4.3.1 Single Robot Selection

In this set of experiments we varied the robots' spatial configuration and tried to select one of them. In every trial, the human operator directly looked towards the desired robot and

[2] `https://youtu.be/I8sJud-OApw`



(a) Selection of one of the side robots

(b) Selection of the middle robot

Figure 4.6: Success rate of selecting an individual robot

uttered *"you"*. To isolate the spatial effects, for this experiment we assume that the voice recognition module works perfectly. The results are shown in Figure 4.6.

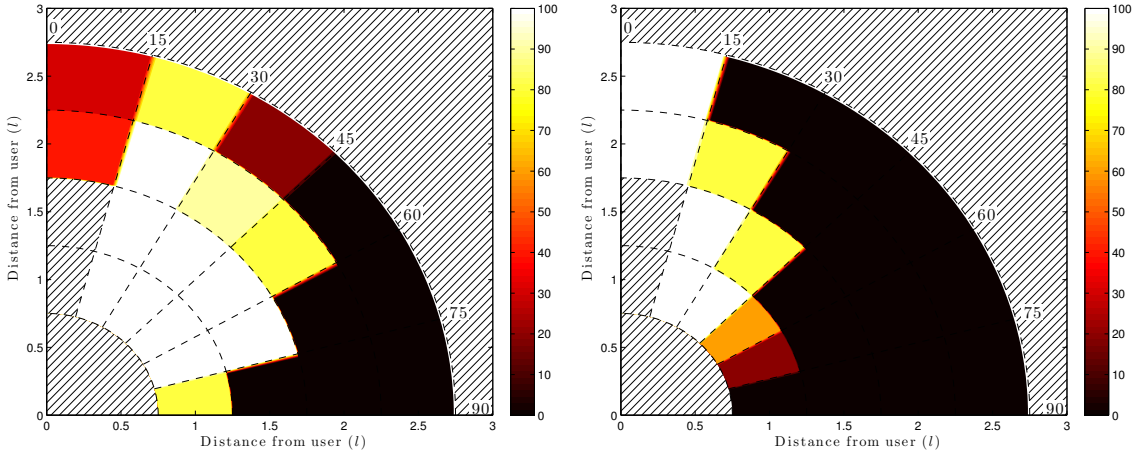
Figure 4.6-a shows the average success rate over five repeats of selecting either of the robots in the left or right side of the user and Figure 4.6-b shows the same measure for selecting the middle robot (i.e. the one in front of the human). The results indicate that when robots are very close to each other or very far from the human operator the success rate of selecting the desired robot decreases. This is due to their face scores becoming similar, so one robot is selected effectively at random. The failure rate is higher when the user tries to choose the middle robot because there are two sources of error (selection of the right or left robot instead of the middle robot).

4.3.2 Multi-Robot Selection

To select a subgroup of robots from a population, we investigate two different ways of making face engagement with multiple robots. One is by looking at the whole group and trying to make face engagement with all of them simultaneously. The other is to select the desired robots one by one. We repeat the experiment above, this time selecting two or three neighboring robots from our population of three. Again we vary the spatial layout to find how sensitive the method is to the spatial layout of the workspace. Results are shown in Figure 4.7 and Figure 4.8.

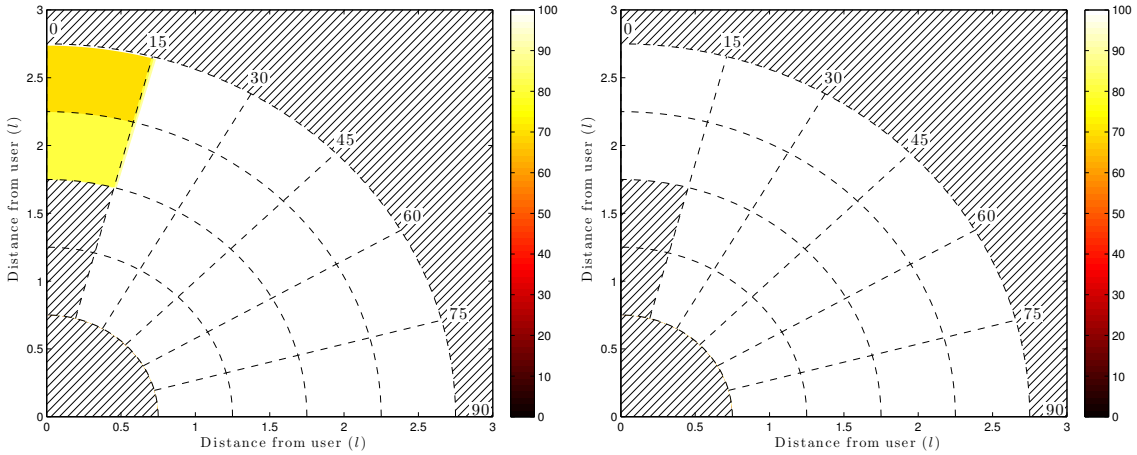
Method 1: Simultaneous Selection

To select a group of robots, one instinctively looks toward them. We call this method *simultaneous selection*. In this mode, the human operator looks toward the whole desired



(a) Selection of two robots (b) Selection of three robots

Figure 4.7: Success rate of *simultaneous* selection of multiple robots



(a) Selection of two robots (b) Selection of three robots

Figure 4.8: Success rate of *incremental* selection of multiple robots.

group of robots and tries to make face engagement with all of them at the same time. The number of robots the user has asked for, will be selected simultaneously by selecting the N robots with the highest face scores.

In this method, our observed success rate varied strongly with the human-robot distance and the angle between robots. Figure 4.7-b shows that when the robots stand less than 15 degrees apart, and all three robots can completely see the user's frontal face, the success rate is very high. As the angle between robots increases, the workspace is limited to shorter distances since the user cannot have face engagement with all of them at the same time. So the spatial workspace of this method is limited.

As in the first experiment selecting one-from-three robots, when selecting two-from-three robots with this method the proximity of the third robot can cause an incorrect selection. Figure 4.7-a shows that with the robots located very close together (θ less than 15 degrees apart and thus with very similar face scores) only a 40% success was observed in this situation.

Method 2: Incremental Selection

To improve on the first method, we devised a second method in which face engagement is iterated over the set of desired robots. We name this method *incremental selection*. In this method, after announcing the desired number of robots (e.g. "*You N*"), the robots with the highest face scores will get selected one after each other in N rounds, with the winner of each round not participating in later rounds.

By incrementally selecting robots, the user can group robots located far from each other because she can have face engagement with each of the desired robots separately. The success rate of incremental selection of multiple robots is illustrated in Figure 4.8. The results indicate that this method of multi-robot selection has wider spatial workspace: it is robust to a broader set of mutual poses. Since the human operator can look individually at all the desired robots for selecting them, their configurations have less effect on the success rate. The only source of failure we saw using this method is when robots are posed very close together, which is in common with the single-robot selection mode. According to Figure 4.8-b we can conclude that the workspace of selecting all three robots incrementally is the whole area in which the face detector works.

4.4 Conclusion

We have described a system which integrates spoken commands and face-engagement to create, modify, and command teams of robots. We introduced two modes of selecting multiple robots and compared them, concluding that iterated election is much more robust to spatial layout compared to the simultaneous selection. This is because in iterated election the user can look around from robot to robot in the team rather than having to look at their center of mass. In Chapter 5 we extend this work to designate teams of robot by name, so we can say "*You three are Red Team*", "*You three join Blue Team*", and "*You switch to Green Team*".

In all this work, we aim for simple, robust methods that are intuitive and easy to use. The data shows that the method is robust as long as the face detector is working and the robots are not too close to each other, but we suggest that the video shows what the data can not: the simple and natural feel of our interaction design.

Chapter 5

"You are Green": A Touch-to-Name Interaction in an Integrated Multimodal Multi-Robot HRI System

Extending our previous work on dynamically creating groups of robots using face engagement and voice commands, we show that we can identify an individual or a group of robots using haptic stimuli, and name them using a voice command (e.g. "You two are Green" or "You two, join Green"). Subsequent commands can be addressed to the same robot(s) by name. We demonstrate this as part of a real-world integrated system in which a user commands teams of autonomous robots in a coordinated exploration task [19].

5.1 The Touch-to-Name Interaction

We have developed a new multimodal interface for human multi-robot systems (HMRS) whereby a user can name individual or teams of autonomous robots from a cooperating population. In the *Touch-to-Name* interaction the user verbally announces the desired number of robots N in the form *"You N"*, (e.g. *"You two"*). If $N = 1$, the number can optionally be omitted, as in Figure 5.1 (left). The user then physically handles the desired set of robots one after the other (Figure 5.1 (middle)). Once the robots are thus selected, the user names the individual or group with a second verbal announcement, of the form *"You are <NAME>"* (e.g. *"You are Green"* as in Figure 5.1 (right)) to create a team or *"Join <NAME>"* or *"Leave <NAME>"* to modify a team. The user can thus create a direct addressing scheme to previously anonymous robots and can command a team of arbitrary size with a single subsequent command (e.g. *"Green! Take off!"*). The team manipulation interactions can happen at any time.

5.1.1 Implementation

In our demonstration, each robot is equipped with a 3-axis accelerometer, a voice recognition system and a wireless communication channel to compare sensor information with its peers. For simplicity, we used a centralized voice recognition system and a Bluetooth microphone worn by the user during our experiments. But our method will also work with audio processing onboard the robots.

To determine which robot is selected by the user, all robots wait for the keyword *"You"*. When it is detected, the robots communicate over the wireless channel to compare their accelerometer readings with each other. The one with the highest acceleration magnitude in a recent time window is considered the one being touched by the user. Figure 5.2 shows the acceleration magnitude of two robots. Robot_1 is the one gently moved by the user and Robot_2 is untouched. This election mechanism avoids the use of a predefined acceleration threshold by assuming that the robot that is being touched by the user has the highest recent acceleration readings.

The robots indicate their current state to the user with bright colored LEDs. We found that compared to vision-based selection methods [21, 74], using accelerometer data is much

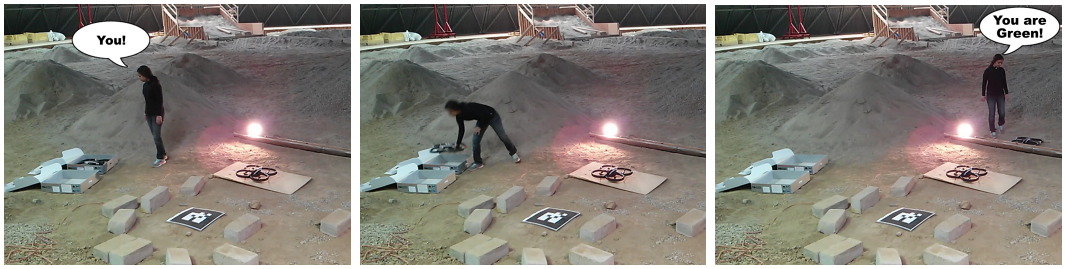


Figure 5.1: The Touch-to-Name interaction: The user first announces the desired number of robots with "You" or "You N" where N is the desired number of robots (left), then handles the intended robot(s) (middle); and finally assigns a name to the selected robot(s) that can subsequently be used to address this robot or team.

faster and computationally less expensive. It requires the human operator to touch the robot platform, but this interaction is simple and straightforward to implement.

Using voice input gives the user the ability to control teams of robots hands-free. Draper et al. [132] showed that using voice commands can significantly improve an operator's ability to manage teams of UAVs, compared to manual controls. While our demonstration shows the human handling each robot to provide the haptic input, the same effect can be achieved by pushing or kicking the robot, nudging a joystick that drives the robot, or moving an actuator that has sensor feedback.

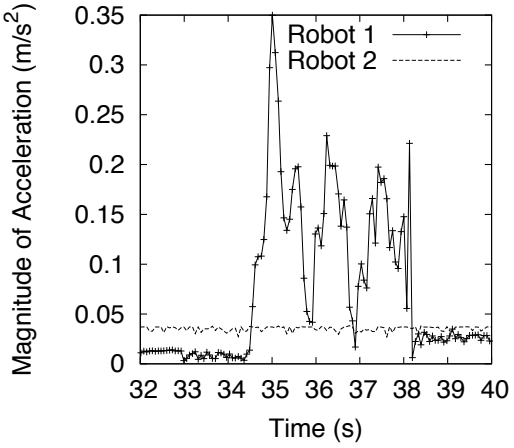


Figure 5.2: Accelerometer readings of two robots during the selection procedure. Robot$_1$ is selected and Robot$_2$ is untouched.

Once the user selects the desired robot(s), she can assign a team name using a different voice command (i.e. *"You are Green"* to create a group, or *"Join Green"* or *"Leave Green"* to modify a team. The user can thus create a direct addressing scheme to previously anonymous robots and can command a team of arbitrary size with a single subsequent command. The team manipulation interactions can happen at any time.

This multimodal HMRS interface allows the user to walk freely among the robots. These interfaces are well suited for systems that require human's supervision for dynamic task allocation, team composition and team re-composition.

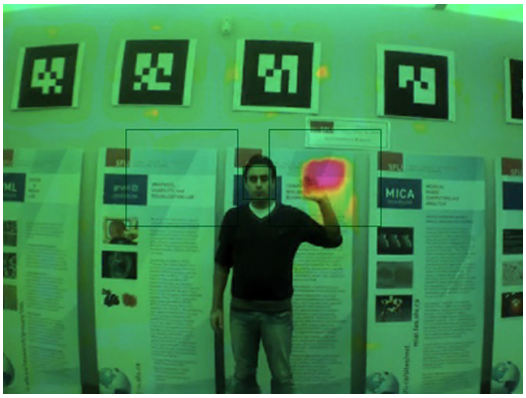


Figure 5.3: Face detection and motion-based hand gestures are used to relocate and command the flying robot [74].

5.2 An Integrated HRI Scenario

We demonstrated the Touch-to-Name interface as part of an integrated system [19] to allow a single user to control a multi-robot system in a coordinated exploration mission in a semi-realistic setting (the Mars Dome facility at the University of Toronto, a 50m diameter circular, covered arena simulating Martian terrain (Figure 5.1). The system can be seen working in the video `http://youtu.be/SxeVZdJFB4s`.

The task of each robot is to autonomously explore and map the world using visual SLAM from its onboard video camera. The robots used were two AR.Drone 2 UAVs, each controlled over WiFi by a laptop. Robots were autonomous at all times. All the user interactions could have been performed wearing a spacesuit, and the user had no instrumentation apart from an earpiece microphone.

In addition to our Touch-to-Name interaction, we added the verbal commands *"Take off"*, *"Come back"*, and *"Land"*, which can be issued to any named robot or team, e.g. *"Green! Take off."*. Robots would verbally confirm their state over a loudspeaker using a voice synthesizer. After take-off and while hovering, each robot looks for a human face in its camera feed. When a face is detected, the robot flies to meet the user at a pre-set distance and angle. The user can move to position the robot as desired. When the robot is suitably positioned, the user tasks it to explore either to the left or right of the user by waving the left or right hand [74]. Informally, we found that (i) all spoken commands were correctly recognized; (ii) the intended robot was always correctly identified; (ii) feedback from LEDs and synthetic voice was helpful to the user. Overall, the proposed concept worked well for this scenario, and we expect it to be widely applicable.

Chapter 6

On the Scalability of Spatially Embedded Human Multi-Robot Interfaces

In this chapter, we focus on assessing embodied, sensor-mediated interfaces for human multi-robot systems (HMRS) while examining how the mechanics of composing teams for concurrent control will affect interface efficiency. We introduce two distinct methods of creating and commanding groups of robots which we name: *Sequential Selection Concurrent Commanding* and *Concurrent Selection Concurrent Commanding*. The amount of time a human requires to interact with multiple robots (*interaction time*) is used as a measure of interface efficiency. We also propose, and demonstrate using experimental evidence, that in real world settings, the interaction time is affected by the spatial configuration of robots with respect to each other and the user.

6.1 Introduction

A key enabler to allowing a human operator to control a large population of robots is the ability to select and/or command multiple robots in parallel. Multiple robots can be selected and identified as a team, and the whole group is controlled with a single interaction. Parasuraman et al. [133] compare individual robot selection and group selection interfaces in a computer game scenario. They show that comparing command of individual robots versus command of groups, the participants won significantly more games without a statistically significant difference in workload.

Another study by Micire et al. [134] shows that for tasks including two robots participants use significantly fewer group selections than individual selections; in tasks with three or more robots, they use considerably more group selections than individual selections: at some critical population size, users appear to switch from favoring individual to group control to reduce their workload. The robot/human ratio is affected by the robots' degree of autonomy, the complexity of the task environment, interface design, human skill level, and workspace constraints.

Olsen and Wood [14] introduced the concept of *Fan-out* which posits a model-based upper-bound on the number of independent homogeneous unmanned vehicles (UVs) that a single human can interact with. This model has been modified to include wait times [135]. Goodrich et al. [136] extended it to the domain of heterogeneous robots and added switching cost. There have been several studies [133, 137–139] analyzing how the human-robot interface and team size can affect system effectiveness and performance, mostly evaluated in simulation with traditional human-computer interfaces without taking into account the challenges presented by real-world settings.

However, spatially embedded, sensor-mediated systems have major scalability constraints. In every case, there will be a practical upper bound of the robot/human ratio due to workspace and sensor limitations. For example, using computer vision the distance and angle of incidence inside which a robot can detect a person is limited by the chosen lens,

camera resolution, algorithms, and a finite number of robots can be physically located inside this space.

Following Olsen and Wood [14], we use the amount of time it takes for a human to interact with a single or a group of robots to indicate the human-robot interface efficiency. The smaller this time, the better. We decompose interaction time into three components: (i) the amount of time required to **select** a single or multiple robots, i.e. to acquire their attention and get them ready to receive control commands; (ii) the time required for the user to **switch** her attention from one robot or group to the next robot or group; and (iii) and the time needed to **command** a robot or team, including acknowledgement from the robot(s). This simple model can be mapped onto most interaction designs and captures enough information to classify models regarding their scalability. We introduce below a taxonomy of HMRS interfaces, distinguished by the growth rate of the individual time components with robot population size. We then classify systems from the literature in these terms and show empirical evidence that the interaction time grows as predicted by the model. Of immediate practical interest is the result that selecting and commanding groups of robots concurrently can be shown to decrease the total per-robot interaction time in our examples, though the degree of concurrency is strictly limited by the available workspace.

To this end, we will discuss two distinct methods of creating groups of robots which we name: *Sequential Selection* and *Concurrent Selection*. These are directly analogous to the familiar CTRL-click and SHIFT-click group selection methods used in desktop computer graphical user interfaces (GUIs). To evaluate these methods, Mizobuchi and Yasumura [140] compared tapping with circling for multi-target selection, regarding accuracy, execution time and shape complexity. In circling, the targets must be surrounded to be selected. In tapping each target must be clicked on to become part of the chosen group. They showed that circling is faster than tapping for highly cohesive targets and it is relatively insensitive to changes in the size of the individual targets. However, tapping selection time is significantly affected by the size and spacing of the targets.

6.2 Interaction Time

Olsen et al. [14] introduced Fan-out, (F)[1] as a measure of the number of robots a human operator can control. Fan-out is defined as the ratio of activity time (A), the time a robot operates autonomously, and interaction time (T), the expected amount of time that a human must interact with one or a group of robots.

$$F = \frac{A}{T} \tag{6.1}$$

[1]For descriptive purposes, we have modified the variables in the Fan-out equation.

Activity time (A) is a function of the robot's degree of autonomy faced with a particular task complexity, while interaction time (T) is proposed as an essential metric for human-robot interaction efficiency [137], where shorter communications are more efficient than longer ones.

