Reasoning About Spatial Patterns of Human Behavior During Group Conversations with Robots

Marynel Vázquez

CMU-RI-TR-17-XX

Thesis Committee:
Aaron Steinfeld, CMU RI (co-chair)
Scott E. Hudson, CMU HCII (co-chair)
Kris Kitani, CMU RI
Brian Scassellati, Yale University

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Robotics Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213

July 2017

Abstract

The goal of this dissertation is to *develop computational models for robots to detect and sustain the spatial patterns of behavior that naturally emerge during free-standing group conversations with people.* These capabilities have often been overlooked by the Human-Robot Interaction (HRI) community, but they are essential for robots to appropriately interact with and around people in many human environments.

In this work, we first develop a robotic platform for studying human-robot interactions, and contribute new experimental protocols to investigate group conversations with robots. The studies that we conducted with these protocols examine various aspects of these interactions, and experimentally validate the idea that people tend to establish spatial formations typical of human conversations with robots. These formations emerge as the members of the interaction cooperate to sustain a single focus of attention. They maximize their opportunities for monitoring one another's mutual perceptions during conversations.

Second, we introduce a general framework to track the lower-body orientation of free-standing people in human environments and to detect their conversational groups based on their spatial behavior. This framework takes advantage of the mutual dependency between the two problems. Lower-body orientation is a key descriptor of spatial behavior and, thus, can help detect group conversations. Meanwhile, knowing the location of group conversations can help estimate people's lower-body orientation, because these interactions often bias human spatial behavior. We evaluate this framework in a public computer vision benchmark for group detection, and show how it can be used to estimate the members of a robot's group conversation in real-time.

Third, we study how robots should orient with respect to a group conversation to cooperate to sustain the spatial arrangements typical of these interactions. To this end, we conduct an experiment to study the effects of varied orientation and gaze behaviors for robots during social conversations. Our results reinforce the importance of communicative motion behavior for robots, and suggest that their body and gaze behaviors should be designed and controlled jointly, rather than independently of each other. We then show in simulation that it is possible to use reinforcement learning techniques to generate socially appropriate orientation behavior for robots during group conversations. These techniques reduce the amount of engineering required to enable robots to sustain spatial formations typical of conversations while communicating attentiveness to the focus of attention of the interaction.

Overall, our efforts show that reasoning about spatial patterns of behavior is useful for robots. This reasoning can help with perception tasks as well as generating appropriate robot behavior during social group conversations.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The success of many robots in unstructured, human environments hinges on their ability to appropriately interact with and around people [42, 127]. Appropriate interactions are especially important for mobile platforms for professional or domestic use. These robots are becoming very popular and their sales are expected to increase in the period of 2016-2019 [131]. This increase follows a generalized trend across the robotics market, where many more robots are manufactured and sold year after year.

Generating socially appropriate robot behavior often requires the ability to detect on-going conversations and their members in human environments. For example, consider Figure 1.1a in which two people talk to each other. Even though the robot is not part of the interaction, it is important for it to know that the conversation is happening to avoid interrupting inappropriately, e.g., by navigating across the members of the group. Now consider Figure 1.1b, where the robot is engaged in a conversation with four other participants. If the robot was able to detect the members of its group, then it could acknowledge when other people joined the interaction, as we normally do in human-human social encounters. The robot could also adjust to people leaving its conversation and potentially react to try to re-engage them.

(a) (b)

Figure 1.1: Group conversations with and without a robot. In (a), two people (outlined in white) converse with one another. The robot (dashed yellow box) is not part of their interaction. In (b), everybody is part of a conversation.

The problem of detecting social conversations and their members has often been overlooked by the HRI community. Many methods have been proposed for important, related problems, including enabling robots to take part in turn-taking processes during conversations [116, 182],

estimating users' engagement and willingness to interact [31, 49, 100, 104, 121, 145, 177], and detecting groups of people that move together in human environments [105, 109]. However, these methods do not address detecting conversations specifically. One reason for this gap is the heavy focus of the community on studying one-on-one human-robot interactions. In these contexts, researchers often assume that the user of interest can only – and is always – interacting with the robot. This simplification has often made the research community ignore the problem of detecting on-going conversations.

This dissertation suggests to embrace dynamic group conversations in HRI. Understanding these type of interactions is essential to enable robots to interact in complex human environments. With this objective in mind, we work towards enabling robots to reason about human spatial behavior typical of group conversations. This ability provides a path for detecting these types of interactions as well as enabling robots to adapt to changes in their social context. More specifically,

> *the goal of this dissertation is to develop computational models for robots to detect and sustain the spatial patterns of human behavior that naturally emerge during free-standing group conversations.*

Even though human spatial behavior is malleable during conversations, it possesses a particular structure that often makes conversations between free-standing people distinctive from other nearby social activities. The structure arises for two main reasons. First, human spatial behavior can effectively convey social expectations and intentions [86]. Second, conversations have an intrinsic communicative purpose that heavily influences the way that people stand with respect to one another during these interactions. Erving Goffman [58] defined conversations as a jointly focused encounter where the participants gather together and openly cooperate to sustain a single focus of attention. This cooperation can be observed in human verbal behavior, e.g., through turn-taking, but also in non-verbal behavior. During conversations, people position and orient themselves to maximize their opportunity to monitor one another's mutual perceptions. The result are spatial organizations where an "eye-to-eye ecological huddle" tends to be maintained [85].

Our efforts build upon prior work on surveillance and automated human behavior analysis from the fields of computer vision and multi-modal interaction. Researchers from these areas proposed the first computational models of human spatial behavior typical of group conversations [40, 54, 74]. They also showed that it is possible to detect conversations by reasoning about this type of human behavior. We validate these ideas in the context of HRI with a new mobile robot that we built as part of this dissertation. Furthermore, we propose a novel framework to detect group conversations between free-standing people and robots. Our main insight is that we can exploit the mutual dependency between two problems for this task. These problems are:

1. detecting group conversations by reasoning about human spatial behavior; and
2. tracking people's lower-body orientation in a scene.

The second problem is very important for the first one because lower-body orientation is a key descriptor of human spatial behavior. But group conversations often bias the spatial behavior of nearby people as well, because conversations are natural social attractors. This means that the location of these interactions can serve as a strong prior for tracking people's lower-body orientation.

Several challenges arise upon studying group conversations in HRI. From a sensing perspective, the more people are around a robot, the more noisy and incomplete sensor data tends to be. In this work, we deal with measurement uncertainty by designing the core elements of our framework for detecting group conversations based on probabilistic methods.

From an experimental perspective, the more people interact with one another, the more dynamic human behavior tends to be. This variability can lead to confounds and unreliable estimates on the effects of behavioral manipulations. For example, in dyadic interactions it is important that both parties are involved in the conversation, or the interaction ends. But in bigger group conversations, people can often leave or join the interaction dynamically. This added flexibility can often induce a *bystander effect* in HRI studies, where people become bystanders or spectators of human-robot interactions, instead of being active participants.[1] This effect can then make it difficult to systematically study human conversations with robots and related phenomena, because it can induce participants to disregard interactions during experiments. Our key methodological insight to reduce the bystander effect in this work was to design experimental protocols that assigned active roles to the participants. These protocols induced people to take part in activities that involve social human-robot interactions, without specifically telling them how to behave – which is essential to observe naturalistic social phenomena during controlled HRI experiments. These protocols can be used to further study various aspects of group conversations beyond the scope of this dissertation.

Reasoning about spatial behavior during conversations raises another important computational challenge because this behavior is inter-dependent. Not only people's pose affect one another, but also robots can influence human spatial behavior and vice-versa. Naively taking into consideration all these different interactions can easily become prohibitively expensive for HRI applications. To address this problem, the framework that we proposed to detect group interactions assumes that people and robots are equal from a social perspective. This leads to a unified viewpoint on spatial behavior that facilitates reasoning about how the motion and pose of social actors relate to one another. We demonstrate this perspective during a user experiment where it was essential for a robot to detect the members of its group conversation in real-time.

We limit the scope of this dissertation to studying spatial behavior typical of group conversations were people are standing and quasi-static, as it often happens during social encounters in public, open spaces. In our HRI experiments, people rarely move during conversations unless there is a good reason for it (e.g., because their current activity requires it, or they engage in other interactions or tasks). Detecting conversations in many other circumstances and through other communication modalities, e.g., speech [45], gaze [84], gestures [171], and posture [39], is an important task that is closely related to our work [27, 46, 70, 107, 136]. However, studying these interactions and other modalities of communication is out of the scope of this dissertation. We discuss future avenues of research in this direction at the end of this dissertation.

[1] To better illustrate the bystander effect, think about a conversation between a robot a multiple people. If the robot addresses the whole group, then any participant can reply. Those who are shy or not very motivated to converse with the robot then have a good excuse to stay quiet and become simple spectators or bystanders. Why should they be active participants? After all, other people can take the lead of the conversation.

We associate the bystander effect with social loafing, a phenomena characterized by a reduction in individual effort when people are in groups [68]. One cause for this phenomena is the diffusion of responsibility amongst the members of a group. People in groups may also perceive a reduced sense of being evaluated as an individual, or no direct link between their personal efforts and success in a task. Moreover, social loafing may occur when the value of a goal is contingent on the members of a group.

1.1 Outline of Approach

We use an iterative research methodology, as depicted in Figure 1.2, to study social group conversations, develop new robot capabilities, and evaluate the effectiveness of the proposed approaches for HRI. This methodology also drives the organization of the rest of this document.

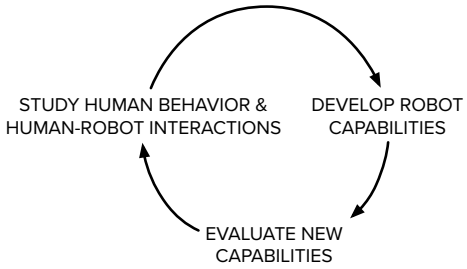


Figure 1.2: Research methodology

(Chapter 2 & 3) Background & Related Work. We start by reviewing relevant background from social psychology, including foundational work on human-human conversations and spatial patterns of human behavior. We then describe related work within HRI and survey computational models for detecting group conversations, especially methods that reason about spatial formations typical of these interactions. These prior efforts informed the development of new robot capabilities in this work.

(Chapter 4) The Furniture Robot Chester and Its Companion Blink. To study human-robot interactions, we designed and built a flexible robotic platform in collaboration with Disney Research. The particular design of the robot was driven by the need to control human expectations for robot characteristics, which is important to avoid issues like the uncanny valley [122]. One of the unique features of the robot is the fact that it can operate as one or two characters simultaneously, depending on the needs of a particular interaction. In this work, we used this platform to conduct three different user experiments, like the one depicted in Figure 1.1. The results from these efforts reinforced prior findings within HRI that suggest that the anthropomorphization of household objects can produce positive engagement effects in users [133, 215]. An overview of the protocols that we devised for these experiments was presented in the Robots In Groups and Teams Workshop at CSCW 2017 [195],

> *"Methods for Studying Group Interactions in HRI,"* M. Vázquez, E. J. Carter, J. Forlizzi, S. E. Hudson, and A. Steinfeld. In Robots In Groups and Teams, CSCW, 2017.

(Chapter 5) Examining Spatial Behavior During Group Conversations. We used our first two user experiments to validate the idea that people tend to establish spatial formations typical of human group conversations with social robots. The first experiment was a field trial, in which we tested our custom robotic platform with groups of children. As part of this effort, we studied various user engagement cues, like spatial behavior, and the effects of a sidekick character in HRI. This experiment was published in the 2014 ACM/IEEE International Conference on Human-Robot Interaction [197],

> *"Spatial and Other Social Engagement Cues in a Child-robot Interaction: Effects of a Sidekick,"* M. Vázquez, A. Steinfeld, S. E. Hudson, and J. Forlizzi. In HRI, 2014.

Because the bystander effect emerged during this experiment, we designed another more controlled protocol to further study social conversations and spatial behavior. This new protocol induced the participants to actively interact with our robot by involving them in a social role-playing game. This game provided an opportunity to study how different roles for the robot could affect human spatial behavior and its interaction with groups of adults. Furthermore, this study served to collect sensor data from which we could start exploring how to enable robots to detect group conversations and their members. A short summary of this work appeared in the 10th Annual ACM/IEEE International Conference on Human-Robot Interaction [198],

> *"Social Group Interactions in a Role-Playing Game,"* M. Vázquez, E. J. Carter, J. A. Vaz, J. Forlizzi, A. Steinfeld, and S. E. Hudson. In HRI (Ext. Abstracts), 2015.

(Chapter 6) Detecting Conversations by Reasoning About Spatial Behavior. Based on our experience conducting the previous experiments, we designed a probabilistic framework to detect group conversations between free-standing people and robots. The core of this framework is an alternating optimization approach that leverages the mutual dependency between two tasks: detecting group conversations based on spatial behavior, and tracking the lower-body orientation of people in a scene. We evaluated this approach in a established dataset from computer vision that is often used to compare methods to detect group conversations between free-standing people[155]. Our results suggest that the proposed framework can help better detect non-interacting people in social environments, like bystanders, without sacrificing group detection performance. This group detection framework was first published in the 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems [199],

> *"Parallel Detection of Conversational Groups of Free-Standing People and Tracking of their Lower-Body Orientation,"* M. Vázquez, A. Steinfeld, and S. E. Hudson. In IROS, 2015.

(Chapter 7) On the Effects of Body Orientation & Gaze. With the proposed group detection framework, we then conducted another user experiment on the effects of varied orientation and gaze behaviors for robots during social conversations. This effort provides a concrete example on how our group detection approach can be implemented for HRI applications, and demonstrates that it can enable the execution of autonomous robot behaviors.

For this third user experiment, we designed another experimental protocol to implicitly induce participants to take part in human-robot interactions. In this case, people were involved in a brainstorming session with a robot, and they tried to solve a problem altogether. In contrast to the role-playing game, the brainstorming activity is less controlled and adversarial in general. It also does not require teaching very specific instructions to the participants.

The results from the brainstorming experiment reinforced the importance of communicative motion in HRI [44, 152, 178]. In addition, they showed that a robot's gaze can influence users' perception of its motion during conversations. But its motion can also influence the perception of its gaze. This mutual dependency suggests that robot gaze and body motion must be designed and controlled jointly, rather than independently of each other. We published these findings in the 2017 ACM/IEEE International Conference on Human-Robot Interaction [196],

> *"Towards Robot Autonomy in Group Conversations: Understanding the Effects of Body Orientation and Gaze,"* M. Vázquez, E. J. Carter, B. McDorman, J. Forlizzi, A. Steinfeld, and S. E. Hudson. In HRI, 2015.

(Chapter 8) Learning to Control Robot Orientation During Conversations. The success of the orientation behaviors that we studied in our last experiment motivated us to explore reinforcement learning techniques to generate socially appropriate orientation behaviors for robots during multi-party conversations. This exploration aimed to reduce the amount of engineering required to enable appropriate spatial behavior with noisy multi-modal sensor data. Our results from tests in a simulated environment suggested that a new state representation that we designed for this problem can be used to find good control policies for mobile robots. Moreover, these policies have the potential to generalize across conversations with different numbers of people. This work was published in the 25th IEEE International Symposium on Robot and Human Interactive Communication [194],

> *"Maintaining Awareness of the Focus of Attention of a Conversation: A Robot-Centric Reinforcement Learning Approach,"* M. Vázquez, A. Steinfeld, and S. E. Hudson. In RO-MAN, 2016.

(Chapter 9) Conclusion. We conclude this dissertation with a brief summary of our contributions and a discussion of future research directions.

Chapter 2

Background

This chapter reviews relevant background from social psychology. This foundation on human gatherings and spatial patterns of human behavior drives the design of computational models of spatial formations typical of group conversations, as later described in Chapters 3 and 6. This background was also essential for our effort to develop appropriate behaviors for robots during group conversations.

2.1 A Social Psychology Perspective on Human Gatherings

Erving Goffman [58] described *gatherings* as a set of individuals who are in one another's immediate presence. Gatherings can be of two types, as illustrated in Fig. 2.1. *Unfocused gatherings* are associated with the management of mere co-presence, e.g., pedestrians on a street, strangers waiting at a bus stop, etc. *Focused gatherings* are instead characterized by people coming together to sustain a single focus of attention.

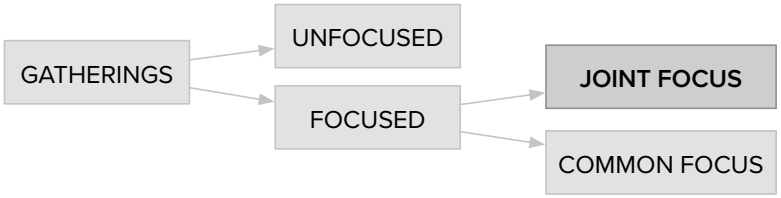


Figure 2.1: Types of human gatherings [58, 85]. Conversations, in particular, fall within the class of jointly focused gatherings.

If there is a joint responsibility between the people in a gathering to cooperate to sustain a focus of attention, then the interaction is said to be a *jointly focused gathering* [85]. This category includes conversations, tennis games, dancing couples, pairs of workers co-operating to solve a task which requires sustained attention, psychotherapy sessions, etc. Other focused encounters where there is no shared cooperation are known as *common focused gatherings*. Typical examples include a platoon on a parade, pupils paying attention to what a teacher says in a classroom, or a guided museum tour.

Some information is *given* voluntarily during gatherings, such as the content of what people say, while other information is *given off* whether the interactants choose to provide it or not. As explained by Kendon [85], the latter is an inevitable and unavoidable product of people's presence and of their actions. For example, people might provide additional information through their manner of talk, accent, gaze, or their body posture. The ecological arrangements that

emerge during conversations are also very informative, as further discussed in the next Section. While these aspects may seem unimportant in comparison to the information that is given voluntarily, they play a crucial role in structuring interactions.

2.2 Spatial Patterns of Human Behavior During Conversations

Situated human conversations are the most common type of jointly focused gatherings. The members of these interactions converse in one another's immediate presence and cooperatively sustain their focus of cognitive and visual attention. They pursue a common line of concern, where the topic is jointly created and sustained. When a participant has the turn to speak but (s)he doesn't or can't, conversations often end.

During conversations among free-standing people, the participants position and orient themselves such that they have equal, direct, and exclusive access to the space between them. People maneuver in relation to one another to create a sort of "no-man's land", and maintain a separate world from their surrounding [85]. The result is a distinct spatial organization, typically known as a face formation or ***F-formation*** within social psychology [86]. This organization maximizes the opportunity of the interactants to monitor one another during conversations and maintains their group as a spatially distinct unit from other nearby interactions.[1]

F-formations begin when the members of a group position themselves such that their *transactional segments* intersect (as in Fig. 2.2a). These segments extend in front of each person and encompass the physical space that they are using for their current activity. Transactional segments are the space into which they look and speak, or into which they reach to handle objects. People will work to maintain their transactional segment free of intrusions for as long as they are engaged in an activity that requires it.

The physical area where the transactional segments of the members of a conversation intersect is known as the *o-space* of the corresponding F-formation [86]. The o-space of dyads standing in a face-to-face arrangement is in-between the participants (as in Fig. 2.2a). During side-by-side or "L" arrangements, the o-space tends to be in front of the members of the conversation (Fig. 2.2b and 2.2c). Bigger groups tend to form semi-circular or circular arrangements with their o-space towards the center of the circle (Fig. 2.2d and 2.2e).

A transition from a conversation into another type of interaction, or vice-versa, is often visible in the spatial organization of the participants. For example, F-formations often transform into a less uniform spatial arrangement when a conversation shifts into a common focus encounter [86, 112]. When the focus of attention becomes a particular person, a separation between this interactant and the rest of the group is often observed due to a difference in social status or role. When no particular spatial arrangement is observed in common focus encounters, the group is said to be organized in a cluster.

[1] Interestingly, similar spatial organizations have been observed in cases where people are seated in an open space and can adjust the position their chairs to hold conversations with one another [19].

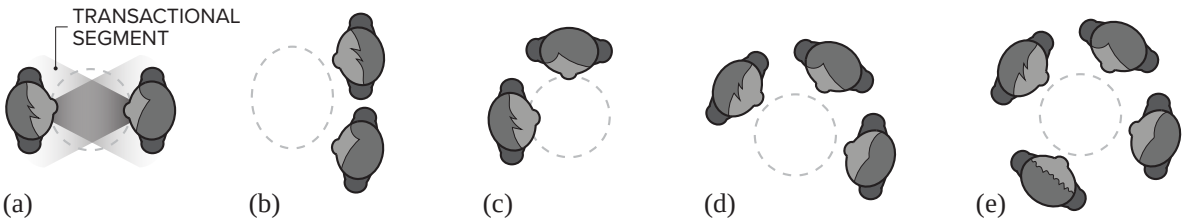


Figure 2.2: Spatial arrangements typical of F-formations. Dashed areas represent o-spaces.

Chapter 3

Related Work

This chapter first reviews related work within the field of Human-Robot Interaction, especially on the social use of space. Then, it describes prior efforts to detect human interactions and to model F-formations computationally. Our efforts build upon this related work.

3.1 Proxemics & Spatial Formations in Human-Robot Interactions

Significant prior research has focused on proxemics or the study of how we, humans, use and perceive space [60]. Many factors can influence human proxemics, including social norms, peoples' familiarity with one another, and to what degree people are interacting [14]. Environmental conditions, such as lighting [2], can alter proxemics as well. In the particular case of HRI, the study of proxemics has become increasingly important as robots become more and more capable of interacting in public settings [47, 77, 96, 160, 183, 189]. Several factors can influence the level of comfort that people have with robots and, thus, the distance that they like to maintain from them. These factors include a robot's voice and height, the direction from which a robot approaches users, mutual gaze, users' previous experience with robots and pets, gender, age, and personality [124, 179, 205, 206, 207, 208].

Proxemics has been a key factor in social robot navigation [95]. Various approaches have been designed to incorporate social conventions into the way that robots move within human environments. One approach to achieve this goal is to model social conventions as social forces that steer a robot as it navigates nearby people [51, 113]. Another approach is to model social conventions as cost functions (or potential fields) in navigation plans [88, 108, 125, 147, 154, 176, 188]. In particular, [147] not only incorporated costs into their planning algorithm to avoid violating personal space, but also added costs to avoid crossing o-spaces corresponding to group conversations. Besides, other work has developed custom strategies for specific scenarios, such as cases where robots give museum tours [28, 219] or stand in line [129].

Other efforts have focused specifically on enabling robots to appropriately handle proxemics upon initiating interactions with users [11, 81, 150, 153, 157]. For example, Althaus et al. [11] adjusted a robot's speed and orientation to make it face towards the middle of the group of people that was going to interact with. Kato et al. [81] instead proposed that robots should mimic service staff in a mall to decide whether or not they should start an interaction with people. That is, robots should try to initiate interactions only if users exhibit an analogous intention. To implement this behavior, the authors created a classifier to estimate whether people intended to interact with a robot based on their motion trajectories. The output of this classifier was then used to decide if a robot should proactively approach nearby people, only

turn its body towards users to signal that it was available to start an interaction, or wait idly without worrying about interacting with anybody.

Proxemics has also been used to estimate users' engagement levels with robots [118, 120, 121, 123, 153, 159]. For example, Michalowski et al. [121] proposed a categorical model of engagement, based on the distance that users kept from a robot and their head pose. Satake et al. [153] instead relied on users' motion trajectory to decide if they were engaged with a robot and accepted interacting with it. Moreover, Mead and Matarić [118, 120] leveraged psychophysical factors that provide a functional sensory explanation to proxemic behavior for detecting the initiation and termination of dyadic and triadic interactions. Inspired by these prior efforts, we also use proxemics in this dissertation as a cue for social engagement.

F-formations have been observed in HRI, specially during dyadic social interactions. For example, Huettenrauch et al. [72] observed that people tend to sustain face-to-face, side-by-side of L-shaped arrangements with a social robot during conversations. Furthermore, Kuzuoka et al. [97] provided evidence that suggests that robots can induce subtle reconfigurations of F-formations by rotating their body. The same result can hardly be achieved by just rotating the head of a robot, which instead induces changes in attention.

To the best of our knowledge, Yousuf et al. [219] proposed the first approach to automatically recognize a specific spatial formation within HRI. Their rule-based method was designed for a robot that explained artworks to two people in a simulated museum within a laboratory environment. If the users did not establish the desired spatial formation with the robot, then it explicitly asked them to "move closer" or "back a little" to better match the target formation.

In agreement with the social psychology background presented in Chapter 2, Karreman et al. [80] posed that the typical social encounters that people have with museum guide robots are common focused gatherings. This implies that the spatial arrangements that often emerge with these robots are not F-formations, but less uniform spatial organizations. Interestingly, an experiment by these authors suggests that people tend to stand farther away from a museum guide robot that faces them when it explains an exhibit, than from a robot that faces the artwork. However, people more easily loose interest in the latter robot than the former one.

These results regarding F-formations are important because they suggest that people not only assign the idea of transactional segments to other people, but also to social robots. Motivated by this observation, we conduct additional HRI experiments as part of this dissertation to validate the idea that people establish spatial arrangements typical of human conversations with robots. Our efforts focus on studying group conversations that have 3 to 5 participants most of the time, instead of dyadic interactions like prior work.

Generating appropriate spatial organizations in HRI is important as well. Shi et al. [157] used a rule-based system to enable a service robot to appropriately start an interaction and position itself with respect to a user. The rules chosen for this task were inspired by results from a study on how a human clerk may greet a customer in a store. Alternatively, social stimuli can also be used to generate appropriate spatial configurations during dyadic interactions [117, 119]. This effort led to a controller that optimizes the pose of a robot relative to a user, subject to maximizing the robot's expectation of hearing the user and seeing his or her body gestures. As an extension, the robot can adjust its speech output levels and gestures based on how well it believes that the user can receive these social signals.

Our work on enabling robots to cooperate to sustain spatial arrangements typical of conversations is related to these prior efforts. However, our focus is on generating appropriate spatial behavior for group conversations with two or more participants. We also leverage social group phenomena for robot control instead of individual stimuli.

3.2 Detecting Interactions: Computational Models of F-formations

The problem of detecting situated human interactions using proxemics has been of interest in various disciplines, including:

Computer Vision:	[5, 10, 32, 33, 35, 36, 40, 48, 56, 99, 134, 135, 144, 155, 156, 187, 193, 220, 222]
Human-Computer Interaction:	[27, 37, 46, 54, 59, 74, 75, 111]
Signal Proc. & Sensor Fusion:	[50, 105, 114]
Systems Engineering:	[20, 132]
Natural Language Processing:	[213]
Robotics:	[109]

As shown in Table 3.1, different fields tend to prefer different sensing modalities for the task of identifying situated group interactions. Useful modalities include cameras, microphones, wireless devices, optical tracking systems, accelerometers and lidars.

The methods listed in Table 3.1 also differ by the particular aspects of human interactions that they leverage for the detection task. Some methods were designed to find situated interactions by estimating shared attention, turn-taking patterns or coordinated gestures. Other methods focused on analyzing human motion trajectories, or the relative distance (and orientation) between people in a particular instant of time. Finally, some methods focused specifically on detecting conversations by identifying F-formations.[1] The next paragraphs provide more information on the latter approaches because they are the closest efforts to our work on enabling robots to detect group conversations. Developing methods that can jointly reason about multiple interacting cues, such as those listed in Table 3.1, is left for future work.

Most approaches to detect F-formations are model-based. They implicitly model the properties of F-formations, e.g., in an affinity matrix that can be used for graph clustering [74], or explicitly encode the transactional segments of people in scene in order to detect o-spaces [40, 54, 111, 144, 155, 156]. The latter methods in particular assume that there is a one-to-one mapping between o-spaces and F-formations, and between F-formations and group conversations. Thus, wherever an o-space is found, a conversation is also said to be detected. We continue the tradition of using model-based approaches to detect F-formation in this work because of the lack of big datasets for this task, especially within HRI. Different to most prior work in this area, though, we opt for computing soft group assignments for people in a scene. As shown by Chang et al. [32], soft assignments can help overcome measurement uncertainty when reasoning about human spatial behavior.

Lower-body orientation is a key descriptor of F-formations. All the prior efforts on detecting F-formations observed that knowing the lower-body orientation of people in a scene can help differentiate between these organizations (due to conversations) and other spatial arrangements that emerge in close proximity. In many cases were estimating people's lower-body orientation is difficult, e.g., from a top-view of a scene, it is better to approximate this orientation than to ignore it when reasoning about spatial behavior. For example, body orientations can be approximated by head orientations, as in [40, 155, 156].

[1] The line between methods that aim to detect shared attention and those that aim to identify F-formations is often blurry. Various prior efforts, e.g., [5, 10, 193], were inspired by Kendon's F-formation system [86] but ended up modeling visual attention to detect interactions. For consistency, Table 3.1 categorizes these methods within the "Shared attention" group.

Table 3.1: Categorization of related work on detecting situated social interactions. All the methods rely on proxemics, but they focus on different aspects of human interactions: shared attention, motion trajectories, turn-taking and coordinated gestures, proximity, spatial arrangements, and f-formations. The sensing modalities identified for each paper correspond to: color or depth cameras in the environment (Env. Cam.); wearable cameras (Wea. Cam.); microphones (Mics.), wireless communication devices (Wireless) such as BlueTooth devices, radio-frequency identification (RFID) systems, or infrared receivers and transmitters; or other modalities (Other) such as optical tracking systems, accelerometers, and lidars. The markers next to the references indicate to the area of the publication venue: computer vision (*), human-computer interaction (†), natural language processing (††), robotics (‡), signal processing & sensor fusion (‡‡), and systems engineering (**).

	Env. Cam.	Wea. Cam.	Mics.	Wireless	Other
Shared attention					
Fathi et al., 2012 [48] *		✓			
Park et al., 2012 [135] *		✓			
Bazzani et al., 2013 [20] **	✓				
Leach et al., 2014 [99] *	✓				
Vascon et al., 2014 [193] *	✓				
Alletto et al., 2014 [10] *		✓			
Park and Shi, 2015 [134] *		✓			
Aghaei et al., 2016 [5] *	✓				
Turn-taking or coordinated gestures during conversations					
Choudhury and Pentland, 2002 [37] †			✓	✓	
Brdiczka et al., 2005 [27] †			✓		
Wyatt et al., 2007 [213] ††			✓		
Hung et al., 2014 [75] †					✓
Motion trajectories					
Ge et al., 2009 [56] *	✓				
Choi et al., 2009 [35] *	✓				
Chang et al., 2011 [32] *	✓				
Luber and Arras, 2013 [109] ‡	✓				✓
Linder and Arras, 2014 [105] ‡‡	✓				
Feng and Bhanu, 2015 [50] ‡‡	✓				
Proximity (relative distance information)					
Eagle and Pentland, 2006 [46] †				✓	
Yu et al., 2009 [220] *	✓				
Zen et al., 2010 [222] *	✓				
Group spatial arrangements (distance + orientation information)					
Olguín et al., 2009 [132] **				✓	
Groh et al., 2010 [59] †					✓
Chen et al., 2011 [33] *	✓			✓	
Matic et al., 2012 [114] ‡‡				✓	✓
Tran et al., 2013 [187] *	✓				
Choi et al., 2014 [36] *	✓				
F-formations specifically (as defined by Kendon [86])					
Cristani et al., 2011 [40] *	✓				
Hung and Kröse, 2011 [74] †	✓				
Marquardt et al., 2012 [111] †	✓				
Gan et al., 2013 [54] †	✓				
Setti et al., 2013 [155] *	✓				
Setti et al., 2014 [156] *	✓				
Ricci et al., 2015 [144] *	✓				

Our insight is that F-formations can also be leveraged to estimate lower-body orientation because the underlying interactions that induced the F-formations have a tendency to bias the spatial behavior of nearby people. Based on this observation, we designed a group detection framework that takes advantage of the mutual dependency between the problems of (1) detecting conversations by reasoning about spatial behavior, and (2) tracking people's lower-body orientation in a scene. Chapter 6 introduces this framework in detail. Interestingly, Ricci et al. [144] proposed to jointly estimate people's head, body orientation and F-formations to detect social conversations. While this work and the present dissertation were developed independently of each other and they are aimed at different applications, they both reinforce the idea that contextual information, like knowledge about social interactions, can aid in the problem of estimating human pose. A similar hypothesis was explored by Yang et al. [216] on the task of classifying different types of dyadic interactions or *touch codes*.