In multi-robot systems, interaction time (T) can be decomposed into three components: i) robot monitoring and selection time (L); ii) switching time (W); and iii) command expression time (C) [141]:

$$T(n) = \sum_{i=1}^{n} (W_i + L_i + C_i) \tag{6.2}$$

where:

- $T(n)$ is the amount of time the operator needs to interact with n robots,
- L_i is the amount of time required to select robot$_i$,
- W_i is the amount of time required to switch to robot$_i$,
- C_i is the amount of time required to issue a command to robot$_i$.

We include time taken to provide feedback to the user in the respective terms.

In real-world settings, with embodied interfaces, W_i, L_i and C_i are functions of the communication method, physical workspace, spatial arrangement of the user and robots, and the amount of time needed by the robot to analyze the input signal. Also, it may be required to repeat the selection and command phases as necessary to compensate for sensing or processing failures in the robot. Therefore a human operator can have substantially different interaction times with individual robots, or the same robot at various times. However, under the assumptions of homogeneous robots and identical relative user-robot positions, Equation (6.2) can be simplified to:

$$T(n) = n \times (W + L + C) \tag{6.3}$$

This interaction mode is *Sequential Selection Sequential Commanding (SSSC)*. Interaction time scales linearly with robot population size, assuming robots are on average evenly distributed in the environment.

We can use *SSSC* and Equation (6.3) as the baseline for the interaction time of an HMRS. Intuitively, interaction designs that add concurrency should scale better than this baseline. In the following sections, we explain how adding the ability to form and command team of robots can achieve this.

6.3 Concurrent Commanding

In *SSSC* the user can control one robot at a time. However, by dynamically forming groups of robots, the operator is able to command all selected robots at once [21, 74, 134, 142, 143]. This reduces the amount of time needed to command n robots from linear time (nC) to constant time (C). As a result, the overall interaction time should decrease. To evaluate how team make-up will improve system efficiency, we will examine two methods of selecting robots using embodied interfaces and study three proposed interaction systems for HMRS.

A Taxonomy

We distinguish 4 classes of interaction that differ in their scalability:

- *SSSC*: Sequential Selection Sequential Commanding
- *SSCC*: Sequential Selection Concurrent Commanding
- *CSCC*: Concurrent Selection Concurrent Commanding
- *CSSC*: Concurrent Selection Sequential Commanding

We found examples of *SSSC*, *SSCC* and *CSCC* classes in the literature, but *CSSC* seems unused, so we omit it. *SSSC* has already been introduced in Section 6.2, so we describe *SSCC* and *CSCC* below.

6.3.1 Sequential Selection Concurrent Commanding

In computer user interfaces, typically there are two methods for selecting multiple files or folders using keyboard and mouse. One way is to hold down the CTRL key and then click each desired item. In this method, the user can select a non-consecutive group of files or folders. Similarly, in HMRS, the user can sequentially select the desired robot and add it to the group. Once the team is formed, the operator can issue a command to the robots to perform a common task. Since commanding is simultaneous for all selected robots, the time required is just C. We call this method *Sequential Selection Concurrent Commanding (SSCC)*. The interaction time is:

$$T_{sscc}(n) = n \times (W_{sscc} + L_{sscc}) + C_{sscc} \tag{6.4}$$

Here, we assume that there is no failure in receiving selection or commanding signals by the robots. In the literature, we identified two interfaces for team make-up using this selection method. Monajjemi et al. [74] used face engagement and gestures to add/remove robots to/from a group. Similarly, in our previous work, where we extended single-robot selection by face engagement [73], we proposed a system [21] which integrates spoken commands and face engagement to dynamically create and modify teams of robots.

For both of these interface designs, C_{ss} is the same for interacting with one robot or with a team. As a result, due to simultaneous commanding, the model predicts a reduction in T_{ss} compared to the baseline.

6.3.2 Concurrent Selection Concurrent Commanding

The second method for selecting a consecutive group of multiple folders or files in computer interfaces is to drag the mouse pointer to create a selection rectangle around the outside of all the items to be included. The selection result is identical to clicking the first item, holding down the Shift key and then clicking the last item. In their work, Milligan et al. [142], Skubic et al. [143] and Micire et al. [134] applied *Concurrent Selection Concurrent Commanding* where the user can make a group selection by circling around the desired group of robots. We call this method *Concurrent Selection Concurrent Commanding (CSCC)*. The interaction time of this interface is composed of the time to switch to a group of desired robots, draw a circle around them and issue a command:

$$T_{cscc}(n) = W_{cscc} + L_{cscc} + C_{cscc} \tag{6.5}$$

Similar to *SSCC*, the operator issues only one command to the determined group. Therefore the predicted interaction time of this interface is smaller than of the baseline. In principle, the selection time L_{cscc} is a function of the number of robots to be selected. To select more robots, the user has to draw a larger circle. However, for most settings, this increase is relatively insignificant, and we, therefore, assume L_{cscc} to be independent of the number of robots.

6.4 Spatial Constraints for Real World HMRS

The preceding discussion has introduced a model of the interaction time with spatially embedded interfaces for HMRS without considering the challenges offered by real robots. Real robots are embodied and have to share their physical space with co-located human operator(s) and other robots; they are also situated, and their abilities to deal with the world are limited by sensors and actuators. We propose that the amount of time a user spends to interact with a group of robots is affected by the spatial configuration of the robots with respect to each other and the user.

There are scenarios where a user cannot communicate with an individual or a group of robots without changing her location, she has to turn or walk through the workspace to be able to interact with a new group of robots and may need to spend some time to switch attention to the robots in the new location. These switching times are significantly different from the ones in Equations (6.4) and (6.5). In this case, we split the workspace into sub-spaces that the user can stand still in and interact with all robots in that sub-space.

Formally, we divide the whole workspace into M sub-spaces. The maximum number of robots that fits in sub-space j is denoted by N_j. Therefore, Equations (6.4) and (6.5) are valid for $1 \leq n \leq N_j$. By summing over all the interaction times and switching times, we can calculate the general interaction time as:

$$T_G = \sum_{j=1}^{M} (T_j^* + W_j^*) \tag{6.6}$$

where:

- T_j^* is the amount of time needed to interact with a group of robots in sub-space j ,
- W_j^* is the amount of time the user needs to change locations to start interacting with a group of robots in sub-space j .

6.5 Experimental Results

We suggest that creating robot teams can improve interaction time and system efficiency compared to a per-robot baseline,

To assess this, we reviewed the available video footage of experimental trials from 3 different HMRS interfaces and measured the components of the interaction time (T) for various robot population sizes. The amount of time needed to switch attention to a robot, select it and receive its feedback, is considered as (W+L). (C) is measured from the moment the user starts expressing the command (by gesture or speech) until the moment she gets feedback from the robots. (T) is the sum of the terms mentioned as in Equations (6.4) and (6.5).

In Experiments A, B, and C, we compare interaction time of different group selection methods with our sequential baseline *SSSC* (Equation (6.3)) and examine how an instance of concurrent commanding affects the interaction time (T). For every method, we measure the interaction time with one robot (T_1), and multiply it by the number of robots in the team n, i.e. the baseline for interacting with n robots will be nT_1. According to Equation (6.3), in interacting with n robots, the human operator will spend n switching and selection times ($W + L$) to select each robot and add it to the group. Also, one command time (C) is required for issuing a command to the newly created group. In Experiment D, the interaction time of an HMRS interface is measured and evaluated in various spatial configurations of the robots with respect to the user and each other.

6.5.1 Experiment A: Face Engagement and Indirect Speech Interface for HMRS

Using the multimodal interface proposed in our previous work [21] (see Chapter 4), we reviewed video footage of experimental trials and measured the components of the interaction

time (T). In this system, we integrate spoken commands and face engagement to create, modify, and command teams of robots. To isolate the group selection method's effects on interaction time, we assume that all interface modules work correctly.

Figure 6.1b shows the average time spent to interact with multiple robots over six trials and compares it with the baseline. The robots were located 2.5m from the user with 30 degrees of separation and the user at the center (Figure 6.1a). As expected, the interaction time (T) increases as the number of robots in the team increases. A statistically significant difference (paired-sample t-test: $df = 5$, $p = 0.0012$ for two robots and $p = 0.0021$ for three robots) exists between the time needed to interact with multiple robots individually as in the baseline case and the interaction time (T) of this *Sequential Selection Concurrent Commanding* method. The main difference comes from the fact that the operator issues the command once for selected group.

Table 6.1 shows the different components of interaction time (T). It can be seen that the command time (C) remains the same for different numbers of robots in line with our model in Equation (6.4). However, switching and selection times ($W + L$) do not increase linearly with the number of robots. This is caused by this particular interface design scheme. In this interface, switching and selection time ($W + L$) are composed of four components:

(a) Face engagement and indirect speech interface for HMRS [21]

(b) The graph compares the interaction time with the baseline ($d \approx 2.5m, \theta \approx 30deg.$) (Sample size = 6)

Figure 6.1: Experiment A

Table 6.1: (Experiment A) Components of interaction time. (Sample size = 6)

No. of Robots	$(W_{sscc} + L_{sscc})$ (Sec) Mean (SD)	C_{sscc} (Sec) Mean (SD)	T_{sscc} (Sec) Mean (SD)
1	5.44 (1.05)	1.82 (0.14)	7.26 (1.03)
2	7.31 (1.77)	1.88 (0.11)	9.21 (1.80)
3	12.77 (1.96)	1.90 (0.15)	14.67 (1.86)

i) time needed to express the spoken command, ii) processing the spoken command, iii) sequentially making the face engagement with all robots in the group and iv) iterative selection of the robot with highest "face score". Since in interacting with multiple robots, the user announces the desired number of robots once, the first two components are not affected by the number of robots. However, the last two elements of switching and selection time ($W + L$) depend on the number of robots being selected. These two components are also sensitive to the spatial arrangement of the robots and the user, which will be examined later in experiment D.

(a) Waving gesture interface for HMRS [74]

(b) The graph compares interaction time with the baseline. (Sample size = 5)

Figure 6.2: Experiment B

Table 6.2: (Experiment B) Components of interaction time. (Sample size = 5)

No. of Robots	$(W_{sscc} + L_{sscc})$ (Sec) Mean (SD)	C_{sscc} (Sec) Mean (SD)	T_{sscc} (Sec) Mean (SD)
1	5.10 (1.09)	4.35 (0.91)	9.45 (1.88)
2	13.54 (1.41)	4.38 (0.67)	17.92 (1.69)
3	21.70 (1.91)	4.64 (0.33)	26.34 (1.76)

6.5.2 Experiment B: Waving Gesture Interface for HMRS

In the second experiment, we also examine the effect of *Sequential Selection Concurrent Commanding (SSCC)* on the interaction time T and system efficiency with another spatially embedded interface for HMRS proposed previously by our group [74]. Using this interface, the user can create a multi-robot team from a group. The user starts the interaction with each robot, by standing in front of it and making a face engagement. When the robot confirms the engagement, the user adds it to the team by a right-hand wave gesture (Figure 6.2a). Waving both hands, he commands the entire group to execute a mission. Similar to the previous *SSCC* interface, and based on Equation (6.4), we expect to see

a reduction in the interaction time T. This is due to the time being saved by sending commands to the whole group at once. To show this reduction in interaction time, we again compare against the baseline.

The average time spent on interaction with multiple robots over five trials, along with the baseline, is illustrated in Figure 6.2b. The result of paired-sample t-test on this small sample size shows that there is no statistically significant difference ($df = 4$, $p = 0.3371$ for two robots and $p = 0.2615$ for three robots) between the individual and group selection. The reason is that the switching time W for changing the workspace is large and it can not be subsumed in the baseline. Table 6.2 shows that the command time C does not depend on the number of robots which agrees with our model in Equation (6.4). The switching and selection time ($W + L$) on the other hand depends on the number of robots since the user has to walk to each robot and perform a waving gesture. The effect of the spatial embeddedness is large in this case because the robots in this experiment are quadcopters which require substantial free space for safe operation. As a result, to model the interaction time of this interface, Equation (6.6) is more suitable.

6.5.3 Experiment C: Circling Gesture Interface for HMRS

In this experiment, we examined the effect of *Concurrent Selection Concurrent Commanding* on the interaction time T. As showed in Equation (6.5), we expect that using $CSCC$, the interaction time T will be reduced to one switching and selection time ($W + L$) and one command time C and it does not depend on the number of robots in the team. To analyze this hypothesis, another method of selecting groups of robots from a population proposed by Milligan et al. [142] was examined. In this interface, multiple robots can be selected and commanded concurrently to perform a task using a vision-based approach. To select robots, this method calls for the user to draw a circle around all robots she wants to select (Figure 6.3a). In this way, robots in the circle get selected and assigned to a common group. The robots can determine whether they are circled by the user by tracking the user's hand and face. After selection, a command is issued to the team using a pointing gesture.

The data from reviewing the video footage of this interface is presented in Table 6.3 and Figure 6.3b. The baseline is measured similar to the previous experiments. For this experiment, only one sample was available for different numbers of robots. Figure 6.3b shows the amount of time the user spends to select and command different size groups.

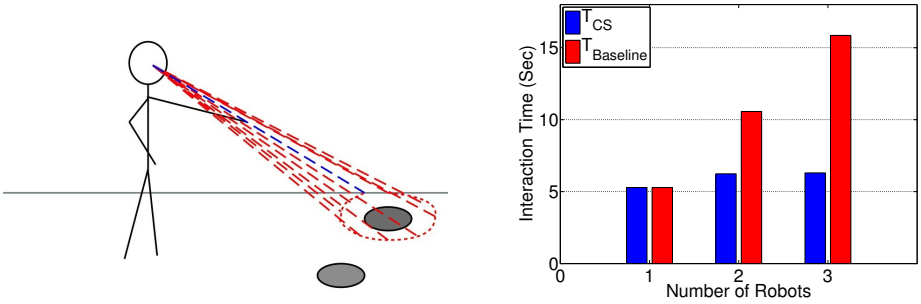
6.5.4 Experiment D: Spatial Configuration

We propose that the amount of time a user spends to interact with a group of robots is affected by the spatial configuration of the robots with respect to each other and the user.

To test this, we use the experimental results of the interface in Experiment A [21], where a user can select multiple robots by announcing the number of desired robots and

sequentially looking at them. The amount of time that the user spends to interact with three robots is measured in various spatial arrangements of robots with respect to the user and each other. In every experiment, robots are located d meters from the user on a circle with θ degrees of separation. d ranges from 1 to 2.5 meters with 0.5 meter steps while θ ranges from 30 to 90 degrees with steps of 15 degrees. Each experiment is repeated five times. The data are presented in Figure 6.4. The results show that the interaction time T is affected by the user-robot distance as well as the angle between the robots. When the distance increases, it takes more time for robots to get selected because the user's face is harder to detect. This is caused by the limitation of camera resolution. The increase in the interaction time T with the robots' separation is because the user has to switch between robots and directly look at them. When the robots are close to each other, the user's face is visible to all robots, so when the user switches between robots, there is no need for robots to start detecting the user's face. However, by increasing the angle between robots, the interaction time T increases because it takes more time for robots to detect the user's face and become selected.

We previously showed that this interaction method has inherent limits on the usable range and bearing between human and robot [21]. This experiment confirms that the spatial arrangement of robots also affects the interaction times.



(a) Circling gesture interface for HMRS [142]

(b) The graphs compares the interaction time with the baseline. (Sample size = 1).

Figure 6.3: Experiment C

Table 6.3: (Experiment C) Components of interaction time. (Sample size = 1)

No. of Robots	$(W_{cscc} + L_{cscc})$ (Sec)	C_{cscc} (Sec)	T_{cscc} (Sec)
1	1.82	3.47	5.29
2	2.59	3.65	6.24
3	2.58	3.72	6.29

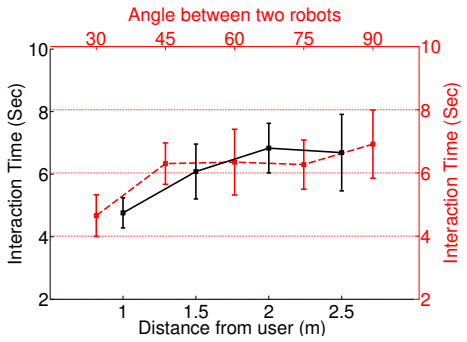


Figure 6.4: (Experiment D) Comparison of interaction times for various configurations of robots with respect to the user and each other. (Sample size = 20)

6.6 Conclusion

This chapter presented a model for calculating the interaction time of spatially embedded HRI designs. We propose that the amount of time required to interact with multiple robots can be reduced by different methods of creating groups of robots and doing concurrent interactions that exploit locality. A useful class of HMRS interfaces provide the ability to dynamically create, modify and command groups of robots, ideally scaling to large populations of robots and overcoming the time limitations a user has when interacting with large numbers of robots. We examined two different classes of group selection and commanding methods: *Sequential Selection Concurrent Commanding (SSCC)* and *Concurrent Selection Concurrent Commanding (CSCC)*.

Our analysis of previous experiments show that the *CSCC* methods exhibits a shorter interaction time compared to the *SSCC* method. This is due to the time reduction in concurrently selecting robots. However, *SSCC* has the advantage that the selection does not have to be consecutive. In *CSCC* method all robots to be selected have to be physically positioned close to each other and the cluster of robots has to be free of unwanted robots. This makes *CSCC* method good for homogeneous groups but limits its application in heterogeneous groups. We also propose, and demonstrate using experimental evidence, that in real world settings the interaction time is affected by the spatial arrangement of the workspace. This will affect the upper-bound of the number of robots that a single human can interact with. We also point out two types of switching cost: directing attention between robots in the local workspace, and moving to attend to another group of robots that were previously out of range.

Chapter 7

A Probabilistic Sensor Fusion Framework for Finding HRI Partners in a Crowd

In this chapter, we present a probabilistic framework for multimodal sensor fusion that allows a mobile robot to reliably locate and approach the most promising interaction partner among a group of people, in an uncontrolled environment. Our demonstration integrates three complementary sensor modalities, each of which detects features of nearby people. The output is an occupancy grid approximation of a probability density function over the locations of people that are actively seeking interaction with the robot. A series of real-world experiments demonstrates the robustness and practicality of the proposed system for controlling robot's attention.

7.1 Introduction

One long-term aim of autonomous robot research is to have robots work with and around people in their everyday environments, taking instructions via simple, intuitive human-robot interfaces. All else being equal, we would prefer that these interfaces require no special instrumentation of the humans and little or no training. In this chapter, we demonstrate such a system, shown in Figure 7.1 and Figure 7.2, whereby a self-contained autonomous robot can reliably detect and approach the person in a crowd that most wants to interact with it.

A prerequisite for a successful natural human-robot interaction is for each party to find an interested counterpart. In scenarios with multiple people, the robot must decide which human (if any) to interact with. We want the robot to be able to automatically recognize potentially interested humans present in its workspace and then evaluate the posture, gesture or other salient features of each person to determine their intent to interact.

While studies on attention control typically focus on close range human-robot distances ($<2m$ separation) [32, 144, 145], mostly on stationary robots [42, 44, 50, 55], our work looks at controlling a mobile robot's attention in distant ($>2m$ separation) multi-human robot interaction.

This is a challenging task. In addition to ordinary sensor noise, other people may be moving around the environment and occlude the subject; people walking by or performing other tasks will change their appearance to the robot's sensors; the robot's ego-motion changes the sensor readings at every sample; sensor false-positives may mislead the robot. We suggest that there is not a single sensor that can reliably serve.

We achieve robustness by employing an array of multimodal human detectors and probabilistically fusing their outputs. As a working example, but without loss of generality, we use a laser range finder to detect legs, an RGB camera to detect human torsos, and a microphone array to detect the direction of sound sources. All of these detectors have very different fields of view, detection ranges, and accuracies, while their different modalities allow them to cover each other's weaknesses. The laser, for example, gives us very precise range and bearing measurements, while the microphone array only provides rough

Figure 7.1: A live demonstration at HRI'15. The mobile robot is able to robustly track people and approach the person of interest, despite the noisy and crowded environment.

directional information. Our fusion method is not limited to these three modalities, but can easily incorporate additional detectors.

To choose sensors and the features they detect, we use our knowledge of simple regularities in human behavior. For example, among a group of bystanders, a person who is standing facing the robot and calling it will have the highest probability of being a potential interaction partner. We have observed this behavior combination is generated spontaneously in untrained human subjects [16]. We fuse two independent sources of body pose information with directional audio, placing greater weight on the audio as an actively-generated signal. No single modality is necessary, but we require two modes to agree to suppress false positives. This differs from previous work in active-speaker detection [7, 49, 146].

The contributions of this work are:

- designing a straightforward but effective method for sensor fusion of human detectors that selects the most engaging person to approach for further one-on-one interaction.
- demonstrating this method as part of an interaction system for controlling a robot's attention in distant multi-human robot interaction through a series of outdoor experiments.
- a ROS-based implementation, freely available online, using widely-available sensors.

7.2 Background

To increase the robustness of real-time human detection and tracking, many approaches integrate more than one source of sensory information such as visual and audio cues [7,147,148],

Figure 7.2: Real-world campus setting for experiment 7.4.3 with five uninstrumented users at arbitrary locations. One person, chosen at random, tries to get the robot's attention, and the robot reliably approaches him. The subjects then change their locations and switch their interaction role and random.

visual cues and range data [46, 149, 150] or vision-based and radio-frequency identification (RFID) data [151].

Associating multimodal information with detected humans allows the robot to selectively initiate the interaction with the person with higher interest. Lang et al. [56] proposed an attention system for the mobile robot BIRON, to estimate the position of the partner of interest and maintain attention during interaction. Each person in the robot's vicinity will be tracked as soon as his or her legs or face are recognized by the system. If a detected person starts talking, that person will be recognized as the person of interest and the robot shifts its attention by turning its camera toward them. If this person stops talking for more than 2 seconds it will lose the robot's attention to another talking person. However, in this system, people have to stand near the robot ($< 2m$) to be considered as a potential communication partner. Also the user must keep talking to maintain the robot's attention. Our system relates these constraints. In a similar approach Lin et al. [6] defined the interested user as the first person, detected by legs and face, appearing in the direction of a detected sound. If all these human features are found in the same direction, he or she will be considered as a potential interaction partner. If the robot doesn't receive validating information from all 3 sensors in 1 second it starts searching for another interested person.

Several authors have worked on enabling a robot to direct its attention to a specific person and/or estimating a user's level of interest in interaction with a robot. Some approaches use distance and spatial relationships as a basis for evaluating engagement. Michalowski et al. [32] and Nabe et al. [145] proposed an approach based on the spatial relationship between a robot and a person to classify the level of engagement. Finke et al. [34] used sonar range data to detect a target person at closer than one meter, based on motion. Muller et al. [152] and Bruce et al. [153] used trajectory information to classify people in the surrounding of

the robot as interested in interaction or not. However in some situations having humans approach the robot is infeasible or undesirable, and it is the robot's responsibility to arrive at the target person for one-on-one interaction.

Okuno et al. [49] developed an auditory and visual multiple-speaker tracking for an upper-torso humanoid robot. Some work has explored different methods to detect and track multiple speakers [146]. However, our experiment suggests that sound alone does not provide reliable performance in dynamic environments with lots of ambient noise. People can speak, shout or clap to get robot's attention, but by using sound only the robot can get attracted to irrelevant sound sources such as bystanders talking.

In most of these studies, the robot's attention is oriented to the target person by head turning, body turning or eye movements. The person of interest can also lose the robot attention when they stop talking. In this work, we consider a more general situation, where the robot and people are outdoors, mobile, surrounded by distracting people and sound sources, and are in arbitrary locations and poses.

7.3 System Design

7.3.1 Multimodal Human Feature Detection

We use a simple probabilistic sensor fusion approach that is easy to understand and implement (Figure 7.3). The idea of fusing multiple occupancy grids is not novel: Elfes' [154] introduction of the method showed multi-sensor fusion. We describe the efficacy of this approach for our HRI-partner-finding task. Here, "*occupancy*" is an estimate of the spatial probability density of finding a partner.

We use three sensors: (i) laser range finder to detect legs, (ii) camera to detect torsos, and (iii) microphone array to detect sound direction. These sensors have different trade-offs in the field of view, range, and accuracy. They also measure different properties of the user. For example, the leg detector gives accurate location data but is ambiguous about whether the person is facing toward or away from the robot. Sound, on the other hand, is something the user actively emits and is a strong signal for attention-getting, as when calling a dog. As we will explain below, we make explicit use of these differences.

Leg Detector

Finding legs in laser range data is a well-explored method for detecting humans. We employed Inscribe Angle Variance (IAV), proposed by Xavier and Pacheco [155] to find legs by analyzing their geometric characteristics, essentially looking for discontinuities with certain properties in the laser scan. This leg detector runs at $50Hz$ and provides highly accurate human location information in the robot's coordinate frame with a wide field of view of 270

degrees. A weakness of the leg detector is a high false positive rate. Unfortunately a lot of objects cause similar sensor readings, e.g. furniture, trees, bushes, trash cans.

Torso Detector

To detect torsos, we use a camera mounted facing forward at the front of the robot. Grayscale images from the camera are processed to obtain Histograms of Oriented Gradients (HOG) features [156]. These features are robustly classified using linear SVMs trained to detect human torsos. In our system, we use the OpenCV implementation [157] which provides fast multi-scale detections using an image pyramid and runs at $5Hz$ on CPU on our mobile-class onboard computer.

To estimate the location of humans, we first compute a bounding box around each torso detection. Given an expected human body size, we use the size and image location of the bounding box to estimate the position of a human in the robot coordinate frame. This detector works well at subject distances of up to $10m$. However, the accuracy is poor in cases of partial occlusions and large deviations of subject height from our median prior.

Directional Sound Detector

To detect directional sound we use the Kinect's microphone array. Audio signals are processed using Multiple Signal Classification (MUSIC) [158] to detect the direction of sound sources in the ground plane of the robot frame. We use an implementation of MUSIC from Kyoto University (HARK) [159]. In contrast to the other modalities, the sound detector only provides direction and no true range information for each sound source (since source intensity is unknown). As our goal is to rendezvous, we can use the direction information and rely on the sensor fusion (see below) to obtain position estimates.

Calling the robot by voice, whistle or clap, is a simple and intuitive interface that needs little or no instruction, so we select the loudest detection found by HARK above a certain threshold as an active attention signal. The weakness of sound as an interaction cue is frequent false positives caused by ambient sounds or even echoes. Our system encountered passing buses, talking passers-by and noisy construction equipment. Loud ambient sounds also cause false negatives as the loud signal overwhelms the sensor's ability to detect human voices. We also found that untrained users tend to call the robot occasionally rather than continuously. To reduce the sparsity of sound signals over time, we latch the most-recently-detected sound for two seconds (informally, we observed that this trick was very important for getting good responses to sparse audio).

7.3.2 Probabilistic Sensor Fusion Framework

Our proposed framework aims at "multiple-sensor multiple-target" tracking, where human percepts detected through various sensor modalities are all associated with the correct

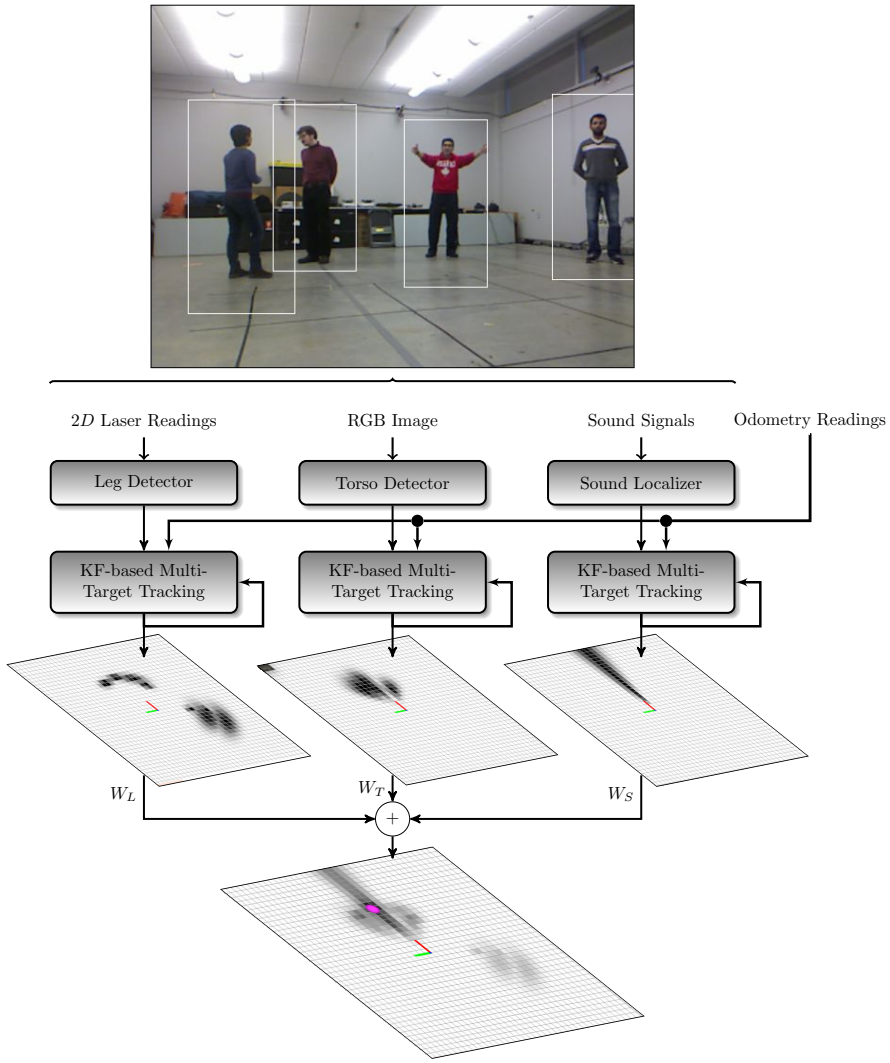


Figure 7.3: An overview of the system: Raw sensor data are filtered separately, then projected into evidence grids. Grids are then fused by weighted averaging into a single integrated grid. Grids show real-world data.

targets. Each of the detectors we have introduced independently tracks different human features and estimates their position relative to the robot frame of reference. The challenge is fusing this data in a way that captures the different characteristics of each sensor, while also being flexible enough to allow new sensors to be added or substituted.

We will address multiple-target tracking at the sensor level through filtering detections into a set of tracks. Afterward, we handle multiple-sensor fusion by converting each sensor's filtered output to a common probabilistic grid format and merging these grids in a weighted average. This adaptive approach allows us to add however many sensors and modalities we want, while also incorporating the characteristics of each sensor through calibrating properties of both the filters and the fusion.

Multiple Target Tracking

For each modality, we independently track detected humans using a bank of Kalman Filters (KFs), allowing us to associate evidence collected over time with a particular "track" feature. It also compensates for the robot's motion by incorporating wheel odometry information but does not model the movement of detected people.

New tracks are spawned when a detection is made beyond a certain threshold distance of any existing track. Otherwise, detections are associated with the nearest neighboring track. Those tracks that do not receive a measurement update, i.e. no associated detection was made, only have the prediction step of the filter performed - retaining the track but increasing uncertainty. Once a track's uncertainty exceeds a threshold, it is removed.

This filtering process provides some robustness against intermittent sensor readings. For example, occlusions, false negatives, and even inconsistent user stimuli (e.g. if a person temporarily stops sending active signals). As each sensor modality has its own filter, it can also have its own thresholds for track association distance and the uncertainty at which it is removed, allowing new modalities to have filters adapted to their particular sensor characteristics.

Informally, we found that filtering the raw sensor data for each modality independently before fusion was much more successful than fusing raw sensor output and tracking the combined signal.

Probabilistic Grids

The middle-step that allows us to fuse the results from different sensor modalities is converting the output of the Kalman Filters into probabilistic evidence grids. These grids are similar to occupancy grids [154] but instead of holding the probability of an obstacle, we store the probability that an attentive subject is at each location.

For this, we compute a location probability distribution for each tracked human feature using a modality-specific sensor model. In our implementation, leg detections are modeled

with a normal distribution. For torso detection, we use a multi-variant normal distribution to reflect the fact that range estimates are not very reliable. Sound detections are modeled using a cone along the measured direction vector. The cone length is limited to $10m$. This is a simple model of the likely distribution of ranges of a user who is calling the robot. The probability distribution for each modality is then discretized into a separate evidence grid.

Sensor Fusion

To compute the integrated probability distribution for all detected humans, a fused evidence grid is calculated as the weighted average of corresponding grid cells from all other grids. Each modality-specific grid is centered over the robot, ensuring that detections from the same human will overlap. Example grids are shown in Figure 7.3. The integration weights for each modality are assigned based on sensor characteristics and uncertainties.

We have some a priori reasoning in our implementation for choosing the *relative* weights: since sound is actively generated it may be more likely to indicate interest, while legs and torsos are possessed by interested and uninterested people alike. Hence, we assigned the highest weight to the (S)ound evidence grid. In our experience, the (T)orso detector exhibits fewer false positives than the (L)eg detector, so we assigned a higher weight to the torso grid than the leg grid. This results in an implicit ordering from most-reliable combinations to least-reliable combinations of [TLS], [TS], [LS], [TL].

This means for example that if two people are calling out, and both have their legs detected, but only one has a visible torso, we prefer the person with visible torso since that person is probably facing the robot and is thus directing her attention to it.

7.3.3 Attention Control and behavior Design

The integrated evidence grid can now be used to generate the robot's behavior. Several methods could be considered, but we chose a very simple and explicable approach to demonstrate the efficacy of the sensor fusion: we simply find the highest probability in the evidence grid and servo the robot towards that location. As the robot moves the evidence grid is continuously updated and the robot corrects the approach vector. This enables the user to move and be followed by the robot and it gives the robot an opportunity to recover from false sensor readings. Once the robot has approached the human to within 2 meters the robot stops. To give the impression that it is ready for a close range interaction, it plays a happy sound. If the person does not respond, the robot gives up, plays a sad sound and turns away looking for another person.

If all probabilities in the evidence grid are below a given threshold its detections are considered unreliable. In this case, the robot turns to sweep its sensors over the environment to find humans. We define detections made by only one sensor modality as unreliable, e.g. leg detections without a torso detection are often caused by furniture and not by people.

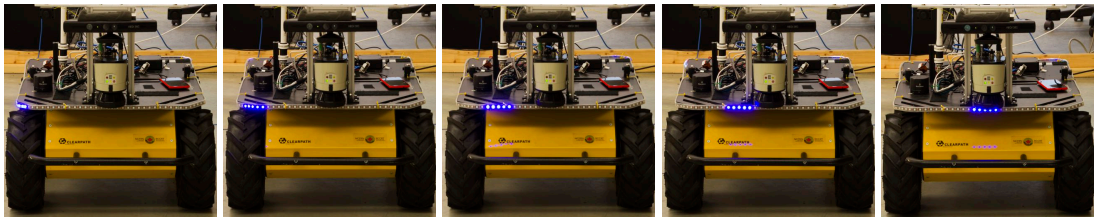


Figure 7.4: An array of coloured LED lights is wrapped horizontally around the robot body to display the direction of the robot's desired heading.

The user and the robot form a tight interaction loop that appears similar to that between a dog and its owner. By observing the robot, the user can deduce if the robot is paying attention to her (approaching) or not. If the robot is not paying attention the user can simply provide more stimuli, e.g. call louder or orient more towards the robot.

7.3.4 Feedback System

While approaching the potential interaction partner, the robot uses an array of programmable coloured LED lights to communicate information about its internal state and actions. The LED strip is wrapped horizontally around the robot body (Figure 7.4). The direction of the robot's desired heading at each moment is projected on the LED strip. During the approach state all but the five pixels signalling that heading are turned off. Once the robot stops in front of the human waiting for close-range interaction, all pixels will be turned green and red. If the robot is too close to an obstacle and enters the emergency stop the LED strip goes red.

7.4 Experimental Results

We implemented the designed system on a typical outdoor mobile robot, Husky by Clearpath Robotics. The robot is equipped with a Kinect providing the RGB Camera and a 4 channel microphone array, and a 2D SICK laser scanner. The sensors have different but overlapping fields of view. Legs can be detected in a 270 degree arc up to a distance of 10 meters. The camera has a 57 degree horizontal FOV and is capable of detecting human torsos at distances up to 8 meters. The microphone array has a detection zone of 180 degrees in front of the robot but only reports bearing and not range.

Three experiments were performed to validate our probabilistic sensor fusion framework's ability to select an interaction partner. In all three experiments, the robot is co-located with a group of people including one who wants to initiate an interaction (the *interactor*). This person will stand facing the robot and occasionally call for it verbally.

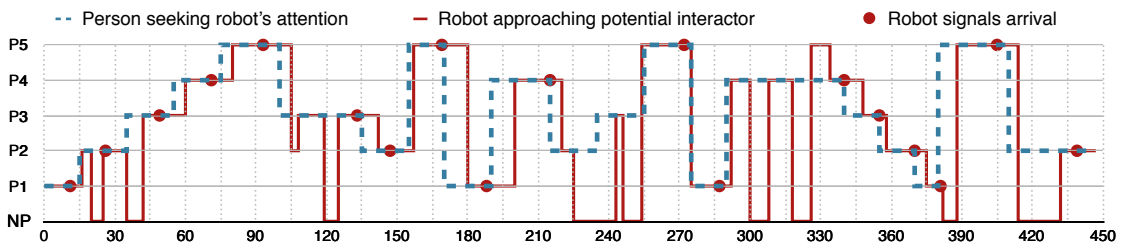


Figure 7.5: Diagram of the robot's responses to rapidly switching the interactive role between five people (P1-P5) at random. The blue dashed line marks the time line of which subject is seeking attention, the red solid line shows which person the robot is paying attention to and the red dots indicate when the robot entered close range interaction state. The robot attends to the correct person 18 out of 20 times.

7.4.1 Experiment A: Framework Only

The first experiment is designed to test the reliability of the sensor fusion by selecting the most promising person seeking robot's attention in an artificial setting. It is not our intended HRI scenario, but rather an exhaustive test of the system's functionality by exposing it to a wide range of possible detections at once. The robot's objective is to pick the interactor from a group of 8 research assistants $7m$ away. Subjects are positioned outdoors in a semi-circle with a $7m$ radius around the robot and approximately $2m$ apart from each other.

We systematically set up distractions by positioning people in a way that each shows a different subset of attractive features. For example, we ask some to cover their legs, some to stay quiet, and some to stand outside the camera/torso-detector field of view. Only the preferred interactor presents the full set of legs, torso and occasional sound to the robot.

The robot is given a 10 second time window to determine the location of the interactor. Actually approaching the selected human for interaction is omitted here in order to focus on the reliability of the attention system.

We call a selection successful if the robot "favours" the interactor during this period. We define favour to mean that the detected interactor position is closest to the true position of the interactor for longer than it is closer to any of the distractors.

Users take turns taking the role of interactor and varying their appearance to the robot according to a predefined script ensuring all permutations were tested. The robot correctly identified the right person on 21 out of 24 trials (87.5%), giving 99% confidence this approach improves on selecting one detected person at random. Failures occurred when ambient sound was coming from the same direction as a distractor person, whose legs and torso were detected (our test location had loud intermittent construction noise in the background).