Chapter 4

The Furniture Robot *Chester* and its Companion *Blink*

As part of this dissertation, we contributed to the design, fabrication and software set up of a flexible robotic platform for human-robot interaction. Mathew Glisson, Braden McDorman, Moshe Mahler, Scott Hudson, Aaron Steinfeld, Jodi Forlizzi, Brian Mizrahi, and Jessica Hodgins also contributed to various aspects of this effort. We thank Adam Stambler and Ken Bolden for providing their voice for Chester.

4.1 Design

We aimed to control human expectations for robot characteristics through the form of the robot. As argued by prior work [137], controlling expectations is important because it can potentially make users more adaptable to systems [139]. If users expect a system to have capabilities that it does not really possess, they may also feel frustrated [209]. Moreover, certain expectations can potentially lead to the problem of the uncanny valley [122].

We chose to build a robotic character that looked like a piece of furniture because we expected users to have limited expectations for this class of objects. Inspired by Disney's Beauty and the Beast movie,[1] we considered various furniture pieces that the robot could look like, including a chest of drawers, a table (as in Fig. 4.1A), and a chiffonier (Fig.

[1] http://www.imdb.com/title/tt0101414/

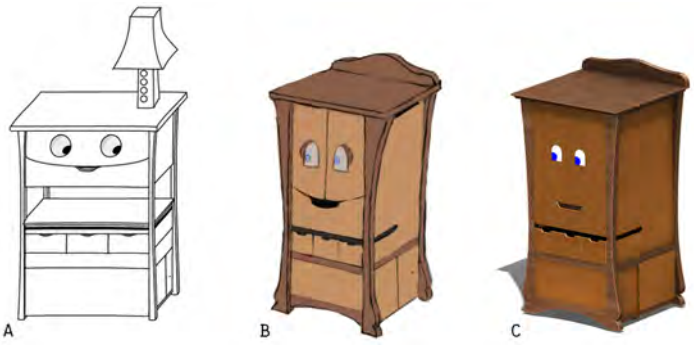


Figure 4.1: A few early designs of the robot: table model (A), and chiffonier (B, C). Figure (B) is an early sketch of the final robot design, while (C) is our first complete CAD model.

4.1B). We ended up choosing the latter model for our robot, based on its friendly appearance, some fabrication constraints, and safety considerations. We named this furniture robot *Chester*.

Once we had chosen a general look for the robot, we carried out further refinements of the model using Computer-Aided Design (CAD) software, as seen in Fig. 4.1C (early model) and Fig. 4.2 (final model). This iterative design process allowed us to reason about the placement of sensors in the interior of the robot, and other important details with respect to how well the design could work for both children and adults. For example, we designed the face of Chester to be simple and appealing. We made the robot's eyes especially big, because this associates them with the look of babies [61]. To facilitate bringing the character to life, we also designed the interior of the robot such that it could host a projector. This projector served to easily animate the robot's eyes and mouth, following animation principles [181] that can facilitate human-robot communication [180, 190].

The whole body of Chester was designed to rest on top of a Pioneer P3-DX robot, by Adept MobileRobots, as depicted in the right image of Fig. 4.2. This particular set up made it easy to control Chester's movement around indoor environments because we could rely on off-the-shelf motion controllers by Adept for this task. One important consideration due to this set up, though, is that it makes our platform a differential-drive robot. It can rotate in place like people do, but it cannot move sideways.

Chester also comes with two actuated drawers, which allow the robot to give (and receive) objects to (from) users. The simulated middle drawer in its front serves to hide a LMS100 laser measurement system by SICK Sensor Intelligence. This lidar can measure distances to nearby objects through a small gap in the front and the sides of the robot's casing.

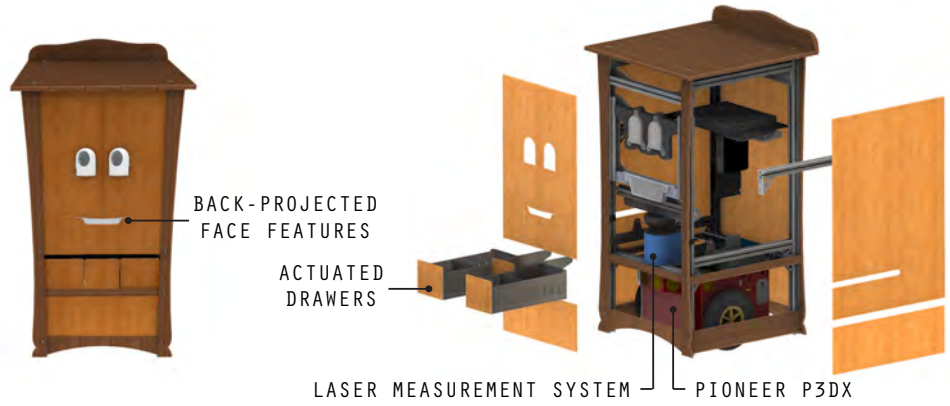
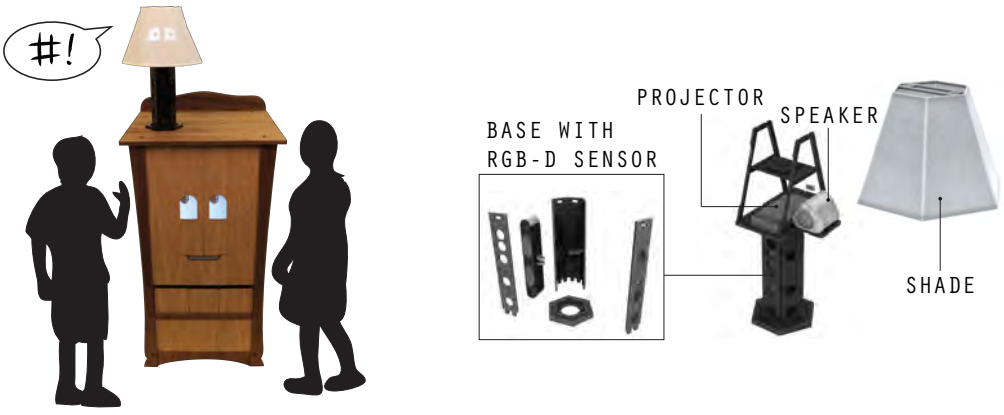


Figure 4.2: Renderings of Chester's main hardware components

The robot can operate as one or two characters simultaneously depending on the needs of a particular interaction. This is a unique characteristic of our platform, which we leveraged to study the effects of a sidekick character in HRI (Sec. 5.1).

The second character on the platform is *Blink*, the lamp on the top of Chester (Fig. 4.3a). The lamp physically holds a number of hidden components. For example, a speaker is inside of the shade for Blink to communicate verbally through non-linguistic utterances. Blink does not have a visible mouth like Chester, but it does have back-projected eyes that can express various emotions, like surprise and anger. A hidden pico projector is used to render the eyes, as depicted in Figure 4.3b. The base of the lamp also holds a hidden Xtion PRO LIVE RGB-D sensor by Asus. The position of this sensor provides a wide vertical field of view for its RGB and depth cameras, thus allowing the robot to sense short and tall people in front of it.



(a) Blink and Chester (b) Lamp hardware

Figure 4.3: The left image shows a photograph of the finished platform (with Blink enabled) in proportion to 5 year old kids. The right image shows a rendering of the main hardware components inside of its the lamp.

4.2 Fabrication

The whole casing of the furniture robot was fabricated at Disney Research Pittsburgh using rapid prototyping techniques. For example, the panels that cover the robot were made of foam-core or layers of acrylic. Foam-core was used in places were we expected users to make little contact with the robot, and was particularly advantageous to reduce the weight of the platform. The layers of acrylic were placed in the top of Chester, where the lamp needed to be supported, as well as in the edges of the front and back faces of the robot. We tried to make these edges as robust as possible, since we worried that the robot could hit nearby objects while developing and testing software. We attached veneer to the foam-core and acrylic panels to make the robot look as if it were made of real wood.

The eyes of Chester were 3D printed, and the mouth was made of clear acrylic. In both cases, we placed a piece of professional back-projection fabric behind them, so that we could use a small-size projector to illuminate them and render custom-made facial expressions.

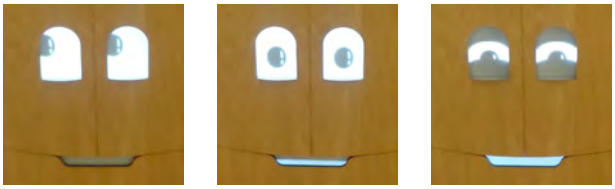
Similarly, the rigid parts of the lamp were made of laser-cut acrylic. The lamp's shade used white back-projection fabric, which worked great for rendering the eyes of Blink even in well lit indoor environments.

4.3 Software Implementation

Our robotic platform contains two computers: one inside of the Pioneer base, and a laptop inside of Chester's wood casing. The first computer is a Versalogic Mamba EBX-37 industrial grade computer with a Dual-Core 2.26 GHz processor. This computer was dedicated to rendering Blink's face and to execute navigation-related tasks. For example, these tasks include interfacing with the Pioneer driver that makes the wheels of the robot move, and gathering measurements from the lidar in the robot, which is often used for robot localization. The second computer is a EON11-S gaming laptop by Origin with an Intel Core i7 - 3610QM processor. This laptop is used to render the face of Chester and for other perception tasks. The various processes that run in these two machines communicate between them using the Robot Operating System (ROS) [140].

4.3.1 Facial Animations

We built custom face controllers for Chester and Blink using OpenFrameworks,[2] an open-source library for creative coding. These controllers are in charge of rendering facial expressions, typically composed of a sequence of animations, and (optionally) playing pre-recorded audio clips. Facial expressions can modify the position of the eyes and the eyelids of the characters, the size of their pupils, and the mouth of Chester. Figure 4.4c shows a few illustrative facial expressions. To enable Chester to look at arbitrary 3D locations in the world, we also implemented a calibration procedure to learn a mapping from these locations to 2D pupil positions within the character's eyes. Appendix A describes this procedure in detail.



(a) (b) (c)

Figure 4.4: Close up photos of Chester's face. The robot has its mouth closed and is looking towards its right side in (a), has its mouth half-opened and is looking in front of it in (b), and has its mouth opened and is squinting in (c).

The mouth of Chester can be opened or closed, and is animated independently of the eyes. To generate the corresponding animations, we first recorded Chester's utterances and processed the audio files to compute the amplitude of their sound waves. This feature sets how open the mouth should be: the louder the sound, the wider the mouth is opened.

4.3.2 Motion Control & Sensing

In general, we use a combination of custom software and pre-existing ROS packages to control the robot and enable it to sense its environment. For example, the drawers of the robot are operated by an Arduino board.[3] This board runs a custom program that senses the state of the drawers using infrared distance sensors, and that handles requests for opening or closing the compartments at a constant velocity.

To manually teleoperate the robot, we have implemented a variety of custom interfaces in ROS that send motion commands to the Pioneer base, and a safety mechanism to override these commands if the lidar in the robot senses obstacles in close proximity. For robot localization, we often build a map of the environment and use open-source algorithms to estimate the pose of the robot as it moves around.[4] For autonomous robot motion, we also rely on ROS's navigation pipeline with layered cost-maps [108].

The furniture robot has a hidden microphone inside its lamp. Due to the focus of this dissertation, we have not developed custom sensing modules for automatically processing audio data. However, our platform could potentially be used for research on natural language processing.

[2] http://www.openframeworks.cc

[3] http://www.arduino.cc

[4] In particular, we often use GMapping (http://wiki.ros.org/gmapping) and ROS's open-source Adaptive MonteCarlo Localization algorithm (hrefhttp://wiki.ros.org/amclhttp://wiki.ros.org/amcl) for these tasks.

4.3.3 Other Implementation Details

For cases in which the characters carry heavily scripted interactions with users, we created a program that handles requests for utterances and face animations, according to the current state of the interaction. This program forwards valid requests to the face controllers (as described in Sec. 4.3.1), and reasons about what the characters can do next based on a directed graph that encodes the script. This graph has a unique source node (with no other nodes pointing to it) that contains information about the state in which our robotic platform should be in right before the interaction starts. The rest of the nodes of the graph correspond to communicative actions that the characters can take, such as utterances or facial expressions. An illustrative example of the graph for an scripted interaction is presented in Fig. 4.5.

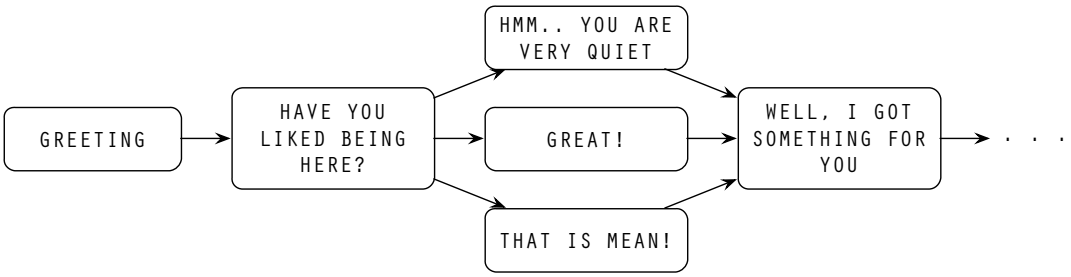


Figure 4.5: Initial part of a graph for a scripted interaction with Chester. In this case, each node had an associated facial expression (not displayed in this image).

Users are often unpredictable. Thus, it is possible that scripts do not encode all the possible states of a social encounter. To prevent interactions from breaking in these cases, we designed a set of fallback actions for the characters. These fallback actions can be requested at any given point of an interaction, e.g., to express that Chester did not understood what a user said, or briefly answer unexpected questions from the interactants.

4.4 Limitations

Our furniture robot platform is not fully autonomous. Even though we have developed various capabilities for our robot based on the particular needs and focus of this dissertation, there are many others that are lacking. For example, we rely on pre-recorded utterances chosen by an operator for generating robot dialog. While this approach can make it easy for Chester to communicate with users, it limits the kind of conversations that it can have, and slows down its response time. For example, this delay can vary from half a second to a few, based on the time that the robot operator takes to choose an appropriate response during a conversation.

While Chester's mobile base was a practical choice for its design, it limits the motion of the robot because it is differential drive platform and cannot move sideways. This constraint can prevent the generation of human-like motion trajectories for Chester and, without care, can easily lead to unnatural and counter-intuitive robot motion.

Chapter 5

Examining Spatial Behavior during Group Conversations

This chapter presents two experiments that we conducted to validate the idea that people establish group formations typical of human conversations with robots, and investigate various aspects of multi-party human-robot interactions. The protocols that we designed for these experiments were presented at the CSCW 2017 Robots In Groups and Teams Workshop [195].

The first experiment in particular was a field trial with groups of children. In this case, we studied various social engagement cues, like spatial behavior, and the effects of a co-located sidekick character in HRI. A publication about this work appeared in the 2014 ACM/IEEE International Conference on Human-Robot Interaction [197].

In the second experiment, we studied human-robot interactions with groups of adults in the context of a social role-playing game. This setting allowed us to explore two interesting scenarios: having the robot interact with groups of participants as a player of the game; and having it moderate the activity. In the former case, the robot had the same role as the participants. In the latter, the robot acted as a leader and was more in control of the group. A short paper about this second experiment was published in the 10th Annual ACM/IEEE International Conference on Human-Robot Interaction [198].

5.1 A Case Study of Child-Robot Interaction: Effects of a Sidekick

This first exploratory study had two objectives: (1) investigate the effect of a sidekick character in HRI, and (2) study spatial behavior and other social engagement cues during group interactions with our platform. A *sidekick* is a character that is closely associated with another primary character, and regarded as a subordinate or partner. Sidekicks are popular in various forms of narrative, where they are often used as comic relief or to introduce an accessible character to increase audience engagement [76, 170, 225]. Likewise, sidekicks can act as a vehicle for raising an obvious concern to the primary character from the audience. For example, a sidekick may yell, "Look out!" to the hero when a villain appears on screen.

In this experiment, we made Chester the main character and Blink its sidekick. We predicted that the addition of Blink as a sidekick would lead children to be more engaged in the interaction and treat Chester in a more sociable manner. We also thought that Blink could help mitigate apprehension and fear in cases where children felt uncomfortable with Chester, e.g., due to its physical appearance and rigid motion.

5.1.1 Related Work

This section briefly describes work that is related to various aspects of our sidekick experiment, and that was not previously mentioned in Sections 2.1, 2.2 and 3.1.

Watching two robots interact can facilitate HRI along a number of dimensions. For example, Kanda et al. [78] provided evidence that suggests that people can better understand robot utterances when observing two robots interact with each other and with their environment. Shiomi et al. [158] noticed a similar effect in a museum exhibition, where the interaction between two robots attracted people's attention, and helped convey information.

Hayashi et al. [64] highlighted the potential of a pair of robots as a *passive-social medium* to communicate information in the context of Japanese Manzai. A follow up work tested robots as a non-interactive communication medium in a Japanese railway station [63]. In this case, the people that observed the two robots interact with each other paid more attention to the content of their dialogues than those who observed one or both robots (with limited interactivity) try to convey the same information directly.

A few prior efforts have explored the idea of companion or subordinate characters in HRI. For instance, the interface of the ProVAR system [191, 204] for assistive manipulation was based on two built-in characters that enhanced user's interaction by leveraging social behavior. One of these characters, Pinocchio, was a down-to-earth robot arm that physically helped with manipulation tasks. Pinocchio was able to communicate fully with the other character, but could only provide physical gestures and audible tones to the user. This other character, Jiminey, played the role of a helpful consultant embedded in the graphical user interface of ProVAR. This character worked as a complementary communication channel for Pinocchio, and provided support during difficulties with the system, e.g., by providing advice to the user, and actively mitigating frustration. Jiminey often blamed Pinocchio's limitations for the problems that arose while ProVAR was in use. Another example are some of the animatronic figures developed by Walt Disney Imagineering, e.g., [69]. Our efforts are inspired by these prior works.

Several factors can affect children's personal space. In psychology, Bailey et al. [16] provided support for the notion that personal space behavior of fifth and sixth graders can be manipulated via the principles of modeling. Children tended to stay close or far from an object person as a function of a confederate's behavior. In HRI, Yamaji et al. [215] examined spatial aspects of the interactions between children and Sociable Trash Box (STB) robots, which are mobile, expressive furniture. These authors found that proxemic distances differed depending on how the robots were behaving. When the robotic trash boxes were moving individually, children exploited two spaces to interact with them: "personal-social" and "public" – in reference to Hall's spatial zones [60]. However, when the robots interacted in groups, as swarms, the authors noted three kind of spaces ("personal", "social", and "public"). There are logical reasons for different proxemics when robots are independent, but it is unclear from this work whether co-location, such as in the case of Chester and Blink, impacts proxemics.

Different cues can be used to estimate user engagement with robots, including gaze, head pose, turn-taking, human-robot kinesics, and spatial information [22, 121, 145, 148]. The latter is particularly popular for social human-robot interaction analyses, as discussed in Section 3.1, because spatial information can easily be compared with Hall's seminal work on proxemics [60].

5.1.2 Method

We conducted the experiment as a field trial during the *2013 Summer Games* event at Disney Research Pittsburgh. In this event, groups of children came to the laboratory to interact with different kinds of characters and participate in a collection of activities. For our experiment, we used a Wizard of Oz arrangement [83] as an attempt to identify appropriate robot behaviors en route to implementing autonomy during group interactions [169].

Participants only experienced one of two conditions: *without sidekick* (C) or *sidekick* (S). Only Chester was active in the control (C) case, while both Chester and Blink (the sidekick) interacted with the participants in (S). The interactions were scripted and designed to be as similar as possible.

5.1.2.1 Participants

Twenty groups of 3 or 4 children interacted with Chester and Blink, for a total of 74 participants. Children were 4 to 10 years old, were accompanied by at least one adult, and some were siblings. Adults were allowed to observe upon request, but were asked to avoid interrupting the activities of the Summer Games event. This included trying to stay as far back as possible from the place where our experiment happened (Figure 5.1).

Ten groups (37 children) experienced the (C) condition, while ten other groups (37 children) experienced the (S) condition. The average age for each group was 6.8 and 6.9 years old, with standard deviations of 2.1 and 2.1, respectively. Ages were split into three categories: A1 for 4-5 years old, A2 for 6-8 years old, and A3 for 9-10 years old. The number of participants per condition and age group was roughly similar. We had 12, 16, and 9 children in the A1, A2 and A3 categories for the control condition (C), and 12, 14, and 11 for (S).

Even though we tried to balance for gender, the proportion of boys with respect to girls was greater in (S) than in (C). We had 23 boys in the *sidekick* condition (62% of that group), but only 18 boys in the *without sidekick* condition (49%).

Some kids expected to see a robot because the Summer Games' recruitment flier said that "we study how children (...) interact with toy, animated, and robotic characters". However, children were unaware of the appearances of our robotic characters prior to the study. Both Chester and Blink were kept out of public sight until they interacted with the participants.

Figure 5.1: Sketch of the environment where the experiment happened. The wizard was seated at the end of the table (1). Chester was at (2) when children started to approach from the conference room at the end of the hall (3). Parents were asked to remain near (3).

5.1.2.2 Procedure

Children first participated in a virtual "mix and match" game during the Summer Games event, where they picked apparel and accessories to change a character's look. Kids were able to take a picture of the character whenever they wanted and, at the end of the game, each got to pick their favorite image. The pictures that were selected by the participants were then printed, and stored inside of Chester's drawers without kids knowledge. Chester's mission was to give these pictures to the participants during our experiment.

The physical space where the interaction occurred is depicted in Figure 5.1. The robot operator, or wizard, was in the same room as the participants due to safety concerns, because this was our first experiment with the platform. She pretended to be working with a laptop at a table nearby ((1) mark in Fig. 5.1) for about 1 hour before the interaction. This allowed participants to familiarize with her presence.

An experimenter brought the kids into a conference room prior to the interaction ((3) mark in Fig. 5.1). Subsequently, the robot was secretly positioned against the wall in the dining area ((2) mark). The experimenter in the conference room then brought the children out and down the hall, with the premise of getting their pictures. The wizard started controlling the characters at this point, using a PlayStation 3 game-pad to surreptitiously command Chester's motions, open and close its drawers, and activate pre-recorded utterances and associated facial expressions for both characters. No participant discovered that the wizard was controlling the robots with the game-pad under the table.

The interaction followed various Phases in the (S) condition:

a) *Acknowledgment.* The participants were acknowledged. Blink and Chester looked towards the end of the hall, and realized that the children were coming. As participants approached, Chester and Blink verbally indicated that they were checking if they had the children's photos.

b) *Greeting.* Chester greeted the participants, and introduced Blink.

c) *How are you?* Chester asked the participants how they were doing and if they liked being at Disney's research facility.

d) *Remember.* Chester asked the children if they remembered the pictures they took during the earlier game. Chester told them that the photos were in its drawers.

e) *Stuck.* Participants experienced the rising action part of the story: Chester realized its drawers were "stuck" and, after conferring with Blink, said that they may need oil.

f) *Bump.* Chester indicted that he thought that bumping into a wall was a good way of fixing the problem, but Blink dissuaded Chester to prevent him from damaging the wall.

g) *Spin.* Chester asked participants to step back and spun around in an attempt to unstick the drawers, but was unsuccessful.

h) *Shaking.* Chester shook, following Blink's advice, and finally got the drawers unstuck.

i) *Opened drawers.* Chester told the participants to "come grab your pictures". The participants then grabbed their pictures or, if they did not want to, the experimenter grabbed them and gave them to the children.

j) *Visit again?* Chester asked the participants if they liked the pictures, if they would come to visit again, and if they had to leave.

k) *Goodbye.* Chester and Blink said goodbye to the children, retreated to a safe location, and closed their eyes.

After the end of the interaction, the experimenter offered the participants stickers of Chester and of the characters from the earlier mix and match game for them to take home if they wanted. Finally, children were brought back to the conference room where they were before the experiment, or to another Summer Games activity.

The interaction in the control (C) condition was similar to the interaction in the (S) condition, except that the lamp on top of Chester was not a character, but just a lamp. Blink's eyes were not visible in (C), and it did not emit any sounds. Since Blink was not there to help, the script was modified such that Chester realized that bumping into a wall was a bad idea by itself. Also, it occurred to Chester (not to Blink) that shaking may unstick the drawers.

The wizard had three special buttons in the controller that scheduled animations to help continue with the flow of the script in case of potential deviations. When one of the buttons was pressed by the wizard, Chester said "No! No! Let me do it myself" in response to situations in which children wanted to open the drawers with their hands. The other two buttons activated animations for "Ouch!" with a sad face and "Don't poke me" with an angry face. These were prepared to prevent very outgoing kids from touching the robot in dangerous ways, e.g., by leaning on its top, or sticking fingers near its lidar.

In general, participants were free to approach our robots as desired during the experiment. The experimenter that brought the kids from the conference room stopped approaching our robots about 4 meters away to reduce potential bias on the children's proxemic behavior.

5.1.2.3 Data collection and coding

The participants were equipped with a wireless microphone, attached to their clothes, for the duration of their participation in the Summer Games, and were recorded throughout the whole interaction. Video was captured from the Xtion Pro Live sensor inside of Blink's base, a Kinect sensor mounted on the ceiling of the dining area, and a standard camcorder positioned on a tripod in the back of the room (next to the sofa in the right side of Figure 5.1).

Two professionals transcribed with ELAN [167] when participants spoke, touched the robot, turned their head away from it, or laughed. At the beginning of the process, the transcriptions were evaluated twice for procedural errors. At the end, inter-coder reliability was computed for 16 participants (22%) that were transcribed by both people. Cronbach's alpha was 0.90 for number of utterances directed at the characters. Cohen's kappa was 0.87 and 0.93 for touching and head turning, respectively. Coders differed only by 1 annotation for laughing.

One transcriber annotated when the participants sat on the ground. She also marked down the participants and the robot's location in the video from the Kinect on the ceiling, as shown in Fig. 5.2. These locations were then converted to 3D coordinates using depth data, and

Figure 5.2: Top view of the scene during three different sessions of the experiment.

projected into the ground plane for 2D spatial analysis. To confirm precision, we computed the distance between Chester's top-front corners (in the ground plane) and compared it with the real width of the robot. The average difference was 2.3cm ($SD = 2.3$cm).

When a participant was not visible from the top-view video stream, then his or her position in the scene was transcribed using the camcorder video that was captured from the back of the room. These locations were mapped to the top view using a homography on the ground plane [62], and used for the spatial analysis presented in the following section.

5.1.3 Results

Our analysis focused on the interactions that happened from the beginning of the experiment up to when Chester gave the pictures to the participants. We did not analyze most sessions beyond this point because a significant number of children got distracted by the photos. Kids typically forgot which pictures they requested and became preoccupied with finding their own.[1]

5.1.3.1 Proxemics

To analyze proxemics, we first mapped the 3D positions of the participants from the top view of the scene to the ground plane. These positions were then transformed from the global frame of reference on the ground to a frame of reference originating from the middle of the front of Chester, as depicted in Figure 5.3.

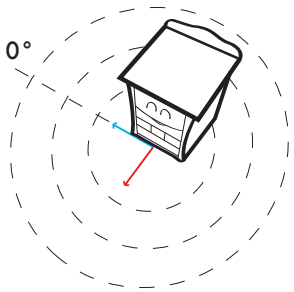


Figure 5.3: Frame of reference with respect to the robot

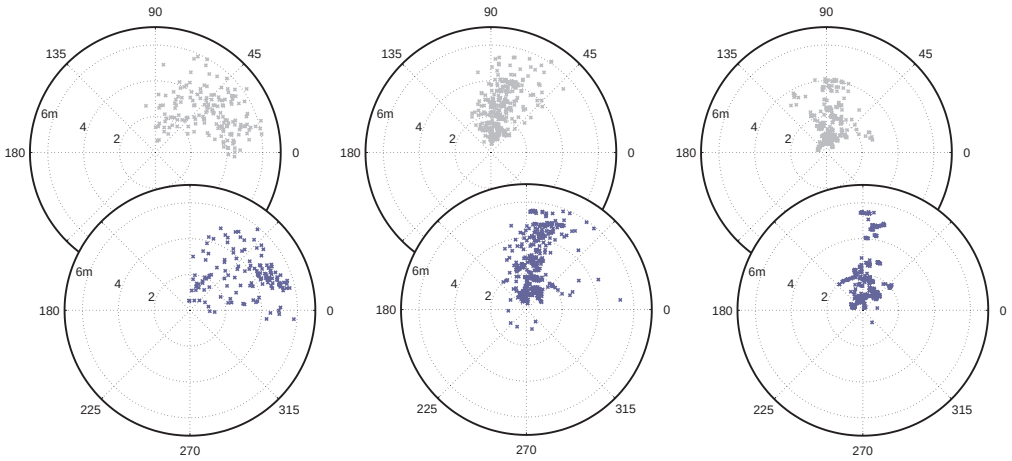
We plotted the distances computed with respect to the robot per interaction phase, since we expected participants' proxemic behavior to change based on activity (Figure 5.4). We considered the first 9 phases of the experiment, up to when the robot handed out the pictures to the participants. The boundaries between these phases was set based on robot utterances. For example, the Greeting phase started when the robot said "Hello".

Angular data showed that participants tended to interact with the robot by standing in front of its face, not to its side, nor behind it. The angular range $[0, 180]$ contained 99.7% of all the angles computed for the participants with respect to the front of the robot, excluding when Chester spun. The distribution of these angles followed a bell curve, with the maximum, central peak near 90 (i.e., *face on*).

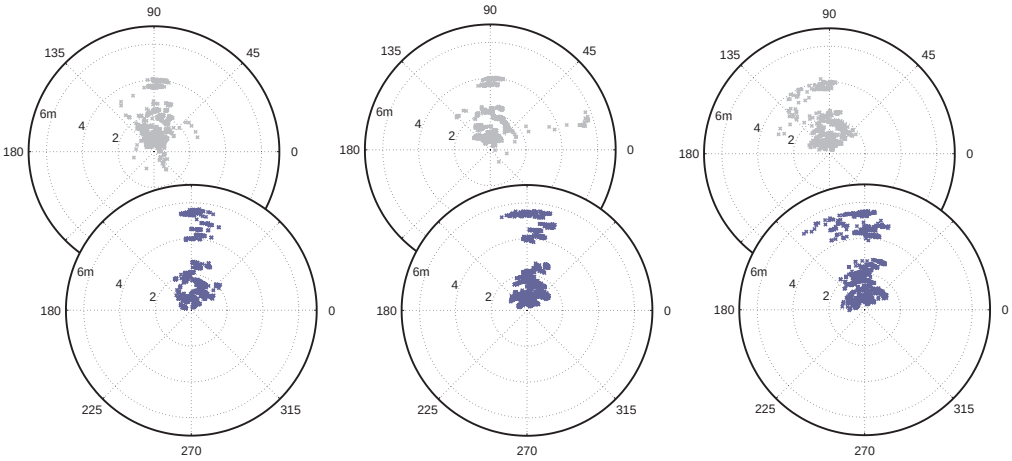
We inspected the distribution of distances between the participants and the robot(s), and noticed that the first two phases of the interaction (Fig. 5.4a and 5.4b) were more chaotic and did not provide as much insight on proxemic behavior as the rest. Further inspection of the data revealed that many children did not realize that Chester was talking during this time, or were still approaching it. Thus, we excluded this data from further analysis, and focused on the phases (c-i).

We found that three normal distributions closely fit the participant distances in logarithmic scale during the time between when Chester said "How are you?" until the pictures were distributed. We converted the distances from meters to log scale in order to reduce the bias of close encounters, as in [192]. We used $f(x) = log(x+1)$ to transform the data, and then followed a standard Expectation Maximization procedure to fit a mixture of Gaussian distributions to the transformed distances (Fig. 5.5). The means and variances of the Gaussians in log scale were $\mu_1 = 0.48, \sigma_1 = 0.11, \mu_2 = 0.94, \sigma_2 = 0.23, \mu_3 = 1.71, \sigma_3 = 0.11$.