7.4.2 Experiment B: Testing Discrimination at Range

We placed two research assistants outdoors at a distance of $7m$ in front of the robot and varied the distance between them in order to test how well the sensor fusion system could

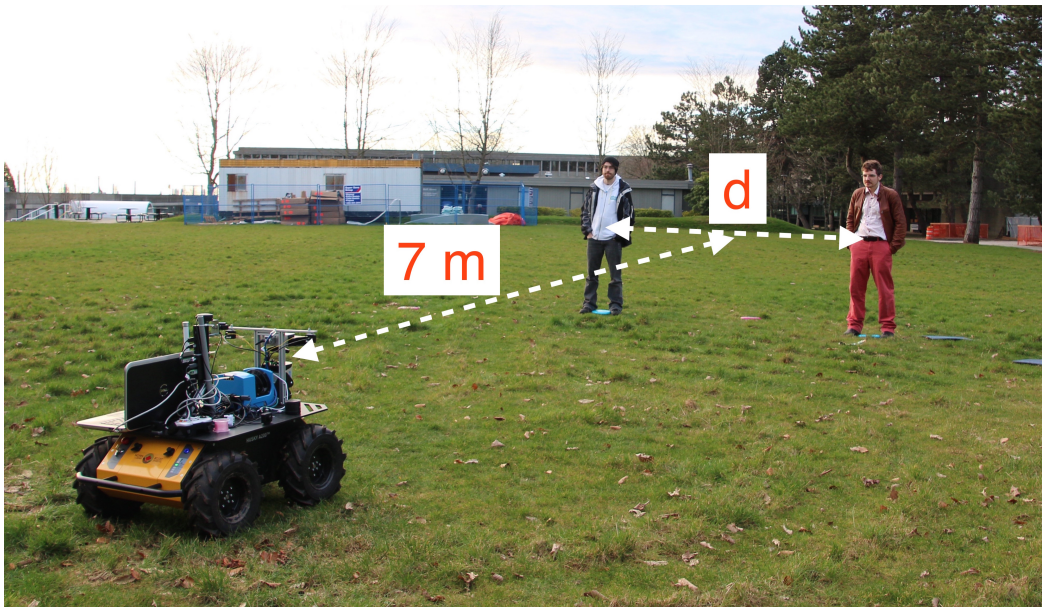


Figure 7.6: Experiment 7.4.2: Two people stand 7 meters away in front of the robot. One person calls the robot. We empirically determine the minimum distance d between the people at which the robot can no longer distinguish the attentive user from the bystander.

discriminate between adjacent humans. The robot was now allowed to approach a detected interactor to also examine interference from motion and changing distances. We measured the success rate and time required for the robot to reach the correct target, where a trial was successful if the robot was facing the correct person when it stopped. Results of 65 trials (5 repeats for each distance) are presented in Figure 7.7.

In trials where the people are standing very close to each other ($<1.5m$), the system has difficulty distinguishing the individual humans. This is mainly due to the relatively large uncertainty in the sound source direction detection. In these cases, the robot approached the centre between the 2 people. For strictness, we declared these outcomes as failures, but for most practical purposes the correct person is now within close interaction range.

At each distance there were some cases where the robot was distracted by the non-interactor participant but recovered when the interactor kept calling the robot. The further apart the two humans were, the more off-course this "wavering" could pull the robot and thus the higher average arrival time and variance. At $12m$ or more, the participants were at the extreme range of our sensors, making initial detection difficult and sometimes drawing the robot too far away to recover.

7.4.3 Experiment C: Playing Tag with Five People

In the third experiment, we examined the robustness and responsiveness of the system in a dynamic environment where the role of interactor would switch over the course of

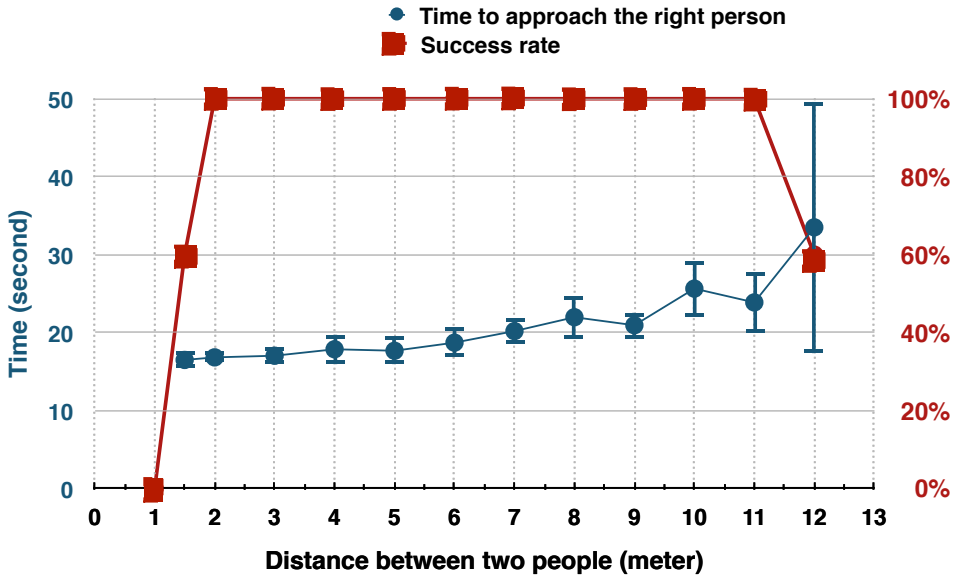


Figure 7.7: Experiment 7.4.2: Success rate and approach time in relation to distance between subjects.

one continuous test. We instructed 5 research assistants to stand in arbitrary positions surrounding the robot (see Figure 7.2).

One person was selected at random to be the first interactor. The interactor stands still and calls the robot in a normal voice, while the other research assistants walk around the vicinity of the robot as bystanders. The robot approaches the strongest fused detection, and when the robot stops directly in front of its chosen person it plays a "happy sound" to indicate its readiness to engage in a one-on-one interaction. If this person is the interactor, she moves away and chooses a new interactor at random. If she is not, she ignores the robot, which times-out and returns to scanning for new interactors. A section of this experiment is shown in the demonstration video: `youtu.be/PLB6OQWqaUs`

The timeline of interactions is shown in Figure 7.5, plotting the time when each of five people (P1-P5) were in the interactor role, and the time when the robot was focused on them or on no-person (NP), and the moment (dots) when the robot correctly announced it was ready for a one-on-one.

In seven and a half minutes, the robot engaged in 20 interactions. In 18 cases, the robot successfully found the interactor and correctly announced its arrival. However, we observed that in two cases between 220 and 260 seconds, the robot would find the target for a short time, but become distracted by another person and did not find the correct interactor.

7.5 Discussion

This work considers human-robot interaction over relatively large distances compared to almost all the literature (>2m separation). We noticed but have not yet exploited the way people interact with the robot varies over the course of an interaction as their mutual distance changes. We expect that the robot's behavior and sensing should also change with mutual distance.

The experiments used to evaluate our system focused on raw performance but do not address improvements to the human-robot interaction experience. In a separate user-study, we recorded other performance and user preference data that will be presented in next chapter.

Failure cases from each experiment suggests some possible improvements. Differentiating between human-sourced audio like words and clapping versus environment noise might usefully improve the reliability of sound as a detection method. Speech recognition might also be useful in distinguishing between active encouragement and discouragement signals.

Our Husky robot platform is not well suited for indoor social environments where its size, appearance and movement are awkward. We showed that the same system works indoors and out, but our indoor work will use a telepresence robot form factor in future.

Environmental factors can also modify human behavior, where the intensity of the interaction signals may increase with the intensity of the social setting - a loud party might cause users to call loudly, or to prefer gestures to calls in a library setting. Adapting sensor fusion parameters to the current setting could be a useful extension.

7.6 Conclusion

We proposed a system which integrates detected human features from multiple modalities for a mobile robot to choose the most likely person interested in a close interaction in a robot-multi-human scenario. Our probabilistic sensor fusion framework combined passive and active stimuli to successfully direct the robot's attention. A series of real-world experiments in outdoor uncontrolled environments and a live demo at HRI '15 in a crowd of hundreds of people all demonstrate the practicality of our approach. Our ROS implementation is freely available online[1].

[1] `https://goo.gl/TFV8y5`

Chapter 8

Finding an Interaction Partner in a Crowd: A User Study

Robots navigating crowded, uncontrolled environments may have to choose the most promising interaction partner among a group of people. This requires a system for measuring interest, identifying attention-seeking signals from an interested party, and approaching them. One solution could be manually controlling the robot and driving it toward someone for close-range one-on-one interaction, however, the user would have to carry a controller and be trained to use it. We propose an alternative autonomous interaction system, which allows any human to start an interaction with a ground mobile robot without instrumentation. For robustness, the system integrates three multimodal human percepts and has the robot to approach the point with the highest probability of an interested human, communicating its current state and intention with audio and light feedback. This chapter presents a study conducted with our functional and publicly available system and a pool of inexperienced users to evaluate its performance both objectively and as perceived by participants. Also, we report on the observation from the users' natural behavior when asked to "make the robot come to you". The study compares our system to 1) a teleoperated alternative and 2) an ideal autonomous robot responsive to all human attention-seeking signals. Both qualitative and quantitative results show our system is preferred by 82% of the users and performs as well or better than its competitor in all measured areas.

8.1 Introduction

As robots enter into more work, social and domestic settings, there is a growing challenge in matching a robot with an interested interaction partner without requiring a human to provide manual control. Whether it is a cocktail party robot looking to deliver drinks to guests or a taxi robot recognizing a potential passenger, autonomously choosing which person to approach would represent a significant step forward for social human-robot interaction.

To achieve this, a robot must be able to autonomously track potential interaction partners from their appearance or behavior. Our objective is to design simple, intuitive and robust interaction systems for autonomous mobile robots to interact directly with the untrained general public in their everyday life. No special environment, instrumentation or movement restrictions should apply, making such systems portable and applicable.

Designing such a system poses a number of sensory perception challenges. In addition to ordinary noise, passing bystanders may occlude the target and subjects engaged in other tasks may be difficult to recognize. The robot's ego-motion also changes the sensor readings at every sample, and false-positive detections may mislead the robot and should not leave it waiting for an interaction forever. No one sensor may be sufficiently reliable.

In this work, we will evaluate our previously-introduced interaction system (presented in Chapter 7) for selecting and approaching the most likely candidate from a crowd. Existing literature on attention control in HRI focuses heavily on stationary sensors and robots or

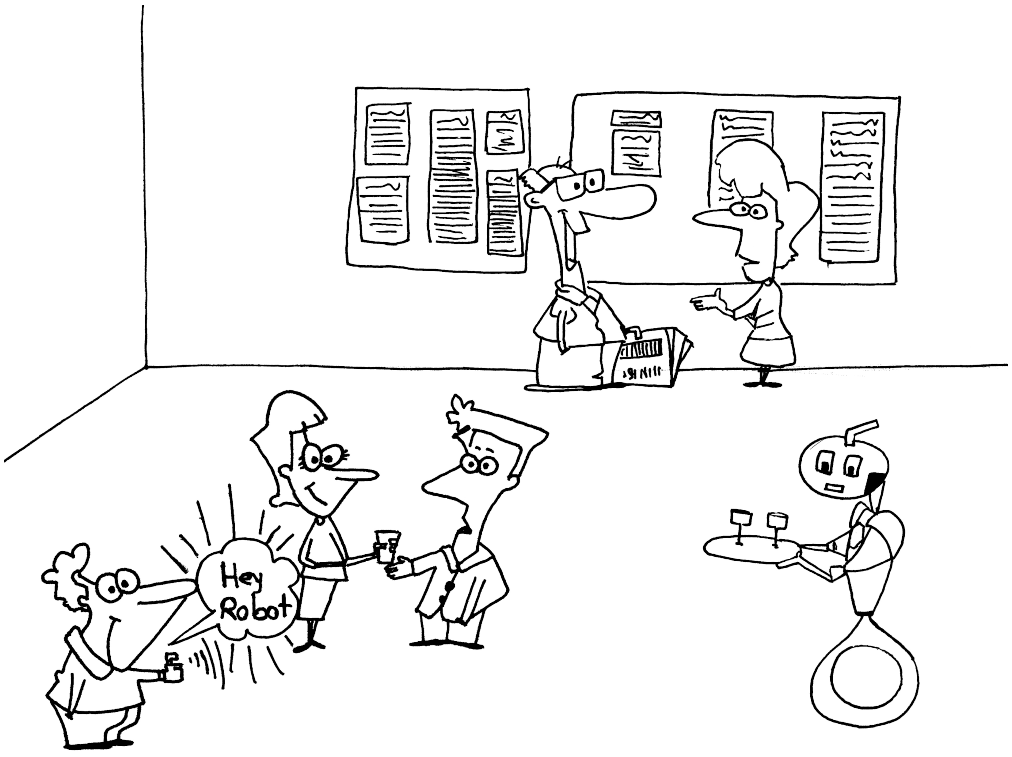


Figure 8.1: A possible application scenario.

at close ranges. By contrast, our multimodal interaction system aims at medium-range (< 10m) multi-human scenarios in semi-real world settings.

The current trend of demonstrative, lab-based methods do not capture all of the challenges of real environments. The robot and people may be outdoors, mobile, surrounded by distracting bystanders and sound sources and in arbitrary locations and poses, meaning that the most interested party might not always be in view or could change over time. People may change their behavior in response to the robot's actions or as their engagement changes over the course of interaction.

Therefore, our proposed system is one where the robot can adaptively change its focus, approaching and initiating a close interaction with the person with the current highest apparent interest. This method also permits robots to convey their current state and intended actions using audio and light feedback. A user study is undertaken to evaluate the system's performance with the untrained general public.

These experiments test our proposed system against manual teleoperation as a baseline interaction method, as well as an ideal "Wizard of Oz" (WOZ) case of near-perfect responsiveness. The study asks novice participants to bring the robot over to them using only their natural behavior in a series of trials. When analyzing the results of these trials, we find support for our claim that the proposed system perform as well or better than the

alternative. We also examine participant behavior and questionnaire feedback according to demographics and characteristics for additional insight relevant to HRI development.

This work has the following novel contributions:

- A formal user study comparing the proposed multimodal interaction system to teleoperation and a WOZ "ideal" interaction system in terms of performance and user experience when tasked with bringing the robot over.
- Investigating what untrained people do when asked "make the robot come to you".
- Investigating the effects of training and light feedback on the system's usability and interaction experience.

8.2 Background

Attention Control in Multi-Human Robot interaction Scenarios

For our scenario, the robot needs to be able to detect and track humans and identify when they intend to initiate communication, then selectively engage in a one-on-one interaction with the person with highest interest among all detected people. There already exist several proposed methods for recognizing humans seeking a robot's attention through communication initiation cues such as eye gaze, hand gesture or stepping into a zone.

In the work by McKeague et al. [42], the robot directs its attention toward the detected person who waves at it. In an approach by Aguirre et al. [47] the degree of the person's attention is detected by analysis of the head pose. Spatial positioning has been also employed as a cue for willingness in interaction, assuming the person that tries to get the robot's attention would come closer to the robot [7, 46, 53–55]. Several authors consider the person that is currently speaking as the target person [6, 49, 50, 56].

Some previously explored scenarios for these methods include, for example, an interactive museum tour guide robot which should be able to turn its head and look at the person of interest [46], or a robot capable of serving drinks in a multi-party or cocktail party setting which should be able to recognize the intentions of people ordering drinks [43, 45]. Another example is the taxi robot [160] whose task is to find the potential passengers and approach them.

Robot Acknowledgment Signaling

After detecting and tracking potential human interaction partners, the next step is how the robot will signal its attention. In most of these studies, the robot's attention is oriented to the target person by head turning [6, 46], body turning [7, 56] or eye movements [48, 54], rather than having the robot approach the target. These methods share the assumption that it is the human's responsibility to approach the robot and initiate interaction.

Other research has explored robots signaling as a way to initiate interactions with humans. In the works by Monajjemi et al. [161] and Bruce et al. [162] the robot approaches the person waving at it from a distance in an outdoor setting. In another example, Saito et al. [160] designed a pedestrian detection system to support a taxi service. If a person waves his arm while facing the robot, it moves in front of him. Our own system will draw from both of these approaches, where the robot will be actively searching for interested parties to signal their desire to interact.

Evaluation Methods

Several methods have been used for reporting the accuracy, robustness and usability of robotic attention-control systems. The most common objective criterion for measuring success in establishing mutual attention is the false positive rate of detecting the target user among distractors in real robot trials [11].

Chen et al. [9] evaluated the performance of their proposed multimodal human interest detection algorithm in over 100 hours of real life experiments. They used precision, recall and F-measure for evaluating the accuracy and robustness of the repeated interest detection algorithm and the average amount of time required for the robot to react to participants' interaction demands for evaluating responsiveness. Some approaches used multiple conditions such as different target objects and users [83], various integration schemes [46], or the accuracy of the sub-systems [6, 34]. Draper et al. [132] and Marge et al. [163] compared their heads-up, hands-free control systems to manual control, showing their system is an effective alternative to robot teleoperation.

Some validated their approach with proof-of-concept demonstrations of an exemplary interaction cycle on real robots [71, 153] or a series of real-world trials [74].

Several proposed attention systems have been evaluated with the help of subjective questionnaires to study how the attention behavior changed people's perception of the robot and how it can improve the social aspects of HRI. Yonezawa et al. [25] used various attentive behaviors of the robot to show the importance and effectiveness of their proposed cross-modal awareness in creating positive impressions of the robot. Kobayashi et al. [43] conducted a Wizard of Oz study to evaluate the effectiveness of the robot's capability to display acknowledgement by turning its gaze toward the person who wants to initiate interaction. The results verified that people feel more satisfied with the robot even if it cannot immediately deal with their request. Similarly, Holthaus et al. [31] showed that a robot that displays its attention and intentions is perceived as more interested in interacting with people compared to having no attentional behavior.

User Studies with Autonomous Real Robots or Non-Wizard-of-Oz Studies

Many researchers evaluated their proposed methods and systems through WOZ as an experimental technique, by bypassing the real behavior of the robot. However, in this setting, the robot controlled by the wizard, is just a proxy for an interaction between humans via robot rather than a true human-robot interaction [164]. In more realistic HRI scenarios, the robot's behavior and the user's experience is highly dependent on its sensory perception rather than the perfect perception of the wizard (or experimenter) [165]. In order to take a step towards autonomous HRI, interaction systems should effort realistic sensor data. Therefore, we believe WOZ is insufficient for evaluating the performance and usability of such systems.

8.3 Method

We evaluated the performance and usability of our proposed multi-human single-robot interaction system in a detailed user study with the general public in a semi-controlled setting. We studied how the robot's level of autonomy and response impacted the user's perception of the robot as well as their interaction experience and perceived workload.

8.3.1 Setup

The participant was asked to make the robot come to them across an 8×5m room without leaving a fixed spot. A Clearpath "Husky" robot was set in one corner of the room. The participant and three lab assistants stood at the far end of the room, with the participant and one assistant facing the robot and two other assistants facing each other and conversing (Figure 8.2).

8.3.2 Scenario

Each participant engaged in five trials (listed below) during the study, each involving a mobile robot setting in one corner of the room. Since each subject participated in all five trials, the study design is within-subject and the interaction system is within-subject variable.

1. **WOZ**: **W**izard **O**f **O**z - The participant perceived the robot as performing the task autonomously, when in reality one of the experimenters was secretly controlling the robot. The participants were not given any description of the robot's capabilities or instructions on how to behave and invited to act freely in bringing the robot over to them. Whatever the participant did, the experimenters secretly teleoperated the robot toward them in emulation of an ideal, perfectly-responsive system. The goal was to understand the participant's natural behavior for attracting a robot's attention. The

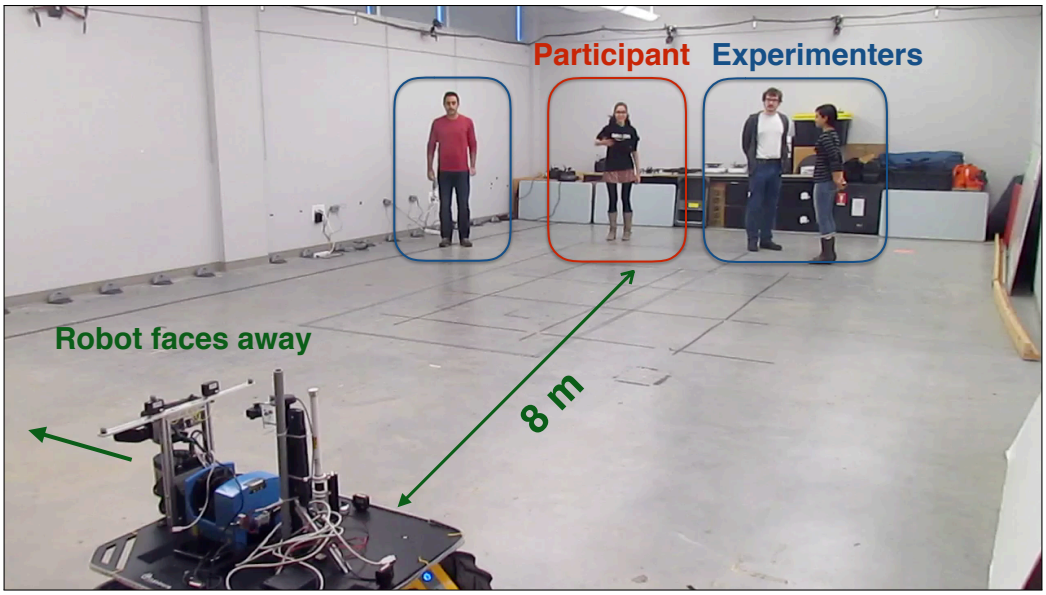


Figure 8.2: The user study setup. The participant stands eight meters from the robot while trying to attract the robot's attention, with two experimenters on their left and one on their right. The robot is facing away such that the participant starts the experiment outside of the field of view of the camera.

details and evaluation results of this trial are presented in [16]. We also compared the questionnaire responses to this trial with responses to the following trials.

2. **TOP**: **T**ele**OP**eration: The participant is given the controller and instructed to bring the robot over. The experimenter showed them how to use the controller and which buttons should be used to drive the robot. The results of this trial were used as a baseline for evaluating the proposed interaction system.

3. **A**: **A**utonomous System without LED feedback or training: Similar to trial one (WOZ), the participant is asked to bring the robot over, without the controller, except that the interaction system is activated and the robot only responds to specific communication signals. Upon detecting the participant, the robot drives toward them.

4. **AF**: **A**utonomous System with LED **F**eedback but no training: This trial was the same as the third one (A), except that the robot provided LED feedback.

5. **AFT**: **A**utonomous System with LED **F**eedback and **T**raining: Before this trial, the system was fully explained to the participant. This means telling them the robot will look for human legs and body and direction of sound to detect the potential interaction partner and works best if they stand facing the robot and call it to get its attention. We also explained that the LED lights around the robot indicate the direction of the robot's attention and they can stop calling the robot once the lights are pointing towards them. They were asked once more to bring the robot over.

WOZ provides an ideal case of an autonomous system with perfect responsiveness and performance, while TOP represents a non-autonomous alternative to the proposed system.

To study the impact of robot state feedback through expressive lights, the ordering of trial three (A) and trial four (AF) was alternated across participants to mitigate any learning effect. The purpose of these two trials, where the system is not explained to the participants, is to investigate the impact of training on system usability compared to trial five (AFT), where participants are given training.

8.3.3 Procedure

The study began by asking the participant to fill out a demographic survey including age, gender and degree of familiarity with robots. We then asked them to stand in a marked place to start their interaction with the robot. After completing each trial, the participant filled out two questionnaires on their interaction experience with the robot and the cognitive workload involved.

Upon finishing trial one (WOZ) and completing the questionnaires, the experimenter explains to the participant that the robot was not autonomous but rather teleoperated by one of the study conductors. Next is trial two (TOP), so experimenters explained how to drive the robot using a controller. Before trial three (A) and trial four (AF), participants are only told the robot will now be looking for their signals. Finally before trial five (AFT), the system and how to gain robot's attention is explained in detail and any questions are answered before proceeding.

Once participants complete all trials they completed a post-experiment survey about their preferred interaction system and provide some additional feedback. Each full test took an average of 20 minutes.

8.3.4 Participants

We recruited 34 participants (23 females, 11 males), ranging in age from 17 to 73, with the median of 28 years old, from around our university (14 undergraduate and 22 graduate students). Participants provided information about their level of familiarity with robots on a five-point scale (1=not familiar at all, 3 = somewhat familiar, 5=extremely familiar). On average ($M = 2.08, SD = 0.93$) participants were slightly familiar with robotics. Participation was voluntary, and not compensated.

8.3.5 Hypotheses

To achieve our evaluation goal, we set five hypotheses:

- H_1: The efficiency of the autonomous multimodal interaction system is higher than teleoperation.

- H_2: The effectiveness of the autonomous multimodal interaction system is higher than teleoperation.
- H_3: The perceived workload of the autonomous multimodal interaction system is lower than teleoperation.
- H_4: People perceive the robot more positively using the autonomous multimodal interaction system than teleoperation.
- H_5: People prefer autonomous multimodal interaction system over teleoperation.

We are also interested in knowing if gender or success in completing the task will influence user's interaction experience or interface preference.

8.3.6 Measures

The main independent variable in this study was the interaction mode the participants used in bringing the robot over (i.e. teleoperation or autonomous).

Observational Measures

Observational measures included *(a)* the *interaction time* and *(b)* the *completion rate* of each interaction mode. Collected video data were used to calculate the interaction time and record the success or failure of attempts to complete the task during each trial.

Interaction Time Interaction time represents the efficiency of the proposed interaction system compared to teleoperating the robot. In trial one (WOZ) and trial two (TOP), interaction time was the amount of time spent manually controlling the robot. When the autonomous system was activated for trials three (A), four (AF) and five (AFT), interaction time was the amount of time it took for the robot to find the right interaction partner and approach them.

Completion Rate Completion rate estimates the effectiveness of the designed interaction system when used by the participants in the latter three trials. This measure also assesses the participant's ability to manually control the robot and complete the task of driving the robot over to them in trial two (TOP). Whenever a user drove the robot toward walls or in the wrong direction, we asked them to repeat the trial from the starting position. In that trial, if the number of failed attempts for a participant was more than 3, we marked it as a failure case. In that trial, participants were given three tries before being marked as a failure case, as the controller needed to be reset in the event of a near collision. The interaction timer was also reset at the start of each attempt.

Questionnaire Measures

Questionnaire data were used to quantify the perceived workload and perception of the robot during different interaction modes. For assessing the workload experienced during each trial, we used a slightly modified version of the NASA Task Load Index (TLX) [166], a subjective survey for assessing the operator's workload during their interaction with machines. This survey derives an overall workload score based on an average of ratings on six subscales including mental demands, physical demands, temporal demands, performance, effort and frustration. Unlike the standard TLX, the participants were asked to make responses on a 5-point scale and the temporal demand was removed since there was no time constraint in finishing the task.

Participants reported their perception of the robot's responsiveness, intuitiveness, expressiveness and perceived intelligence using a 5-point Likert scale. Finally, we polled participants about their general preference between the two different interaction systems.

8.4 Results

8.4.1 Observations

Interaction Time

Hypothesis H_1 predicts that the efficiency of the proposed autonomous multimodal interaction system is higher than teleoperation. The efficiency metric is the time taken for the robot to reach the participant. Interaction time was used to evaluate this hypothesis. We consider only the successful attempts for each trial, our reasoning being that in failure cases the robot did not successfully reach its target so we could not mark the end of the interaction. In trial one (WOZ) all attempts were successful since robot's behavior is perfect, however in trial two (TOP), only trials where the participant successfully drove the robot to its destination are counted, and in trials three (A), four (AF) and five (AFT), we only count where the robot successfully detected and approached the interaction partner.

Since the interaction time data were not normally distributed ($p < 0.05$, Shapiro-Wilk test) and thus violated the assumption for parametric statistical tests, they were analysed by nonparametric Kruskal-Wallis one-way ANOVA, followed by posthoc pairwise comparisons by the Mann-Whitney U test. The results supported Hypothesis H_1. The interaction mode used in each trial had a significant main effect on interaction time ($H(4) = 16.01, p < 0.005$, Kruskal-Wallis). Trial two (TOP) has significantly longer interaction time than trial three (A), trial four (AF) and trial five (AFT) ($p < 0.05$, Mann-Whitney test) and also trial one (WOZ) ($p < 0.01$, Mann-Whitney test). Figure 8.3 shows the interaction time of different interaction modes. Box plots provide details about interaction time distributions and the red line shows the mean.

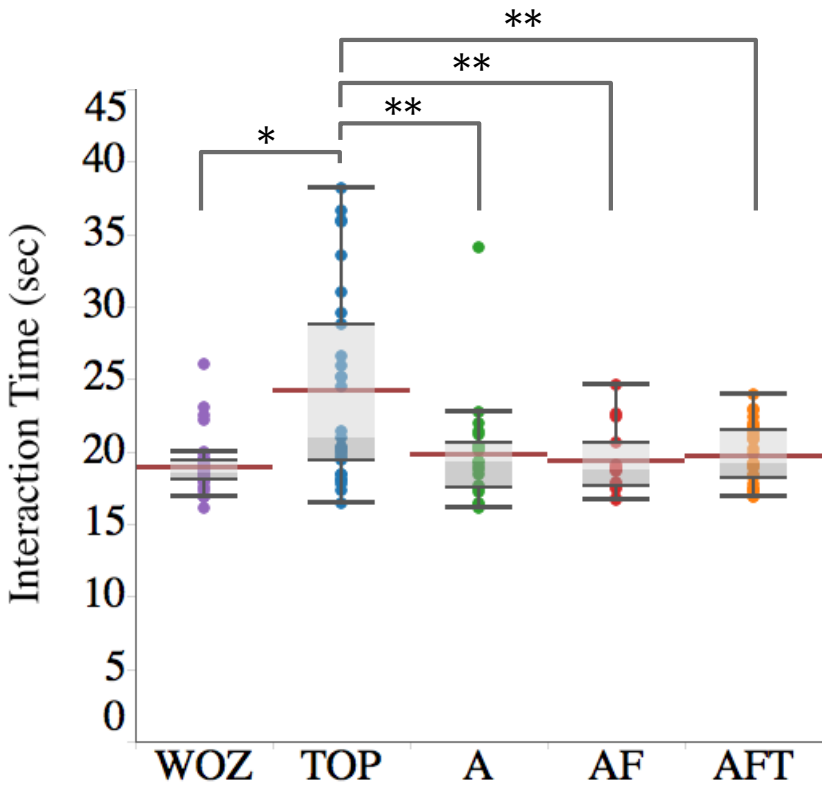


Figure 8.3: Interaction time for each trial. The data are filtered by the completion results, which keeps only successful trials. ($*\, p < 0.01, **\, p < 0.05$)

Completion Rate

Hypothesis H_2 predicts that the designed autonomous multimodal interaction system is more effective than teleoperation. The effectiveness metric is the percentage of trials that was completed successfully. The data show there is a significant difference ($p < 0.001$, CochranâĂŹs Q test) between the completion rates of different interaction modes. Applying posthoc pairwise comparisons showed that both trial two (TOP) and trial five (AFT) have significantly higher completion rate than trial three (A) ($p < 0.05$) and trial four (AF) ($p < 0.001$).

The results do not support Hypothesis H_2. However, since there is no significant difference between the completion rate of trial two (TOP) and trial five (AFT), we might claim that our proposed interaction system performed at least as well as the alternative. The completion rate of the autonomous interaction system was lower when the participants were not trained, although participants also had training for how to teleoperate the robot before that trial. Figure 8.4 provides details.

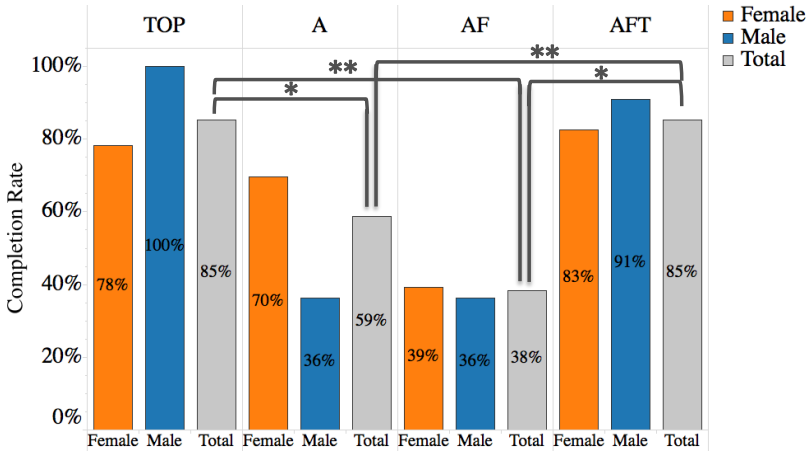


Figure 8.4: Completion rate for each trial broken down by gender. Color shows details about interaction mode and gender. ($*\, p < 0.05, **\, p < 0.001$)

8.4.2 Questionnaire Responses

Post-Trial

First, we analysed the survey data by performing an internal consistency reliability analysis. Cronbach's α for five items task load and four items perception of the robot were 0.80 and 0.76, respectively. Cronbach's α is a tool for assessing the reliability of summated scales [167]. The α coefficient ranges from 0 to 1, where the higher the score the more reliable the scale is. A commonly-accepted rule of thumb is that α values higher than 0.7 are considered acceptable reliability coefficient [168].

Task Load Hypothesis H_3 predicts the perceived workload of using the autonomous multimodal interaction system will be less than teleoperation. The data show that interaction mode had significant effects on the overall task load score ($H(4) = 22.27, p < 0.001$, Kruskal-Wallis H test). Follow-up pairwise comparisons shows that, as expected, the overall perceived workload of trial one (WOZ) is significantly less than all other interaction modes ($p < 0.01$, Mann-Whitney U test). However there is no significant difference between the autonomous modes and teleoperation. Figure 8.5 shows these results. While this does not support our hypothesis that perceived workload would be lower, it is again no worse than the alternative.

Perception of the Robot We inquired as to the participants' perceptions of the robot's *responsiveness*, *expressiveness*, *intuitiveness* and *perceived intelligence* (Table 8.1). The overall perception is estimated by averaging these four scores. Hypothesis H_4 predicts that people perceive the robot more positively using the autonomous multimodal interaction system rather than teleoperation.

Table 8.1: Interaction Experience Questionnaire – (ranked on 5 Likert scale ranging from 1 = strongly disagree to 5 = strongly agree for the first three questions and from 1 = unintelligent to 5 = intelligent for Perceived Intelligence)

Metric	Question
Responsiveness	The robot reaction was quick and appropriate.
Expressiveness	The robot behaviour was understandable.
Intuitiveness	The interaction with the robot was natural and intuitive.
Perceived Intelligence	Please rate your impression of the robot on these scales.

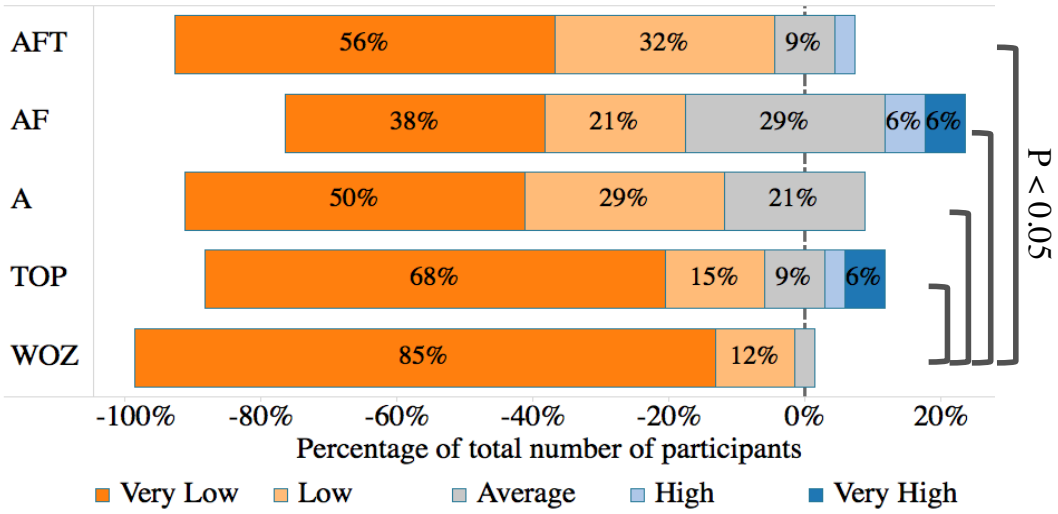


Figure 8.5: The overall task load for each trial. The graph is shown using the diverging stacked bar chart. Color shows details about task load ratings. Size of bars shows the percentage of participants that vote for each rate. Having lower task load is evidence in favour of an interface.

Kruskal-Wallis H tests showed that interaction mode had significant effects on perceived "responsiveness" ($H(4) = 16.5, p < 0.001$),"expressiveness" $H(4) = 23.21, p < 0.001$), "intuitiveness" ($H(4) = 13.31, p < 0.01$),"intelligence" ($H(4) = 23.19, p < 0.001$) and "overall perception of the robot" ($H(4) = 24.9, p < 0.001$). On average, participants rated the autonomous multimodal interaction system in trial five (AFT) more positively than teleoperation in trial two (TOP), supporting Hypothesis H_4, while they didn't rate trial five (AFT) significantly different from trial one (WOZ). We summarized these results in Figure 8.6.