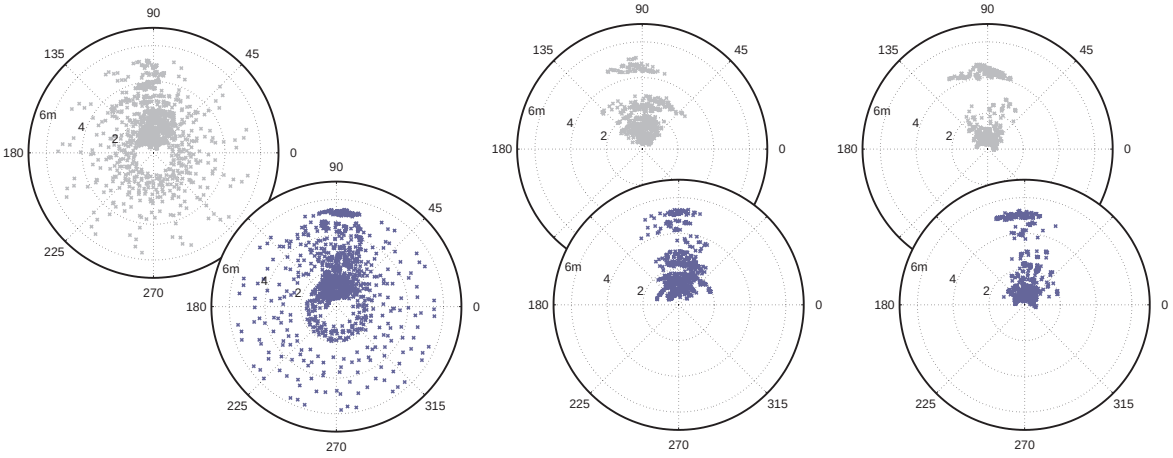
[1] Experience with the mix and match game did not suggest that children cared about which pictures were theirs, but it mattered in our study. For similar protocols, we recommend giving the same object to the participants.



(a) Acknowledgment (b) Greeting (c) How are you?



(d) Remember (e) Stuck (f) Bump



(g) Spin (h) Shaking (i) Opened drawers

Figure 5.4: Polar plots of the position (in meters) of the participants with respect to the robot. Data is grouped by interaction phase, with the frame of reference set with respect to the front face of Chester (Fig. 5.3). The gray marks indicate the position of the participants in the (C) condition, and the purple marks indicate the position in (S).

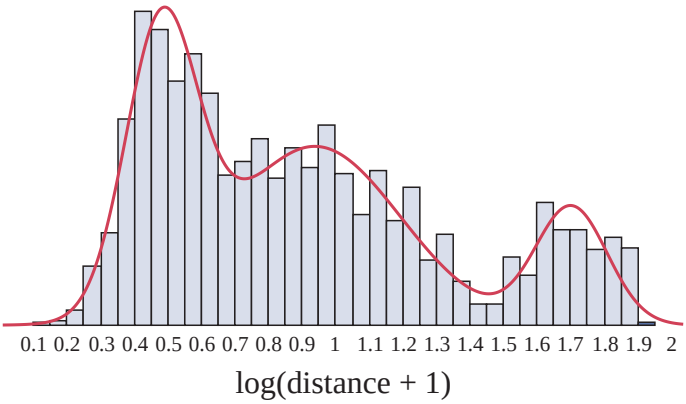


Figure 5.5: Histogram of the distances with respect to the robot in logarithmic scale. The red line shows the approximation found by fitting a mixture of Gaussians.

Table 5.1 shows the spatial zones obtained after converting the classification boundaries of the mixture of Gaussians back to meters. The first spatial zone, from 0.1 to 1.1 meters, encompasses Hall's intimate and personal spaces [60]. This ranges from 0.15 to 1.2 meters, in principle, though variations may typically occur due to culture and activity type. The second zone we found for Chester ranged from 1.1 to 3.3 meters, and was similar to Hall's social space (1.2 to 3.7 meters). Finally, our third zone extended beyond 3.3 meters. We believe this is similar to Hall's public space, which starts at 3.6 meters. Interestingly, the boundary between our zones 1 and 2 was close to the boundary between the distance clusters that emerged for STBs as they moved towards children in [215].

We computed the proportion of time spent in each of Chester's spatial zones, based on participant and interaction phase group. Phase group distance distributions showed similarity within logical activity sequences. Phase group P1 included "How are you?", "Remember", "Stuck" and "Bump" (before the participants were asked to step back). Phase group P2 included "Spin" and "Shaking", while P3 was just "Opened drawers".

A regression with Condition (S, C), Age Group (A1: 4-5 years old, A2: 6-8, A3: 9-10), spatial Zone (1, 2, or 3), and Phase Group (P1, P2, P3) showed significant differences for Zone on proportion of zone occupancy, $F[2, 663] = 54.71$ ($p < 0.001$). Occupancy occurred significantly more in Zone 1 ($M = 0.49$, $SE = 0.03$) than in the other Zones. Likewise, occupancy was significantly larger in Zone 2 ($M = 0.36$, $SE = 0.03$) than in Zone 3 ($M = 0.14$, $SE = 0.02$). The interaction between Zone and Age Group was also significant, with $F[4, 661] = 12.81$, $p < 0.001$. A Tukey HSD post-hoc revealed that younger participants (A1,A2) spent significantly more time in Zone 1, while older participants (A3) spent more time in Zone 2, compared to the rest. Finally, the interaction between Phase Group and Zone was significant as well, $F[4, 661] = 29.16$, $p < 0.001$. Phase Group 1 had significantly more occupancy in Zones 1 and 2, than in Zone 3. As instructed by the robot, the majority of the participants

Table 5.1: Spatial zones found for Chester, corresponding social distances by Hall [60], and distance clusters found for the Sociable Trash Boxes (STBs) that moved towards children [215].

Zone	**Range (m)**	**Hall**	**STB**
Zone 1	0.1 - 1.1	Intimate, Personal	Cluster 1
Zone 2	1.1 - 3.3	Social	Cluster 2
Zone 3	3.3 or more	Public	Cluster 2

then moved to Zone 2 in Phase Group 2 (when the robot spun around and shook) and moved back to Zone 1 in Phase Group 3 (when the pictures were given to them). Condition was not significant, nor any interactions with Condition.

5.1.3.2 Reactive Behavior

We computed how far the participants moved away from Chester when he said "step back", right before spinning around. The average distance participants moved back per Condition was $M = 2.67$ meters ($SE = 0.36$) for (C), and $M = 3.41$ meters ($SE = 0.47$) for (S). Further inspection of the data showed that the distribution of these distances in the control condition looked unimodal and skewed towards small distances. However, the distribution in the sidekick condition looked bimodal with a gap close to 4 meters. A logistic regression on whether participants stepped back more than 4 meters with Condition and Age Group as main effects showed significant differences for Condition only, $\chi^2(1, 74) = 8.18$, $p = 0.004$. The proportion of participants that stepped back more than 4 meters was 13.5% in (C), and 38% in (S). The interaction between Condition and Age Group was not significant.

5.1.3.3 Group arrangements

We measured the spread of spatial arrangements during the interaction based on (*i*) the angle spanned by children in front of the robot, and (*ii*) the average and (*iii*) the standard deviation of the distances between the participants per frame (1Hz). We grouped the data by Phase Group, once again, based on the distance distribution of the interaction phases.

A REstricted or REsidual Maximum Likelihood (REML) analysis [138, 172] on the above measures per frame with Condition (C or S) and Phase Group (P1, P2, or P3) as main effects, and participants' Group as Random effect, provided significant results for Phase Group, $F[2, 2667] = 406.67$ ($p < 0.001$). A post-hoc test for angle span based on Phase Group showed that the span was significantly higher when participants grabbed their pictures (P3), which is logical since they tended to group around the robot when reaching into the drawers. Average angle span was also significantly reduced when the robot spun and shook (from P1 to P2). This was expected since Chester said "step back" right before spinning.

The interaction results for Condition and Phase Group on the average and standard deviation of interpersonal distance were interesting. Both of these interactions were significant, with $F[2, 2667] = 13.18$ ($p < 0.001$) and $F[2, 2667] = 6.8$ ($p = 0.001$), respectively. However, the average interpersonal distances showed no functional differences, as they were small enough to be attributable to possible measurement error.

A Tukey HSD post-hoc test for the standard deviation of interpersonal distances revealed that participants in (C) varied their interpersonal distances significantly more in P1 ($M = 0.39$, $SE = 0.02$) and P2 ($M = 0.38$, $SE = 0.02$), compared to P3 ($M = 0.25$, $SE = 0.01$). The latter difference was not observed for (S), with $M = 0.28$, $M = 0.27$ and $M = 0.24$ for P1, P2, P3, respectively, and with standard errors below 0.01. This tells us that the spatial arrangement of participants in the Sidekick condition tended to be more uniformly spread (i.e., with similar distances between participants) compared to the Control condition. Even when the participants stood apart from the robot during P2, the standard deviation of interpersonal distances did not significantly change in (S).

We saw some evidence of F-formations [86], especially with older participants. However, children's position variability and impulse control made labeling difficult. At times, spatial arrangements seemed to be based on other factors than the conversation with the robot. This is not surprising given the age range of our participants. Children frequently adopt postures and positions not used by adults, and the lack of an authority figure during the experience may

have increased the amount of impulsive body motions. We also investigated several methods for quantifiably classifying participants into formation types described in the F-formation literature. None of these approaches proved tractable due to challenges with edge conditions, but we feel such approaches are worth exploring further and may be easier with adult participants.

Children often exhibit hiding behavior and defensive positioning when encountering new things, so we examined how often a participant was occluded by a fellow participant for Condition and Phase Group. This analysis was inconclusive and there were no significant differences.

5.1.3.4 Physical contact

We did not find significant differences on how much participants touched the robot between Conditions, but there were differences between Age Groups. We found that the proportion of participants who grabbed pictures from Chester's drawers significantly increased with age, $\chi^2(2, 74) = 7.47$ ($p = 0.02$). In particular, 62%, 77% and 95% of A1, A2 and A3 grabbed pictures. Interestingly, very different proportions were found for touching Chester's face, above the drawers, by Age Group, $\chi^2(2, 74) = 10.50$ ($p < 0.01$). No participant in A3 touched Chester's face, while 17% and 30% of A1 and A2 did. Further inspection of when participants first touched the robot using interaction phase as ordinal data (1 to 9) showed the first touch for A3 ($M = 8.6$) happened significantly later than the first touch for A1 ($M = 7.06$) and A2 ($M = 7.08$), $\chi^2(2, 60) = 6.38$ ($p = 0.04$). Participants in A1 and A2 appeared to be more exploratory and less inhibited than their older peers (A3).

5.1.3.5 Focus of Attention

As mentioned earlier, we annotated when the participants oriented their head away from the characters. These annotations were labeled as "Participant", "Experimenter", or "Other" based on the target that they focused their attention on. In general, participants did not turn their heads away for long: 11% of the turn away annotations ended in less than 1sec, 74% ended in less than 5sec, and 14% lasted for longer.

A regression on the length of the turn aways (in seconds) with distraction Target, Condition, Age Group, and interaction Phase Group provided significant differences. As expected, participants were looking away from the robot for significantly longer time during phase group P3, $F[2, 619] = 12.29$ ($p < 0.001$). The post-hoc test on the interaction between Phase Group and distraction Target showed that participants turned away their heads for a significantly longer time at some "other" target during P3, $F[4, 617] = 5.48$ ($p < 0.001$). This was not surprising since many participants were curious about others' pictures and, sometimes, there were arguments about which pictures belonged to whom. Moreover, the interaction between Phase Group and Age Group revealed that participants of age 6-8 (A2) turned their heads away from the robot significantly more time in P3 than in P1 and P2, $F[4, 617] = 2.7$ ($p = 0.03$). The latter difference was not significant for participants in A1 and A3. Finally, there was an interaction between distraction Target and Condition, $F[2, 619] = 3.13$ ($p = 0.044$). A Student's t post-hoc test showed that, on average, the participants in the (C) condition turned their heads towards the experimenter for significantly shorter periods of time ($M = 1.7$, $SE = 0.2$) than the participants in the (S) condition ($M = 3.4$, $SE = 0.8$).

5.1.3.6 Audio analyses

We used audio transcription to count participants' utterances and laughter, and performed logistic regressions with Condition and Age Group as main effects on these metrics. The number of participants with at least one utterance directed to the characters was significantly different

for Age Group ($\chi^2(2, 74) = 7.02$, $p = 0.03$). Only 75% of the children in the youngest age group (A1) spoke to the robot, while 97% and 85% of A2 and A3 did. The interaction between Age Group and Condition was also significant ($\chi^2(2, 74) = 7.01$, $p = 0.03$). The participants in the age group A1 made fewer utterances to the characters when the sidekick was present ($M = 0.67$ versus $M = 0.83$), while those in A3 talked more ($M = 1.0$ versus $M = 0.67$).

We also found that the number of participants that laughed at least once was significantly greater in (S) than in (C) ($\chi^2(1, 74) = 4.98$, $p = 0.03$). The average percentage of participants that laughed was 46% ($SE = 0.08$) and 22% ($SE = 0.06$), respectively.

5.1.3.7 Other Findings

About 19% of the participants sat on the ground near Chester while interacting ($N = 7$ for each condition). These children tended to stay on the ground for long periods ($M = 74.3$ seconds, $SE = 11.4$), suggesting that they felt comfortable in close proximity with the robot.

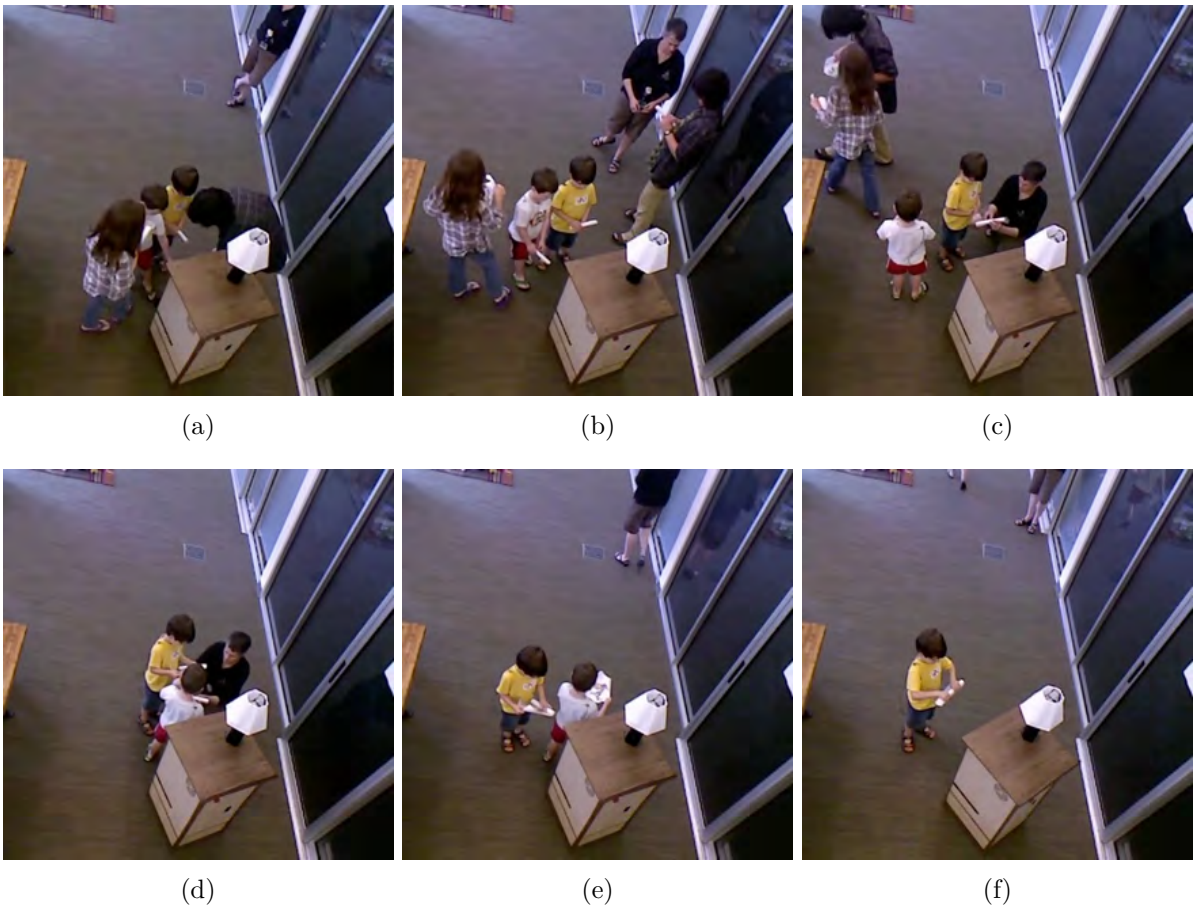
Additional analysis of the participants' utterances revealed interest in the sidekick. For example, one participant said the following when Chester was about to turn: *"I feel bad for the lamp. I hope you are going to be OK"*. After the experiment, another children told Chester: *"Oh, by the way, your friend (Blink) kind of sounds like R2-D2"*. While this data is sparse, it reinforces earlier findings showing greater engagement when the sidekick was present.

5.1.4 Discussion

The experiment was limited in several ways. Our characters sometimes fell short in responding adequately to children due to their limited verbal abilities. The beginning and the end of the interaction were often chaotic, because the participants were not expecting to interact with the robot and frequently got distracted with their pictures (a typical example is depicted in Fig. 5.6). This limited our spatial analysis, and reduced user engagement at times. Also, results were obtained with a co-located sidekick, and further testing is needed to confirm these findings in other settings.

The sidekick had effects on the interaction. While we found that our co-located sidekick did not alter proximity, it seemed to increase attention to spoken elements of the interaction. Differences were found in verbal utterances, laughter, visual attention, and reactive behavior. For example, more participants moved way back in (S) than in (C) when Chester said "step back". Blink clearly had a positive entertainment effect, resulting in twice as many participants laughing at least once during the experiment session. The sidekick relationship in the literature and in entertainment media often creates comic relief. Our evidence suggests that this effect can be translated to HRI, even when the robots are co-located.

There were no differences for spatial behavior between conditions. Our data supported three spatial zones with respect to the front of Chester, reinforcing earlier findings on proxemics with the Sociable Trash Boxes [215]. We relied on radial distance measurements for our spatial categorization because participants rarely stood on the sides or the back of the furniture robot. However, we expect these spatial zones to change as people approach the platform from directions other than its front [65]. In these situations, we suggest measuring distances with respect to the closest point on the casing of the robot, instead of with respect to its front face [192]. While it was difficult to systematically label F-formations due to children's position variability and impulse control, we observed these types of spatial organizations naturally emerge during the interactions with our robot.



(a) (b) (c)

(d) (e) (f)

Figure 5.6: Sequence of frames from a session of the experiment. These frames illustrate a typical situation where the participants got distracted with their pictures after the robot opened its drawers. Figure (a) shows the time when the participants grabbed their pictures from Chester. One of the kids then approached the experimenter looking for some help to unroll the picture that he had gotten from the robot (b). This participant then started a conversation to check another girl's picture, while the experimenter tried to help the child with yellow shirt unroll his gift (c). An instant later, the participant with white shirt asked the experimenter for help (d), effectively excluding the robot from their interaction. Once the experimenter had left, the child with a yellow shirt repositioned spatially, and continued the conversation with Chester and Blink (e). Finally, the participant with the white shirt walked away, leaving the robot in a dyadic interaction (f).

Our work has design implications. An early design goal was to create a robot and experience that was friendly and interesting to children. In this regard, our results show excellent engagement in general. The participants routinely entered Hall's Intimate and Personal zones [60], positioned themselves square with Chester, and spoke to and laughed at the characters. While some children maintained a healthy distance from the robots, the overall appearance and behavior of Chester and Blink were positive. These findings reinforce the STB results showing furniture to be a good robot design for children [215]. We are also able to generalize Osawa et al.'s [133] findings that the anthropomorphization of household objects can produce positive engagement effects.

5.2 Group Interactions in a Social Role-Playing Game

We performed another Wizard of Oz experiment [83] to continue studying spatial behavior in the context of HRI. In contrast to our prior work, though, our focus here was on studying group interactions between a robot and adults in a more structured scenario. The motivation for these changes was twofold. First, we wanted to study conversations in HRI with a different group of participants that were not as impulsive as children. Second, we wanted to prevent the participants from getting distracted or becoming passive spectators of human-robot interactions during the experiment. We achieved this goal by assigning active roles to the participants by means of the experimental protocol.

In this second experiment, small groups of adults played an established social game, *"Mafia"*, with our furniture robot. In this semi-structured activity, the players were secretly assigned to teams (villagers or mafia) and were involved in group discussions. The villagers sought to identify the mafia before they were all killed in the game, while the mafia hid his/her identity.

An advantage of this role-playing game scenario is that it let us study two perspectives of the interaction: with the robot as a player, or as the moderator of the game. We expected the interaction with the robot to change based on its role, because being a member of the same social category can potentially increase group identity [94].

Before explaining in detail this second experiment, this section describes related work on the use of games for studying human interactions, especially in the context of HRI. Then, it documents a set of pilot sessions that we ran with small groups of people to test the Mafia game as a viable scenario for this experiment. Our methodology, findings, and conclusions from this effort are presented thereafter.

5.2.1 Related Work on Studying Human-Robot Interactions in Gaming Contexts

Games are advantageous for studying HRI for several reasons. First, games take into account existing social practices, and can generate new ones. This makes games a good scenario for unexplored human-robot interactions [214]. Second, games are restricted to a particular domain and, thus, can be very practical for research. Third, games are often fun and engaging. This makes them interesting to participants. It is important, though, to tailor games to reflect on the HRI experience. As expressed by Xin and Sharlin [214], *"game rules can and should be altered in order to allow the robots and the human to interact in a manner that will inform on the HRI design question"*. We follow this principle in our own work.

Games are commonly used as a platform for Robotics and HRI research. For example, the well-known RoboCup competition is focused on enabling robots to play soccer [90]. The main goal of RoboCup is to develop a team of humanoid robots that can play soccer against the best human team by 2050. Several of the technologies that have been developed for RoboCup have been used in other domains, like search and rescue. In a related effort, Argall et al. [13] explored human-robot coordination within *"Segway Soccer"*, a research domain built upon RoboCup robot soccer.

Two games that have been used for HRI research are *"Mastermind"* and *"Rock-Paper-Scissors"*. The former game was used by Bartneck et al. [17] to investigate if humans are more hesitant to switch off an intelligent robot compared to a less intelligent machine. The latter game was used by Short et al. [161] to examine the degree to which variations in robot behavior resulted in attributions of mental state and intentionality. In this experiment, the participants played the game with a humanoid robot which, in some cases, made ambiguous errors that could be interpreted as malfunctioning or cheating. A follow-up work used the same scenario to investigate adversarial cheating [106].

Leite et al. [102] studied how children perceive and interact with an empathetic social robot in the context of "Chess". Their results suggested that empathic robot behaviors can impact positively children's perception of a robotic character. A long-term evaluation of an empathetic robot model in this scenario suggested that, in general, children can feel supported by the robot and by their peers to a similar extent while playing the game [103].

In addition, games have been used to study turn-taking with interactive characters. For example, Lehman [101] describes an experience with a mix-and-match game to study children's communicative behavior. More recently, Al Moubayed and Lehman [6] designed a collaborative game to further study turn-taking, user engagement, and addressee identification in HRI.

The success of prior efforts investigating human interactions with Mafia inspired us to use this game to study HRI. For example, Mafia was used by Park et al. [136] to collect data on human primary gaze behavior. In addition, Batcheller et al. [18] studied the effects of physical presence on Mafia. In one condition of their experiment, the participants played collocated; in the other, they played over video. The results from this experiment suggest that people can have similar levels of satisfaction, fun, and frustration playing over video in comparison to the collocated scenario.

Finally, Hung and Chittaranjan [73] used a popular version of Mafia, called Werewolf, to create an audio-visual corpus for social signal processing. The video feed and audio recordings that were captured during these games were then used to study human deception [142]. On a related note, Zhou et al. [223] tested an online version of Mafia to study how well people could detect deception.

5.2.2 Piloting the Game

Before we decided the full details of our experimental protocol, we piloted various versions of Mafia with small groups of people. This experience helped us tune the game and adapt it our time constraints.

In total, we ran 3 pilot tests with 4 people each. Eight of the participants were female, and four were male. Their average age was 37.67 years old (SE = 4.76), and all except for two participants were born in the U.S. (one was born in Bosnia, another one was from Korea). All of the participants were fluent in English, and one of them was deaf. She was able to communicate with a hearing aid, though. Only two participants were acquainted by having participated in another experiment together prior to our pilot; the rest were complete strangers.

The experimenter moderated two games per pilot session. Each of these games followed a specific set of rules:

Version 1. Conventional Mafia game with no special roles. In order to assign the roles of mafia or villager, the participants picked playing cards from the top of Chester, who acted as a non-interactive piece of furniture for the purposes of the pilot. The game had 3 villagers and 1 mafia player. This information was not revealed to the participants.

The experimenter moderated the game starting with the *night phase*. In this phase, the mafia player secretly indicated a villager to "kill". Next, the *day phase* started, and the moderator revealed who was killed during the night. This person stepped back from the group and stayed quiet as if he/she were dead for the rest of the game. The remaining players then discussed who they thought was part of the mafia, and subsequently convicted a player upon agreement. The convicted player then left the game like the person that was killed by the mafia previously. This sequence of night and day phases continued until the mafia player was identified (villagers won) or only two players remained in the game (mafia won).

Version 2. Same as version 1 with a special (secret) role: one of the villagers was a "doctor". During the night, this person got the chance to save a player from the mafia.

Version 3. Same as version 2, but without reducing the number of participants during the day phase. Once the participants agreed upon a player who they thought was part of the mafia, the moderator acted as "police officer" and checked the role of the accused player. If this person was not part of the mafia, he or she stayed in the game. Otherwise, the game ended as the villagers correctly identified the only member of the mafia that they were playing with.

At the beginning of each game, the rules were briefly explained to the participants. Afterwards, we solicited their opinion about the particular version of the game that they had just played. At the very end of the session, we also asked them to rank the different versions of the Mafia game that they experienced. Figure 5.7 shows illustrative images from a pilot session.

(a) (b) (c)

Figure 5.7: Example session from the pilots. To secretly assign roles to the participants, the experimenter placed role cards face-down on top of the robot (a). People then chose one of these cards to get a team (villagers or mafia) and a specific role in the game, depending on the version that participants were playing. The experimenter then moderated the activity (b). When the participants discussed who was part of the mafia, the experimenter stepped out of the group (c). The robot was not interactive in the pilots; it just served as furniture.

Table 5.2 presents the versions of Mafia that were played per session of the pilot. The numbers inside the parenthesis indicate the duration of the day phase in minutes, which was chosen based on the feedback from the participants. First we tried 5 min – which was perceived as very long – and then 3 minutes. Finally, we tried reducing the time from 3 minutes to 2 in the last pilot. The participants responded positively to this change.

Table 5.2: Versions of the Mafia game that were tested in each pilot session. The numbers in between parenthesis indicate the duration of the day phases in each case.

Pilot Session	Game 1	Game 2
1	Version 1 (5 min)	Version 2 (3 min)
2	Version 1 (3 min)	Version 2 (3 min)
3	Version 2 (3 min)	Version 3 (2 min)

The overall feedback that we got from the pilot sessions was positive. The participants did not seem to feel uncomfortable playing Mafia with strangers, liked the social dynamics of the game, and most expressed that they would like to play again. We did notice, though, that most of the participants did not seem to know what to say to each other or how to figure out who was part of the mafia at the beginning of the first day phase. We tried improving the presentation of the instructions of the game in the real experiment to reduce this effect.

In terms of rankings, the second version of the game was the favorite among the first two pilot sessions (5 vs. 1 votes, excluding two people who reported picking their favorite version based on whether they were killed by the mafia). In the third pilot session, the version 3 of Mafia was preferred by 3 out of 4 participants. Their reasons included the way that the interaction went along and the new role of the moderator during the day phases.

5.2.3 Method

We designed the protocol for the experiment under a Wizard of Oz arrangement based on our experience from the pilots. This protocol was approved by our Institutional Review Board.

5.2.3.1 Participants

Participants had at least 18 years of age and were recruited through a local participation pool. English was their native language.

Ten groups of four adults played Mafia with our robot. In total, 22 women and 18 men participated, and their average age was 28.4 years old (SE = 2.1). Three sessions had balanced gender, three had 3 men, three had 3 women, and one was all women. All participants except for one were born in the United States.

In general, most participants did not know each other before the experiment. Only two people in two different sessions acknowledged knowing each other. Another person reported knowing one more participant in his session but did not say who. Most participants reported on a 7-point Likert scale using a computer daily (M = 6.9, SE = 0.08) and were not very familiar with robots (M = 3.05, SE = 0.24).

5.2.3.2 Procedure

Before beginning the experiment, the wizard hid in a room next to our laboratory and positioned the robot with its eyes closed within the environment where the interaction happened. When the participants entered this environment, they were given colored badges. These badges were used to identify the participants during the game, and to facilitate addressee identification when the robot communicated with them.

The participants began the experiment by completing a demographics survey, and watching an instructional video about the Mafia game.[2] The experimenter then woke up Chester and introduced it to the group. Subsequently, the participants played Mafia twice with the robot.[3] Each of these games was a different experimental condition:

(G1) In the first game, the experimenter moderated the activity, while the rest of the participants played with Chester. The rules of this game followed version 2 of the pilot, as described in Sec. 5.2.2. The roles were secretly and randomly assigned using cards (Fig. 5.8A), but the game was rigged such that Chester was always a villager. This allowed the interaction to continue even when the robot was erroneously convicted.

(G2) The second game was similar to (G1) but Chester served as the moderator. This reduced the number of players by one because the experimenter that ran (G1) did not participate in (G2). In this condition, Chester also played a "cop" who investigated the role of the accused players and revealed it at the end of the day phases (Fig. 5.8C). Unlike in (G1), accused villagers kept playing in (G2), as in the third version of Mafia that we tested in the pilot.

[2]The video was a modified version of "How to Play Mafia" (www.youtube.com/watch?v=75BMDrtpVtA). It explained how to play the second version of the game that we tested in the pilot.

[3]In general, we limited the day phases to 1.5 min maximum in both games.

Figure 5.8: Participants playing Mafia. (A) Players took a role card from Chester. (B) Chester played a game. (C) Chester moderated a game and checked the roles of the accused players.

We did not balance the condition order because we wanted to use Chester as a player who could break the ice in the first day phase without biasing participants' proxemic behavior. Since we feared proxemic bias due to the moderator of (G1) as well, she stepped away from the group of players when the game could continue without her (e.g., during discussions).

After each game, the participants completed a post-condition survey where they rated a few 7-point Likert scale items about the interaction (details about this survey are provided in the Results section). At the end of the experiment, people answered a final survey that queried their opinion about Chester's performance in the Mafia game and their overall experience. Lastly, we debriefed the participants about the presence of the wizard, and our interest in studying human spatial behavior in HRI.

5.2.4 Results

We analyzed the Mafia games, participant's perception of the robot, and their spatial behavior. The sensor data that was collected during the study helped inform our approach to enable robots to detect conversations by reasoning about spatial behavior (Chapter 5).

5.2.4.1 Gameplay

Both games lasted a few minutes on average (G1: M=295 secs, SE=36; G2: M=256 secs, SE=16). Chester was killed on the first night by one participant and on the second night by another. Moreover, the robot was typically convicted early in error. In particular, he was convicted 5 times at the end of the first day and twice at the end of the second day. It appeared that this sometimes happened because Chester started accusing players to break the ice, thereby generating suspicion. Another reason could be that the participants thought that it was easier to attribute blame to the robot than to another player in the game.

Overall, villagers won 3 times in G1 and 6 times in G2. It was easier to identify the mafia in G2 with fewer players and without incorrect convictions.

5.2.4.2 Post-Condition Survey

As shown in Table 5.3, the participants enjoyed Mafia in general. The ratings for "(b) Chester made the game fun" further suggest that the robot had a positive entertainment effect.

We conducted REML analyses on all post-condition items (a)-(d), (f) & (g) that queried perceptions of the robot or the interaction. For these analyses, we used Game (G1/G2), Participant Team (villager/mafia), Won (1/0 if the player got to the final phase and his/her team won/lost), and Cared About Winning (1/0 if the response to (e) was above/below 4) as main

Table 5.3: Post-condition ratings. Both conditions were used to compute average ratings and standard errors.

Statement	Avg. Rating	Std. Err.
a) I enjoyed this game of Mafia	5.36	0.14
b) Chester made the game fun	5.74	0.14
c) The interaction was enjoyable	5.80	0.12
d) I would have preferred to be part of the other team	3.63	0.20
e) I cared about winning the game	3.88	0.22
f) I would have liked to play longer	4.56	0.18
g) I liked the social dynamics of the game	5.13	0.13

effects, and Participant ID as a random effect nested within Session. We found significant differences for (f) in terms of Cared About Winning ($F[1, 78] = 5.11$, $p = 0.028$). As expected, a post-hoc t-test showed that the participants who cared about winning were more interested in playing for longer ($N = 35$, $M = 4.8$, $SE = 0.26$) relative to the rest ($N = 45$, $M = 4.38$, $SE = 0.24$). There was also a trend for higher (c) ratings when the players cared about winning ($F[1,78] = 3.2$, $p = 0.08$).

5.2.4.3 Spatial Behavior

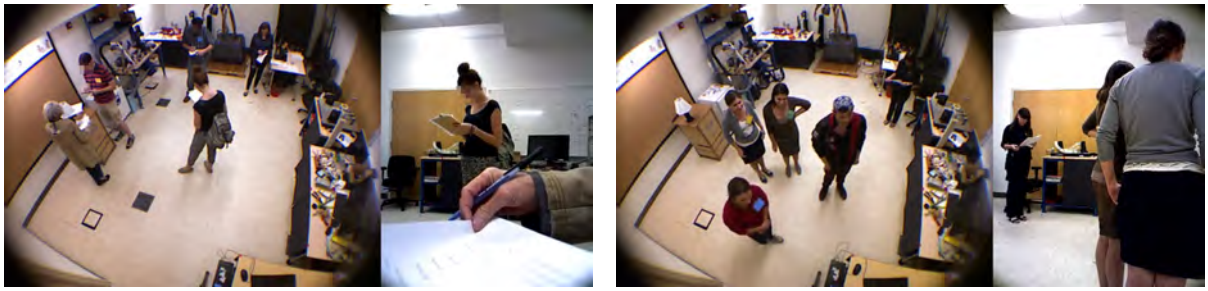
For proxemics, we annotated the positions of the players at 1Hz using laser measurements from the robot. The participants were 1.8m from Chester on average ($SE = 0.003$) during the phases of the games, which is within the typical range for human social interactions [60].

We further analyzed the average distances between the participants and the robot during the day phases of the games, when the robot actively interacted in both conditions. A Least Squares regression for Distance Type (to the robot or inter-participant) and Game showed a significant statistical differences for Distance Type ($F[1,38]=12.25$, $p < 0.01$). A post-hoc Student's t-test showed that the average inter-participant distance computed for the day phases ($N=20$, $M=1.46$m, $SE=0.07$) was significantly smaller than the average distance between the participants and the robot ($N=20$, $M=1.76$m, $SE=0.05$). Moreover, we noticed that the average distance to the robot increased proportionally from G1 to G2 ($N=10$, $M=1.69$m, $SE=0.06$ vs. $N=10$, $M=1.82$m, $SE=0.07$), but this difference lacked functional meaning. While we suspect that Chester's role could have slightly induced this variation in proxemics, it was small and may be influenced by the lack of counterbalancing.

We observed F-formations [86] during Mafia, as in Figure 5.8B and 5.8C. Circular arrangements often emerged when the games started and were sustained for most of the interaction. When the robot stepped out of the group in G1, we often observed spatial re-configurations (e.g., the players closer to Chester changed their orientation to subtly exclude the robot). Face to face spatial arrangements were sometimes initiated by the robot, e.g., when it accused players (G1) or announced deaths (G2).

5.2.4.4 Functional vs. Social

Until introduced by the experimenter, Chester was silent with closed eyes and many people did not notice that it was a robot (Fig. 5.9). Before Chester spoke, 6 participants stood close to its face with their backs to it (significantly blocking some of its sensors), 2 participants used it as a table for writing, and 2 other participants did both. These behaviors were not observed again after Chester was introduced, suggesting different use models based on user's perceptions.



(a) Demographics Survey (b) Instructional Video

Figure 5.9: Many people did not notice that Chester was a robot before the interaction started. This happened while the participants were filling the demographics survey at the beginning of the experiment (a), and when they were watching the instructional video about Mafia (b). The left images in each case show a top view of the scene. The right images were captured from the camera inside the lamp of the robot.

5.2.4.5 Chester's role

We asked the participants which role they preferred for the robot. Twenty-three participants (57.5%) selected moderator, sixteen (40%) selected player and one said that it was equal. Several factors supported their preferences, including interaction time with Chester (the longer they could interact with the robot the better), entertainment (how funny and engaging it was), role skills (e.g., *"good at organizing the group"* as moderator), how mechanical Chester seemed (e.g., *"more machine-like as moderator"*), its value to the game (e.g., *"helped (as player) because not all participants were very vocal"*), social inclusion (G1 *"makes Chester more part of the human crowd"*), perceived intelligence, and trust.

5.2.5 Discussion

The experiment had limitations. As in the previous study, our robot had limited verbal capabilities since its dialog was scripted. This constrained the set of responses that it could provide as a player in Mafia, and made it look repetitive at times.

When the robot's face had no animation scheduled,[4] we biased the gaze of the robot towards human faces. If there were more than one face in the view of the RGB-D camera inside of Chester's lamp, the robot looked towards the face that was closer to the middle of the image. As a result, the wizard had to orient the robot towards the person that it was addressing. In a few opportunities, though, this approach generated confusing gaze behaviors because of missed face detections and limitations with the speed of the robot as people moved around it.

Our experience with Chester suggests that people may have conflicting use models for furniture robots. On one hand, they may want to use them as objects, for utilitarian purposes. After all, they are a piece of furniture. On the other, they may assign human-like attributes to these machines, or treat them as pets, as they move around or behave socially. These attributions can then restrict their typical use. As noted by Sirkin et al. [165], this is a case of *mixed metaphors*. It is possible that robotic furniture will become its own genre in the future, making it easier for users to make sense of the actions executed by this class of robots.

[4]See Chapter 4 for a detailed description of how the robot operates.

The spatial behavior of the participants around the robot suggests that they treated it as a social agent once it was introduced to the group. The bigger separation between the robot and the participants that we observed in the experiment could be attributed to the fact that Chester is different – it is a robot, not a person. This fact may have made the participants more cautious around it, or made them perceive the social status of the robot as significantly different than that of the rest of the group. An alternative explanation is that Chester's face is planar. This form makes it difficult to observe the robot's eyes and facial expressions from its sides and, thus, encourages frontal interactions.

Most of our results did not suggest significant differences between the games (G1 and G2). Part of the problem could have been the lack of counter-balancing, and small sample size. Nonetheless, we believe that the experiment was useful for exploring group interactions with robots, and collecting data from which to start developing a method to reason about spatial patterns of behavior (Chapter 6). Different to the previous experiment, all participants in this case were adults, and had already learned the social conventions that typically guide our behavior in public settings. Nobody in this case violated Chester's personal space once it was introduced to the group, nor the space of the other participants.

We observed an increased separation between the participants and Chester during the second game of Mafia, in comparison to the first. Unfortunately, this difference was small enough, that we thought it had no functional meaning. We do believe, though, that a robot's role in a group can potentially be reflected in the way people stand with respect to it. Moreover, it may be possible to accentuate this role by manipulating its distance to users, because group leaders often separate themselves slightly from the rest of the members of a social interaction [86, 112].

The Mafia game is an interesting scenario to further investigate group interactions with robots. As we saw in our experiment, the game is engaging and entertaining for participants. It also provides opportunities to explore new aspects of HRI, such as the effect of the role of a robot within a group. Because the interaction is always bounded by the rules, Mafia is a practical choice to implement and design robot behaviors for. This does not mean, though, that Mafia is inflexible. Rules are often added to the game and modified. This feature provides significant leeway for tailoring the game towards specific aspects of a research agenda. In our opinion, the only big drawback of Mafia for studying HRI is that the participants need to understand the rules of the game well to play at their fullest potential. Thus, time must be spent in explaining and clarifying the instructions to the players.

5.3 Summary

This chapter presented two experiments that we conducted with our robotic platform to study spatial behavior and group interactions in two scenarios. In the first experiment, we used both Chester and Blink to study various social engagement cues and examine the effects of a sidekick character in a child-robot interaction. Our results suggested that the addition of a co-located sidekick has potential to increase user engagement, without altering proxemic behavior. In particular, children interacted with our robot from three spatial zones, which were typically occupied based on their activity. We also observed some evidence of F-formations during the experiment, but it was hard to annotate them because the young participants often did not cooperate to sustain their spatial arrangements. Rather, their spatial behavior was substantially variable and seemed impulsive at times. In addition, our findings reinforce the idea that furniture is a good robot design for children.

The limitations of our platform with respect to handling the natural dynamics of free standing conversations became especially evident during the first experiment. In this case, our robot

failed to react to cases in which the group of children with whom it was interacting became engaged with something else. The robot did not have any mechanism to detect these situations, nor effective strategies to regain children's attention. This limitation motivated us to work towards enabling robots to automatically detect nearby conversational groups.

In our second experiment, we used Chester to examine group interactions in the context of a social, role-playing game. The participants of this experiment interacted with the robot as one more player of the game and as the moderator of the activity. In both cases, we noted that the average inter-participant distance during the day phases of the games tended to be smaller than the average distance between the participants and the robot. Various reasons could potentially explain this result, such as the fact that Chester is a robot and its social status could have been perceived differently than that of the participants. We also noticed during the experiment that the separation between the participants and the robot was higher when it was moderating the game than when it was a player. Even though this difference was small and could have been influenced by the lack of counterbalancing, we suspect that Chester's role could have induced this result. In fact, some of the participants expressed preferring the robot as a player than as a moderator because being a player made it feel more part of their group. Other factors that influenced their preferences for Chester's role included interaction time, its value to the game, trust, and how skilled the robot seemed under each condition.

The second experiment was interesting from a design perspective. First, it provided evidence that suggests that people may have conflicting use models for furniture robots. Even though they may want to use them as objects for utilitarian purposes, they may assign human-like attributes to them that restrict their typical use as furniture. Second, the feedback from the second experiment was very positive overall, suggesting that social, role-playing games can be a good scenario for further investigating multi-party human-robot interactions.

Overall, we observed F-formations emerged during group interactions with our furniture robot. This validation is important because sets the foundations for using human spatial behavior as a mechanism to detect social group conversations in HRI, and improve robot perception. The next chapter describes our efforts in this direction.

Chapter 6

Detecting Group Conversations by Reasoning About Spatial Behavior

This chapter describes our efforts towards enabling robots to detect free-standing group conversations and their members. This ability is essential for these machine to adapt to the natural dynamics of social interactions and appropriately operate in human environments.

Different types of information can be used to detect group conversations, as discussed in Sec. 3.2. For example, one can rely on speech cues to detect small group configurations [27], based on synchronized turn-taking during group conversations. Another approach – which we follow in this dissertation – is to reason about spatial patterns of behavior that emerge during free-standing conversations [40, 54, 74, 155, 156, 218]. These methods can be applied in loud, public environments, and can be used to detect interactions in close proximity or from afar.

Here, we introduce an alternating optimization procedure to leverage the dependency between two problems to detect group conversations. One problem is estimating the lower-body orientation of people in a scene; the other one is detecting F-formations. Our efforts build upon prior work on reasoning about spatial human behavior, and our experience studying group conversations within HRI as described in Chapter 5. The proposed approach was first published in IROS'15 [199].

6.1 Problem Statement

As in prior work [40, 54, 74, 155, 156], we frame of the problem of detecting conversational groups in a scene as a clustering problem with an unknown number of clusters. More specifically, at any time t, our goal is to estimate a set of conversational groups $\mathcal{G}_t = \{G_1, G_2, \ldots\}$ by reasoning about F-formations in a scene. We model each group as a set G that holds the numeric identifiers of its members. For example, if the first group has three members, then $G_1 = \{a, b, c\}$, where $a, b, c \in \mathbb{N}$ are the identifiers of the interactants.

In general, we assume that the 2D position of the people of interest is given (e.g., as output by a person tracker), and that we can measure people's lower-body orientations (yaw angle) to some degree. In ideal circumstances, we can directly observe this orientation and use it to model people's transactional segments and find F-formations in a scene. In other circumstances, we may only have access to related observations, like measurements of peoples' head orientations.

Note that the above problem formulation is agnostic to particular sensing modalities. This is important for human-robot interaction applications where robots might be able to gather information about their environments with cameras, lidars, and various other sensors.

6.2 Prior Work

Chapters 2 and 3 of this dissertation present important background on human conversations and related work on detecting social interactions. For completeness, this section describes in more detail some important aspects of prior work on modeling and detecting F-formations.

Most existing methods to detect F-formations focus on finding o-spaces in a scene. Figure 6.1 describes their typical processing pipeline.

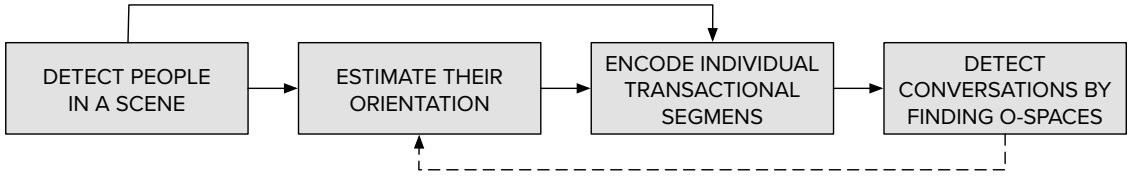


Figure 6.1: The solid arrows connect the typical steps involved in detecting F-formations and their corresponding conversations. In this work, we expand this model by leveraging F-formations to improve lower-body orientation tracking (dashed arrow).

For example, Cristani et al. [40] encoded the transactional segment of a given person i with his or her *o-space proposal* $\mathbf{o}^i = [x^i + d \cdot cos(\theta^i),\ y^i + d \cdot cos(\theta^i)]^T$, where $[x^i, y^i]^T$ corresponds to the position of the person, θ^i is her or her orientation, and d is a model parameter that controls how far away the proposal is from the person's body. To find F-formations, the authors then devised a Hough voting scheme to find the places in a scene where the proposals from different people intersect one another. The F-formation detection approach that we propose in this dissertation is inspired by this prior work. However, instead of computing hard group assignments as in[40], we focus on computing soft assignments to help overcome measurement uncertainty [32].

By definition, transactional segments are directed by people's lower-body orientation [86]. However, most prior methods to detect F-formations use people's head orientations or an estimate of their focus of attention to encode these segments [40, 74, 155, 156]. The reason is twofold: head orientations and the direction of people's focus of attention approximate lower-body orientation, and the former orientations are easier to estimate automatically than the latter.[1] This approximation, though, makes prior F-formation detection methods prone errors due to the inherent variability of people's attention span. To compensate for this variability, prior F-formation detection methods have often become very inclusive. They tend to group people together more often than they should. We discuss this problem further in the evaluation section of the present chapter.

A key insight of our work on detecting group conversations, is that we can leverage information about where these interactions are happening to better track the lower-body orientation of people in a scene. Not only people's orientation is important for detecting F-formations, as emphasized by prior work, but F-formations can also help estimate people's orientation. This idea is illustrated by the dashed arrow in Fig. 6.1.

[1] Two important exceptions are [218] and [54]. In the case of [218], a laser scanner was used to measure users' upper-body orientation with respect to a robot and detect triadic F-formations. This approach required direct view of the chests from the sensor, limiting generalization to other situations. In [54], body skeletons from multiple Kinects were used to detect F-formations. However, orientation measurements needed manual correction when frontal and backward skeletons were mislabeled. This situation could be improved by tracking orientations based on body measurements and contextual data, as proposed in this work.

6.3 Parallel Group Detection & Lower-Body Orientation Tracking

In this work, we propose a general framework to detect free-standing conversational groups and track the lower body orientation of people in a scene. An schematic representation of the framework is presented in Fig. 6.2.

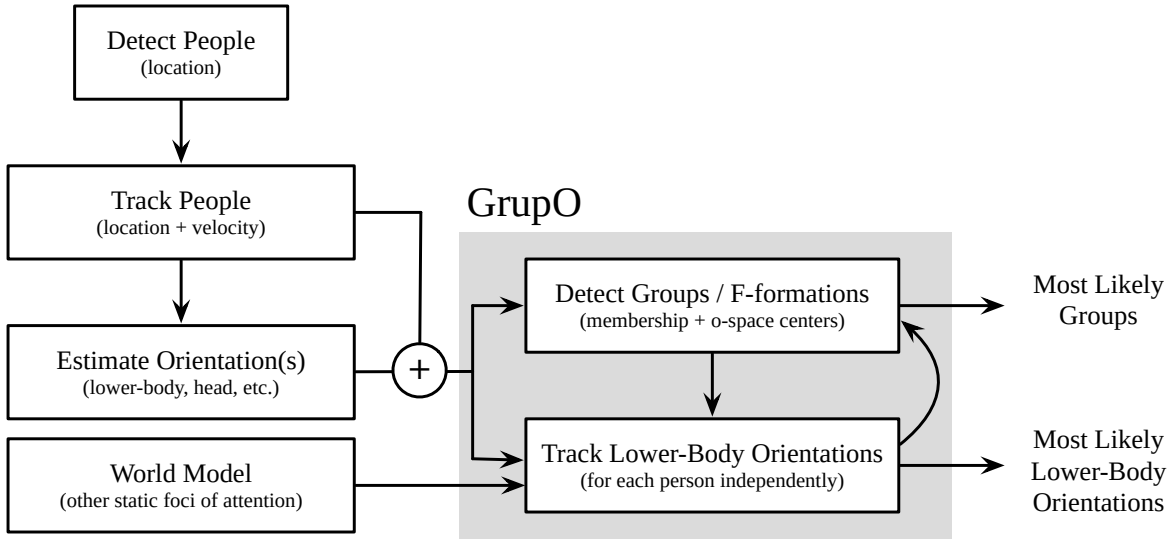


Figure 6.2: Schematic representation of our group detection framework. We propose to detect free-standing conversational groups and to track lower-body orientations in parallel.

At the core of the framework is an alternating optimization procedure that we named GrupO, for Group detection and Orientation tracking. At any given time, GrupO performs these two tasks to compute the most-likely group configuration G_t in a scene and the lower-body orientation of each person of interest (outputs in Fig. 6.2). First, GrupO estimates likely conversations by finding F-formations. These spatial organizations are detected by looking for o-spaces, based on the last positions and likely lower-body orientations that were estimated for the people in the scene. Then, the location of the o-spaces that were found and the probability of people belonging to them are used to update a lower-body orientation tracker per person. Each of these trackers considers not only group information, but also knowledge about nearby foci of attention in the scene to overcome noisy orientation measurements.

The next sections describe the group detection method and orientation tracker that we propose to use for GrupO. Particular implementation details and experimental results are presented afterwards in Section 6.4.

6.3.1 Model-Based F-formation Detection

We propose a new method for detecting F-formations and their members with GrupO. This method was inspired by the Hough voting scheme of [40] and, different to prior work, reasons about nonparametric lower-body orientation distributions. Thus, the proposed method can cope with situations of high uncertainty with respect to the orientation of people in a scene. Moreover, the proposed F-formation detection approach computes soft o-space assignments, which help lower-body orientation trackers recover from group detection errors in GrupO. Another difference between the proposed approach and the Hough voting scheme of [40] is that the proposed algorithm operates in a continuous space. This makes our method faster and more accurate than earlier voting approaches.

Our F-formation detection approach is detailed in Algorithm 6.3.1. This algorithm has four main steps:

1) Generating proposals for the location of the o-spaces in the scene. For each person i, we first generate N proposals for their o-space centers based on their position $\mathbf{p}^i$ and their nonparametric lower-body orientation distribution $\Phi^i = \{\phi^i[j] \in [0, 2\pi] \,|\, 1 \leq j \leq N\}$. The proposals are modeled as normal distributions $\mathcal{N}(\mu_j^i, \Sigma_j^i)$, one per sample j in Φ^i,

$$\mu_j^i = \mathbf{p}^i + R \begin{bmatrix} d \\ 0 \end{bmatrix}, \quad \Sigma_j^i = R \begin{bmatrix} (\sigma_x)_j^i & 0 \\ 0 & (\sigma_y)_j^i \end{bmatrix} R^T, \quad \text{and } R = \begin{bmatrix} \cos(\phi^i[j]) & -\sin(\phi^i[j]) \\ \sin(\phi^i[j]) & \cos(\phi^i[j]) \end{bmatrix} \tag{6.1}$$

where d is the *stride* of the model, and represents the expected separation between the o-space of the person and his or her body. The vector $\mathbf{p}^i$ corresponds to the position of the person i, and $(\sigma_x)_j^i$ and $(\sigma_y)_j^i$ are parameters that control the shape of the Gaussian. An example of possible values for these parameters is later provided in Section 6.4.[2]

2) Finding o-space centers. To find likely o-space centers in a scene, we combine everybody's proposals into a Gaussian mixture: $p(\mathbf{x}) = (1/NP) \sum_{i=1}^{P} \sum_{j=1}^{N} \mathcal{N}(\mathbf{x}; \mu_j^i, \Sigma_j^i)$ where μ_j^i and Σ_j^i come from equation (6.1). We then consider the local maxima of this mixture as likely o-space centers in the scene given people's spatial configuration.

To find the maxima, we use the iterative fixed-point algorithm of [29], starting from the means of the components. The function *fixedPointLoop* in line 11 of Alg. 6.3.1 corresponds to the "fixed point iteration loop" of [29] (see their Fig. 3 for more details). As in the latter work, we decide in line 12 whether a sample point $\mathbf{x}$ reached a local maxima based on the Hessian of the mixture distribution at that point. In general, it is possible that finding all the models of the mixture requires exhaustive search. In our experience, though, starting to search for the maxima from the means provides good results in practice with a reduced computational load.

We finally group the modes that are within τ meters from each other (line 14 of Alg. 6.3.1), and keep track of which component converged to which mode in the process. When this grouping happens, we set the mode with highest mixture probability as the most-likely o-space center in its vicinity. In this manner, the parameter τ helps coping with noise in human motion, as well as in our estimates of people's lower-body orientation.

3) Computing soft group assignments. Once the likely o-space centers are found, we count for each person how many of their mixture components converged per center, and compute their soft group assignment scores by normalizing this count (line 41 of Alg. 6.3.1). Note that this count only considers the o-space centers that are directly visible for people (line 35). We perform this visibility check by modeling individuals as circles with a fixed radius of 0.2m. We then use ray-casting to compute if any person occludes a likely o-space center for anybody else. The resulting soft group assignment scores are passed to the orientation tracker of the corresponding person in GrupO.

4) Computing hard group assignments. To obtain hard group assignments, we proceed in a greedy fashion and pick the likely o-space center with the highest score per person as his or her most probable o-space. A group is then set to be found whenever a likely o-space center has the highest score for two or more people (line 53 of Alg. 6.3.1).

[2]This model for o-space proposals is similar to those used in [40, 155, 156]. The difference with prior work, though, is that because we do not use a single (most-likely) orientation per individual, we generate more than one o-space proposal per person. The number of proposals that we generate per person depends on the number of samples in their nonparametric orientation distribution.

Algorithm 6.3.1: Detect F-formations by mode-finding

Input: Position $\mathbf{p}^i$ and nonparametric lower-body orientation distribution $\Phi^i = \left\{\phi^i[1], \ldots, \phi^i[N]\right\}$ of every person i in the scene $(1 \leq i \leq P)$

Output: Groups $\mathcal{G}$, list $\mathcal{M}$ of possible o-space centers, and lists $\mathcal{S}^i$ of o-space scores for every person

$\mathcal{X} = \emptyset$ // set of mixture components
$w = 1/PN$ // components' weight
for $i = 1$ *to* P **do**
  for $j = 1$ *to* N **do**
    $(\mu_j^i, \Sigma_j^i) = ospaceProposal(\mathbf{p}^i, \phi^i[j])$
    $\mathcal{X} = \mathcal{X} \cup \{(\mu_j^i, \Sigma_j^i, w)\}$
  end
end
$\mathcal{M} = [\,]$ // modes (possible o-spaces)
for $(\mu_j^i, \Sigma_j^i, w_j^i)$ *in* $\mathcal{X}$ **do**
  $\mathbf{x} = fixedPointLoop(\mu_j^i, \mathcal{X})$ // hill climb from the mean [29]
  if $\mathbf{x}$ *is local maxima* **then**
    $(idx, dist) = closestMode(\mathbf{x}, \mathcal{M})$
    if $dist < \tau$ **then** // group modes?
      if $p(\mathcal{M}[idx]; \mathcal{X}) < p(\mathbf{x}; \mathcal{X})$ **then**
        // x has higher probability
        $\mathcal{M}[idx] = \mathbf{x}$
      end
      $k = idx$
    else // add new mode
      add $\mathbf{x}$ to $\mathcal{M}$
      $k = |\mathcal{M}|$
    end
    $mode\_idx_j^i = k$ // bookkeeping
  end
end
// compute soft assignment scores
for $i = 1$ *to* P **do**
  $\mathcal{S}^i = [\,]$
  for $k = 1$ *to* $|\mathcal{M}|$ **do** // initialization
    $n_k^i = 0$
    add 0 to $\mathcal{S}^i$
  end
  for $j = 1$ *to* N **do**
    if $isset(mode\_idx_j^i)$ **then** // reached local maxima
      $k = mode\_idx_j^i$
      if $visible(\mathcal{M}[k], \mathbf{p}^i)$ **then**
        $n_k^i = n_k^i + 1$
      end
    end
  end
  if $\sum_k n_k^i > 0$ **then**
    for $k = 1$ to $|\mathcal{M}|$ **do** $\mathcal{S}^i[k] = n_k^i / \sum_k n_k^i$ **end**
  end
end
// greedy hard group assignment
$\mathcal{G} = \emptyset$
for $k = 1$ *to* $|\mathcal{M}|$ **do**
  $G = \emptyset$
  for $i = 1$ *to* P **do** // get the most-likely o-space
    $idx = \arg\max_m \mathcal{S}^i[m]$
    if $\mathcal{S}^i[idx] > 0$ *and* $k == idx$ **then**
      $G = G \cup \{i\}$
    end
  end
  if $|G| \geq 2$ **then** // found group / F-formation
    $\mathcal{G} = \mathcal{G} \cup \{(G, \mathcal{M}[k])\}$
  end
end

In general, our group detection approach assumes that we are dealing with open, public spaces, and that the configuration of this space does not generate occlusions that prevent people from interacting with one another. However, if there were walls or other static, big objects that could prevent interactions from happening in a given environment, our method could reason about them as it dealt with people occluding o-spaces (line 35 of Alg. 6.3.1). For example, we could run additional verification steps to check that o-space centers are not occluded by a wall or other big physical elements. As discussed in [117], reasoning about these occlusions can be crucial in complex environments.

6.3.2 Tracking Lower Body Orientations

When we track lower body orientations, we assume that people are standing at all times, as it happens during free-standing conversations. In addition, we assume that we are given the location of group conversations (encoded by the location of their o-space centers) and the likelihood of people belonging to these interactions (group assignment scores), e.g., as output by the group detection algorithm of Sec. 6.3.1. We also assume people's position and some measurement indicative of their true body orientation are provided. This measurement can be a noisy observation of their lower body orientation or of their head pose.

We propose to track body orientations using particle filters,[3] which can keep track of multi-modal orientation distributions. The latter property is advantageous for GrupO because it reduces the chances of getting stuck in local solutions. Furthermore, we propose to use one particle filter per person. This approach renders the orientation estimations of the people in a scene independent of each other, given measurements of their pose, and contextual information like the likelihood that they belong to nearby conversations.

In general, particle filters approximate posterior distributions with a finite number of samples $\mathcal{X}_t = \{x[1], \ldots, x[N]\}$, each of which is a concrete instantiation of the hidden state X tracked by the filter at time t. In GrupO, this state at least includes the lower-body orientation ϕ of a person. It could potentially also include other body features, such as head orientation.

The specific evolution of states, controls and observations of the filters is left open to particular implementations because they depend on the state variables and available measurements (examples are provided in Section 6.4 and Chapter 7). In general, though, we expect these filters to evolve as the Bayes network below:

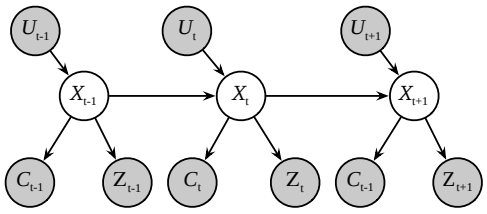


Figure 6.3: Proposed Bayes network that characterizes the evolution of the controls (U), observations (C and Z) and states (X) of the particle filters.

where C represents contextual observations (including group information and known foci of attention in the environment) and Z are orientation measurements (e.g., sensed head or body orientations). For control (U), one could use people's instantaneous velocity to predict how their lower-body orientation might change as they move, e.g., similar to [53].

[3] A friendly introduction to particle filters can be found in [184].

The belief $bel(X_t)$ for the network representing the evolution of the particle filter can be factored as:

$$bel(X_t) = p(X_t|U_{1:t}, Z_{1:t}, C_{1:t}) = \eta\, p(Z_t|X_t)p(C_t|X_t) \int p(X_t|X_{t-1}, U_t)bel(X_{t-1})dX_{t-1} \tag{6.2}$$

where η is a normalization term, $p(X_t|X_{t-1}, U_t)$ is the *state transition probability*, and the product $\big(p(Z_t|X_t)p(C_t|X_t)\big)$ corresponds to the *importance factor* of the particles. The next section provides a specific example on how this type of particle filter can be implemented.

6.4 Evaluation on the Cocktail Party Dataset

This section presents an evaluation of GrupO on a standard computer vision benchmark for group detection. First, we describe the dataset that we used for this evaluation. Then, we provide implementation details to illustrate how the group detection approach (Sec. 6.3.1) and the particle filters that we proposed for GrupO (Sec. 6.3.2) can be adapted to the dataset of interest. Finally, we describe our evaluation criteria, present the results, and briefly discuss our findings.

6.4.1 Dataset

We evaluate GrupO on the "Cocktail Party" dataset. This is a public dataset [155] that is often used to compare F-formation detection approaches within the field of Computer Vision.

The dataset consists of a sequence of more than 24000 images (recorded at 15Hz). The images show six people interact in an instrumented room, as shown in Fig. 6.4. For each frame, the dataset provides the location of each person and their head orientation, as computed by a custom person tracker. Furthermore, group annotations by an expert are given roughly every 5 seconds, for a total of 320 frames.

Figure 6.4: Images from the Cocktail Party dataset [155].

We collected annotations for the lower body orientation of the people captured in the dataset to complement the original ground truth. In particular, we collected annotations for people's lower-body orientation on the 320 images that had group annotations. This ground truth was collected using an interface similar to the one that was used in [110] to gather body orientations.

6.4.2 Implementation Details

We implemented GrupO to take advantage of the measurements provided in the Cocktail Party dataset. These measurements are the position ($\mathbf{p}^i$, $1 \leq i \leq N$) of the people in the scene and their head orientation (θ^i). We further integrated the positions of people to estimate their instantaneous linear velocities ($\mathbf{v}^i$).

6.4.2.1 Detecting Groups

To detect conversational groups, we used the method described in Sec. 6.3.1 to identify F-formations. Based on a small validation set, we used $\tau = 0.75$ in Alg. 6.3.1, and implemented the o-space proposals of eq. (6.1) with the following stride:

$$d = base\_stride + f(abs(\mathbf{d1}^T \mathbf{v}^i)) \tag{6.3}$$

where *base_stride* is the expected distance between given person i and his or her o-space center when this person is standing still while conversing. Based on [40], we use $base\_stride = 0.7$ in this evaluation.[4] The function $f(x) = 2\sigma(x) - 1$, with $\sigma(x) = 1/(1 + \exp(-x))$, adjusts the stride of based on the person's instantaneous velocity and the direction $\mathbf{d1} = [\cos(\phi^i)\ \sin(\phi^i)]^T$ of his or her body orientation. When the person moves forwards or backwards, his or her o-space moves further away up to 1 meter. When (s)he walks sideways, the o-space moves little because $abs(\mathbf{d1}^T \mathbf{v}^i)$ approaches zero in these cases. Also, $d = base\_stride$ for someone who is not moving.

For the covariance Σ_j^i of the o-space proposals in eq. (6.1), we set

$$(\sigma_x)_j^i = (base\_stride/s)^2 + f\Big(\frac{1}{2} abs(\mathbf{d1}^T \mathbf{v}^i)\Big) \tag{6.4}$$

$$(\sigma_y)_j^i = \lambda (base\_stride/s)^2 \tag{6.5}$$

with $\lambda = 0.25$, and $f, base\_stride$ the same as in eq. (6.3). We often set s to a value between 1 and 2.25 (more information about the influence of s is provided in Sec. 6.4.4.3). Figure 6.5 illustrates the flexibility of this o-space proposal model. In addition, Figure 6.6 shows example group detections with this approach in the Cocktail Party dataset.