The responsiveness of the autonomous system was rated significantly lower than in trial one (WOZ). However, when participants are trained with the system in trial five (AFT), it is not perceived as less responsive than teleoperation in trial two (TOP). Participants rated the expressiveness of the robot as significantly lower in trial three (A) and trial four (AF),

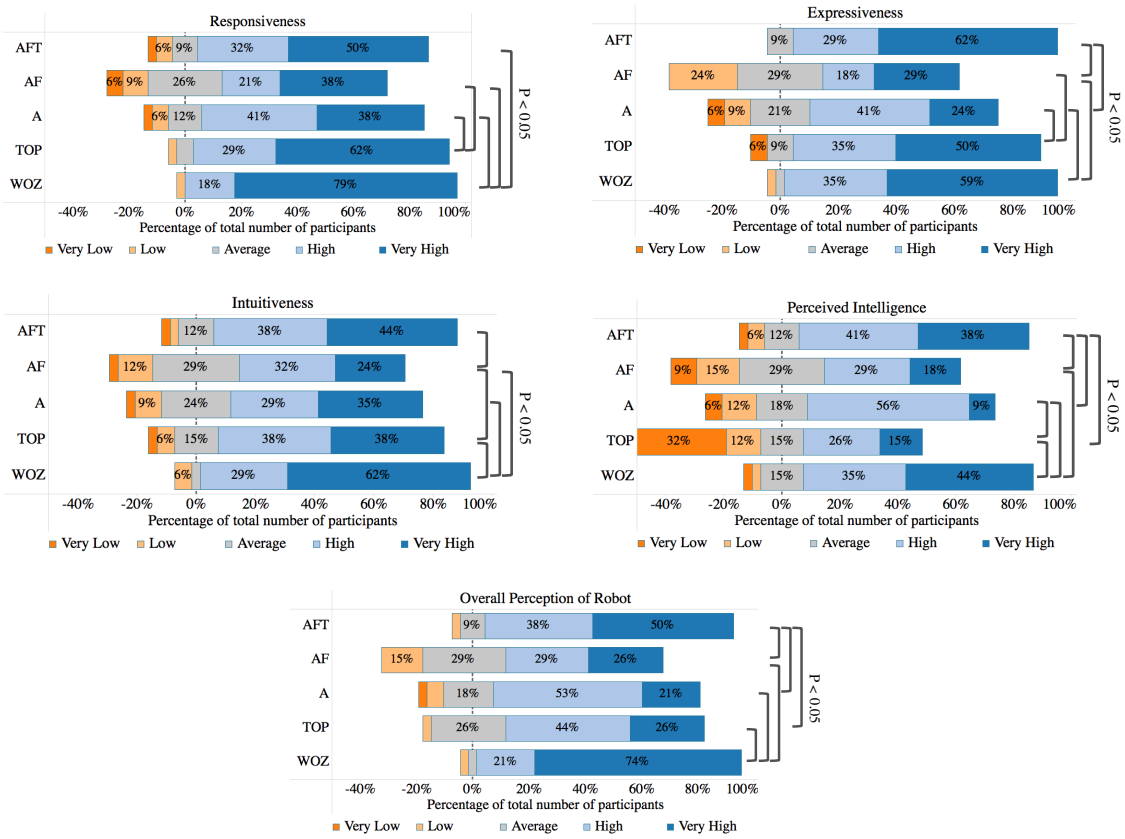


Figure 8.6: The perception of the robot during each trial. The graphs are shown using the diverging stacked bar charts. Color shows details about ratings. Size of bars shows the percentage of participants that vote for each rating. The bars are labeled by percentage. A higher rating is better for the perception of the robot.

where they have not learned how the system works, than the other three trials. In trial two (TOP), teleoperation was rated significantly less intuitive than trial one (WOZ). However, there was not a significant difference between trial one (WOZ) and trial five (AFT) in terms of perceived intuitiveness. Similarly, perceived intelligence in trial two (TOP) was scored significantly lower than in trial five (AFT) and trial one (WOZ), while there was no significant difference between those two trials.

A Spearman's correlation coefficient was computed to assess relationship between task load and perception of the robot. The result shows a moderate negative correlation, which was statistically significant, $\rho(168) = -0.5406, p < 0.001$. As expected this signifies that the greater the effort participants put into interacting the lower perception they will have of the robot.

Post-Study

Hypothesis H_5 predicts that people would favor the autonomous multimodal interaction system over teleoperation. We tested this hypothesis using a post-experiment questionnaire. The questionnaire asked participants to choose, between TOP and AFT, which interface they preferred, which interface was more natural and intuitive and which interface they found more comfortable and easy to use.

The results confirm this hypothesis at the five percent level of significance (Figure 8.7). For all three questions, participants chose the autonomous interaction system over teleoperation in terms of preference ($82\%, p < 0.001$), naturalness and intuitiveness ($64\%, p < 0.05$) and ease of use ($82\%, p < 0.001$).

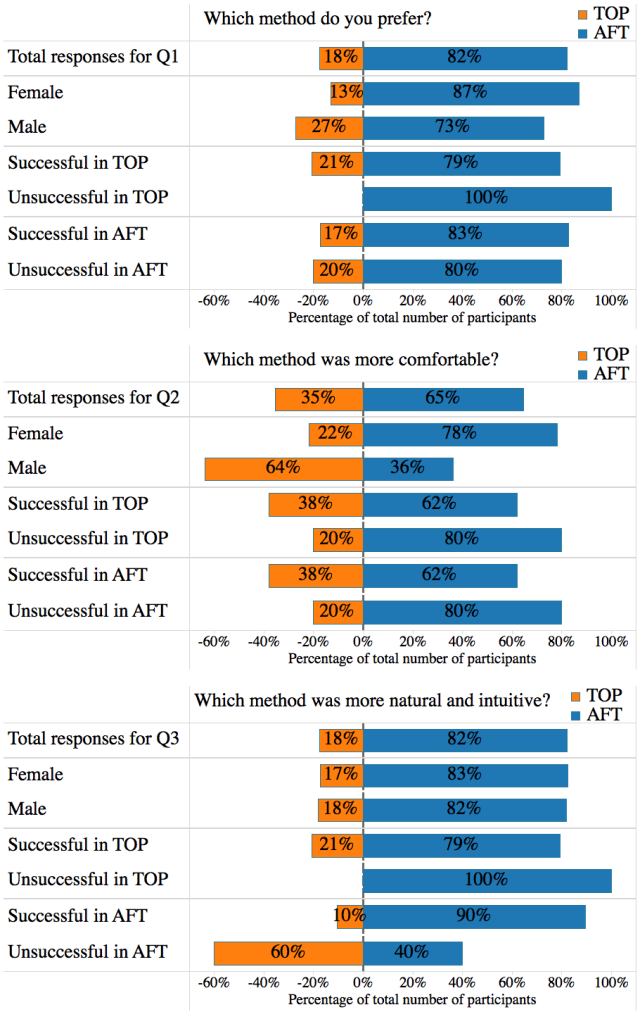


Figure 8.7: The post-study survey. Aside from the overall result, the breakdown for each question's responses are grouped according to gender, success in trial two (TOP), and success in trial five (AFT).

We were interested in determining whether interaction mode preference is related to gender. The Pearson Chi-square test shows no significant association between gender and whether one favored one interaction mode to the other. Chi-square tests of independence were also performed to examine the relation between success in using an interface and preferring it. The result showed success in using an interface is not related to favoring it over the other one. Figure 8.7 summarizes the result of post-experiment questionnaires.

8.5 Observations from Wizard-of-Oz Trial

This section reports on the observations of the Wizard-of-Oz portion of the experiment where participants were invited to act freely in bringing a robot over to them. The goal was to understand what participants did when instructed only to "make the robot come to you" without knowing that whatever they chose would succeed.

The means by which participants tried to communicate to the robot can be classified into either sound signals or gestural cues. Voice commands were the dominant sound signal, and we classified them as one of **deictic** (referenced to the user, e.g. *come here*), **directive** (referenced to the robot, e.g. *go forward*) or **addressed** to the robot (e.g. *hello, robot!*). **Clapping** (which is also counted as a gesture) was the most common non-verbal sound, with one case of finger-snapping and two of thigh-slapping. Gestures were grouped into **waving, beckoning, clapping, reaching** or **null** (i.e. no gesture). Examples of each gesture are given in Figure 8.8. All gestures apart from reaching incorporate periodic motion. Waving and reaching significantly change the user's outline seen from the robot, while beckoning and reaching modify the depth seen by the robot. Beckoning included extending hands toward the robot and moving all or part of the forearm back and forth in an inviting motion.

Figure 8.8: Participants gesturing to attract or maintain the robot's attention. From left: clapping, waving, beckoning, reaching and no gestures

Participants appeared to perceive two different phases to their interaction; they changed their behavior distinctly at one point in the experiment, so we have presented the results in two phases. In the first phase, the robot is facing away from the participant, who tries to *attract* its attention. Once the robot turns toward the participant and begins to approach, they shift focus to *maintain* the attention. Despite being told in advance that the robot

would stop automatically, 37% of participants tried signaling the robot to stop, which might constitute a third phase but was not analyzed here.

Tables 8.2 and 8.3 give the frequency of each pairing of a gesture and a sound during each observed phase. Note that the totals for each table do not sum to the number of participants, as some participants would exhibit multiple behaviors over the course of the trial such as switching between waving and clapping.

For sound signals, deictic and addressing were most common during the *attract attention* phase, while the *maintain attention* phase saw addressing drop off in favor of directives, particularly affirmations. For example, at the start of a trial, when the robot is not facing toward the user, the participant would say "Hey robot! Come here!". When the robot turns towards them, the user switches to "Yes. This way. Good job.". The frequency of participants making no sound was lower during *attract attention* and spiked during *maintain attention*. Making no gesture was common in both phases, although *maintain attention* also has participants doing nothing but watch the robot. Waving and beckoning were the most common gestures, fluctuating only slightly between phases, and were predominantly paired with deictic commands or silence. Reaching was only observed during the *maintain attention* phase, and then only rarely and silently.

Table 8.2: Sound and Gesture Frequencies in *Attract Attention* Phase

Gesture Sound	Beckoning	Waving	Clapping	Reaching	No Gesture
Deictic	6	4	1	-	7
Directive	-	-	-	-	1
Addressing	2	4	-	-	6
Clapping	-	-	6	-	-
Other	-	-	-	-	2
No Sound	2	6	-	-	-

Table 8.3: Sound and Gesture Frequencies in *Maintain Attention* Phase

Gesture Sound	Beckoning	Waving	Clapping	Reaching	No Gesture
Deictic	8	3	-	-	6
Directive	2	1	-	-	6
Addressing	-	-	-	-	1
Clapping	-	-	3	-	-
Other	-	-	-	-	4
No Sound	4	7	-	2	3

8.5.1 Continuous vs. Corrective

Another way to analyze participant behavior is whether they continued to signal to the robot throughout the interaction or stopped once they believed they had the robot's attention, only signaling again to correct perceived errors.

Since the robot was teleoperated, its behavior rarely required correcting, so after appearing to engage the robot's attention participants behaving purely correctively would signal infrequently or stop altogether. The proportions of participants engaged in each behavior are given in Table 8.4.

Table 8.4: Continuous vs. Corrective

Interaction Phase	Continues Gesture	Continuous Sound	Both Continuous	Purely Corrective
Maintain Attention	20%	3%	17%	60%

8.6 Discussion

Comparison with Teleoperation as an Alternative

Of our first three hypotheses concerning the interaction system compared to teleoperation, our results only showed improvement in terms of efficiency. However, while our system did not show gains in effectiveness or perceived task load, it also showed no significant decline in these areas. Furthermore, support for our fourth and fifth hypotheses suggests a better user experience through improved perception of the robot itself and overall user preference.

Among the minority who preferred teleoperation, some mentioned their experience playing video games and their existing familiarity with controllers, one participant claiming "...for me it is effortless to control [the robot] with the controller" or "My preference for using the controller may be related to time spent playing Mario Kart". One participant also mentioned that while the interaction system was easier to use, if they wanted to guarantee success they would use the controller. To provide additional insight, we conducted various supplementary tests to assess if gender or success had any influence on observed task load, perception of the robot or interface preference.

Analysis of completion rate and perception of the robot showed both measures to be independent of gender. Interaction time, however, showed a significant effect for gender in trial two (TOP) ($p < 0.01$, Mann-Whitney U test), where men teleoperated the robot significantly faster than women. By comparison, trial five (AFT) showed no significant difference between male and female interaction times. Also, while gender did not affect overall interface preference, women were significantly more likely to find trial five (AFT) more comfortable and easier to use than men ($X^2(1, N = 34) = 3.92, p < 0.05$, Chi-square test). Similarly, for task load, females perceived significantly higher task load in trial two (TOP) than male participants ($p < 0.05$, Mann-Whitney U test).

Interface preference and robot perception were independent from a participant's success or failure in each trial, with the exception of task load for trial two (TOP), which saw significantly higher task load for participants who failed to drive the robot over ($p < 0.01$, Mann-Whitney U test). While H_2's prediction on effectiveness was not supported, this result may suggest failure for trial two (TOP) has a larger impact on perceived task load.

Comparison with Trial One (WOZ) as an Ideal Case

The Wizard-of-Oz trial can be considered as an error-free, fully responsive case of the proposed interaction system, where the robot quickly reacts to all attention-seeking signals and successfully approaches the interaction partner.

Our results showed the efficiency of the proposed system is not significantly different from trial one (WOZ) (Figure 8.3). However, on average the task load in trial five (AFT) is reported as significantly higher than in trial one (WOZ), (Figure 8.3), suggesting that the participants put more effort in getting the robot's attention. As one the participants stated, "I wish saying "*come here, quick*" should be enough". This is in line with the significantly lower score for responsiveness of the robot during trial five (AFT), although the overall perception of the robot is not significantly different from the ideal autonomous system.

Effect of LED Feedback

The goal of having both trial three (A) and trial four (AF) was to investigate the effects of LED feedback on the interaction experience and system performance. Specifically, we were interested to know whether enabling the feedback would help participants to guess what signals would be effective for acquiring the robot's attention.

Unexpectedly, light feedback was found to decrease the perceived responsiveness, expressiveness and intelligence of the robot. Task load, overall perception of the robot, efficiency and effectiveness were not significantly different between the two trials. We speculate that this is because of poor design for the feedback system, as it might have confused the participants despite our intention of communicating the robot's state. One participant said "I didn't like the light feedback, I think it is more confusing. Without light feedback, I can tell where the robot is going by just looking at, but with feedback I was trying to understand what the light means." Another mentioned that once it was explained what the blue light meant "it felt more intuitive", suggesting some warmed to the feedback method after training. One alternative to this feedback design could be an animated or mechanical pair of eyes that look in the direction of the target while approaching or rotating towards it. Changing the form of the system could provide more familiar, anthropomorphic clues as to its purpose.

Effect of Training

While the level of participant familiarity with controllers varied widely, all participants began the study completely untrained in the use of our proposed system and with no obvious starting point apart from the natural human activity of calling someone over. The results for trial five (AFT) in comparison to trial three (A) and trial four (AF) showed significant improvements in performance ($p < 0.05$) from as little as two rounds of practice and a brief explanation, which suggests both that the system was easy to learn and that with time and practice could become as familiar to general users as controllers are today. One participant noted “Once explained, calling the robot made much more sense".

Additional Observations

Participants would commonly ask if the robot had a name to call it by, though in practice almost all participants defaulted to “robot" when addressing it. Some participants had difficulty recognizing the front or “head" of the robot, which could cause problems in trial two (TOP) when attempting to teleoperate the robot backward. On the other hand, some participants also described the form of the robot as “dog-like", influencing their choice of gestures or sounds. Most significantly for future design, many participants continued to gesture to the robot in trial five (AFT) even after having been told that the robot will not track gestures, purely from habit. This indicates an opportunity for robots to explore these frequently-occurring gestures.

Limitations

One limitation of the study setup was the ordering of the trials, where only trial three (A) and trial four (AF) were alternated for different participants. This ordering was the result of necessity as trial one (WOZ) was meant to capture first reactions and untrained trials could not be completed once the participant was trained for trial five (AFT), but it nevertheless may have impacted the results through participant fatigue or other factors. On post-study review, it was considered that trial two (TOP) could have been alternated with the block of three interaction systems trials, and the failure to do so may have been a weakness of this study.

There were also environment limitations the user study took place in an indoor lab environment, as well as a potential mismatch with the shape and appearance of the robot as the Husky is designed for outdoor use. A more diverse study involving different scenarios, locations and even robots would provide a more thorough test for the proposed interaction system.

8.7 Conclusion

We presented a probabilistic sensor integration approach for controlling a mobile robot's attention using multimodal people detection and tracking. We designed a user study to assess the performance and usability of the proposed system when used by the general public and compared it to manual control. In an experiment with 34 participants, we observed that they favored the proposed multimodal interaction system over teleoperation. In addition, they had a more positive perception of the robot and a more efficient interaction with the multimodal interaction system than when they teleoperate it. We also showed that the proposed system is as effective as teleoperation. We can conclude that the multimodal interaction system is a better alternative to manual control.

Chapter 9

Conclusion and Future Work

In this dissertation, we proposed three multimodal, sensor-mediated interaction systems that regulate a direct interaction between humans and robots. First, we defined attention systems whereby robots can correctly identify human attention-drawing signals such as gaze, body posture, gestures, referential words and/or a combination of these, and respond to them accordingly. Next, we provided an overview of approaches to designing attention systems and recognizing human intentions in initiating an interaction and creating mutual attention.

Our survey showed there is limited research on designing attention systems for scenarios where there are multiple humans and/or multiple robots in the environment. This requires each party to first find an interested counterpart among others and then get its attention to initiate a communication. Moreover, many of the proposed methods simplified the task of human interest detection with stationary robots or static sensors assumptions (e.g. overhead cameras) or constraints on the user's pose. However, we believe that in everyday use, the convenience of using an interface is an important aspect of its performance. Therefore, we have designed interfaces that are simple, easy-to-use and require no special instrumentation of the human user or environment.

In Chapter 3, we presented an integrated system by which multiple humans and robots interact robustly using a combination of sensing and signaling. We showed that reaching toward a robot - a specialization of pointing - can be used to designate a particular robot for subsequent one-on-one interaction. The system employs multiple phases of human-detection with timeouts, retries, and fallback behavior to contribute to robustness. The robot provides carefully-designed audio feedback on its interaction state changes, to quickly reassure users that their waving is working. A series of real-world trials demonstrates the practicality of our approach.

In this design, we used a very small set of discrete gestures. As future work, the gesture set could be extended to allow a user to point to *any* arbitrary place in the environment. This has been done for a single robot system (e.g. [125, 126]); however, an interesting extension would be to exploit multiple robots to jointly estimate the vector given the system's ability to simultaneously capture images of the user from multiple angles.

In Chapter 4, we described a multimodal system for dynamically creating, modifying and commanding groups of robots from a population by face engagement and spoken commands. We evaluated its performance in different spatial configurations of robots and user. The results showed that success rate varies with the human-robot distance and the angle between robots. As the angle between robots increases, the workspace is limited to shorter distances. Two modes of selecting multiple robots were introduced and compared in with different layouts, concluding that iterated election has wider workspace and performs much more robust compared to the simultaneous election in our setting. Chapter 5 presented an extension of this system, where instead of face engagement, one can identify an individual or a group of robots using haptic stimuli, and name them using a voice command (e.g. "You

two are Green" or "You two, join Green"). Subsequent commands can be addressed to the same robot(s) by name.

In Chapter 6, we analyzed the efficiency of spatially embedded human multi-robot interaction systems regarding the amount of time it takes for a human to interact with a group of robots. We proposed, and showed, using experimental evidence, which the time required to interact with multiple robots can be reduced by different methods of creating groups of robots and doing concurrent interactions that exploit locality.

Next, in Chapter 7, we presented a probabilistic approach for controlling a mobile robot's attention using multimodal people detection and tracking. To achieve robust operation, the system integrates three multimodal human percepts and directs the robot to approach the location with the highest probability of an engaged human. A series of series real-world experiments in outdoor environments were performed to validate our probabilistic sensor fusion framework's ability to select an interaction partner.

Finally, in Chapter 8, we reported on a user study to assess the performance and usability of the proposed system when used by the general public and compared it to manual control. The results of this study showed the proposed system is preferred by 82% of the users and performs as well or better than a teleoperated alternative. Also, the overall human's perception of the robot, when running this system, is not significantly different from the ideal error-free autonomous system. The data presented in this chapter provide evidence that inexperienced users were able to reliably obtain the robot's attention and call it over for interaction using only their instinctive behavior and without any instrumentation (i.e. they carried no special equipment or clothing). This is particularly important for robots deployed in public settings, as untrained and non-technical users can engage in an interaction or call the robot's attention with little or no instruction. We also reported on the observations of the Wizard-of-Oz portion of the user study where participants were invited to act freely in bringing a robot over to them, aiming to understand what untrained people do when asked "make the robot come you" without knowing that whatever they chose would succeed. The results show a variety of calls and gestures made to the robot, which changed over time.