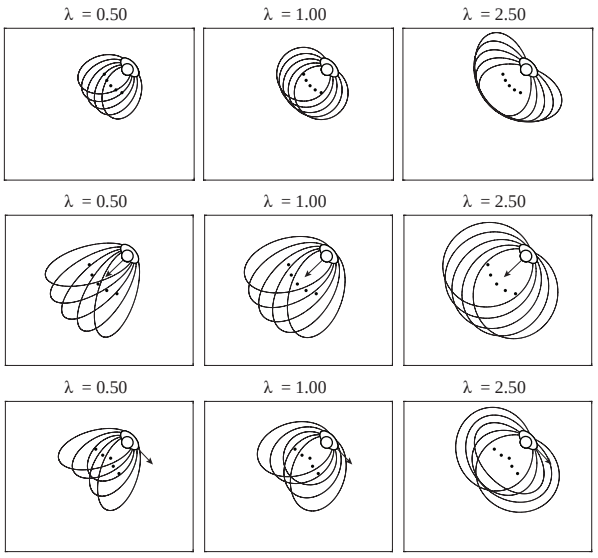
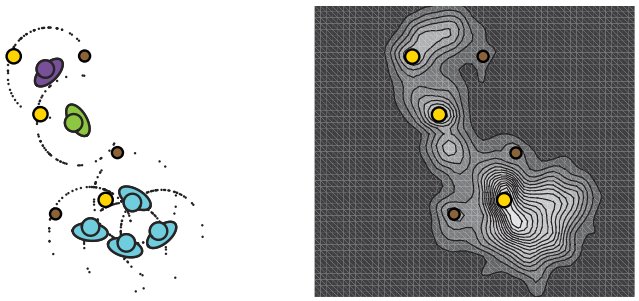


Figure 6.5: Example o-space proposals for 5 orientations at $0, \pm 0.25, \pm 0.5$ radians from the direction of the lower body of an individual. The person's velocity (indicated by an arrow) was zero for the first row, was aligned with the lower-body direction in the second one, and was perpendicular to it in the third. The black dots represent the means of the Gaussian distributions of eq. (6.1) and the ellipses represent their covariances at 99% confidence. The column show how the proposals vary based on λ in eq. (6.5).

[4] Appendix B describes an optimization approach that can be used to validate this baseline stride.



(a) Spatial configuration (b) Contour plot

Figure 6.6: O-space proposals for a frame of the Cocktail Party dataset. *Left:* Means μ of each person's o-space proposals (small black dots) and groups (by color) found in this scene by Alg. 6.3.1. *Right:* Mixture distribution of o-space proposals. The modes (in yellow and brown) were found for $\tau = 0.75$m (line 14 of Alg. 6.3.1). Yellow modes were the most likely o-space centers for at least one person.

6.4.2.2 Tracking Orientations

We implemented the lower-body orientations trackers used in this evaluation based on the following observations: (1) people tend to orient their lower body towards other people or objects of interest while standing still, (2) people often orient their head in the same direction as their lower body, (3) people can turn their heads (temporarily) to attend to visible targets in the scene other than their main focus of attention, and (4) people tend to orient their lower body towards their direction of motion while walking.

Consider a person i in a scene at any time t. We estimated a probability distribution for his or her lower-body orientation ϕ_t^i using the dynamic Bayesian Network of Fig. 6.7. This inference used estimates of the person's velocity $\mathbf{v}^i$, position $\mathbf{p}^i$, and head orientation θ^i. Moreover, it used contextual information C^i, which included:

– the position $\mathbf{p}^j$ (where $j \neq i$) of the other people in the scene;

– a set $\mathcal{O}$ with the locations of the nearby objects that people my interact with;[5] and

– the o-space centers $\mathcal{M}$ and corresponding assignment scores $\mathcal{S}^i[k]$, for $1 \leq k \leq |\mathcal{M}|$, as output by GrupO.

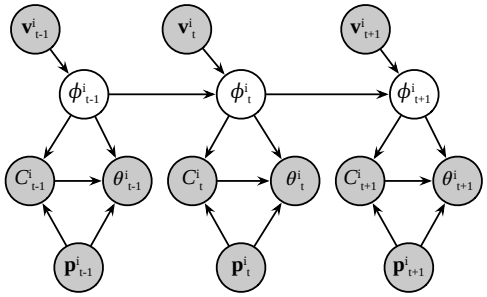


Figure 6.7: Bayes network that characterizes the evolution of the lower body orientation ϕ^i of a person i, based on his or her position $\mathbf{p}^i$, linear velocity $\mathbf{v}^i$, head orientation measurement θ^i, and contextual information C^i.

[5]For the cocktail dataset, $\mathcal{O}$ was composed of two points that represented the table in the room where the interaction happened (this table can be seen in Fig. 6.4).

As in Sec. 6.3.2, we formulated the belief $bel(\phi_t^i)$ at time t recursively:

$$bel(\phi_t) = p(\phi_t|\mathbf{v}_{1:t}, \theta_{1:t}, C_{1:t}, \mathbf{p}_{1:t}) = \eta\, p(\theta_t|\phi_t, C_t, \mathbf{p}_t) p(C_t|\phi_t, \mathbf{p}_t) \int p(\phi_t|\phi_{t-1}, \mathbf{v}_t) bel(\phi_{t-1}) d\phi_{t-1} \quad (6.6)$$

where we have dropped momentarily the superscript i for simplicity. In this factorization, η is a normalization term, $p(\theta_t|\phi_t, C_t, \mathbf{p}_t)$ is the *head measurement probability*, $p(C_t|\phi_t, \mathbf{p}_t)$ is the *context probability*, and $p(\phi_t|\phi_{t-1}, v_t)$ is the *state transition probability*. As suggested in Sec. 6.3.2, we used a particle filter to approximate the posterior $bel(\phi_t)$ with a finite number of samples $\Phi_t = \{\phi_t[1], \ldots, \phi_t[N]\}$, which we initialized from a uniform von Mises distribution $\mathcal{VM}(0, 0)$.[6] We followed a standard particle filter algorithm, as presented in Algorithm 6.4.1, to update the belief $bel(\phi_t^i)$. In practice, we used low variance sampling [184] for the last step of the algorithm (lines 7-10).

Algorithm 6.4.1: Particle filter for lower-body orientation

Input: $\Phi_{t-1}, \mathbf{v}_t, C_t, \theta_t$
Output: Φ_t
1 $\bar{\Phi}_t = \Phi_t = [\,]$
2 **for** $j = 1$ *to* N **do**
3 | sample $\phi_t[j] \sim p(\phi_t|\phi_{t-1}[j], \mathbf{v}_t)$
4 | $w_t[j] = p(\theta_t \mid \phi_t[j], C_t, \mathbf{p}_t)\, p(C_t \mid \phi_t[j], \mathbf{p}_t)$
5 | add $(\phi_t[j], w_t[j])$ to $\bar{\Phi}_t$
6 **end**
7 **for** $j = 1$ *to* N **do**
8 | draw k with probability $\propto w_t[j]$
9 | add $\phi_t[k]$ from $\bar{\Phi}_t$ to Φ_t
10 **end**

The following paragraphs detail the motion and measurement models used for our evaluation of GrupOon the Cocktail Party dataset:

Motion Model. For any person i, we propagated his or her lower-body orientation ϕ^i from time $t-1$ to t as follows:

$$\phi_t^i = \phi_{t-1}^i + \omega(\mathbf{v}_t^i, \phi_{t-1}^i)\Delta T + q \quad (6.7)$$

The angular velocity function $\omega(\mathbf{v}_t, \phi_{t-1})$ in eq. (6.7) controls the rate of rotation of the lower body, ΔT is the time difference from $t-1$ to t, and $q \sim \mathcal{N}(0, r)$ is a small perturbation. In particular, the function ω makes the body rotate towards the direction of motion of the person. As the linear velocity of the person opposes the direction of his or her lower body, ω becomes small in order to prevent sudden body rotations of 180°.

Context Model. We defined the probability of the context C_t^i as a mixture of three other probabilities:

$$\begin{aligned} p(C_t^i|\phi_t^i, \mathbf{p}_t^i) = {} & w_{\text{group}}\, p_{\text{group}}(C_t^i|\phi_t^i, \mathbf{p}_t^i) + w_{\text{eng}}\, p_{\text{eng}}(C_t^i|\phi_t^i, \mathbf{p}_t^i) + \\ & (1 - (w_{\text{group}} + w_{\text{eng}}))\mathcal{VM}(0; \phi_t^i, 0) \end{aligned} \quad (6.8)$$

with the sum $(w_{\text{group}} + w_{\text{eng}})$ of the non-negative weights in $[0, 1]$.

[6] We use von Mises distributions ($\mathcal{VM}$) for tracking the lower body orientations because they naturally model angular distributions [52]. In particular, $\mathcal{VM}(a; \mu, \kappa) = \exp(\kappa \cos(a - \mu))/2\pi I_0(\kappa)$, with $I_0(\cdot)$ the modified *Bessel* function of order zero. The parameters μ and κ are analogous to the mean and the inverse of the variance in the normal distribution. When $\kappa = 0$, the von Mises distribution becomes uniform.

The first component,

$$p_{\text{group}}(C_t^i|\phi_t^i, \mathbf{p}_t^i) = \sum_{k=1}^{|\mathcal{M}_t|} \mathcal{S}_t^i[k]\mathcal{VM}(\beta_k; \phi_t^i, \kappa_{\text{group}}) \; + (1 - \sum_{k=1}^{|\mathcal{M}|} \mathcal{S}_t^i[k])\mathcal{VM}(0; \phi_t^i, 0) \tag{6.9}$$

corresponded to the probability of the person belonging to an o-space given his or her spatial configuration. The angle β_k in eq. (6.9) was the direction of the vector $\mathcal{M}[k] - \mathbf{p}_t^i$. The spread κ_{group} controlled shape of these von Mises distributions.

The second component was another mixture probability,

$$p_{\text{eng}}(C_t^i|\phi_t^i, \mathbf{p}_t^i) = \sum_{v=1}^{V} e_v \, \mathcal{VM}(\beta_v; \phi_t^i, \kappa_{\text{eng}}) \tag{6.10}$$

which represented the likelyhood of engagement with another individual or object v within a field of view of 180°. Here, the angle β_v represented the direction towards this other individual or object of interest from the position of person i.

Finally, the last component $\mathcal{VM}(0; \phi_t^i, 0)$ in eq. (6.8) was a uniform distribution that represented the likelihood of filing to explain the context of the person with his or her orientation.

Head Measurement Model. Similar to the context model, we defined the head measurement model as a mixture of probabilities:

$$\begin{aligned} p(\theta_t^i|\phi_t^i, C_t^i, \mathbf{p}_t^i) = w_{\text{front}} \, p_{\text{front}}(\theta_t^i|\phi_t^i) \; + \; w_{\text{focus}} \, p_{\text{focus}}(\theta_t^i|\phi_t^i, C_t^i, \mathbf{p}_t^i) + \\ (1 - w_{\text{front}} - w_{\text{focus}}) \, \mathcal{VM}(\theta_t; 0, 0) \end{aligned} \tag{6.11}$$

where the sum of the non-negative weights is also in $[0, 1]$. The first component of the mixture accounted for frontal headings,

$$p_{\text{front}} = \mathcal{VM}(\theta_t^i; \phi_t^i, \kappa_{\text{front}}) \tag{6.12}$$

The second component,

$$p_{\text{focus}}(\theta_t^i|\phi_t^i, C_t^i, \mathbf{p}_t^i) \propto \max_{v=1\ldots V} \{\mathcal{VM}(\theta_t^i; \beta_v, \kappa_{\text{focus}})\} \tag{6.13}$$

was proportional to the maximum likelihood of orienting the head towards a (non-occluded) person, object of interest, or most likely o-space center within a 180 deg field of view in front of person i. The third component accounted for unexplained head orientation measurements.

6.4.3 Group Detection Criteria

We adopted the two criteria in [40, 155, 156] for analyzing group detection results versus ground truth annotations. One criteria considered a group to be detected if at least $\lceil (2/3)|G| \rceil$ of its members were identified and no more than $1 - \lceil (2/3)|G| \rceil$ of false subjects were found, where $|G|$ is the cardinality of the group. The other criteria considered a group to be detected if all its members were identified correctly and no false members were found. Precision, recall and F1 scores were computed using these criteria, summing true positives, false positives, and false negatives over all the frames with group annotations.

6.4.4 Results

We compared the performance of the proposed F-formation and tracking algorithms against the state-of-the-art approach of [156]. We used their open-source implementation[7] to generate the results for their method in this evaluation.

As part of this evaluation, we study the performance of the proposed group detection method (Alg. 6.3.1) with head measurements and lower body annotations directly (i.e., without GrupO). To compute these results, we generated an artificial (non-parametric) orientation distribution Φ with $N = 30$ samples. This distribution was generated by sampling $\mathcal{N}(\phi, q)$, where the mean ϕ corresponded to the head measurement or true body orientation of the person. The variance q was a small number that controlled the spread of the samples. In particular, we used $q = 0.07$ when we evaluated the performance of the algorithm with ground truth annotations, and $q = 0.13$ when we used head measurements directly.

6.4.4.1 Orientation Estimation

To verify that GrupO was working, we analyzed the most-likely lower-body orientations that it was estimating for every person in the scene. We compared these values against the head measurements that were provided as part of the Cocktail Party dataset, and the lower-body orientation annotations that we collected for 320 frames. For this test, we ran GrupO with Alg. 6.3.1 and set $s = 2$ in eq. (6.5) and (6.4), which we found to work well in practice (as laser discussed in Sec. 6.4.4.3).

As expected, GrupO tended to better approximate ground truth lower body orientations than the head orientation measurements. Figure 6.8 shows superimposed histograms of the absolute angular difference between lower body orientation annotations and head measurements, and between the annotations and the estimated lower body directions on a typical run of GrupO. On average, the head measurements were 0.59 radians ($\sim 34°$) off from the body annotations (SE=0.013). Using GrupO, the estimated lower body orientations were 0.38 radians ($\sim 22°$) on average from the annotations (SE=0.008).

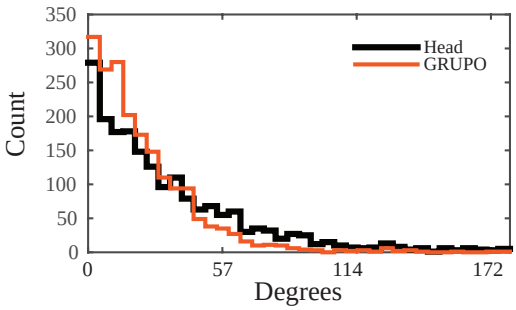


Figure 6.8: Angular difference to lower body annotations

6.4.4.2 Detecting Groups with Graph Cuts [156]

We ran GrupO with the group detection method of [156], which iteratively applies graph-cuts to find F-formations and cluster together the people in a scene. The soft o-space scores used by our particle filters were set to binary $\{0, 1\}$ values depending on the detected groups. This was necessary because the method of [156] only provides hard group assignments.

Table 6.1 shows the best results obtained with [156]. Our exploration of the MDL parameter that controls the behavior of graph-cuts within this approach is documented in [199]. In summary, we found that no single MDL parameter worked best for all input orientations. In fact, the more noise, the higher the MDL should be. But there is a trade-off: the higher MDL, the more inclusive the graph-cuts approach becomes. In other words, high MDL values induce graph-cuts to group people together more often than not.

[7] http://profs.sci.univr.it/~cristanm/ssp/

Table 6.1: Group detection results using the graph-cuts F-formation detection approach of [156]. Results for GrupO were averaged over 5 runs (std. errors were equal to or less than 0.005).

Criteria	Orientation	MDL	Precision	Recall	F1
$\lceil(2/3)\|G\|\rceil$	Lower-Body Annotations	14000	0.84	0.84	0.84
	Head Measurements	30000	0.82	0.81	0.82
	GrupO	14000	0.82	0.80	0.81
$\|G\|$	Lower-Body Annotations	14000	0.69	0.68	0.69
	Head Measurements	30000	0.62	0.61	0.61
	GrupO	14000	0.61	0.60	0.61

Not surprisingly, the graph-cuts approach of [156] works best with lower body ground truth annotations. Our intuition as to why GrupO with graph-cuts does not improve the results over using head measurements directly is that the approach of [156] only outputs hard group assignments. These groups often include false members, and these errors can easily propagate within GrupO.

6.4.4.3 Detecting Groups with Algorithm 6.3.1

Table 6.2 provides the precision, recall and F1 scores for the F-formation detection method proposed in Algorithm 6.3.1. Each row shows the best parameter s for eq. (6.5) and (6.4). In general, the smaller s, the more spread the o-space proposal distributions were.

Table 6.2: Group detection results using Alg. 6.3.1. The parameter s corresponds to eq. (6.5) and (6.4). Results were averaged over 5 runs (std. errors were equal to or less than 0.003).

Criteria	Orientation	s Param	Precision	Recall	F1
$\lceil(2/3)\|G\|\rceil$	Lober-Body Annotations	2	0.86	0.83	0.85
	Head Measurements	1.25	0.81	0.80	0.81
	GrupO	2	0.82	0.80	0.81
$\|G\|$	Lower-Body Annotations	2	0.71	0.69	0.70
	Head Measurements	1.25	0.60	0.59	0.60
	GrupO	2	0.65	0.63	0.64

We found that Algorithm 6.3.1 was as powerful as the graph cuts approach of [156] when we used lower body annotations to orient the transactional segments of the people in the scene. The results for using head measurements directly were slightly lower in this case, but GrupO tended to perform better under the full group detection criteria. Figure 6.9 shows a few illustrative group detection results for GrupO with Algorithm 6.3.1.

6.4.4.4 Individual Interaction Detection

We examined how well [156] could infer if people were interacting or not versus GrupO. Table 6.3 shows the results from this binary classification task, where accuracy is $(TP + TN)/(TP + FP + TN + FN)$, the true positive rate is $TP/(TP + FN)$ and the true negative rate is $TN/(TN + FP)$, with TP the number of true positives, TN the true negatives, FP the false positives, and FN the false negatives. While GrupO and the method of [156] have similar accuracy at the individual level, GrupO is able to double the true negative rate of [156], without any additional computer vision processing.

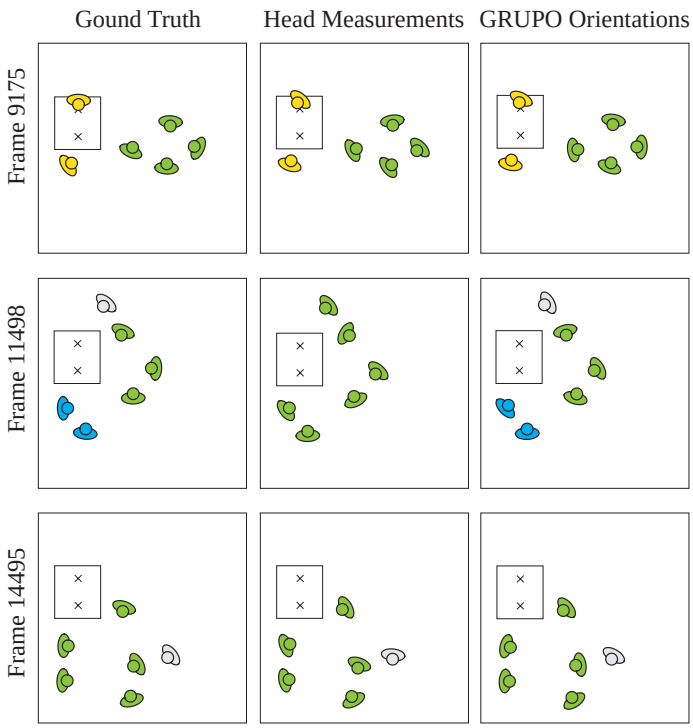


Figure 6.9: Qualitative results for GrupO on the Cocktail Party dataset. The first column shows ground truth groups (by color) and lower body orientations. The second one shows group detections using head measurements. The third uses estimated lower body orientations.

Table 6.3: Individual interaction classification results. We used MDL= 30000 for [156] and $s = 2$ (eq. (6.5) and (6.4)) for Alg. 6.3.1. GrupO results were averaged over 5 runs.

Metric	GC (Head) [156]	GrupO
True Positives	1739	1707.4 (SE = 2.0)
False Positives	140	97.4 (SE = 1.1)
True Negatives	39	81.6 (SE = 1.1)
False Negatives	2	33.6 (SE = 2.0)
Accuracy	0.93	0.93 (SE < 0.01)
True Pos. Rate	1.00	0.98 (SE < 0.01)
True Neg. Rate	**0.22**	**0.46** (SE < 0.01)

6.4.5 Discussion

Our results on the Cocktail Party dataset suggest that GrupO can help better detect non-interacting people, without sacrificing group detection performance. This is particularly important for social robots in human environments. For instance, detecting nearby bystanders effectively can help robots adapt to changes in the members of their conversations. Moreover, detecting bystanders can provide opportunities to start new social interactions.

In general, any F-formation detector and lower-body orientation tracker can potentially be used with our group detection framework. However, our experiments suggest that GrupO works better with soft clustering methods than with traditional approaches that compute hard group assignments. Furthermore, we believe that it is beneficial to reason about non-parametric orientation distributions in GrupO. Avoiding committing to the most-likely set of groups and the most-likely lower-body orientations for the people in a scene can prevent propagating errors through the proposed alternating optimization procedure.

Chapter 7

Understanding the Effects of Body Orientation and Gaze

In the first two experiments presented in Chapter 5, the orientation of the robot was manually controlled by a wizard during the whole interaction. Moving forward, we wanted to examine the possibility of controlling the orientation of the robot automatically during a group conversation with three or four people. How should robots cooperate to sustain F-formations while engaged in a conversation? One strategy is to mimic human behavior. Robots can orient their body towards the middle of their group, as illustrated in Fig. 7.1a. This approach was previously proposed by Althaus and colleagues [11] and was associated with more positive perceptions of the behavior of a telepresence robot [203]. However, this strategy is not the only reasonable one for mobile, low degree-of-freedom (DoF) robots. For robots with a fixed head, such as Chester, FROG [47], SPENCER [189], or any of the the CoBots [200], it may be better to orient towards the focus of attention of the conversation, e.g., the speaker, as illustrated in Fig. 7.1b. This behavior could help establish common ground [87, 168], convey attentiveness to the interaction, and make users perceive robots as more active or responsive.

To further our understanding of robot positioning during group conversations, we conducted another experiment where our mobile robot interacted with small groups of people. In this case, we manipulated the orientation of the robot during conversations, but also its gaze. Even though our robot has a fixed head, the direction of its eyes can still communicate mental states and attention [3, 14]. Thus, we expected gaze to affect the perception of the orientation of the robot.

We designed a new protocol for this experiment to continue studying spatial behavior in HRI and avoid the bystander effect that we observed in our first study (Sec. 5.1). In the

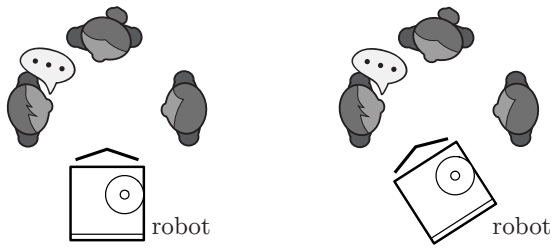


Figure 7.1: Strategies to orient a robot with respect to the members of its group conversation. In (a), the robot orients towards the center of the group. In (b), it orients towards the speaker.

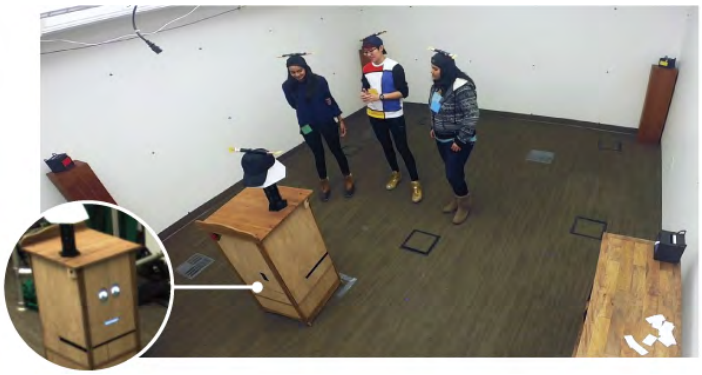


Figure 7.2: HRI experiment where we tested orientation and gaze behaviors for our robot.

new protocol, the robot asks the participants to help it solve a problem in a brainstorming session in the laboratory (Fig. 7.2). Even though this is not a public setting, our design makes the interaction naturalistic. Participants are free to move in the environment as desired and, periodically, are induced to leave the robot's conversational group to document their ideas. This dynamic creates a variety of group formations on a frequent basis, thereby generating numerous instances for studying multi-party interactions. For example, groups with one to four people emerged during our experiment as a result of the flow of the activity. In addition, the proposed brainstorming protocol does not require us to provide the participants with specific instructions on their roles. This property leads to increased interaction time during experiments in contrast to the prior role-playing game protocol described in Section 5.2.

The body and gaze behaviors that we tested for our robot were controlled automatically by a multi-modal perception system during most of the the brainstorming activity. This system relied on several off-the-shelf sensors and data fusion techniques to (1) track the users and the robot, (2) detect conversational groups by reasoning about spatial behavior, and (3) detect the current speaker in the environment. Although the core components of this system were developed previously, their integration is valuable. First, it allowed us to run parts of the experiment in an automated fashion. Second, it allowed us to collect a corpus of human spatial behavior with and around our robot, which helped us validate further the idea that people tend to establish spatial arrangements typical of human conversations with robots. This system and the experiment were published in HRI'17 [196].

The next section presents prior work that was not mentioned previously in this dissertation. Then, we describe in detail the behaviors that we evaluated during our experiment and the multi-modal perception system that we used to control them. The sections that follow present our experiment methodology and results. Finally, this chapter concludes with a brief discussion of our findings, their implications, and the limitations of this work.

7.1 Prior Work

Important background on spatial behavior typical of group conversations was previously presented in Chapter 2. Also, Chapter3 described relevant prior work in HRI on proxemics. Here, we focus on describing other important prior efforts on robot orientation, social gaze, users' sense of groupness in HRI, and multi-modal perception.

The middle orientation behavior that we study in this work was proposed earlier for social navigation [11]. Karreman and colleagues [80] implemented this behavior on a museum guide robot that gave short tours to visitors. Turning the robot towards

visitors led to increased interest in the platform in contrast to turning towards points of interest, like art pieces. There is also evidence that suggests that orienting a telepresence robot towards the center of a group makes people comfortable [203].

There is significant work in social eye gaze for human-computer and human-robot interaction [3, 151]. Related to our work, Garau and colleagues [55] found that synchronizing an avatar's head and eye animations with turn-taking patterns could improve its communication with humans in comparison to a random gaze behavior in which its head and eye animations were unrelated to conversational flow. As in our experiment, random gaze behaviors were also used in prior efforts to study robot gaze. For example, Yoshikawa and colleagues [217] compared a random gaze behavior versus three other gaze behaviors on a Robovie-R2 platform. Their experiment suggests that responsive robot gaze, e.g., gaze that communicates shared attention, induces stronger feelings of being looked at on users in comparison to non-responsive gaze. In addition, Skantze and colleagues [166] studied a random gaze behavior versus a human-inspired gaze behavior on a Furhat robot. This robot has back-projected eyes like the platform used in this work.

Other research has focused on analyzing gaze duration and frequency. For example, prior work [4] suggests that short, frequent fixations by a robot can give an observer stronger feelings of being looked at versus longer, less frequent stares. Also, a robot that looks towards users more often may be perceived as more extroverted than to one that looks more towards the task space [12]. Note that gaze can also influence people's roles in a conversation with a robot [89, 128] and their attitudes towards these machines [79]. Some gaze behaviors may work better than others, depending on the type of conversation [34].

Several efforts within HRI have investigated how much people perceive themselves as part of a group [66, 115, 128, 141]. Similar to prior work, we follow the approach of Mutlu and colleagues [128] to measure interpersonal closeness to our robot with the "Inclusion of Other in Self" (IOS) scale [15]. We use the survey by Williams and colleagues [212] to measure feelings of groupness and ostracism.

Our perception system was inspired by prior work in multi-modal sensing [23, 24, 82, 98, 130, 162, 163, 186] and is an alternative to other approaches meant to enable human-robot interactions in controlled settings. In particular, our system estimates users' positions and body orientations by fusing ultra wide-band tracking information and skeleton data output by a Kinect. Even though prior work used ultra wide-band localization systems to track people [21, 57] or the Kinect to enable interactions [7, 71, 120, 221], we are the first to fuse these types of data for HRI to the best of our knowledge. The fusion offers key advantages: operation beyond the Kinect's range, better occlusion handling, and simple user identification.

Our perception system also builds on advances in localization [184] and human spatial analysis [40, 86, 199]. While recent efforts to detect social interactions based on spatial behavior have focused on analyzing users only [105, 109, 143], we opt to jointly reason about the users' and our robot's spatial configurations under a unified perspective.

7.2 Orientation And Gaze Behaviors

We studied two orientation and two gaze behaviors during group conversations with our furniture robot Chester. As detailed in Chapter 4, this robot has a differential drive base, a fixed face, and back-projected eyes. Even though the robot's design led to specific decisions for the orientation and gaze behaviors, they can be easily adapted to other mobile platforms with expressive eyes. We detail our implementation to facilitate future explorations in this direction.

For the following explanations, assume that the robot has started a conversation and that we know its position $\mathbf{r} = [r_x\ r_y]^T$ and orientation ρ (yaw angle) on the ground. Assume that we also know the position $\mathbf{p}^i$, the lower body orientation, and the velocity of any person i near the robot, so that we can detect its conversational group by reasoning about F-formations (e.g., using the methods described in Chapter 6). Finally, assume that we know who is speaking in the robot's conversation. Data collection methods are later described in Sec. 7.2.3.

7.2.1 Body Orientation Behaviors

For any member i in the robot's conversation, let $\mathbf{u}^i = [u_x^i\ u_y^i]^T = \mathbf{p}^i - \mathbf{r}$ be the direction from the robot to this person, and $\gamma^i = \text{atan2}(u_y^i, u_x^i)$ the corresponding angle. We used this angle to orient the robot as described below.

7.2.1.1 Middle Orientation Behavior (MO)

The robot oriented towards the middle of its conversational group G using the *mean direction* $\bar{\theta}$ of all γ^i [52]:

$$\bar{\theta} = \text{atan2}\left(\sum_{i \in G} \sin(\gamma^i), \sum_{i \in G} \cos(\gamma^i)\right) \tag{7.1}$$

7.2.1.2 Attentive Orientation Behavior (AO)

If the robot was speaking, it biased its orientation towards its addressee; otherwise, it biased its orientation towards the current speaker in its conversational group. Let γ^i be the orientation towards the speaker or the addressee, and $\bar{\theta}$ be the middle orientation in the group, as in eq. (7.1). At any given time, the orientation $\hat{\rho}$ of the robot was set as follows:

$$\hat{\rho} = \begin{cases} \bar{\theta} - \tau & \text{if minAngDiff}(\gamma^i, \bar{\theta}) < -\tau \\ \bar{\theta} + \tau & \text{if minAngDiff}(\gamma^i, \bar{\theta}) > \tau \\ \gamma_i & \text{otherwise} \end{cases} \tag{7.2}$$

where minAngDiff returns the signed minimum difference between two angles, and τ is a parameter that controls how much the robot rotates away from the middle orientation $\bar{\theta}$ (Fig. 7.3). In particular, we set $\tau = 60°$ for our robot so that it would not turn its back to group members to its side.

If the robot was not addressing anyone and nobody had spoken for a significant time (10s), the platform's orientation was set towards the middle direction as in the MO behavior. This also happened when the robot conversed with a single user, given that $\hat{\rho}$ in eq. (7.2) became $\bar{\theta}$.

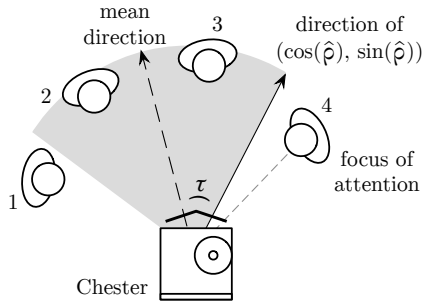


Figure 7.3: Geometric relations for the AO behavior. The $\wedge$ mark denotes the robot's front. The middle direction corresponds to eq. (7.1) and $\hat{\rho}$, τ to eq. (7.2).

7.2.2 Gaze Behaviors

We tested simple gaze behaviors to complement the effects of our orientation manipulation. These behaviors serve as a baseline for future investigations on the relationship between body motion and complex gaze patterns, e.g., involving discourse structure or fixations on the environment [30, 126].

As described in Chapter 4 and Appendix A, we calibrated the gaze of the robot using a projective mapping from 3D world coordinates to 2D pupil positions. We used the mapping for both pupils of the robot, making their lines of sight parallel. While this constraint prevented vergence eye movements, it worked well in practice because the robot's eyes look cartoonish and have a slight curvature. This design makes users forgiving of gaze patterns that do not fully mimic human gaze and induces the Mona Lisa gaze effect [8]: users perceive mutual gaze more often than intended.

7.2.2.1 Random Gaze Behavior (RG)

The robot executed several pre-defined eye animations that helped communicate ideas while it spoke. For example, referential gaze was used at times with verbal utterances to convey spatial information. When no pre-defined animation was scheduled for the eyes, they blinked occasionally or their pupils moved randomly at small intervals.

Our specific implementation of eye blinks was inspired by human blinking activity [43]. The duration of inter-blink intervals followed a normal distribution $\mathcal{N}(5.2, 3^2)$ in seconds.

Gaze shifts were scheduled by sampling time intervals in seconds from the uniform distribution $Unif(1.8, 3)$. When the timer triggered and no blink was set to occur, the pupils moved a small amount horizontally $d_x = eye\_width * \epsilon_1$ and vertically $d_y = eye\_height * \epsilon_2$, based on the size of the eyes. The values ϵ_1 and ϵ_2 were sampled uniformly in a small interval. Any displacement (d_x, d_y) that rendered the pupils outside the limits of the eyes was considered invalid and was re-computed by sampling new values. Furthermore, we prevented Chester from fixating significantly downwards, towards the ground, so that it would not look extremely introverted.

7.2.2.2 Attentive Gaze Behavior (AG)

The robot used the same blinking pattern and pre-defined eye animations as in RG. When no animation was scheduled, the robot attempted to establish mutual gaze with the person who was the focus of attention. That is, the person that the robot addressed in particular, the current speaker if the robot was quiet, or anybody who moved with a speed of at least 0.5 m/s in the group when everybody was silent.

Once the robot gazed towards someone, gaze shifts were sampled as often as in RG but were biased towards the head of the focus of attention. If $\mathbf{q}$ is the 3D position of the head, then the new, biased positions for the pupils were set as:

(x,y) = lookAt($\mathbf{q}$) `// pupils position towards` q
$r \sim Unif(0, 1)$
if $r < 0.2$ **then** `// add noise` 20% `of the time`
    $x = x + eye\_width * \alpha_1$ with $\alpha_1 \sim \mathcal{N}(0, \sigma^2)$
    $y = y + eye\_height * \alpha_2$ with $\alpha_2 \sim \mathcal{N}(0, \sigma^2)$

where lookAt returned the 2D location of the pupils that made the robot look towards the desired direction, and σ controlled the amount of variation in gaze shifts. After 10 seconds of silence and no significant motion in the group, gaze shifts continued without the bias as in RG.

7.2.3 Multi-Modal Perception System

We implemented a real-time system to control the robot's orientation and gaze based on human behavior, as well as to collect data during the experiment. The system required instrumenting the environment with ultra wide-band (UWB) localization beacons[1] and a Kinect. Each participant wore an instrumented baseball cap with two UWB radio beacons for tracking and identification (Fig. 7.2). The robot also wore a cap to make it look like the participants.

7.2.3.1 System Components

Figure 7.4 shows the main components of the system. Grey boxes denote modules that ran on the robot; the rest executed on external computers. The boxes with thicker edges correspond to modules that were in charge of the manipulated behaviors. Note that the robot's speech was controlled by a hidden operator, as detailed in Sec. 7.3.1.

The system processed data as follows. First, the position of the UWB beacons carried by the participants was smoothed with a Kalman filter ("Filter" module in Fig. 7.4). The smoothed values were then aggregated to estimate the position and orientation of each hat ("Hat Pose" module) and fused with the skeleton output of a Kinect ("User Tracker" module). This fusion step output estimates of the position and orientation of each participant, taking advantage of both sensing modalities. The Kinect reduced localization error, which ranged up to 30 cm on average for the hats. The UWB data provided continuous tracking information throughout the environment as well as participants' identities.

While users were localized, the "Robot Localization" module estimated Chester's pose using an on-board laser scanner and a map of the environment [184]. The "Aggregator" program then combined all this information and passed it to the "Group Detector" and "Speaker Detector" modules. The former module reasoned about conversational groups based on F-formations, as in Chapter 6 and illustrated in Fig. 7.5. The latter module was in charge of identifying the current speaker based on the interactants' positions, sound detections (output by "K2 Audio"), and information from Chester's dialog engine ("Mouth Animation" module). If Chester's mouth was moving, the robot was identified as the current speaker. Otherwise, the speaker was the person closest to the Kinect's audio beam (within 1 m) or nobody when no sound was localized.

[1] We used DWUSB sensors by Ciholas, Inc. Appendix C describes these wireless sensors in detail.

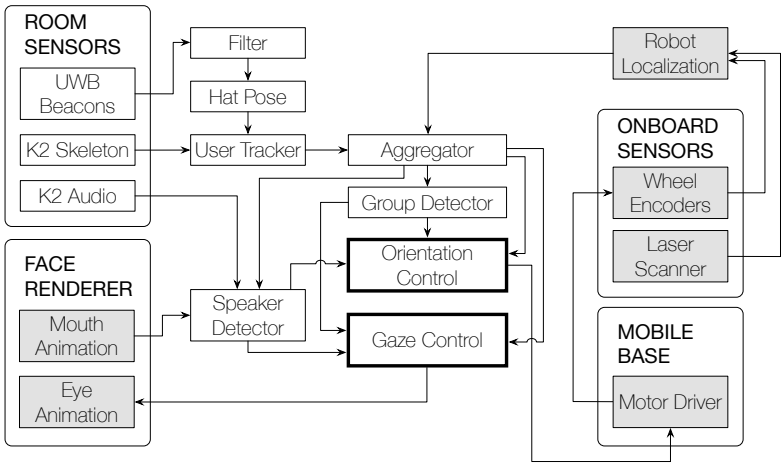


Figure 7.4: System used to control the body orientation and gaze of the robot. "UWB" stands for ultra wide-band and "K2" stands for Kinect v2.

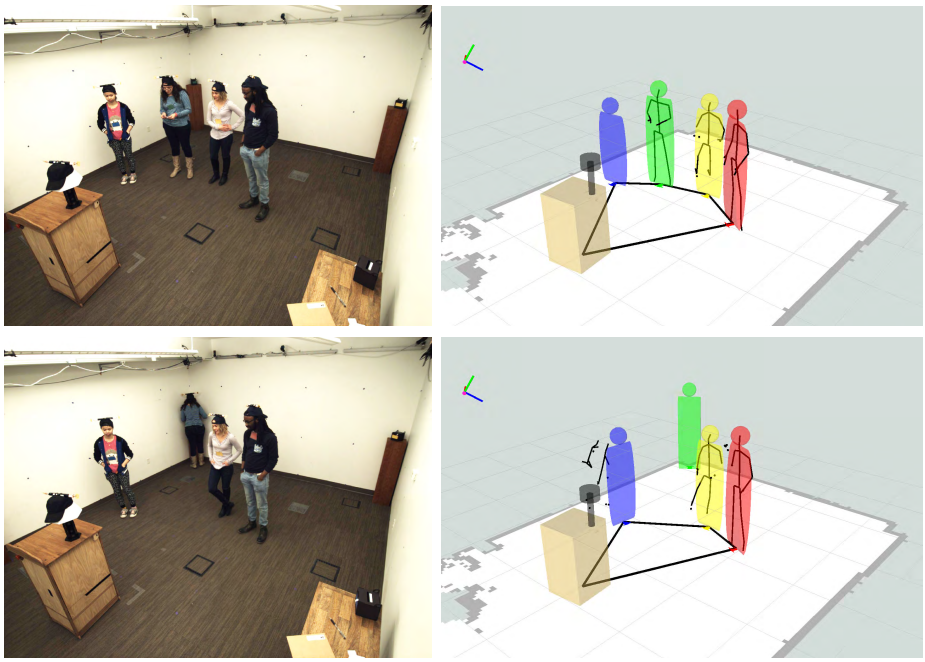


Figure 7.5: Experiment (left) and outputs of our perception system (right). A Kinect in the left corner of the room output skeleton data (shown in black near the participants). Colored markers denote participants' pose as output by the "User Tracker" module. The black lines on the ground connect the estimated members of the robot's conversational group.

Finally, the locations of the participants, the conversational groups, and the identity of the speaker were sent to the "Orientation Control" and "Gaze Control" modules. These programs output motion and gaze commands for the robot.

7.2.3.2 Limitations

Our system is a practical contribution of this work because it can enable human-robot interactions with little human intervention. However, it does not solve all perception problems in HRI, e.g., because it requires instrumentation and this may be impossible or undesirable in some cases.

Two types of errors due to shortcomings of the underlying technologies were the main factors that influenced the system's performance. First, tracking errors were common at the edges of the room due to the Kinect's limited range and field of view as well as a noticeable bias that affected UWB localization in these regions. As discussed in Sec. 7.3.6, these errors rarely affected group detection and the robot's orientation during the experiment because the participants were usually in the middle of the space. Second, sound localization errors were typically caused by simultaneous speech. These events were also infrequent in our experiment as interactants respected turn-taking.

7.3 Method

7.3.1 Study Design and Setup

We designed a 2×2 between-subjects experiment to test orientation (middle vs. attentive) and gaze (random vs. attentive) behaviors. The experiment followed a *Wizard with Oz* arrange-

ment [169] in which the manipulated behaviors were autonomous, but the sequencing of events within the study and the robot's speech were managed by a hidden operator or "wizard". In a few instances, the wizard also re-configured the robot spatially with respect to the participants, as detailed in Sec. 7.3.6. The experiment was approved by our Institutional Review Board.

During the experiment, the robot led and participated in a brainstorming session with a small group of participants. Each session was performed under one of four conditions:

MO+RG condition. The robot oriented towards the middle of its conversational group and randomized its gaze.

AO+RG condition. The robot biased its orientation towards the focus of attention and randomized its gaze.

MO+AG condition. The robot oriented towards the middle and tried to establish mutual gaze with the person who was the focus of attention.

AO+AG condition. The robot biased its orientation and gaze towards the focus of attention.

Given these conditions, we hypothesized that:

H1. The gaze behaviors would affect the perception of the robot's motion, with AG increasing perceived naturalness.

H2. For the AO behavior, participants would find the robot more attentive and responsive than MO.

H3. The AO behavior would make the participants feel like the robot was more of a part of their group than MO.

H4. The AO+AG condition would lead to reduced feelings of ostracism or increased feelings of inclusion compared to MO+RG.

The experiment was conducted in a room with a free space of 4.4×4.4 meters (Fig. 7.2). A table was placed adjacent to a wall for the participants to write down the brainstormed ideas on slips of paper. These slips then had to be deposited in different boxes in the room, according to the author.

The room was equipped with a UWB sensor network, a Kinect v2 and four RGB cameras near the ceiling. The UWB sensors and the Kinect were used to localize the participants, identify them, and detect speakers, as described in Sec. 7.2.3. The cameras recorded the interaction from multiple views and allowed the wizard to monitor the experiment remotely.

7.3.2 Participants

We recruited 20 groups (5 per condition) of 3 or 4 people using a participant pool, word of mouth, and fliers. The participants were at least 18 years of age, fluent in English, and had grown up in the U.S. The last restriction was imposed to reduce the effects of cultural biases in spatial behavior.

Table 7.1 shows details of the 69 participants that interacted with our robot. In general, most participants were university students, and their average age was 24.8 years old (SE = 1.0). In 7 sessions, two or more participants knew each other.

Before the interaction, the participants indicated how often they used a computer and their familiarity with robots on a 7 point Likert responding format (1 being lowest). Most participants used computers daily (M = 6.97, SE = 0.02) but were not very familiar with robots (M = 3.38, SE = 0.20).

Table 7.1: Participant characteristics per condition. "G", "F", "M", and "P" are used to abbreviate groups, female, male, and participants, respectively.

Condition	#G	#F	#M	#P	Age (Std Err)
MO+RG	5	8	10	18	22.2 (0.8)
MO+AG	5	9	9	18	23.5 (0.8)
AO+RG	5	11	5	16	24.4 (1.3)
AO+AG	5	9	8	17	29.3 (3.6)

7.3.3 Procedure

First, an experimenter gave a colored badge to each participant for identification purposes and administered a demographics survey. She then asked the participants to wear instrumented baseball caps with UWB beacons, and explained that each of them had a box in the room with their same color identifier. The experimenter introduced the robot, gave it an instrumented cap to make it look like the participants, and stepped away. The robot opened its eyes, and started a semi-scripted conversation with three phases:

1. *Introduction.* Chester presented itself to the group. The robot explained that the laboratory wanted to retire him, but people might keep him around if they found him useful. Chester encouraged the participants to think of how it could help in the lab and explained its sensors and capabilities. To facilitate brainstorming, the robot provided a first example and explained how it delivered souvenirs to lab visitors in the past. Chester then opened the floor to new ideas.

2. *Brainstorming.* The robot encouraged the group to brainstorm tasks that it could do in the lab for 6 min. Chester replied favorably to useful ideas and requested that authors write them on a slip of paper and deposit the slip in their corresponding box. The robot also asked for more details or discouraged unrealistic and complicated tasks. When people ran out of ideas, Chester provided more suggestions.

3. *Closing.* Chester asked a participant to count the ideas in the boxes and write the color of the box on each slip to help keep track of them. Meanwhile, the robot asked other people about their favorite ideas and gave his opinion. Chester thanked everybody for helping and said good-bye.

Finally, the experimenter adminstered a post-test survey, paid the participants, and debriefed them about the wizard. During debriefing, the experimenter also explained that the requests to deposit paper slips on boxes were an excuse to induce people to leave the robot's conversation and re-enter in natural ways. These requests were motivated by our prior experience, where we found that we had little chance of observing varied spatial behaviors without a task like this one.

7.3.4 Dependent Measures

We considered subjective and objective measures. The post-test survey asked people about their impressions of:

– the robot's motion and gaze;
– closeness to the robot using the IOS scale [15];
– the robot's and the participants' feelings of belongingness and ostracism in the brainstorming group [212];

– Chester with respect to a set of attributes, e.g., perceived intelligence, responsiveness, and entertainment value;
– Chester's ability to lead the brainstorming session and whether it should be decommissioned; and
– any unusual behavior for the robot [161].

Objective measures included the distance that the participants kept from the robot, the participants' membership in the robot's conversational group, and the number of paper slips collected during the brainstorming activity.

7.3.5 Pilot Sessions

Before starting the experiment, we recruited 35 people to conduct two types of pilot sessions. First, we ran 3 human-only sessions to evaluate the dynamics of the brainstorming activity and collect example tasks for the robot. Second, we ran 8 human-robot pilot sessions to test the Chester's dialog and the manipulated behaviors. During these sessions, we also simplified the wizard's teleoperation interface and the protocol of the experiment to avoid confusing procedures.

We considered studying a random orientation behavior for our robot as a baseline. However, the pilot sessions quickly showed that people are highly sensitive to inappropriate or unexpected orientations. These motions often halted interactions because people did not know how to interpret them.

7.3.6 Confirmation of Autonomy and Behaviors

The robot moved autonomously for most of the interaction as defined by the experimental condition. The exceptions were (1) when the robot started conversing, (2) when it said good-bye, and (3) during a handful of situations due to technical difficulties. In the first case, the wizard reconfigured the robot to show that it could move and tacitly induce an F-formation. In the second, the wizard moved Chester away to end the interaction. In the third, the wizard corrected for slight undesired changes in the robot's orientation, e.g., because of people-tracking failures in our perception system. During the brainstorming phase – the main part of the experiment – sporadic reconfigurations of this sort happened in 16 sessions out of 20. In these sessions, total teleoperation time while brainstorming was 9.37 sec on average (SE = 2.04), which represented only 2.4% of the duration of this phase (M = 383.53 sec, SE = 5.43, N = 16). REstricted or REsidual Maximum Likelihood (REML) analyses [138, 172] on the number of teleoperation events and teleoperation time while brainstorming showed no significant differences for the effects of Orientation (Attentive, Middle) and Gaze (Attentive, Random).

To confirm that the robot oriented as expected during the brainstorming phase, two coders annotated the members of the robot's conversation.[2] Using this ground truth and the logs from our perception system, we then computed the *ideal* middle orientation of the robot at 1 Hz. As expected, the absolute angular difference between the robot's orientation and this ideal middle direction was smaller for MO (M = 7.04°, SE = 0.12, N = 3686) than for AO (M = 14.33°, SE = 0.27, N = 3644). Note that these differences were induced in part by the robot's motion planner and a small bias in the robot's orientation towards the table in the room. The planner prevented Chester from jittering by ignoring turns of 5° or less. The bias (1.63° on average) was generated by occasional user tracking errors that made Chester believe that some people were

[2]Inter-coder reliability was computed for 4 sessions (20%). Two annotations were misaligned; Cohen's kappa for the other 75 annotations was 1.0, indicating perfect reliability.

still conversing with it when they left to write paper slips. Interestingly, the motion induced by these errors was interpreted as though Chester was checking that participants were following its instructions.

We transcribed when people spoke in the robot's group and its specific addressees in 2 sessions per condition. We then used the data to check when Chester adjusted its orientation towards these people. As expected, the robot turned more towards these foci of attention with AO (47% of 280 annotated events) than with MO (25% of 319). The robot did not move in many cases because the target was within 5° of its orientation (22% of the events for AO; 21% for MO).

We also inspected Chester's eye fixations during the experiment to confirm that the gaze behaviors worked as expected. As can be seen in Fig. 7.6, the positions of the pupils were less concentrated for RG than for AG because the robot tried to establish mutual gaze with the focus of attention in the latter case. Also, the robot had a tendency to look forward because several of our pre-defined eye animations positioned the pupils towards the middle of the eyes.

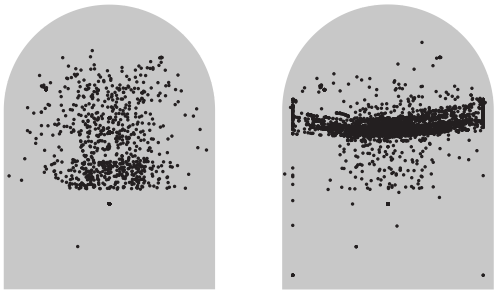


Figure 7.6: Chester's eye fixations in the 20 sessions of the experiment.

7.4 Results

We first analyse survey results and the spatial behavior of the participants around our robot. Then, we discuss the implications of our findings in terms of our hypotheses.

7.4.1 Survey Results

We ran REML analyses to evaluate survey responses. Unless noted, analyses used Participant as a random effect nested by Session, and Orientation (Attentive, Middle), Gaze (Attentive, Random), and Gender as main effects. Student's t-tests and Tukey HSD tests (with significance thresholds of $p < 0.05$) were used for post-hoc analyses of two sample and multiple pair-wise comparisons, respectively. Ratings were on 7-point Likert responding formats and responses were grouped only when Cronbach's alpha was above 0.7.

Robot's gaze. In general, Chester's gaze looked natural to the participants (M = 4.93, SE = 0.16). They did not feel like Chester was staring at them (M = 2.91, SE = 0.17) nor avoiding looking at them (M = 2.01, SE = 0.12). These results had no significant main effects.

The robot's orientation led to significant differences on how much the participants felt that Chester looked at them ($F[1, 68] = 7.47$, $p < 0.01$). As shown in Fig. 7.7, the AO behavior (M = 4.72, SE = 0.18) had significantly higher ratings than the MO behavior (M = 4.11, SE = 0.15) in this respect. The fact that the results were not significantly different for Gaze may be explained by the Mona Lisa gaze effect and the tendency of the robot to look forward.

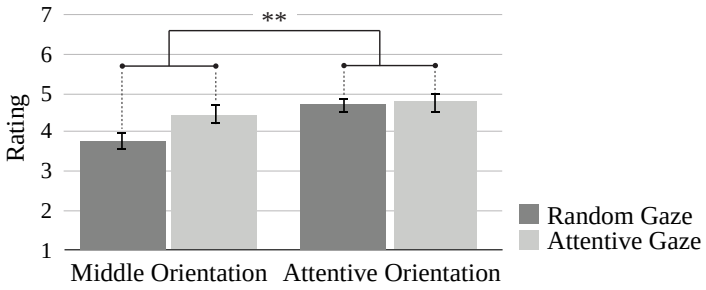


Figure 7.7: Ratings for how much the participants felt that Chester looked at them during the experiment. (**) denotes $p < 0.01$.

Robot's motion. Gaze had a significant effect on users' ratings for "Chester's motion looked natural during the interaction" ($F[1, 68] = 4.08$, $p = 0.05$). As shown in Figure 7.8, the Attentive Gaze behavior elicited significantly higher agreement with the statement relative to the Random Gaze behavior ($M = 4.71$, $SE = 0.28$ vs. $M = 4.00$, $SE = 0.23$). No significant differences were found for "Chester's motion was distracting" ($M = 2.03$, $SE = 0.13$), "I felt confident that the robot was not going to hit me" ($M = 6.42$, $SE = 0.18$), nor "Chester's motion made me anxious" ($M = 1.57$, $SE = 0.12$).

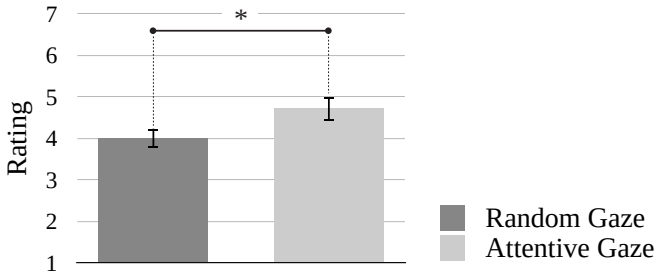


Figure 7.8: Ratings for how natural the robot's motion looked like during the experiment based on its gaze behavior. (*) denotes $p < 0.05$.

Robot's attentiveness. Participants rated how much they thought that Chester paid attention to what they said ($M = 5.33$, $SE = 0.16$) and to what the other participants said ($M = 5.48$, $SE = 0.14$). A REML analysis with Participant as random effect nested by Session, and Orientation, Gaze, Gender, and Speaker (Me, Others) as main effects showed significant differences for the interaction between Orientation and Gender only ($F[3, 66] = 4.94$, $p = 0.03$). The post-hoc test then showed no significant pair-wise differences, but the tendency was interesting: male participants thought that the robot paid more attention with AO than with MO ($M = 5.73$, $SE = 0.20$ vs. $M = 4.92$, $SE = 0.25$).

Inclusion and ostracism: IOS scale [15] ratings indicated that the participants did not feel close to Chester ($M = 2.57$, $SE = 0.14$). However, they thought that both they ($M = 5.03$, $SE = 0.17$) and the robot ($M = 5.33$, $SE = 0.19$) belonged to the brainstorming group.

We found low perceptions of being ignored or excluded by the robot ($M = 1.57$, $SE = 0.12$) or the other participants ($M = 1.42$, $SE = 0.09$). REML analyses for these results resulted in no significant differences, but Orientation was close for the former ($p = 0.06$). The trend suggested that MO could lead to higher feelings of ostracism from the robot than AO ($M = 1.77$, $SE = 0.17$ vs. $M = 1.33$, $SE = 0.14$).

Table 7.2: Ratings for the factors resulting from factor analysis. Machine-like was reversed (R) for the analysis and for computing Chronbach's alpha.

Attribute	Mean (SE)	Cronbach's α	Factor
Responsive Interactive	5.33 (0.11)	0.786	I
Useful Knowledgeable Intelligent Competent	4.62 (0.13)	0.791	II
Entertaining Funny	5.47 (0.15)	0.846	III
Lifelike Machine-like (R)	4.14 (0.18) 4.20 (0.15)	0.623	-

Other perceptions of the robot. The participants generally thought that Chester was a good leader for the brainstorming activity (M = 5.00, SE = 0.17) and had significantly different impressions of how much the robot and the other participants liked them ($F[1, 68] = 4.98$, $p = 0.03$). In particular, the participants thought that the robot liked them significantly more than did the other people in the experiment (M = 5.16, SE = 0.14 vs. M = 4.93, SE = 0.13).

Chester was not perceived as anti-social (M = 1.58, SE = 0.09). The only trend in this respect ($p = 0.06$) suggested that RG could make the robot look more anti-social than AG (M = 1.76, SE = 0.16 vs. M = 1.4, SE = 0.09).

Table 7.2 shows a factor analysis on a series of additional attributes for the robot. Factor I was associated with interactivity, Factor II with competence, and Factor III with entertainment. These factors explained 18.3%, 26.4%, and 20.3% of the variance, respectively. Their ratings were positive in general with no significant main effects of condition.

Only 8 participants of 69 indicated that Chester should be decommissioned in the post-survey. Their responses were typically associated with the robot's usefulness (e.g., *"I can't see a practical use for it, but the robot was entertaining"*).

Interaction: In general, the interaction with Chester was enjoyable (M = 5.45, SE = 0.14). Desire to brainstorm for longer was correlated with the number of paper slips written per session ($r(67) = 0.48$, $p< 0.01$), which tended to be just a few, or ten or more. This result motivated a REML analysis on the ratings for wanting to brainstorm for longer with Slip Count (1 if ten or more slips, 0 otherwise), Orientation, Gaze, and Gender as main effects, and Participant as random effect within Session. Not surprisingly, Slip Count had a significant effect ($F[1, 68] = 15.09$, $p< 0.01$). Ratings in sessions with many slips were significantly higher than the rest (M = 4.65, SE = 0.28 vs. M = 3.00, SE = 0.24). Also, the interaction between Gender and Slip Count was significant ($F[3, 66] = 6.78$, $p= 0.01$). Male participants wanted to brainstorm significantly more when there were at least ten slips (M = 5.27, SE = 0.37) than in other cases (M = 2.53, SE = 0.30). Female ratings were more uniform and neutral.

Note that the robot did not use a balancing criteria to ask people for ideas during the brainstorming activity. A REML analysis on the number of ideas proposed by the participants did not result in any significant differences for the main effects of Orientation, Gaze, and Gender, suggesting that this aspect of the protocol did not generate a confound. Moreover, all but one of the 69 participants proposed ideas. The only person that stayed quiet during the brainstorming phase took part in the activity towards the end, when the robot asked him to count the number of slips in the boxes.

7.4.2 Human Spatial Behavior

We analyzed proxemics during the brainstorming phase, when the participants often moved to write their ideas. For the analyses, we used the spatial information output by our perception system (sampled at 1Hz) and the group membership annotations described in Section 7.3.6.

When the participants conversed with the robot in the brainstorming phase, their average separation from Chester was typical of social encounters (M = 2.15m, SE = 0.04, N = 69) [60]. Because people often adjusted their position as they became familiar with the robot and the activity, we decided to further analyze proxemics during the last minute of the brainstorming part of the experiment. We performed a REML analysis on the distance between the robot and the participants during this period, considering Orientation, Gaze, and Gender as main effects and Participant as random effect nested by Session. Gaze was significant ($F[1, 68] = 5.67$, $p = 0.02$): participants stood significantly farther away from the robot with RG (M = 2.29, SE=0.05) than with AG (M = 2.09, SE = 0.06). The interaction between Gaze and Orientation was also significant ($F[3, 66] = 4.27$, $p = 0.04$). The members of the robot's group were significantly farther away from it with MO+RG than with MO+AG (M = 2.40, SE=0.07 vs. M = 2.03, SE=0.10), as shown in Fig. 7.9.

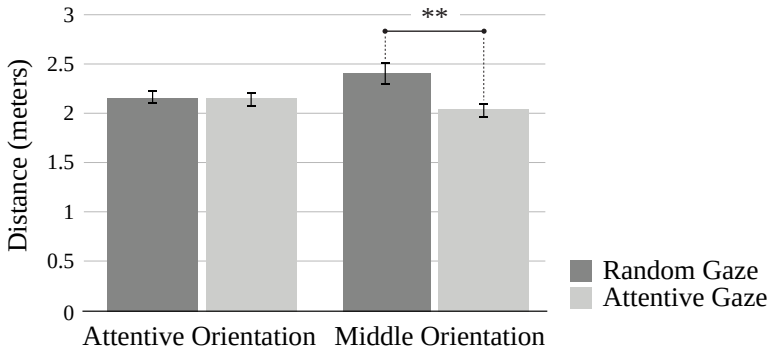


Figure 7.9: Distance to Chester during the last minute of brainstorming. (**) denotes $p < 0.01$.

Throughout the experiment, we observed qualitatively circular or side by side spatial arrangements. In a few cases in which the robot engaged in a dyadic interaction and it was not oriented as expected, participants proactively changed their positions to stand in front of it. These efforts to establish appropriate spatial arrangements suggest that people may be willing to collaborate with robots to establish F-formations and adapt to unforeseen circumstances.

7.4.3 Hypotheses Support and Implications

Attentive Gaze made the participants think that Chester's motion looked more natural in comparison to Random Gaze. This result supported our first hypothesis (H1) and is related to prior findings on the influence of gaze on the perception of a robot's head motion [107]. Furthermore, the orientation behaviors also affected the perception of the robot's gaze. With Attentive Orientation, the participants perceived that the robot looked at them more. These outcomes suggest that robot gaze and body motion should be designed and controlled jointly, rather than independently of each other.

We expected the Attentive Orientation behavior to make the robot seem more attentive and responsive than the Middle Orientation behavior (H2). While we did not find that the robot's orientation altered how responsive it looked, participants thought that the robot gazed at them more with AO than with MO, as mentioned before. There was also a trend that suggested that male participants thought that the robot paid more attention to what people said with AO.

Opinions on how close the participants felt to the robot and whether they perceived it as part of their group were not significantly affected by the orientation behaviors, as hypothesized in H3. Interestingly, the distance between the participants and the robot varied significantly with MO based on the robot's gaze, but did not vary as much with AO. This finding might indicate more subtle effects of the manipulation than can be gleaned from questionnaires. Also, the lack of support opens up possibilities for developing more complex orientation behaviors and fulfilling other non-social tasks during multi-party interactions. Given that both MO and AO were acceptable and did not affect the perception that the robot was part of the group, both behaviors could be used by robots depending on other factors besides the interaction. For example, robots could switch between MO and AO to reduce uncertainty about the environment.

In terms of H4, the AO+AG condition did not reduce feelings of ostracism or increase feelings of inclusion relative to MO+RG. Nonetheless, a trend suggested that MO could lead to higher feelings of ostracism than AO. This finding should be explored further in future research.

Finally, we learned from the pilots that people are sensitive to inappropriate or unexpected robot orientations. If users do not understand why a robot moves, interactions can easily be disrupted. This outcome is related to prior work on communicative robot motion [44, 152, 178].

7.5 Discussion

The experiment described in this chapter was limited in several ways First, Chester's dialog was scripted and, thus, it could not respond appropriately in all circumstances. Second, the physical appearance and capabilities of our robot could have influenced our results and biased some aspects of the design of the behaviors under consideration. For example, the differential drive base of the robot constrained the complexity of its spatial behavior and, in turn, this could have affected the perception of its motion. Third, the perception system that we implemented for the experiment required instrumentation. While this system enabled autonomous robot behaviors, we are now interested in shifting towards on-board computation. This includes improving robots' capabilities so that they can reason about social contexts using on-board sensors only and, therefore, interact more casually.

The brainstorming protocol could be used to study other aspects of HRI. Overall, the perception of the brainstorming activity used in the experiment was positive. The protocol successfully created opportunities for changes in conversation group size, which allowed us to study the behaviors under consideration in different social contexts. In the future, this protocol could be used to study turn-taking patterns and collaboration in HRI. Similar to social games [198], brainstorming activities are customizable (e.g., the topic of the conversation can be easily adapted) and can be conducted with groups of strangers. In contrast, brainstorming sessions are less adversarial and do not require teaching very specific instructions.

Robot gaze and orientation can affect users' perception of these behaviors. The gaze of the robot affected the participants' perception of its motion and its motion affected the perception of its gaze. This dependency implies that robots should reason about and control their gaze and body motion jointly. Furthermore, some trends implied that the Attentive Orientation could be preferred over the Middle Orientation (e.g., AO could make the robot look more attentive and less anti-social). However, these behaviors led to similar feelings of inclusion and belonging to the group, suggesting that both AO and MO could be used as primitives for more complex orientation behaviors.

Overall, our experiment showed that reasoning about spatial behavior is not only important for robots to understand who is part of their conversation, but is also essential for them to co-operate to sustain the spatial arrangements that are typical of these interactions.

Chapter 8

Learning to Control Robot Orientation During Conversations

As we saw in Chapter 7, it is important for robots to co-operate to sustain F-formations. We achieved this cooperation in the previous experiment by relying (1) on external sensing mechanism, and (2) on hand-crafted rules for the robot to orient one way or another. While this approach allowed us to systematically investigate the effects of robot orientation and gaze in the study, it can hardly scale to real-world interactions. The same problem affects tele-operation approaches to control robot motion [80, 203] or other rule-based methods [219].