The best opportunity for improvement a future system could make building on this work is additional support for gesture recognition. The results from the WOZ experiments of participant behavior both showed an robust tendency on the part of participants to signal the robot via gestures and provided data on the diversity of gestures and approaches used. An advantage of our proposed multimodal approach to detecting interested partners is that a gesture recognition module could be smoothly integrated into the existing system alongside the other detectors. The study data also suggest some refinements that could be made to sound-source localization. Differentiating between human-sourced audio like words and clapping versus environment noise could improve the reliability of sound as a detection method. Certain simple verbal commands being widely popular among participants also makes a case for speech recognition. The poor performance of the LED feedback system

and the limitations of the chosen form factor suggest more work remains matching the interaction system with a suitable platform. The study made clear the robot's overall performance is not only the result of the interaction system but also how well it meshes with the other components. More avenues of feedback, like complex sounds or gestures, may also warrant exploration.

This thesis is a contribution to knowledge of human-robot interaction. We have described and demonstrated several interaction methods (offering gesture, "You-two" and Touch-to-Name), provided a framework for describing the scalability of human multi-robot interaction, implemented a robust system for finding an interaction partner in a crowd of people, and finally evaluated it in a user study.

Bibliography

[1] Dennis Perzanowski, Alan C. Schultz, William Adams, Elaine Marsh, and Magda Bugajska. Building a multimodal human-robot interface. *IEEE Intelligent Systems*, 16(1):16–21, 2001.

[2] Chrystopher L. Nehaniv, Kerstin Dautenhahn, Jens Kubacki, Martin Haegele, Christopher Parlitz, and Rachid Alami. A methodological approach relating the classification of gesture to identification of human intent in the context of human-robot interaction. In *IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 371–377, August 2005.

[3] Haibin Yan, Marcelo H. Ang Jr, and Aun Neow Poo. A survey on perception methods for human–robot interaction in social robots. *International Journal of Social Robotics*, 6(1):85–119, 2014.

[4] Michael A. Goodrich and Alan C. Schultz. Human-Robot Interaction: A survey. *Foundations and trends in human-computer interaction*, 1(3):203–275, 2007.

[5] Jijun Wang. *Human control of cooperative robots*. Doctoral Dissertation, The School of Information Sciences, University of Pittsburgh, 2008.

[6] Chia-How Lin, Chia-Hsing Yang, Cheng-Kang Wang, Kai-Tai Song, and Jwu-Sheng Hu. A new design on multi-modal robotic focus attention. In *IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 598–603, August 2008.

[7] Hans-Joachim Böhme, Torsten Wilhelm, Jürgen Key, Carsten Schauer, Christof Schröter, Horst-Michael Groß, and Torsten Hempel. An approach to multi-modal human–machine interaction for intelligent service robots. *Robotics and Autonomous Systems*, 44(1):83–96, 2003.

[8] Candace L. Sidner, Christopher Lee, Cory D. Kidd, Neal Lesh, and Charles Rich. Explorations in engagement for humans and robots. *Artificial Intelligence*, 166(1):140–164, 2005.

[9] Jingdong Chen and William J. Fitzgerald. Continuous multi-modal human interest detection for a domestic companion humanoid robot. In *the 16th IEEE International Conference on Advanced Robotics (ICAR'13)*, pages 1–6, November 2013.

[10] Cynthia Breazeal and Brian Scassellati. A context-dependent attention system for a social robot. In *Proceedings of the 16th International Joint Conference on Artificial*

Intelligence, IJCAI '99, pages 1146–1153, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc.

[11] Joao Filipe Ferreira and Jorge Dias. Attentional mechanisms for socially interactive robots - A survey. *IEEE Transactions on Autonomous Mental Development*, 6(2):110–125, 2014.

[12] Lars Kopp and Peter Gärdenfors. *Attention as a minimal criterion of intentionality in robots*. Citeseer, 2001.

[13] Erving Goffman. *Behavior in public places*. Simon and Schuster, 2008.

[14] Dan R. Olsen Jr and Stephen Bart Wood. Fan-out: Measuring human control of multiple robots. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'04)*, pages 231–238. ACM, April 2004.

[15] Shokoofeh Pourmehr, Jack Thomas, and Richard Vaughan. Finding an interaction partner in a crowd, a robust interaction system and user study. *Manuscript in preparation*, 2016.

[16] Shokoofeh Pourmehr, Jack Thomas, and Richard T. Vaughan. What untrained people do when asked "make the robot come to you". In *Proceedings of the ACM/IEEE International Conference on Human-robot Interaction (HRI'16) (Late-Breaking Abstract)*, March 2016.

[17] Shokoofeh Pourmehr, Jake Bruce, Jens Wawerla, and Richard T. Vaughan. A sensor fusion framework for finding an HRI partner in crowd. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA'17), (Submitted)*.

[18] Shokoofeh Pourmehr, Jens Wawerla, Richard T. Vaughan, and Greg Mori. On the scalability of spatially embedded human multi-robot interfaces. In *Proceedings of the 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI'15) Workshop on Human-Robot Teaming*, Portland, USA, March 2015.

[19] Valiallah Mani Monajjemi, Shokoofeh Pourmehr, Seyed Abbas Sadat, Fei Zhan, Jens Wawerla, Greg Mori, and Richard T. Vaughan. Integrating multimodal interfaces to command UAVs. In *Proceedings of the 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI'14)*, pages 106–106, March 2014.

[20] Shokoofeh Pourmehr, Valiallah Mani Monajjemi, Seyed Abbas Sadat, Fei Zhan, Jens Wawerla, Greg Mori, and Richard T. Vaughan. You are Green: a Touch-to-Name interaction in an integrated multimodal multi-robot HRI system. In *Proceedings of the 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI'14)*, pages 266–267, March 2014.

[21] Shokoofeh Pourmehr, Valiallah Mani Monajjemi, Richard T. Vaughan, and Greg Mori. "You two! Take off!": Creating, modifying and commanding groups of robots using face engagement and indirect speech in voice commands. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'13)*, pages 137–142, November 2013.

[22] Shokoofeh Pourmehr, Valiallah Monajjemi, Jens Wawerla, Richard T. Vaughan, and Greg Mori. A robust integrated system for selecting and commanding multiple mobile robots. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA'13)*, pages 2874–2879, May 2013.

[23] Hideki Kozima, Cocoro Nakagawa, and Hiroyuki Yano. Attention coupling as a prerequisite for social interaction. In *IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 109–114, August 2003.

[24] Hideki Kozima. Infanoid. In *Socially Intelligent Agents*, pages 157–164. Springer, 2002.

[25] Tomoko Yonezawa, Hirotake Yamazoe, Akira Utsumi, and Shinji Abe. Evaluating crossmodal awareness of daily-partner robot to user's behaviors with gaze and utterance detection. In *Proceedings of the 3rd ACM International Workshop on Context-Awareness for Self-Managing Systems*, pages 1–8. ACM, 2009.

[26] Hirotake Yamazoe, Akira Utsumi, Tomoko Yonezawa, and Shinji Abe. Remote gaze estimation with a single camera based on facial-feature tracking without special calibration actions. In *Proceedings of the Symposium on Eye tracking research and applications*, pages 245–250. ACM, 2008.

[27] Mani Monajjemi, Jake Bruce, Seyed Abbas Sadat, Jens Wawerla, and Richard T. Vaughan. UAV, do you see me? Establishing mutual attention between an uninstrumented human and an outdoor UAV in flight. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'15)*, pages 3614–3620, September 2015.

[28] Jake Bruce. Finding tiny people: Long-range outdoor sensing for establishing joint attention in human-robot interaction. Master's thesis, School of Computing Science, Simon Fraser University, December 2015.

[29] Shogo Nabe, Takayuki Kanda, Kazuo Hiraki, Hiroshi Ishiguro, Kiyoshi Kogure, and Norihiro Hagita. Analysis of human behavior to a communication robot in an open field. In *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction (HRI'06)*, pages 234–241. ACM, 2006.

[30] Patrick Holthaus, Ingo Lütkebohle, Marc Hanheide, and Sven Wachsmuth. Can I help you? A spatial attention system for a receptionist robot. In *Social Robotics*, pages 325–334. Springer, 2010.

[31] Patrick Holthaus, Karola Pitsch, and Sven Wachsmuth. How can I help? *International Journal of Social Robotics*, 3(4):383–393, 2011.

[32] Marek P. Michalowski, Selma Sabanovic, and Reid Simmons. A spatial model of engagement for a social robot. In *the 9th IEEE International Workshop on Advanced Motion Control*, pages 762–767, 2006.

[33] Marek P. Michalowski, Selma Sabanovic, and Reid Simmons. A spatial model of engagement for a social robot. In *the 9th IEEE International Workshop on Advanced Motion Control, 2006.*, pages 762–767. IEEE, 2006.

[34] Markus Finke, Kheng Lee Koay, Kerstin Dautenhahn, Chrystopher L. Nehaniv, Michael L. Walters, and Joe Saunders. Hey, I'm over here - How can a robot attract people's attention? In *IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 7–12, August 2005.

[35] Edward T. Hall, Ray L. Birdwhistell, Bernhard Bock, Paul Bohannan, A. Richard Diebold Jr, Marshall Durbin, Munro S. Edmonson, JL. Fischer, Dell Hymes, Solon T. Kimball, et al. Proxemics [and comments and replies]. *Current anthropology*, pages 83–108, 1968.

[36] Rachel Gockley, Allison Bruce, Jodi Forlizzi, Marek Michalowski, Anne Mundell, Stephanie Rosenthal, Brennan Sellner, Reid Simmons, Kevin Snipes, Alan C. Schultz, et al. Designing robots for long-term social interaction. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'05)*, pages 1338–1343, August 2005.

[37] Yusuke Kato, Takayuki Kanda, and Hiroshi Ishiguro. May I help you?: Design of human-like polite approaching behavior. In *Proceedings of the 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI'15)*, pages 35–42, March 2015.

[38] Richard Kelley, Alireza Tavakkoli, Christopher King, Monica Nicolescu, Mircea Nicolescu, and George Bebis. Understanding human intentions via hidden markov models in autonomous mobile robots. In *Proceedings of the 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI'08)*, pages 367–374, March 2008.

[39] Jake Bruce, Jens Wawerla, and Richard T. Vaughan. Human-robot rendezvous by co-operative trajectory signals. In *Proceedings of the 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI'15), Workshop on Human-Robot Conference on Human-Robot Interaction Workshop on Human-Robot Teaming*, March 2015.

[40] Rowel Atienza and Alexander Zelinsky. Intuitive human-robot interaction through active 3D gaze tracking. In *the 11th International Symposium of Robotics Research*, pages 172–181. Springer, 2005.

[41] Sethu Vijayakumar, Jörg Conradt, Tomohiro Shibata, and Stefan Schaal. Overt visual attention for a humanoid robot. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'01)*, volume 4, pages 2332–2337, October 2001.

[42] Stephen McKeague, Jindong Liu, and Guang-Zhong Yang. Hand and body association in crowded environments for human-robot interaction. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA'13)*, pages 2161–2168, May 2013.

[43] Yoshinori Kobayashi, Masahiko Gyoda, Tomoya Tabata, Yoshinori Kuno, Keiichi Yamazaki, Momoyo Shibuya, Yukiko Seki, and Akiko Yamazaki. A considerate care robot able to serve in multi-party settings. In *IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 27–32, August 2011.

[44] Rainer Stiefelhagen, Jie Yang, and Alex Waibel. Tracking focus of attention for human-robot communication. In *IEEE/RAS International Conference on Humanoid Robots (Humanoids)*, 2001.

[45] Andre Gaschler, Sören Jentzsch, Manuel Giuliani, Kerstin Huth, Jan De Ruiter, and Alois Knoll. Social behavior recognition using body posture and head pose for human-robot interaction. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'12)*, pages 2128–2133, October 2012.

[46] Peter Poschmann, Sven Hellbach, and Hans-Joachim Böhme. Multi-modal people tracking for an awareness behavior of an interactive tour-guide robot. In *International Conference on Intelligent Robotics and Applications*, pages 666–675. Springer, 2012.

[47] Eugenio Aguirre, Miguel Garcia-Silvente, Antonio González, Rui Paúl, and Rafael Munyoz. A fuzzy system for detection of interaction demanding and nodding assent based on stereo vision. *Journal of Physical Agents*, 1(1):15–26, 2007.

[48] Jonas Ruesch, Manuel Lopes, Alexandre Bernardino, Jonas Hornstein, José Santos-Victor, and Rolf Pfeifer. Multimodal saliency-based bottom-up attention a framework for the humanoid robot iCub. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA'08)*, pages 962–967, May 2008.

[49] Hiroshi G. Okuno, Kazuhiro Nakadai, Ken Ichi Hidai, Hiroshi Mizoguchi, and Hiroaki Kitano. Human-robot interaction through real-time auditory and visual multiple-talker tracking. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'01)*, volume 3, pages 1402–1409, October 2001.

[50] Hiroshi G. Okuno, Kazuhiro Nakadai, and Hiroaki Kitano. Social interaction of humanoid robot based on audio-visual tracking. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pages 725–735. Springer, 2002.

[51] Yosuke Matsusaka, Tsuyoshi Tojo, Sentaro Kubota, Kenji Furukawa, Daisuke Tamiya, Keisuke Hayata, Yuichiro Nakano, and Tetsunori Kobayashi. Multi-person conversation via multi-modal interface - A robot who communicate with multi-user. In *EUROSPEECH*, volume 99, pages 1723–1726, 1999.

[52] Yosuke Matsusaka, Shinya Fujie, and Tetsunori Kobayashi. Modeling of conversational strategy for the robot participating in the group conversation. In *Interspeech*, volume 1, pages 2173–2176, 2001.

[53] Maren Bennewitz, Felix Faber, Dominik Joho, Michael Schreiber, and Sven Behnke. Towards a humanoid museum guide robot that interacts with multiple persons. In *5th IEEE/RAS International Conference on Humanoid Robots (Humanoids'05)*, pages 418–423, December 2005.

[54] Maren Bennewitz, Felix Faber, Dominik Joho, Michael Schreiber, and Sven Behnke. Integrating vision and speech for conversations with multiple persons. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'05)*, pages 2523–2528, August 2005.

[55] Tsuyoshi Tasaki, Shohei Matsumoto, Hayato Ohba, Mitsuhiko Toda, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno. Dynamic communication of humanoid robot with multiple people based on interaction distance. In *IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 71–76, August 2004.

[56] Sebastian Lang, Marcus Kleinehagenbrock, Sascha Hohenner, Jannik Fritsch, Gernot A. Fink, and Gerhard Sagerer. Providing the basis for human-robot interaction: A multi-modal attention system for a mobile robot. In *Proceedings of the 5th International Conference on Multimodal interfaces*, pages 28–35. ACM, 2003.

[57] Shuyin Li, Marcus Kleinehagenbrock, Jannik Fritsch, Britta Wrede, and Gerhard Sagerer. "BIRON, let me show you something": Evaluating the interaction with a robot companion. In *IEEE International Conference on Systems, Man and Cybernetics (SMC'04)*, volume 3, pages 2827–2834, October 2004.

[58] Axel Haasch, Sascha Hohenner, Sonja Hüwel, Marcus Kleinehagenbrock, Sebastian Lang, Ioannis Toptsis, Gernot A. Fink, Jannik Fritsch, Britta Wrede, and Gerhard Sagerer. BIRON - the Bielefeld robot companion. In *Proceedings of International Workshop on Advances in Service Robotics*, pages 27–32. Citeseer, Fraunhofer IRB Verlag, 2004.

[59] Liyuan Li, Qianli Xu, and Yeow Kee Tan. Attention-based addressee selection for service and social robots to interact with multiple persons. In *Proceedings of the Workshop at SIGGRAPH Asia*, pages 131–136. ACM, 2012.

[60] Hiroaki Kitano, Hiroshi G. Okuno, Kazuhiro Nakadai, Iris Fermin, Theo Sabisch, Yukiko Nakagawa, and Tatsuya Matsui. Designing a humanoid head for robocup challenge. In *Proceedings of the 4th International Conference on Autonomous Agents*, pages 17–18. ACM, 2000.

[61] Jannik Fritsch, Marcus Kleinehagenbrock, Sebastian Lang, Thomas Plötz, Gernot A. Fink, and Gerhard Sagerer. Multi-modal anchoring for human–robot interaction. *Robotics and Autonomous Systems*, 43(2):133–147, 2003.

[62] Thorsten Spexard, Axel Haasch, Jannik Fritsch, and Gerhard Sagerer. Human-like person tracking with an anthropomorphic robot. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA'06)*, pages 1286–1292, May 2006.

[63] David Payton, Mike Daily, Regina Estowski, Mike Howard, and Craig Lee. Pheromone robotics. *Autonomous Robots*, 11(3):319–324, 2001.

[64] Mike Daily, Youngkwan Cho, Kevin Martin, and Dave Payton. World embedded interfaces for human-robot interaction. In *Proceedings of the 36th Annual Hawaii International Conference on System Sciences (HICSS'03)*. IEEE Computer Society, 2003.

[65] Dennis Perzanowski, Alan C. Schultz, William Adams, Magda Bugajska, Elaine Marsh, J. Gregory Trafton, Derek Brock, Marjorie Skubic, and Myriam Abramson. Communicating with teams of cooperative robots. In *Multi-Robot Systems: From Swarms to Intelligent Automata*, pages 185–193. Springer, 2002.

[66] Gordon Briggs and Matthias Scheutz. Multi-modal belief updates in multi-robot human-robot dialogue interactions. In *Proceedings of the AISB/IACAP Symposium on Linguistic and Cognitive Approaches to Dialogue Agents (LaCATODA)*, pages 67–72, 2012.

[67] Tian Xu, Hui Zhang, and Chen Yu. Cooperative gazing behaviors in human multi-robot interaction. *Interaction Studies*, 14(3):390–418, 2013.

[68] Gaëtan Podevijn, Rehan O'Grady, Youssef SG. Nashed, and Marco Dorigo. Gesturing at subswarms: Towards direct human control of robot swarms. In *Conference Towards Autonomous Robotic Systems*, pages 390–403. Springer, 2013.

[69] Javier Alonso-Mora, Stefan Haegeli Lohaus, Philipp Leemann, Roland Siegwart, and Paul Beardsley. Gesture based human-multi-robot swarm interaction and its application to an interactive display. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA'15)*, pages 5948–5953, May 2015.

[70] Ayberk Özgur, Stéphane Bonardi, Massimo Vespignani, Rico Möckel, and Auke J. Ijspeert. Natural user interface for roombots. In *IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 12–17. IEEE, 2014.

[71] Michael Lichtenstern, Martin Frassl, Bernhard Perun, and Michael Angermann. A prototyping environment for interaction between a human and a robotic multi-agent system. In *Proceedings of the 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI'12)*, pages 185–186, March 2012.

[72] Brian Milligan. Selecting and commanding groups of robots using a vision-based natural user interface. Master's thesis, School of Computing Science, Simon Fraser University, 2012.

[73] Alex Couture-Beil, Richard T. Vaughan, and Greg Mori. Selecting and commanding individual robots in a vision-based multi-robot system. In *Proceedings of the IEEE Canadian Conference on Computer and Robot Vision (CRV'10)*, pages 159–166, May 2010.

[74] Valiallah Mani Monajjemi, Jens Wawerla, Richard T. Vaughan, and Greg Mori. HRI in the sky: Creating and commanding teams of UAVs with a vision-mediated gestural interface. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'13)*, pages 617–623, November 2013.

[75] Jawad Nagi, Alessandro Giusti, Luca M. Gambardella, and Gianni A. Di Caro. Human-Swarm Interaction using spatial gestures. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'14)*, pages 3834–3841, September 2014.