Inspired by the prior success of Reinforcement Learning (RL) in a wide array of robotics tasks [92], this chapter explores RL techniques to find good policies to control the orientation of a robot during simulated group conversations. In this effort, we suppose to be the case that the robot can identify its conversational group (e.g., using the framework introduced in Chapter 6), and focus our efforts on the problem of learning good orientation policies. In general, we assume that the *correct* behavior for the robot is to turn towards the speaker of its conversation. As we discussed in the previous Chapter, this type of behavior can help convey attentiveness to this person and maintain awareness of the focus of attention of the interaction. Different to our prior experiment, though, the robot is not in an instrumented environment in this case. It must learn to orient towards the speaker of the conversation using its onboard sensors only.

We approach the problem of controlling the orientation of a robot during group conversations using the Oz of Wizard methodology [169]. Our efforts are focused on evaluating RL approaches in simulated group conversations (Fig. 8.1) as a precursor to future work with real users. The simulation offers the opportunity to systematically study the effect of sensing noise on the performance of the robot as well as the generalization of learned policies to other group interactions. This is a first step towards reducing the amount of engineering required to generate appropriate spatial behavior for robots during situated conversations with users.

Our main contribution in this work is a robot-centric state representation for the RL task that is agnostic to the number of people in the conversation. This means that the same representation can be used to control the orientation of a robot while it interacts with 2, 3, 4 or more people. Moreover, the policies learned with this state representation can potentially generalize across these different scenarios. This work was first published in RO-MAN'16 [194].

The rest of this chapter is organized as follows. The next section describes our general approach to control the orientation of a robot with Reinforcement Learning. Section 8.2 introduces our simulated environment and details the interaction dynamics and sensing mechanisms that we modeled for this work. Section 8.3 and 8.4 then describe our empirical evaluation and results. Finally, Section 8.5 discusses our results and future research directions.

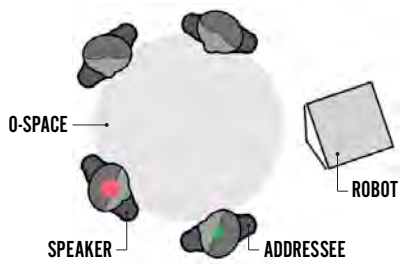


Figure 8.1: Simulated group conversation between a robot and four people. The red and green circles on top of the agents identify the speaker and addressees, respectively. The big gray circle represents the o-space of the group's F-formation.

8.1 General Approach

We model our motion control problem as a sequential decision-making process. At any time-step t, the robot (or agent) receives some representation of the environment *state* $\mathbf{s}_t$ and executes an *action* a_t. Executing this action triggers a transition to a following state, represented by $\mathbf{s}_{t+1}$, and results in an immediate *reward* r_{t+1}. The goal of the robot is to choose actions that maximize the discounted total reward that it receives while it interacts with the world. That is, maximize $\sum_{t=0}^{\infty} \lambda^t r_{t+1}$ with $\lambda \in [0, 1]$ a discount rate.

Out of the many RL techniques that exist, we focus on evaluating popular online approaches that estimate an *action-value function* $Q(\mathbf{s}, a)$ to try to find solutions to the motion control problem. The function

$$Q^{\pi}(\mathbf{s}, a) = E_{\pi}\Big[\sum_{k=0}^{\infty} \gamma^k r_{t+1+k} | \mathbf{s}_t = \mathbf{s}, a_t = a\Big]$$

is an estimate of how good it is to choose action a in a given state $\mathbf{s}$ and then follow policy π. Readers interested in more details about this function are encouraged to refer to [174].

Online approaches are advantageous for our task because they improve as the robot interacts with people and they can adapt quickly to specific interaction dynamics. The latter is particularly advantageous when the social context changes, e.g., when some of the members of the conversation leave or others join the interaction.

8.2 Group Simulation

We developed a simulated environment to test control policies for a robot. The simulation was inspired by models from social psychology that explain human spatial behavior during free-standing group conversations [58, 86]. These models were previously introduced in Chapter 2.

Our simulation modeled a free-standing group conversation with an established F-formation, as shown in Figure 8.1. To start, the simulation placed the interactants in a circular arrangement and randomly chose the first speaker and his or her addressee(s). The addressee(s) were either one person or the whole the group. Then, the simulation repeatedly executed its main loop, as detailed in Alg. 8.2.1. The set of actions that the agent could execute at the beginning of this loop were angular velocity commands that changed the orientation of the robot with respect to the group. The state representation provided to the agent at the end of the loop is later described in Sec. 8.3.

The simulated speakers turned their heads towards their addressee or towards the center of the o-space if they spoke to the whole group. The other people in the conversation turned their

Algorithm 8.2.1: Main loop of our simulation

```
while the simulation is running do
    Update the simulated clock
    Receive the last action that was taken by the RL agent
    Update the state of the robot with the last action taken
    if the previous speaker finished talking then
        Choose a new speaker and addressee
        Set desired head orientation for the speaker
        Set desired head orientation for the listeners
    Update the states of all the people in the conversation
    Compute the reward r for the robot
    Update the robot's internal representation s of the state of the conversation
    Provide r and s to the RL agent so that it can choose a new action
```

heads towards the speaker, as described in the next section. In general, heads rotated at a fixed angular velocity towards their respective target, typically during multiple simulation steps.

For this work, we never allowed the robot to take a turn to speak. As a result, it had to adapt to the flow of the conversation set by the rest of the group.

8.2.1 Main Simulation Parameters

The main parameters that controlled the spatial arrangement of the group and the dynamics of the interaction were:

Number of interactants: The number of people in the conversation, including the robot.

O-space center: The location of the center of the o-space in the world-coordinate frame of the simulation.

Stride: The expected distance between the center of the o-space and the interactants.

Time step: The time elapsed between simulation updates.

Robot actions: List of angular velocity commands that could be executed by the robot. In particular, we used the set $[-15.0, -7.5, 0.0, 7.5, 15.0]$ (in deg/sec) for this work.

Speaking time distribution: Normal distribution that modeled how long a person typically spoke for. In general, we used $\mathcal{N}(5.0, 2.0)$, but prevented sampled values from being smaller than 0.5 secs to avoid very short speaking times.[1]

Look-at noise: When a person i in the simulation looked at the speaker, his or her head was set towards the angle β_i:

$$\beta_i = \arctan(d_y/d_x) + \varepsilon_i \text{ with } \varepsilon_i \sim \mathcal{N}(0, \sigma_i^2) \tag{8.1}$$

where $\mathbf{d} = [d_x \; d_y]^T$ denotes the direction towards the speaker from person i. The noise term ε_i in eq. (8.1) controlled how accurately the person looked at the speaker, depending on the standard deviation σ_i. Note that when σ_i made the head turn more than 90 degrees from the front of the person, we clamped β_i to prevent the head from turning backwards.

[1] We acknowledge that this model is a crude approximation of real group conversations because people often speak for significantly longer than 5 secs. We opted for short speaking times, though, because longer speeches simplify the control problem by reducing the number of speakers in any given interaction.

8.2.2 Robot Perception

We modeled the robot as a platform with a small number of degrees of freedom, similar to Chester, the CoBots [200] or Frog [47]. The robot had a camera and a microphone array fixed to the front of its body. The camera could be used to detect the position of people, as well as their head and body orientations. The microphone array provided the angular directions toward nearby speakers from the robot's perspective.

The sensors had configurable fields of view. People within these fields of view could be sensed with some probability; those outside were not detected at all. The specific values that we used for these parameters are provided in Sec. 8.4.3.

8.2.3 Reward

In this work, we assumed that the correct behavior for the robot was to turn towards the speaker. Consequently, the reward r_{t+1} that the environment provided to the robot for taking an action a_t was:

$$r_{t+1} = \exp(-\varphi^2) + b \tag{8.2}$$
$$\text{with } b = \begin{cases} 1 & \text{if abs}(\varphi) <= \tau \text{ and } a_t == 0.0 \\ 0 & \text{otherwise} \end{cases}$$

The angle $\varphi \in [-\pi, \pi]$ was the difference between the orientation of the robot and the angle representing the direction towards the speaker from the robot's position. The bonus b was given to reward zero angular velocity commands when the difference φ was small. In general, we used 10 degrees for τ to prevent oscillatory motions.

8.2.4 Limitations

Even though the simulation was useful to explore reinforcement learning techniques for motion control, it is by no means a perfect model of the real world. Our simulation did not capture all the complexity and variability of human behavior during group conversations nor sensing noise. Nonetheless, our efforts are an important first step towards testing RL techniques for the problem under consideration. The policies learned from our simulation can be considered as prior knowledge for learning better behaviors with real users. Furthermore, the simulation allowed us to explore the sensitivity of several methods to particular types of noise, something that is hard to accomplish during human-robot interaction experiments [169].

8.3 State Representation

Our key contribution in this work is a state representation that is well suited to solve our motion control problem with RL techniques. This representation was composed of six features:

f1. Continuous feature in $[-\pi, \pi]$ representing the rotation that the robot needed to execute to direct its body towards the speaker. This value is the same as φ in eq. (8.2) if the speaker was within the field of view of the robot's microphone array and the sensor detected the audio signal coming from this person. Otherwise, *f1* was set to zero by convention.

f2. Binary feature indicating if *f1* is valid or not. This feature was zero when no audio signal was detected by the microphone array; otherwise, *f2* was one.

f3. Continuous feature in $[-\pi, \pi]$ representing the rotation that the robot needed to execute to direct its body towards the location of maximum *social saliency* induced by the visible people in its group. Social saliency encoded gaze concurrences and was estimated using the primary gaze rays of the people detected by the robot's camera, as described in [135] and illustrated in Fig. 8.2. When multiple locations were socially salient and had equal contribution from the primary gaze rays of the visible people, ties were broken randomly and only one location was used to compute *f3*. When a single person was detected by the camera and no gaze concurrence could be computed, we uniformly sampled possible social saliency locations along the primary gaze ray of this person and used their average for *f3*. If nobody was visible, then *f3* was set to zero by convention.

f4. Binary feature indicating if *f3* is valid. This feature was zero when social saliency could not be computed because nobody was detected through the robot's camera. Otherwise, *f4* was set to one.

f5. Continuous feature in $[-\pi, \pi]$ representing the rotation that the robot needed to execute to orient its body towards the center of the o-space of its conversational group. We computed an estimate $\mathbf{c}$ of the true o-space location using an exponential moving average of center proposals:

$$\mathbf{c} = (1.0 - n * a)\mathbf{c} + a \sum_{i=1}^{n} \underbrace{\mathbf{p}_i + s * \mathbf{u}_i}_{\text{center proposal}} \tag{8.3}$$

where n was the total number of people visible through the robot's camera, a was a small number that controlled the contribution of every proposal under the constraint $n * a \in [0, 1]$, s was an expected value for the stride of the o-space, $\mathbf{p}_i$ was the position of the i-th person that was visible, and $\mathbf{u}_i$ was a unitary vector pointed in the same direction as the front of the body of person i. The model used to generate o-space center proposals ($\mathbf{p} + s * \mathbf{u}$) was inspired by prior work [40, 155, 156, 199] and was used to initialize $\mathbf{c}$ with the proposal corresponding to the first person detected by the robot when the simulation started. Before any person was detected, *f5* was zero by convention.

f6. Binary feature indicating if *f5* is valid. If no o-space center had been estimated, then *f6* was zero. Otherwise, *f6* was one.

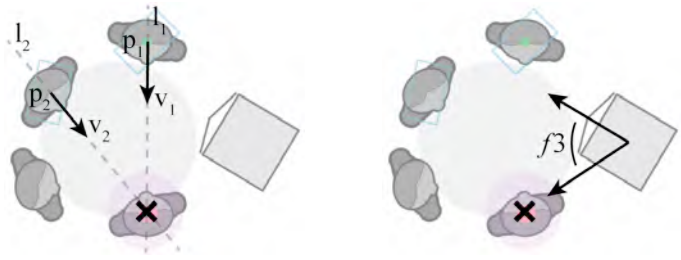


Figure 8.2: Example of the primary gaze rays ($\mathbf{l_1}$ and $\mathbf{l_2}$) used to compute the point of maximum social saliency (a) and the resulting feature *f3* of our state representation (b). The point of maximum social saliency is marked with a cross ($\times$) and surrounded by a light-colored circle. The primary gaze rays were estimated only for persons 1 and 2, who were visible through the robot's cameras, using their position ($\mathbf{p}$) and head direction ($\mathbf{v}$).

8.3.1 Properties

Our state representation is agnostic to the size of the group. This means that the same features can be used to describe conversations with a few people or with more interactants.

Because the features are computed from the perspective of the robot, their descriptive power depends on the performance of the robot's sensors. For example, the more people are detected at any given time, the more the estimate of the o-space center converges to the true value. The more people are detected, the closer the point of maximum social saliency is to the place where most gaze directions converge. This place is typically the location of the speaker.

8.4 Evaluation

We performed a series of experiments to evaluate several RL agents in our simulated environment. Our goal was not to prove the superiority of one method, but rather to evaluate empirically what kind of approach may be better suited for our particular task and whether the proposed state representation generalized as expected. More precisely, our experiments focused on addressing:

1. whether a robot could quickly learn reasonable policies for the orientation task in our simulated environment;
2. how the performance of the RL agents under consideration degraded with noisy measurements;
3. how their performance could be affected by atypical human behavior; and
4. whether learned policies could generalize to conversations with more or fewer interactants.

To address the first goal above, we studied the amount of reward that several agents received as a function of time, as well as how quickly their performance saturated. This test included agents that estimated action values for discrete versions of the state space or that approximated them using linear regression. For the second and third goals, we studied how measurement noise and variability in human motion affected the performance of the agents that, on average, learned good policies faster. For the last goal, we tested policies that were learned from interacting with 4 people in other group conversations.

8.4.1 Agents

We considered several agents for our evaluation. First, we decided to test model-based RL methods because they tend to be more sample efficient than model-free approaches when good transition and reward models can be learned quickly. These methods included TEXPLORE [67], which was designed for the robotics domain, and DYNA-2 [164], which can leverage prior experience while learning. Because the latter architecture uses Sarsa [174] to estimate the action-value function Q, then we also decided to test Sarsa by itself as a model-free method. Brief descriptions of the specific versions of the agents that we considered in our evaluation are provided below for completeness.

Sarsa(λ): Baseline on-policy learning agent [174]. This implementation discretizes the continuous features of the state space (Sec. 8.3) and estimates the action-value function Q using a tabular representation. With any new tuple $(\mathbf{s}_t, a_t, r_t, \mathbf{s}_{t+1}, a_{t+1})$,

$$\begin{aligned} Q_{t+1}(\mathbf{s}, a) &= Q_t(\mathbf{s}, a) + \alpha \delta_t e_t \text{ for all } (\mathbf{s}, a) \\ \text{with } \delta_t &= r_{t+1} + \gamma Q_t(\mathbf{s}_{t+1}, a_{t+1}) - Q_t(\mathbf{s}_t, a_t) \end{aligned} \tag{8.4}$$

where α is the learning rate, γ is the discount factor, and e_t is the (accumulating) eligibility trace [175]:

$$e_t = \begin{cases} \gamma\lambda e_{t-1} + 1 & \text{for } (\mathbf{s}_t, a_t) \\ \gamma\lambda e_{t-1} & \text{for all other state-action pairs} \end{cases} \tag{8.5}$$

which represents the credit assigned to state-action pairs for subsequent errors in evaluation.

For action selection, this agent uses the common ϵ-greedy policy, which chooses the best actions $\arg\max_a Q(\mathbf{s}, a)$ with a probability of $1 - \epsilon$. Otherwise, it selects a random action.

Sarsa(λ) with tile coding and adaptive learning rate: Sarsa agent with linear function approximation [146]:

$$Q(\mathbf{s}, a) = \phi(\mathbf{s}, a)^T \theta \tag{8.6}$$

where ϕ is a function that transforms the state-action pair to a large binary vector with tile coding [173] and θ is a collection of weights. The weights get updated by the rule $\theta_{t+1} = \theta_t + \alpha\delta_t e_t$, with the eligibility traces being $e_t = \gamma\lambda e_{t-1} + \phi(\mathbf{s}_t, a_t)$ and α being updated automatically according to [41]. This agent also uses an ϵ-greedy policy.

DYNA-2: RL architecture that combines sample-base learning with sample-based planning [164], as described in Algorithm 8.4.1. The agent has two "memories" that encapsulate all of the features and parameters used to estimate the value function. The *permanent* memory (ϕ, θ) is updated from real-world experiences and is used to compute the best overall estimate of the action-value function $Q(\mathbf{s}, a) = \phi(\mathbf{s}, a)^T\theta$ ("learn" procedure of Alg. 8.4.1). The *transient* memory $(\bar{\phi}, \bar{\theta})$ is updated during simulations to track a local correction to the permanent memory ("search" procedure). This update is achieved with the combined action-value function $\bar{Q}(\mathbf{s}, a) = \phi(\mathbf{s}, a)^T\theta + \bar{\phi}(\mathbf{s}, a)^T\bar{\theta}$.

Algorithm 8.4.1: Main steps of DYNA-2

Procedure learn *($\mathbf{s}_t$, a_t, r_t, $\mathbf{s}'_{t+1}$)*
  Store ($\mathbf{s}_t$,a_t,r_t,$\mathbf{s}_{t+1}$) to update dynamics model
  search ($\mathbf{s}_{t+1}$) and pick next action $a_{t+1} = \pi(\mathbf{s}, \bar{Q})$
  `// update permanent memory (Sarsa)`
  $\theta_{t+1} = \theta_t + \alpha[r_t + \gamma Q(\mathbf{s}_{t+1}, a_{t+1}) - Q(\mathbf{s}_t, a_t)]e_t$
  return (a_{t+1})
Procedure search *($\mathbf{s}$)*
  for $k = 1$ **to** *num_rollouts* **do**
    Initialize $\bar{e}_0 = \mathbf{0}$ and $\mathbf{s}_0 = \mathbf{s}$
    Pick action $a_0 = \bar{\pi}(\mathbf{s}_0, \bar{Q})$
    for $t = 0$ **to** *max_steps* $- 1$ **do**
      $(\mathbf{s}_{t+1}, r_t)$ = queryDynamicsModel$(\mathbf{s}_t, a_t)$
      Pick next action $a_{t+1} = \bar{\pi}(\mathbf{s}_{t+1}, \bar{Q})$
      `// update transient memory`
      $\bar{\theta}_{t+1} = \bar{\theta}_t + \bar{\alpha}[r_t + \gamma\bar{Q}(\mathbf{s}_{t+1}, a_{t+1}) - \bar{Q}(\mathbf{s}_t, a_t)]\bar{e}_t$

The main difference between Alg. 8.4.1 and its description by Silver and colleagues [164] is that we apply DYNA-2 to a non-episodic scenario with discounted returns ($\gamma < 1$). Moreover, we use Sarsa(λ) with tile coding and an adaptive learning rate to estimate θ and $\bar{\theta}$. For the transition and reward models, we use regression forests, as described in the next paragraphs for TEXPLORE. Whenever enough samples are collected to learn a new model, we clear DYNA's transient memory so that it quickly adapts to the new dynamics.

TEXPLORE: Model-based architecture that uses sample-based planning [67]. In particular, this architecture uses a regression forest to estimate the transition and reward functions from real experience. Every time a query is made for planning, a random tree from the forest is chosen to make the prediction. This tree can be considered as one possible hypothesis of the true model of the domain. For planning, this architecture uses the UCT(λ) algorithm with discretized states-action spaces. Sample actions are selectively chosen using Upper Confidence Bounds [93].

In contrast with the original description of TEXPLORE [67], our implementation does not generalize action-values across depths in the search tree nor runs the act, plan, and model threads in parallel. The first modification was made because generalizing values resulted in poor performance in our particular domain. The second change simplified our implementation. Even though speed is important for robotics applications, such as ours, it was not a crucial factor for the present work.

8.4.2 Other Implementation Details

We used RL-Glue[2] as the interface between our simulated environment with the agents under consideration. For the Sarsa(λ) agent with tile coding, we used PyRL's implementation.[3] For the rest, we used our own implementation in Python, as described in the previous section. Moreover, we used the same model approximator for Dyna-2 and TEXPLORE. The approximator was a collection of independent regression forests for each of the dimensions of the state space and for the reward function, as proposed by Hester [67]. Our code extended the functionalities of the regression forest model of the Scikit Learn library[4] with an option to query the prediction of a randomly-chosen tree in the forest.

8.4.3 Results

Unless otherwise noted, the results presented in this section are averages over 10 runs of 18000 steps (equivalent to 1 hour of interaction time) and actions were executed every 0.2 seconds. We set the field of view and detection probability of the microphone array on the robot to 100 degrees and 0.9, respectively. For the camera, we used 80 degrees and 0.75. These values were set based on our prior experience with off-the-shelf sensors of this kind.

The simulations had a total of 5 interactants (including the robot) which were arranged in a circular formation with a stride of 1.25 meters, as illustrated in Fig. 8.1. In general, future rewards were discounted with $\gamma = 0.7$.

Whenever the state space was discretized by an agent, we used 24 bins for each of the continuous angular features. For Sarsa(λ) and the ϵ-greedy policies, we set $\lambda = 0.4$ and $\epsilon = 0.1$, respectively. For the discrete Sarsa agent, $\alpha = 0.7$; for the continuous version, we used 64 tiles to approximate the Q function. In the case of DYNA-2 and TEXPLORE, we used regression forests with 15 trees to estimate a model of the dynamics. This model was updated every 300 samples (i.e., once a minute). For sample-based planning, we used 16 rollouts with 3 look-ahead actions. These parameters provided good results in our simulated environment.

8.4.4 Learning to Orient

First, we studied how quickly the RL agents under consideration learned to orient towards the speaker. We considered the same look-at noise distribution $\mathcal{N}(0, \sigma^2)$ in eq. (8.1) for all the

[2] http://glue.rl-community.org
[3] https://github.com/amarack/python-rl
[4] http://scikit-learn.org

people in the simulation, where σ was set to $0.0, 0.3, 0.6, 0.9$, or 1.2 radians (which is equivalent to 0.0, 17.2, 34.3, 51.5, and 68.8 degrees). In the particular case where $\sigma = 0$, no noise was added to the head orientations.

In general, all the agents converged to good policies within 1000 to 2000 steps (i.e., within 3.3 to 6.6 minutes) except for the baseline version of Sarsa with a tabular Q representation. The poor performance of this version of Sarsa with respect to the other agents can also be observed in Figure 8.3. This plot shows the number of speakers towards whom the robot failed to orient with an angular velocity of 0.0 rad/sec and with $abs(\varphi)$ in eq. (8.2) less than 10 deg. These can be considered speakers the the robot failed to acknowledge properly. The fact that all the agents but the baseline version of Sarsa found good policies quickly suggests that generalizing Q estimates across states is beneficial for our task.

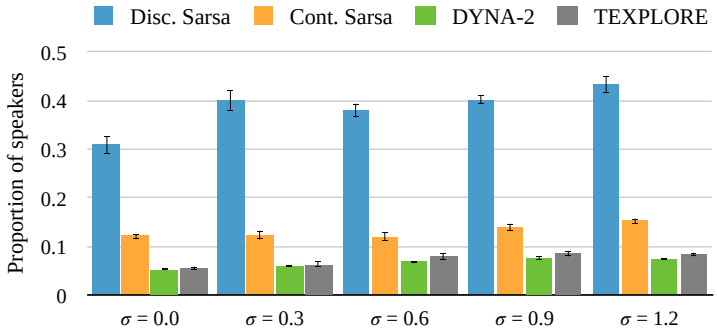


Figure 8.3: Proportion of speakers towards whom the agents failed to orient as a function of the look-at noise (lower is better). “Disc. Sarsa" is the discrete version of Sarsa(λ) with tabular Q; “Cont. Sarsa" is the continuous version with tile coding. Error bars represent standard errors. (Best viewed in color)

Figure 8.4 shows the proportion of steps in which the agents achieved good behavior and received a bonus $b = 1.0$ as part of their reward (see eq. (8.2)). In general, model-based agents performed the best in terms of orienting properly towards the speakers. This result reinforces the idea that random forests are sample efficient when it comes to estimating dynamics models [67]. Furthermore, Fig.8.4 shows that the more people look away from the speaker, the worse the agents tend to perform. This reduction in performance happens because higher σ values lead to fewer gaze concurrences, which is precisely what social saliency tries to estimate. One option to counter-act this effect is to estimate σ for every person during the conversation and

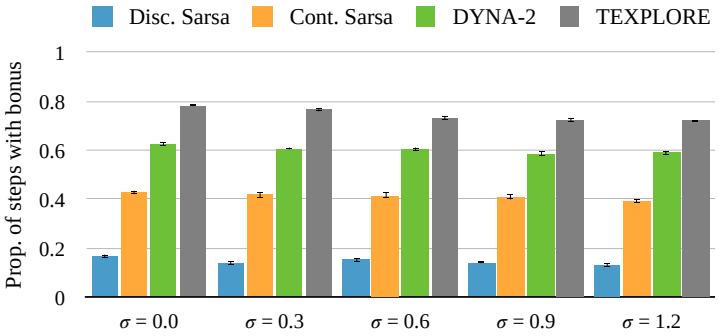


Figure 8.4: Proportion of steps (out of 18000) in which the agents received a reward with a positive bonus. Results are groups by look-at distribution. Error bars represent standard errors. (Best viewed in color)

incorporate this information into the estimate of social saliency (see Section 3.2 of [135]).

Figure 8.5 shows the cumulative reward that the agents obtained during the 18000 steps of interaction time. TEXPLORE clearly outperformed the other agents in this respect, likely because of its better policy in comparison to ϵ-greedy.

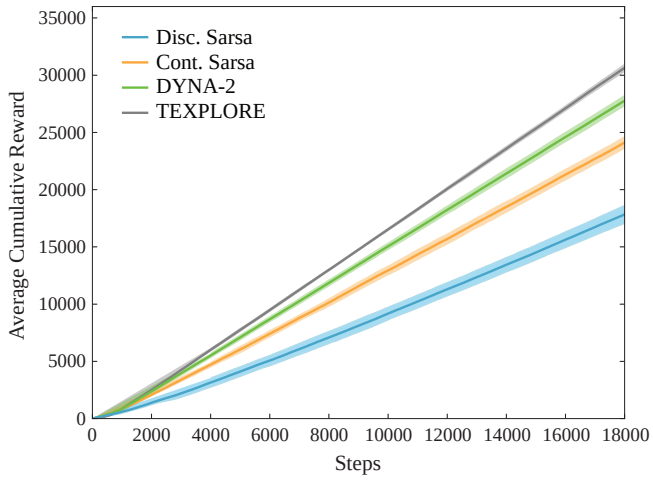


Figure 8.5: Average cumulative reward for $\sigma = 0.0$. The shaded areas around the curves represent the standard error. Similar trends were obtained for the other look-at noise distributions. (Best viewed in color)

8.4.5 Sensitivity to the Detection Probabilities

Because multiple factors can affect robot perception, such as background audio or illumination, we decided to further investigate how different detection probabilities for the robot's sensors affected its performance. In particular, we focused on evaluating DYNA-2 and TEXPLORE in this experiment, given that they performed the best previously.

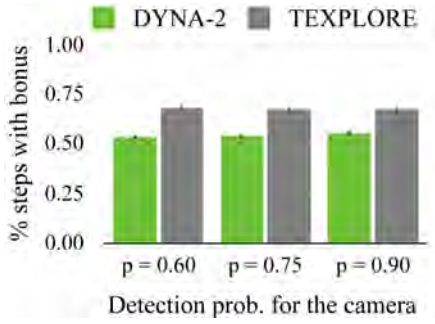
We evaluated two detection probabilities for the microphone array (0.75 and 0.90) and three for the camera (0.60, 0.75 and 0.90). As before, this experiment was repeated 10 times per agent until it completed 18000 steps. We used $\mathcal{N}(0, 0.6^2)$ for the look-at noise distribution of all the people.

Figure 8.6 shows the proportion of steps in which the robot received a positive reward with the different detection probabilities. The results were affected by the detection probability of the microphone array: the lower the probability, the fewer times the robot received a positive bonus. Interestingly, the results did not vary much with lower detection probabilities for the camera.

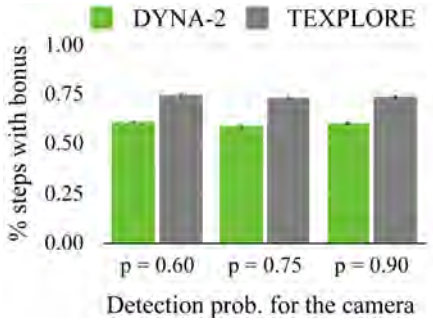
These findings are encouraging for future tests in real human-robot interactions because audio detection using microphone arrays tends to be more reliable than people detection with computer vision approaches. It is worth noting, though, that the wide field of view of the microphone array that we modeled for the robot and the lack of false positive detections likely influenced these outcomes.

8.4.6 Individual Behaviors

So far, we have considered situations in which the people in the conversation are all affected by the same look-at noise distribution and do not move. Of course, this not realistic. People often exhibit individual behaviors that differentiate them from others. To study these types of



(a) Mic. detection prob. of 0.75



(b) Mic. detection prob. of 0.90

Figure 8.6: Proportion of steps (out of 18000) in which the agents received a positive bonus as part of their reward. The left/right plots shows the results when the detection probability of the microphone array was 0.75/0.90.

situations, we investigated the performance of DYNA-2 and TEXPLORE when one person was affected by more look-at noise than the other people and when people slightly adjusted their position with respect to the rest of the group.

8.4.6.1 Non-Uniform Look-At Noise

For this test, we set the look-at noise distribution of one person to $\mathcal{N}(0, 1.2^2)$ and the rest to $\mathcal{N}(0, 0.6^2)$. Because there were four people in the simulation, we tested all four combinations with one outlier.

We found that the outlier look-at noise distribution did not affect the performance of the agents; the results were similar to those obtained for the experiment of Sec. 8.4.4. In particular, the proportions of steps in which the robot received a bonus reward were 0.58 (STE $<$ 0.01), 0.58 (STE $<$ 0.01), 0.59 (STE = 0.01) and 0.59 (STE = 0.01) with DYNA-2. With TEXPLORE, the proportions were 0.72 (STE = 0.01), 0.73 (STE $<$ 0.01), 0.73 (STE $<$ 0.01), and 0.73 (STE = 0.01). This result is not surprising given that the agents relied more on audio detections than visual information, as discussed in Sec. 8.4.5.

8.4.6.2 Changes in Location

We modified our simulation to induce small re-configurations of the people in the conversation. Every time a new speaker was selected, as described in Sec. 8.2, we flipped a coin with a success probability p_t for every other person in the conversation. If the outcome of a flip was a success, we set a desired new position $\mathbf{p}'_i$ for the corresponding person i and updated his or her position towards this location at a constant velocity. In particular,

$$\mathbf{p}'_i = \begin{cases} \mathbf{p}_i + \mathbf{t}_i & \text{if } \|\mathbf{p}_i - \mathbf{p}_i^{ini}\| < 0.5 \text{ meters} \\ \mathbf{p}_i^{ini} & \text{otherwise} \end{cases} \tag{8.7}$$

with $\mathbf{p}_i^{ini}$ the initial location of person i at the beginning of the simulation, $\mathbf{p}_i$ their previous location, and $\mathbf{t}_i$ a translation drawn from a 2D normal distribution with mean $\mathbf{0}$ and covariance $[0.01\ 0.0; 0.0\ 0.01]$. In this manner, equation (8.7) induced controlled re-configurations of the spatial arrangement while preventing dissolving the group's F-formation.

For the test with translational motion, we considered four success probabilities p_t (0.1, 0.2, 0.3, and 0.4). In general, we used a constant look-at noise distribution of $\mathcal{N}(0, 0.6^2)$ and set the detection probabilities of the microphone array and the camera to 0.9 and 0.75, respectively.

Even though DYNA-2 and TEXPLORE learned good policies with translational motion, we found that this motion slightly reduced the learning speed of DYNA-2 in comparison to using $p_t = 0.0$ (as in Sec. 8.4.4). This outcome is illustrated in Figure 8.7. Each of the plots of this figure show the absolute angular difference between the robot and the direction towards the speaker from its location ($abs(\varphi)$ in eq. (8.2)). The baseline DYNA-2 (without translational motion) converged to an absolute offset of about 0.2 radians (11.5°) after 12000 steps, whereas DYNA-2 took longer to converge with $p_t > 0.0$. TEXPLORE seemed to perform slightly better with translational motions, likely because its policy was better suited for our problem.