[76] Jawad Nagi, H. Ngo, Luca M. Gambardella, and Gianni A. Di Caro. Wisdom of the swarm for cooperative decision-making in human-swarm interaction. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA'15)*, pages 1802–1808, May 2015.

[77] Brian Scassellati. Imitation and mechanisms of joint attention: A developmental structure for building social skills on a humanoid robot. In *Computation for Metaphors, Analogy, and Agents*, pages 176–195. Springer, 1999.

[78] Aaron P. Shon, David B. Grimes, Chris L. Baker, Matthew W. Hoffman, Shengli Zhou, and Rajesh PN Rao. Probabilistic gaze imitation and saliency learning in a robotic head. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA'05)*, pages 2865–2870, April 2005.

[79] Matthew W. Hoffman, David B. Grimes, Aaron P. Shon, and Rajesh PN. Rao. A probabilistic model of gaze imitation and shared attention. *Neural Networks*, 19(3):299–310, 2006.

[80] Andrea Lockerd Thomaz, Matt Berlin, and Cynthia Breazeal. An embodied computational model of social referencing. In *IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 591–598, August 2005.

[81] Zeynep Yucel, Albert Ali Salah, Cetin Meriçli, and Tekin Meriçli. Joint visual attention modeling for naturally interacting robotic agents. In *the 24th International Symposium on Computer and Information Sciences, (ISCIS'09)*, pages 242–247. IEEE, September 2009.

[82] Charles Rich, Brett Ponsler, Aaron Holroyd, and Candace L Sidner. Recognizing engagement in human-robot interaction. In *Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI'10)*, pages 375–382, March 2010.

[83] Rowel Atienza and Alexander Zelinsky. Intuitive interface through active 3D gaze tracking. In *Proceedings of International Conference on Active Media Technology (AMT'05)*, pages 16–21. IEEE, 2005.

[84] Beatriz Oliveira, Pablo Lanillos, and Joao Filipe Ferreira. Gaze tracing in a bounded log-spherical space for artificial attention systems. In *Robot 2015: Second Iberian Robotics Conference*, pages 407–419. Springer, 2016.

[85] Yukie Nagai, Koh Hosoda, Akio Morita, and Minoru Asada. A constructive model for the development of joint attention. *Connection Science*, 15(4):211–229, 2003.

[86] Yukie Nagai. The role of motion information in learning human-robot joint attention. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA'05)*, pages 2069–2074, April 2005.

[87] Hidenobu Sumioka, Koh Hosoda, Yuichiro Yoshikawa, and Minoru Asada. Acquisition of joint attention through natural interaction utilizing motion cues. *Advanced Robotics*, 21(9):983–999, 2007.

[88] Brian Scassellati. Theory of mind for a humanoid robot. *Autonomous Robots*, 12(1):13–24, 2002.

[89] Boris Schauerte, Jan Richarz, and Gernot A Fink. Saliency-based identification and recognition of pointed-at objects. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'10)*, pages 4638–4643, October 2010.

[90] Nils Hofemann, Jannik Fritsch, and Gerhard Sagerer. Recognition of deictic gestures with context. In *Pattern Recognition*, pages 334–341. Springer, 2004.

[91] David Droeschel, Jörg Stückler, Dirk Holz, and Sven Behnke. Towards joint attention for a domestic service robot - person awareness and gesture recognition using time-of-flight cameras. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA'11)*, pages 1205–1210, May 2011.

[92] David Droeschel, Jörg Stückler, and Sven Behnke. Learning to interpret pointing gestures with a time-of-flight camera. In *Proceedings of the 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI'11)*, pages 481–488, March 2011.

[93] Jesus Suarez and Robin R. Murphy. Hand gesture recognition with depth images: A review. In *IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 411–417, August 2012.

[94] Michael Van den Bergh, Daniel Carton, Roderick De Nijs, Nikos Mitsou, Christian Landsiedel, Kolja Kuehnlenz, Dirk Wollherr, Luc Van Gool, and Martin Buss. Real-time 3D hand gesture interaction with a robot for understanding directions from humans. In *IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 357–362, August 2011.

[95] Dan Xu, Yen-Lun Chen, Chuan Lin, Xin Kong, and Xinyu Wu. Real-time dynamic gesture recognition system based on depth perception for robot navigation. In *the 2012 IEEE International Conference on Robotics and Biomimetics (ROBIO'12)*, pages 689–694. IEEE, 2012.

[96] Mohamad Bdiwi, Alexey Kolker, Jozef Suchỳ, and Alexander Winkler. Segmentation of model-free objects carried by human hand: Intended for human-robot interaction applications. In *the 16th IEEE International Conference onAdvanced Robotics (ICAR'13)*, pages 1–6, November 2013.

[97] Soohwan Kim, Dong Hwan Kim, and Sung-Kee Park. On-line object segmentation through human-robot interaction. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'10)*, pages 1734–1739, October 2010.

[98] Masato Ito and Jun Tani. Joint attention between a humanoid robot and users in imitation game. In *Proceedings of the International Conference on Development and Learning (ICDL)*, June 2004.

[99] Axel Haasch, Nils Hofemann, Jannik Fritsch, and Gerhard Sagerer. A multi-modal object attention system for a mobile robot. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'05)*, pages 2712–2717, August 2005.

[100] Yoshinori Kuno, Michie Kawashima, Keiichi Yamazaki, and Akiko Yamazaki. Importance of vision in human-robot communication understanding speech using robot vision and demonstrating proper actions to human vision. *Intelligent Environments*, pages 183–202, 2008.

[101] Frank Lömker and Gerhard Sagerer. A multimodal system for object learning. In *Pattern Recognition*, pages 490–497. Springer, 2002.

[102] Osamu Sugiyama, Takayuki Kanda, Michita Imai, Hiroshi Ishiguro, and Norihiro Hagita. Natural deictic communication with humanoid robots. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'07)*, pages 1441–1448, October 2007.

[103] Osamu Sugiyama, Takayuki Kanda, Michita Imai, Hiroshi Ishiguro, and Norihiro Hagita. Three-layer model for generation and recognition of attention-drawing behavior. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'06)*, pages 5843–5850, October 2006.

[104] Cynthia Breazeal, Guy Hoffman, and Andrea Lockerd. Teaching and working with robots as a collaboration. In *Proceedings of the 3rd International Joint Conference on Autonomous Agents and Multiagent Systems*, pages 1030–1037. IEEE Computer Society, 2004.

[105] Robin R. Murphy and Debra Schreckenghost. Survey of metrics for human-robot interaction. In *Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI'13)*, pages 197–198, March 2013.

[106] Astrid Weiss, Regina Bernhaupt, Michael Lankes, and Manfred Tscheligi. The USUS evaluation framework for human-robot interaction. In *Proceedings of the Symposium on New Frontiers in Human-Robot Interaction (AISB'09)*, volume 4, pages 11–26, 2009.

[107] Edward Clarkson and Ronald C. Arkin. Applying heuristic evaluation to human-robot interaction systems. In *Flairs Conference*, pages 44–49, 2007.

[108] Jennifer Casper and Robin R. Murphy. Human-robot interactions during the robot-assisted urban search and rescue response at the World Trade Center. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 33(3):367–385, 2003.

[109] Aaron Steinfeld, Terrence Fong, David Kaber, Michael Lewis, Jean Scholtz, Alan Schultz, and Michael Goodrich. Common metrics for human-robot interaction. In *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction (HRI'06)*, pages 33–40. ACM, 2006.

[110] Brian R. Duffy. Anthropomorphism and the social robot. *Robotics and Autonomous Aystems*, 42(3):177–190, 2003.

[111] Douglas Adams. *The Hitchhiker's Guide to the Galaxy.* Pan Macmillan, 1979.

[112] Rainer Stiefelhagen, C. Fugen, R. Gieselmann, Hartwig Holzapfel, Kai Nickel, and Alex Waibel. Natural human-robot interaction using speech, head pose and gestures. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'04)*, volume 3, pages 2422–2427, October 2004.

[113] Xiao Wang, Xavier Clady, and Consuelo Granata. A human detection system for proxemics interaction. In *Proceedings of the 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI'11)*, pages 285–286, March 2011.

[114] James McLurkin, Jennifer Smith, James Frankel, David Sotkowitz, David Blau, and Brian Schmidt. Speaking Swarmish: Human-robot interface design for large swarms of autonomous mobile robots. In *AAAI Spring Symposium: To Boldly Go Where No Human-Robot Team Has Gone Before*, pages 72–75, 2006.

[115] Jun Kato, Daisuke Sakamoto, Masahiko Inami, and Takeo Igarashi. Multi-touch interface for controlling multiple mobile robots. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'09), Extended Abstracts*, pages 3443–3448. ACM, 2009.

[116] David Payton, Regina Estkowski, and Mike Howard. Pheromone robotics and the logic of virtual pheromones. In *International Workshop on Swarm Robotics*, pages 45–57. Springer, 2004.

[117] Amir M. Naghsh, Jeremi Gancet, Andry Tanoto, and Chris Roast. Analysis and design of human-robot swarm interaction in firefighting. In *IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 255–260, August 2008.

[118] Shengdong Zhao, Koichi Nakamura, Kentaro Ishii, and Takeo Igarashi. Magic cards: A paper tag interface for implicit robot control. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'09)*, pages 173–182. ACM, 2009.

[119] Anqi Xu, Gregory Dudek, and Junaed Sattar. A natural gesture interface for operating robotic systems. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA'08)*, pages 3557–3563, May 2008.

[120] Antoine Deleforge and Radu Horaud. The cocktail party robot: Sound source separation and localisation with an active binaural head. In *Proceedings of the 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI'12)*, pages 431–438, March 2012.

[121] Brian Milligan, Greg Mori, and Richard T. Vaughan. Selecting and commanding groups in a multi-robot vision based system. In *Proceedings of the 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI'11) (Video Session)*, pages 415–415, March 2011.

[122] Sushmita Mitra and Tinku Acharya. Gesture recognition: A survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 37(3):311–324, 2007.

[123] Stefan Waldherr, Roseli Romero, and Sebastian Thrun. A gesture based interface for human-robot interaction. *Autonomous Robots*, 9(2):151–173, 2000.

[124] Matthew M. Loper, Nathan P. Koenig, Sonia H. Chernova, Chris V. Jones, and Odest C. Jenkins. Mobile human-robot teaming with environmental tolerance. In *Proceedings of the 4th ACM/IEEE International Conference on Human-Robot Interaction (HRI'09)*, pages 157–164. ACM, 2009.

[125] David Kortenkamp, Eric Huber, and R Peter Bonasso. Recognizing and interpreting gestures on a mobile robot. In *Proceedings of the National Conference on Artificial Intelligence*, pages 915–921. AAAI Press/The MIT Press, 1996.

[126] Christian Martin, Frank-Florian Steege, and Horst-Michael Gross. Estimation of pointing poses for visually instructing mobile robots under real world conditions. *Robotics and Autonomous Systems*, 58(2):174–185, 2010.

[127] Zahar Prasov. Shared gaze in remote spoken HRI during distributed military operations. In *Proceedings of the 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI'12)*, pages 211–212, March 2012.

[128] Morgan Quigley, Ken Conley, Brian Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, and Andrew Y Ng. ROS: An open-source Robot Operating System. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA'09), workshop on open source software*, May 2009.

[129] Gary Bradski and Adrian Kaehler. *Learning OpenCV: Computer vision with the OpenCV library.* " O'Reilly Media, Inc.", 2008.

[130] Paul Viola and Michael J Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.

[131] David Huggins-Daines, Mohit Kumar, Arthur Chan, Alan W. Black, Mosur Ravishankar, and Alexander I. Rudnicky. Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices. In *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing*, volume 1. IEEE, 2006.

[132] Mark Draper, Gloria Calhoun, Heath Ruff, David Williamson, and Timothy Barry. Manual versus speech input for unmanned aerial vehicle control station operations. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 47, pages 109–113. SAGE Publications, 2003.

[133] Raja Parasuraman, Scott Galster, Peter Squire, Hiroshi Furukawa, and Christopher Miller. A flexible delegation-type interface enhances system performance in human supervision of multiple robots: Empirical studies with RoboFlag. *IEEE Transactions on Systems, Man, and Cybernetics - part A: Systems and Humans*, 35(4):481–493, 2005.

[134] Mark Micire, Munjal Desai, Amanda Courtemanche, Katherine M. Tsui, and Holly A. Yanco. Analysis of natural gestures for controlling robot teams on multi-touch tabletop surfaces. In *Proceedings of the ACM International Conference on Interactive Tabletops and Surfaces*, pages 41–48. ACM, 2009.

[135] Mary L. Cummings and Paul J. Mitchell. Predicting controller capacity in supervisory control of multiple UAVs. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 38(2):451–460, 2008.

[136] Michael A. Goodrich, Morgan Quigley, and Keryl Cosenzo. Task switching and multi-robot teams. In *Multi-Robot Systems. From Swarms to Intelligent Automata Volume III*, pages 185–195. Springer, 2005.

[137] Jacob W. Crandall and Mary L. Cummings. Identifying predictive metrics for supervisory control of multiple robots. *IEEE Transactions on Robotics*, 23(5):942–951, 2007.

[138] Curtis M. Humphrey, Christopher Henk, George Sewell, Brian W. Williams, and Julie A. Adams. Assessing the scalability of a multiple robot interface. In *Proceedings of the 2nd ACM/IEEE International Conference on Human-Robot Interaction (HRI'07)*, pages 239–246. IEEE, 2007.

[139] Peter Squire, Greg Trafton, and Raja Parasuraman. Human control of multiple unmanned vehicles: Effects of interface type on execution and task switching times. In *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction (HRI'06)*, pages 26–32. ACM, 2006.

[140] Sachi Mizobuchi and Michiaki Yasumura. Tapping vs. circling selections on pen-based devices: evidence for different performance-shaping factors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'04)*, pages 607–614. ACM, 2004.

[141] Dan R. Olsen Jr and Michael A. Goodrich. Metrics for evaluating human-robot interactions. In *Proceedings of the NIST Performance Metrics for Intelligent Systems Workshop*, 2003.

[142] Brian Milligan, Greg Mori, and Richard T. Vaughan. Selecting and commanding groups of robots in a vision based multi-robot system. In *Proceedings of the 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI'11) (Video Session)*, 2011.

[143] Marjorie Skubic, Derek Anderson, Samuel Blisard, Dennis Perzanowski, and Alan Schultz. Using a hand-drawn sketch to control a team of robots. *Autonomous Robots*, 22(4):399–410, 2007.

[144] Vivian Chu, Kalesha Bullard, and Andrea L Thomaz. Multimodal real-time contingency detection for HRI. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems(IROS'14)*, pages 3327–3332, September 2014.

[145] Shogo Nabe, Takayuki Kanda, Kazuo Hiraki, Hiroshi Ishiguro, Kiyoshi Kogure, and Norihiro Hagita. Analysis of human behavior to a communication robot in an open field. In *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction (HRI'06)*, pages 234–241. ACM, 2006.

[146] Masamitsu Murase, Jean-Marc Valin Shun'ichi Yamamoto, Jean-Marc Valin, Kazuhiro Nakadai, Kentaro Yamada, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno. Multiple moving speaker tracking by microphone array on mobile robot. In *INTERSPEECH*, pages 249–252, 2005.

[147] Benjamin Kühn, Boris Schauerte, Kristian Kroschel, and Rainer Stiefelhagen. Multimodal saliency-based attention: A lazy robot's approach. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'12)*, pages 807–814, October 2012.

[148] Keng Peng Tee, Rui Yan, Yuanwei Chua, Zhiyong Huang, and Somchaya Liemhetcharat. Gesture-based attention direction for a telepresence robot: Design and experimental study. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'14)*, pages 4090–4095, September 2014.

[149] Nicola Bellotto and Huosheng Hu. Multisensor-based human detection and tracking for mobile service robots. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(1):167–181, 2009.

[150] Christian Martin, Erik Schaffernicht, Andrea Scheidig, and Horst-Michael Gross. Multi-modal sensor fusion using a probabilistic aggregation scheme for people detection and tracking. *Robotics and Autonomous Systems*, 54(9):721–728, 2006.

[151] Thierry Germa, Frédéric Lerasle, Noureddine Ouadah, Viviane Cadenat, and Michel Devy. Vision and RFID-based person tracking in crowds from a mobile robot. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'09)*, pages 5591–5596, October 2009.

[152] Steffen Muller, Sven Hellbach, Erik Schaffernicht, Antje Ober, Andrea Scheidig, and Horst-Michael Gross. Whom to talk to? Estimating user interest from movement trajectories. In *IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 532–538, August 2008.

[153] Jake Bruce, Jens Wawerla, and Richard T. Vaughan. Human-robot rendezvous by cooperative trajectory signals. In *Proceedings of the 10th ACM/IEEE International Conference on Human-Robot Interaction, Workshop on Human-Robot Teaming*, March 2015.

[154] Alberto Elfes. Occupancy grids: A stochastic spatial representation for active robot perception. *arXiv preprint arXiv:1304.1098*, 2013.

[155] João Xavier, Marco Pacheco, Daniel Castro, António Ruano, and Urbano Nunes. Fast line, arc/circle and leg detection from laser scan data in a player driver. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA'05)*, pages 3930–3935, April 2005.

[156] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893. IEEE, 2005.

[157] Gary B Bradski. *Open source computer vision library*. Springer, 2004.

[158] Ralph Schmidt. Multiple emitter location and signal parameter estimation. *IEEE transactions on antennas and propagation*, 34(3):276–280, 1986.

[159] Kazuhiro Nakadai, Hiroshi G. Okuno, Hirofumi Nakajima, Yuji Hasegawa, and Hiroshi Tsujino. An open source software system for robot audition HARK and its evaluation. In *8th IEEE/RAS International Conference on Humanoid Robots (Humanoids'08)*, pages 561–566, December 2008.

[160] Manabu Saito, Kimitoshi Yamazaki, Naotaka Hatao, Ryo Hanai, Kei Okada, and Masayuki Inaba. Pedestrian detection using a LRF and a small omni-view camera for outdoor personal mobility robot. In *the IEEE International Conference on Robotics and Biomimetics (ROBIO'10)*, pages 155–160. IEEE, 2010.

[161] Mani Monajjemi, Sepehr MohaimenianPour, and Richard T. Vaughan. UAV, come to me: End-to-end, multi-scale situated HRI with an uninstrumented human and a distant UAV. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'16)*, October 2016.

[162] Jake Bruce, Valiallah Mani Monajjemi, Jens Wawerla, and Richard T. Vaughan. Tiny people finder: Long-range outdoor HRI by periodicity detection. In *2016 Canadian Conference on Computer and Robot Vision, (CRV'16)*, 2016.

[163] Matthew Marge, Aaron Powers, Jonathan Brookshire, Trevor Jay, Odest C Jenkins, and Christopher Geyer. Comparing heads-up, hands-free operation of ground robots to teleoperation. *Robotics: Science and Systems VII*, 2011.

[164] Astrid Weiss. *Validation of an evaluation framework for human-robot interaction: The impact of usability, social acceptance, user experience, and societal impact on collaboration with humanoid robots.* PhD thesis, University of Salzburg, 2010.

[165] Nikolas Martelaro. Wizard-of-Oz interfaces as a step towards autonomous HRI. In *2016 AAAI Spring Symposium Series*, 2016.

[166] Sandra G. Hart and Lowell E. Staveland. Development of NASA-TLX (task load index): Results of empirical and theoretical research. *Advances in Psychology*, 52:139–183, 1988.

[167] Lee J Cronbach. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3):297–334, 1951.

[168] Jum Nunnally. Psychometric Theory. *New York: McGraw-Hill*, 1978.