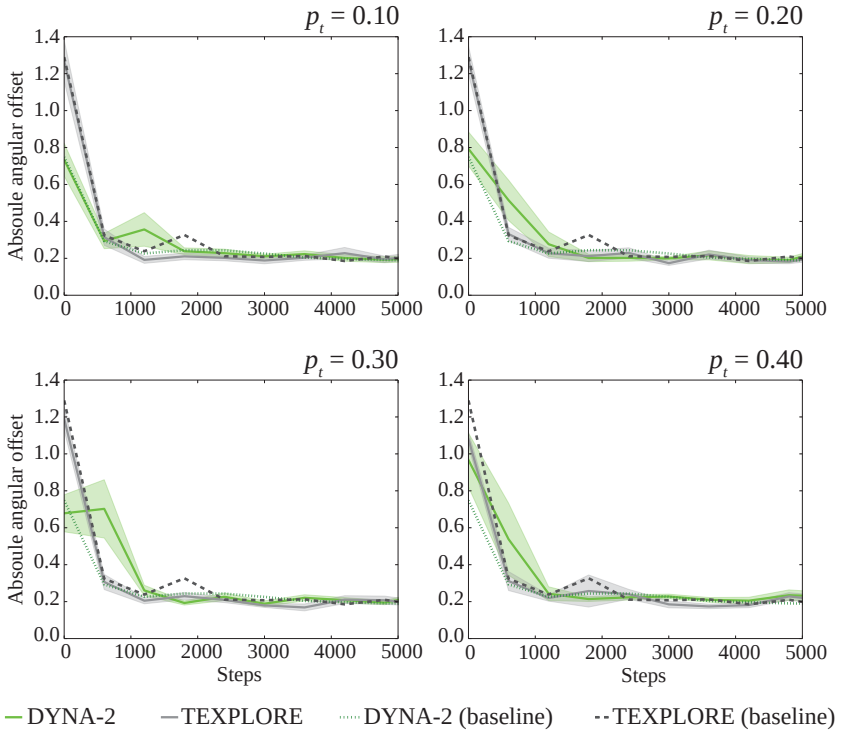


Figure 8.7: Absolute angular offset (in radians) between the robot's direction and the direction towards the speaker. Results were averaged over windows of 600 contiguous steps and over 10 runs. The shaded areas behind the DYNA-2 and TEXPLORE lines represent std. errors. Baseline results correspond to $p_t = 0.0$, i.e., no translational motion. (Best viewed in color)

8.4.7 Generalization To Other Groups

Finally, we decided to test how well pre-trained agents performed in other conversations with different numbers of people. For this test, we exposed the DYNA-2 and TEXPLORE agents that were trained with 4 people for the experiment of Sec. 8.4.4 to interactions with 2, 3, 5 and 6 people. For interactions with less than 4 people, we used a stride of 1.25 meters, as in the other experiments. When more people conversed with the robot, we increased the stride to 1.5 meters to accommodate the extra participants. For this experiment, we also let the agents adjust their Q value estimates in an online fashion. Each run lasted a total of 3000 steps (10 minutes of interaction time).

Figure 8.8 shows the average number of steps for which the agents received a bonus reward in the new environment. As expected, the agents that were pre-trained outperformed agents that started to learn from scratch. This result shows the generalization power of our state representation.

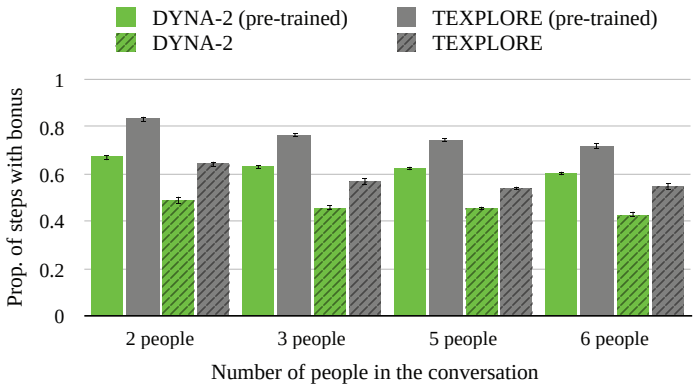


Figure 8.8: Proportion of steps (out of 3000) in which the robot got a reward with a positive bonus. Pre-trained agents interacted with 4 people prior to this test; other agents learned from scratch. Error bars represent std. errors.

8.5 Discussion

We explored reinforcement learning methods to control the orientation of a robot during simulated multi-party conversations. Our main assumption throughout this work was that the robot should turn towards the speaker to convey attentiveness and maintain awareness of the focus of attention.

The proposed state representation for the RL task encodes the likely direction of the speaker from the perspective of the robot, as captured by a camera and a microphone array on the platform. This representation is agnostic to the number of people in the conversation and, thus, can be used to generalize learned behaviors across social contexts.

Our work was limited by our simulated environment. The environment was inspired by models from social psychology that describe spatial behavior during group interactions, but did not fully capture all the complexity of real world conversations nor sensing noise. For example, only one person spoke at a time in the simulation, even though people sometimes speak simultaneously during conversations. Furthermore, the sensors that we simulated either detected people correctly or didn't detect them at all. Real sensors, however, typically provide detection scores and suffer from false positive detections.

Our findings suggest that RL has potential to succeed in motion control tasks during social conversations. In comparison to manually designing a policy to control the orientation of a robot, RL methods tend to scale more easily as the dimensionality of the state space increases. That is, RL methods can incorporate more easily additional aspects of the interaction in the decision making process. As we move forward towards controlling a real platform, this increase in dimensionality will likely be necessary to deal with the added complexity of real interactions. For example, we expect that using the detection probability of the microphone array in the state representation, rather than using a binary feature for whether or not a detection was successful, would lead to more robust policies. Similarly, incorporating a score for how many people gaze towards the point of maximum social saliency would help identify important social saliency locations from minor foci of attention. Of course, we pay a price for this flexibility: the more complex the state is, the more data RL methods need to learn from. It might be possible, though, to leverage simulations to reduce the amount of data needed to learn good policies in practice, e.g., as in [185, 224].

Chapter 9

Conclusion

This dissertation presented our first steps toward a research agenda where we aim to make robot perception cognizant of social behavior for HRI. Providing robots with mechanisms to perceive and reason about social group phenomena, like spatial behavior, is essential for them to appropriately interact with and around multiple of people in dynamic social contexts.

In particular, we focused on enabling robots to reason about spatial patterns of human behavior typical of social conversations. The main take-home messages of this dissertation are:

1. F-formations naturally emerge in HRI. Prior work had shown that this happens in dyadic interactions [72, 97]. We validated this idea further in the context of small group conversations. This validation is important because it sets the foundations for using spatial behavior as a mechanism to improve robot perception.

2. Reasoning about F-formations is useful for robots. First, it can help with perception tasks, such as tracking the lower-body orientation of nearby people. Second, it can enable robots to detect group conversations, and to cooperate to sustain them.

9.1 Summary of Contributions

This dissertation has several contributions:

- **The Furniture Robot Platform.** We built a new platform for social HRI research (Chapter 4). This platform was designed to interact with both children and adults, and is flexible in that it can operate as one or two characters simultaneously. The author contributed to designing and building parts of the robot, as well as setting up and integrating most software components.
- **Methods for Studying Group Conversations with Robots.** We designed new protocols for studying group interactions with robots (Section 5.1, Section 5.2 & Chapter 7). These protocols can be used to further study various aspects of group conversations in HRI, including spatial behavior, turn-taking patterns, social influence, and collaboration.
- **Further our Understanding of Group Conversations with Robots.** With the proposed new protocols, we performed a series of experiments to study various aspects of group interactions with robot(s), including spatial patterns of behavior that naturally emerge during conversations. These efforts informed the design of the perception and control algorithms that we proposed for these encounters in Chapters 5 and 8.

- **Framework for Group Detection:** We proposed a general framework for detecting conversational groups by reasoning about spatial behavior. This framework takes advantage of the mutual dependency between two tasks: detecting F-formations, and tracking the lower-body orientation of people in a scene (Section 6.3). As part of this work, we proposed a model-based approach to detect F-formations and estimate soft group assignment probabilities (Sec. 6.3.1). Also, we designed a particle filter for lower-body orientation tracking.
- **Methods for Robots to Cooperate To Sustain Spatial Arrangements:** We studied two orientation behaviors for robots to sustain spatial arrangements typical of group conversations (Chapter 7). We also explored Reinforcement Learning techniques to facilitate generating appropriate spatial behavior for robots in these situations (Chapter 8).

9.2 Limitations and Avenues for Future Research

While our work showed the benefits of enabling robots to reason about F-formations, it is by no means without limitations. The next sections discuss several of these limitations and present corresponding avenues for future research.

9.2.1 Data-Driven Approaches for Spatial Reasoning

So far, our efforts have focused on building model-based approaches to enable robots to reason about human spatial behavior typical of group conversations. While these efforts were an important first step towards understanding F-formations in HRI, the models that we proposed to detect these spatial organizations are limited to conversations in open environments, where F-formations are often very regular. But conversations happen in a variety of other conditions and people's spatial behavior is malleable enough to adapt to these situations. For example, people often adapt to crowded environments by interacting closer to one another as needed. Likewise, people's spatial behavior typically changes inside an elevator because the space is tight or narrow.

Moving forward, it would be interesting to explore data-driven approaches to enable robots to reason about human spatial behavior. For this task, one could build upon recent advancements in computer vision for markerless motion capture [210] and motion prediction [9]. Provided that one can collect sufficient data to learn from, data-driven approaches have the potential to be more scalable than model-based methods. They could be used to more efficiently represent the inherent malleability of F-formations across social contexts, and to account for differences in spatial behavior due to particular robot designs. Morevoer, data-driven approaches may lead to measurable properties for these spatial organizations. This quantitative information could then complement the original, qualitative models of F-formation that were proposed within social psychology, e.g., by Kendon [86]. Ultimately, these efforts could help us better understand human conversations.

9.2.2 Localization in Crowded Environments

In this dissertation, we assumed that our robot could localize during conversations with small groups of people. In our laboratory experiments, we leveraged the lidar in our robot and advances robot localization and mapping to accomplish this task [184]. Nonetheless, using a lidar for localization with these methods may not always work in crowded environments. In these situations, people will likely occlude the sensor and prevent the robot from perceiving its surrounding.

Robot localization in crowded environments might be enabled by other kind of sensing modalities, like wireless localization systems, in combination with recent machine learning techniques for sequential prediction and filtering [201, 202]. To explore this possibility, we have been conducting a series of tests with the ultra-wideband (UWB) localization system that we tried for the last user experiment described in this dissertation (Chapter 7). Our initial results suggest that the Predictive State Inference Machine filter proposed by Venkatraman et al. [202] can help better fuse UWB localization information with odometry measurements in comparison to using an Extended Kalman Filter [1] for this task. Future work should explore the problem of robot localization in crowded environments further to unlock human-robot interactions in busy human gatherings.

9.2.3 Interactive Perception

Even though we designed our experiments to be naturalistic, they were conducted in controlled laboratory spaces that were not as complex as unstructured human environments. From a perception point of view, this increase in complexity can often make it more difficult – sometimes impossible – for robots to fully perceive their surroundings and the social signals provided by their interactants. The result is an inevitable increase in perception uncertainty that can lead to inappropriate or undesired robot behaviors.

One approach to deal with perception uncertainty in more complex settings is designing perception systems that can leverage the interactive capabilities of robots. For example, imagine being at a loud party. When someone talks to us and we can't hear what the person said, we often move closer to the person and ask to repeat again what s(he) said. Why can't robots do the same and take advantage of their social environments to solve difficult perception problems?

A few methods have been proposed recently to enable robots to ask for help with spatially situated tasks [149] and manipulation tasks [91], or to ask for clarifications of user commands [211]. Building on these approaches, future research efforts should aim at generalizing these ideas to incorporate *interactive perception* mechanisms on all sort of perception tasks for HRI, specially those that are necessary during multi-party interactions. We expect these efforts to lead to adaptable and robust robot behavior in unstructured environments.

Appendices

Appendix A

Chester's Gaze Calibration

In order to make the eyes of Chester gaze towards an arbitrary (X, Y, Z) position in the world, we estimated a mapping from 3D world coordinates to 2D (x, y) pupil positions within the eyes. For this task, we constrained the line of sight of the two eyes of the characters to be parallel. This means, that the two pupils are always in the same position relative to their corresponding eye, and that we only need to learn one mapping for both of them. While this assumption is not very realistic – our robot is unable to reproduce vergence – it worked well in practice. One reason is that the eyes of the characters have a cartoonish look, which makes users forgiving to gaze patterns that do not fully match the properties of our physical world. Another reason is that the eyes of our characters are (almost) planar and, thus, suffer from the Mona Lisa gaze effect [8]. That is, users feel like the characters establish mutual gaze with them, more often than not.

Because the two pupils move together, we compute their mappings from 3D world coordinates to local 2D pupil positions using a simulated, third eye that is centered in-between the real pupils. As illustrated in Figure A.1, the position (x, y) of the pupil in this third "middle eye" passes through a ray that connects the center of projection of the eye and an observed (X, Y, Z) point in the world. Mathematically, this projection mapping [225] can be formulated as:

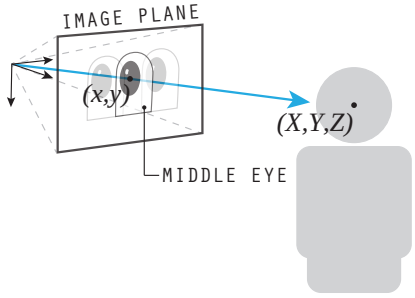


Figure A.1: Gaze calibration

$$\begin{bmatrix} x' \\ y' \\ w' \end{bmatrix} = \begin{bmatrix} f_x & 0 & 0 \\ 0 & f_y & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \implies \begin{array}{l} x = x'/w' = f_x X/Z \\ y = y'/w' = f_y Y/Z \end{array}$$

We estimated the intrinsic parameters f_x and f_y of this model by solving over-determined systems with multiple $(X, Y, Z) - (x, y)$ correspondences. To obtain the correspondences, we asked a person to stand in various places towards the front of the robot (as illustrated in Fig. A.1). As the person moved from place to place, we ran an off-the-shelf face detector on the RGB-D stream from the sensor inside the robot's lamp to gather the (X, Y, Z) position of her head. Meanwhile, we asked the person to control the gaze of the robot using a wireless gamepad. The person moved the eyes in each case until the robot established mutual gaze with her, effectively providing (x, y).

Once this 3D $\rightarrow$ 2D mapping was estimated, there was nothing in the model that prevented the pupils from going outside the scleras of the eyes. Thus, we constrained the resulting (x, y) pupil positions to reasonable limits in practice.

Appendix B

Optimizing the Stride of the O-Space Proposals

Equation (6.1) introduced a parametric o-space proposal model, similar to the one used in [40, 155, 156]. This model served to estimate the center of the o-spaces generated by social conversations in a scene, given people's position and lower-body orientation. Here, we present an optimization approach that can be used to fit the stride of this o-space proposal model using a dataset of conversational groups, like the Cocktail Party dataset [155].

For completeness, let us first re-introduce the o-space proposal model. Without loss of generality, consider a particular person i in a social environment, and let the position and lower-body orientation of this person be $\mathbf{p}_i \in \mathbb{R}^2$ and $\theta_i \in [0, 2\pi]$, respectively. If this person is standing still and conversing with other people, his or her o-space will likely be centered at:

$$\mathbf{o}_i = \mathbf{p}_i + d \begin{bmatrix} \cos(\theta_i) \\ \sin(\theta_i) \end{bmatrix} \tag{B.1}$$

where d is known as the *stride* of this parametric model. The stride represents the expected distance between the person and his or her o-space center.

What stride should we use for the proposals?

Assume that you have access to a dataset of conversational groups G, where each group $g \in G$ has P_g participants. In addition, assume that this dataset provides accurate pose information, i.e., the position $\mathbf{p}$ and orientation θ of every person in the groups is well known. Then, we pose that a good stride s for the o-space proposals in eq. (B.1) is the distance that induces the proposals to be as close as possible to the average $\bar{\mathbf{o}}_g = \frac{1}{P_g} \sum_{j=1}^{P_g} \mathbf{o}_j$ in their group. Note that this average proposal has been used by prior work to model the true location of o-spaces in a scene [156].

The above objective can be formalized as,

$$\arg\min_s L(s), \text{ with } L(s) = \sum_{g=1}^{G} \sum_{i=1}^{P_g} \left\| \bar{\mathbf{o}}_g - \mathbf{o}_i \right\|^2 \text{ and } s \in (0, \infty) \tag{B.2}$$

This is a convex problem (see the next section of this Appendix for the proof). Any local minimum must be a global minimum.

To solve for the best stride, let us first simplify the loss function $L(s)$ in eq. (B.2):

$$\begin{aligned} L(s) &= \sum_{g=1}^{G}\sum_{i=1}^{P_g} \left\| \bar{\mathbf{o}}_g - \mathbf{o}_i \right\|^2 = \sum_{g=1}^{G}\sum_{i=1}^{P_g} \left\| \frac{1}{P_g}\sum_{j=1}^{P_g} \left(\mathbf{p}_j + s \begin{bmatrix} \cos(\theta_j) \\ \sin(\theta_j) \end{bmatrix} \right) - \left(\mathbf{p}_i + s \begin{bmatrix} \cos(\theta_i) \\ \sin(\theta_i) \end{bmatrix} \right) \right\|^2 \\ &= \sum_{g=1}^{G}\sum_{i=1}^{P_g} \Big\| \underbrace{\left((\frac{1}{P_g}\sum_{j=1}^{P_g} \mathbf{p}_j) - \mathbf{p}_i \right)}_{\mathbf{a}_i^g} + s \underbrace{\left((\frac{1}{P_g}\sum_{j=1}^{P_g} \begin{bmatrix} \cos(\theta_j) \\ \sin(\theta_j) \end{bmatrix}) - \begin{bmatrix} \cos(\theta_i) \\ \sin(\theta_i) \end{bmatrix} \right)}_{\mathbf{b}_i^g} \Big\|^2 \\ &= \sum_{g=1}^{G}\sum_{i=1}^{P_g} \left\| \mathbf{a}_i^g + s\mathbf{b}_i^g \right\|^2 \end{aligned} \tag{B.3}$$

Then, we can find the minimum of L by differentiating with respect to s,

$$\begin{aligned} \frac{dL}{ds} &= \sum_{g=1}^{G}\sum_{i=1}^{P_g} (\mathbf{b}_i^g)^T \left(\mathbf{a}_i^g + s\mathbf{b}_i^g \right) + \left(\mathbf{a}_i^g + s\mathbf{b}_i^g \right)^T \mathbf{b}_i^g \\ &= \sum_{g=1}^{G}\sum_{i=1}^{P_g} 2\Big((\mathbf{b}_i^g)^T \mathbf{a}_i^g + s(\mathbf{b}_i^g)^T \mathbf{b}_i^g \Big) \end{aligned}$$

and making the derivative equal to zero,

$$\sum_{g=1}^{G}\sum_{i=1}^{P_g} (\mathbf{b}_i^g)^T \mathbf{a}_i^g + s \sum_{g=1}^{G}\sum_{i=1}^{P_g} (\mathbf{b}_i^g)^T \mathbf{b}_i^g = 0$$

The optimal stride then becomes:

$$s = \frac{-\sum_{g=1}^{G}\sum_{i=1}^{P_g} (\mathbf{b}_i^g)^T \mathbf{a}_i^g}{\sum_{g=1}^{G}\sum_{i=1}^{P_g} (\mathbf{b}_i^g)^T \mathbf{b}_i^g} \tag{B.4}$$

which can be computed analytically using a dataset of group conversations with annotated interactions and pose information. For example, in the case of the Cocktail Party dataset, this optimization approach resulted in a stride of 0.72 meters using ground truth body orientation annotations that we collected for this dataset.

Proof of convexity

To prove that the minimization problem of eq. (B.2) is convex [26], we need to show that the domain of the problem is convex (Lemma B.0.1) and that the function $L(s)$ is convex as well (Lemma B.0.2).

Lemma B.0.1. *The set $(0, \infty)$ is a convex set.*

Proof. Let $\lambda \in [0, 1]$ and x_1, $x_2 \in (0, \infty)$. We need to show that $\lambda x_1 + (1 - \lambda)x_2$ is also in $(0, \infty)$.

The derivation is the same as for any other open interval in $\mathbb{R}$ [26]. First, assume that $x_1 \leq x_2$ without loss of generality. Then, $\lambda x_1 + (1 - \lambda)x_2 \geq \lambda x_1 + (1 - \lambda)x_1 = x_1$. Similarly, $\lambda x_1 + (1-\lambda)x_2 \leq \lambda x_2 + (1-\lambda)x_2 = x_2$. Thus, the convex combination $\lambda x_1 + (1-\lambda)x_2 \in [x_1, x_2]$ and because x_1 and x_2 are in $(0, \infty)$, then $\lambda x_1 + (1 - \lambda)x_2 \in (0, \infty)$ as well. □

Lemma B.0.2. *The function $L(s)$ in eq. (B.2) is convex.*

Proof. The sum of convex functions in a common convex domain is also convex [26]. Thus, to show that the function $L(s)$ in eq. (B.2) is convex, it suffices to show that all the terms $\|\mathbf{a}_i^g + s\mathbf{b}_i^g\|^2$ that are summed together in its simplified form in eq. (B.3) are indeed convex.

Without loss of generality, let $e(s) = \|\mathbf{a}_i^g + s\mathbf{b}_i^g\|^2$ be the term in eq. (B.3) that corresponds to any group $g \in G$ and individual $i \in P_g$. We can express $e(s)$ as the composition of two functions $(f \circ g)(s)$, where $g(x) = \|\mathbf{a}_i^g + x\mathbf{b}_i^g\|$ and $f(x) = x^2$. We can now prove that e is convex by showing that g and f are convex and that f is non-decreasing in the range of g, because any such composition of convex functions is convex as well (see [26], Sec. 3.2.4).

First we prove that g is convex by showing that $g(\lambda x + (1-\lambda)y) \leq \lambda g(x) + (1-\lambda)g(y)$ for any $\lambda \in [0,1]$ and $x,\ y \in (0,\infty)$. In the proof below, we use two key properties of the ℓ_2 norm: homogeneity ($\|k\mathbf{v}\| = |k|\|\mathbf{v}\|$ for any scalar k and vector $\mathbf{v}$), and the triangle inequality ($\|\mathbf{v}+\mathbf{u}\| \leq \|\mathbf{v}\| + \|\mathbf{u}\|$ for $\mathbf{v},\mathbf{u} \in \mathbb{R}^2$) [25].

$$\begin{aligned}
\lambda g(x) + (1-\lambda)g(y) &= \lambda\|\mathbf{a}_i^g + x\mathbf{b}_i^g\| + (1-\lambda)\|\mathbf{a}_i^g + y\mathbf{b}_i^g\| \\
&= \|\lambda(\mathbf{a}_i^g + x\mathbf{b}_i^g)\| + \|(1-\lambda)(\mathbf{a}_i^g + y\mathbf{b}_i^g)\| \\
&\geq \|\lambda(\mathbf{a}_i^g + x\mathbf{b}_i^g) + (1-\lambda)(\mathbf{a}_i^g + y\mathbf{b}_i^g)\| \\
&= \|\mathbf{a}_i^g + (\lambda x + (1-\lambda)y)\mathbf{b}_i^g\| \\
&= g(\lambda x + (1-\lambda)y)
\end{aligned}$$

The function $f(x) = x^2$ is non-decreasing because $f'(x) = 2x \geq 0$ for all x in the range $[0,\infty)$ of g. The function $f(x)$ is also convex because $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y)$ for $\lambda \in [0,1]$ and $x, y \in [0,\infty)$:

$$\begin{aligned}
& f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y) \\
\iff & (\lambda x + (1-\lambda)y)^2 \leq \lambda x^2 + (1-\lambda)y^2 \\
\iff & \lambda^2 x^2 + 2\lambda(1-\lambda)xy + (1-\lambda)^2 y^2 \leq \lambda x^2 + (1-\lambda)y^2 \\
\iff & 0 \leq (1-\lambda)\lambda x^2 + (1-(1-\lambda))(1-\lambda)y^2 - 2\lambda(1-\lambda)xy \\
\iff & 0 \leq (1-\lambda)\lambda x^2 + \lambda(1-\lambda)y^2 - 2\lambda(1-\lambda)xy \\
\iff & 0 \leq (1-\lambda)\lambda(x-y)^2
\end{aligned}$$

The last inequality $0 \leq (1-\lambda)\lambda(x-y)^2$ is true because $(1-\lambda)$, λ, and $(x-y)^2$ are all greater or equal to zero. □

Appendix C

Tracking with Ultra Wide-Band Wireless Sensors

We used Ciholas' Ultra Wide-Band (UWB) sensors [38] to track participants in our third user experiment (Chapter 7). These wireless sensors come in a compact, lightweight format and incorporate DecaWave's DW1000 transceiver chip for real-time localization (Fig. C.1).

To use these sensors, we set up a set of them as *anchors* in the environment. These anchors were then used for localizing another set of sensors, known as *tags*, using multilateration techniques. This localization information was output by Ciholas' software with respect to a global coordinate frame.

Figure C.1: UWB sensor

Calibrating UWB Position Measurements

The raw position measurements output by Ciholas' software suffered from near field effects at close range. These effects became apparent as a constant, but spatially-varying, bias in UWB position measurements.

We used machine learning to compensate for this bias and get more accurate 2D position information for the UWB tags. To that end, we built a dataset $\mathcal{D} = \{(\mathbf{z}_i, \mathbf{x}_i) \,|\, 1 \leq i \leq N\}$ of 2D UWB position measurements $\mathbf{z}$ and corresponding ground truth values $\mathbf{x}$. We modeled the spatially-varying bias as a deterministic function $f : \mathcal{R}^2 \rightarrow \mathcal{R}^2$, such that $\mathbf{x} = \mathbf{z} + f(\mathbf{z})$. While this function could be learned with any (non-linear) supervised learning technique, we used two gaussian processes in this work (one for the x direction and another one for the y direction on the ground plane of the environment). Gaussian processes worked well as they allowed us to compute a smooth bias functions with a small set of manual measurements.

Tracking Instrumented Caps

To track people in our laboratory, we instrumented baseball caps with two UWB tags each (Fig. C.2). The 3D measurements provided by Cihola's software for each of these tags was then filtered with a Kalman filter. The hidden state of this filter was $\mathbf{x}_t = [x_t \; y_t \; z_t \; \dot{x}_t \; \dot{y}_t \; \dot{z}_t]^T$, where $(x, y, z)_t$ denoted the position of the wireless tag being tracked at time t, and

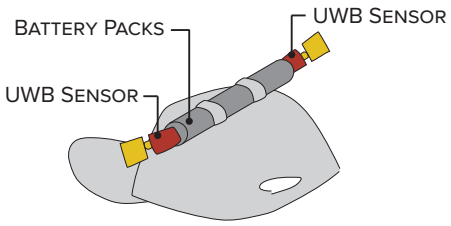


Figure C.2: Cap

$(\dot{x}, \dot{y}, \dot{z})_t$ was its instantaneous velocity. This 6D state evolved under a constant velocity model as follows:

$$\mathbf{x}_{t+1} = \begin{bmatrix} 1 & 0 & 0 & \Delta T & 0 & 0 \\ 0 & 1 & 0 & 0 & \Delta T & 0 \\ 0 & 0 & 1 & 0 & 0 & \Delta T \end{bmatrix} \mathbf{x}_t + \mathbf{w}_t \tag{C.1}$$

where $\mathbf{w}_t \sim \mathcal{N}(0, Q)$ was the process noise, and ΔT corresponded to the time difference between t and $t-1$. The observation model of this filter was:

$$\mathbf{z}_t = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix} \mathbf{x}_t + \mathbf{v}_t \tag{C.2}$$

where $\mathbf{z}_t$ was the position measurement of the UWB tag at time t, as output by Ciholas' system. The vector $\mathbf{v}_t \sim \mathcal{N}(0, R)$ was the observation noise. We tuned the parameters of this filter based on experimental observations.

Once the position of the UWB tags was filtered, we converted their 3D location to 2D by projecting the tag's (x, y, z) state on the ground plane of the room, i.e., making their vertical coordinate $z = 0$. We disregarded the vertical measurements because they were significantly noisy.

We finally combined the measurements of the two tags on a cap. Let $[x_1\ y_1]^T$ and $[x_2\ y_2]^T$ be the 2D positions of the left and right tags on a given cap after projecting on the ground and compensating for the spatially-varying bias described in the previous section. The cap's 2D position was set to the average $\mathbf{p} = 0.5 * \sum_{i=1}^{2} [x_i\ y_i]^T$, as depicted in Figure C.3 (yellow circle). The orientation of the cap was computed as the yaw angle $\theta = \arctan2(y_2 - y_1, x_2 - x1)$.

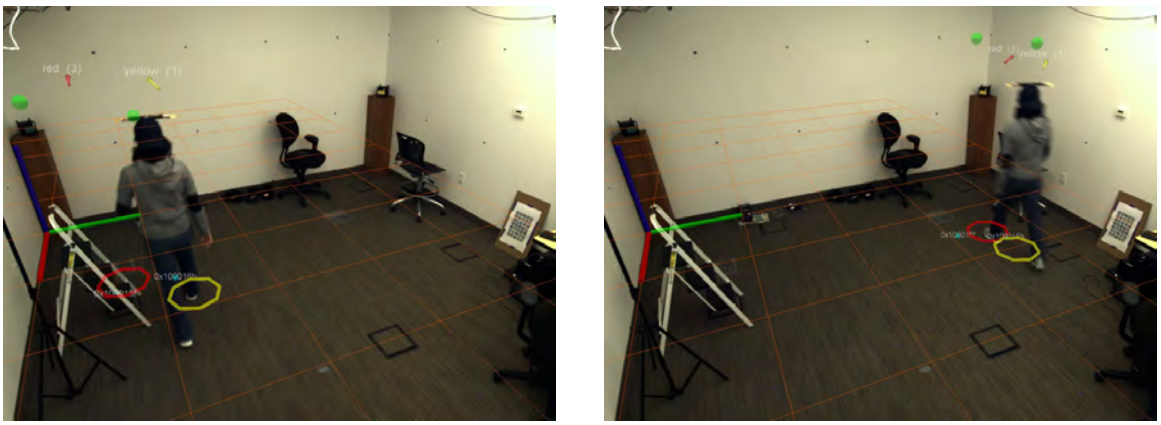


Figure C.3: Tracking of a baseball cap with two UWB tags. The green spheres indicate the filtered 3D positions of the two tags on the person's cap. Because the vertical estimates of the tags were very noisy, we projected their 3D position on the ground plane and processed their 2D location only. The center of the red circle corresponds to the average 2D position of the tags after being projected. The center of the yellow circle was computed like the center for the red one, but the projected 2D positions of the tags was pre-processed to remove the spatially-varying bias that we computed for the room.

Even though we tried tracking people's lower-body orientation with these sensors as well, we found that heavy occlusions generated more noise than we could handle in this case. Ciholas' software is evolving, though. It might be possible to use these sensors to track additional body features in the future.

Bibliography

[1] Pieter Abbeel, Adam Coates, Michael Montemerlo, Andrew Y Ng, and Sebastian Thrun. Discriminative Training of Kalman Filters. In *Robotics: Science and systems*, volume 2, page 1, 2005.

[2] Leslie Adams and David Zuckerman. The effect of lighting conditions on personal space requirements. *The journal of general psychology*, 118(4):335–340, 1991.

[3] Henny Admoni and Brian Scassellati. Social Eye Gaze in Human-Robot Interaction: A Review. *Journal of Human-Robot Interaction (JHRI).*, 2017.

[4] Henny Admoni, Bradley Hayes, David Feil-Seifer, Daniel Ullman, and Brian Scassellati. Are You Looking at Me?: Perception of Robot Attention is Mediated by Gaze Type and Group Size. In *Proc. of the 8th ACM/IEEE Int'l Conference on Human-robot Interaction*, HRI '13, pages 389–396, 2013.

[5] Maedeh Aghaei, Mariella Dimiccoli, and Petia Radeva. Towards Social Interaction Detection in Egocentric Photo-streams. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2016.

[6] Samer Al Moubayed and Jill Lehman. Design and Architecture of a Robot-Child Speech-Controlled Game. In *Proc. of the Tenth Annual ACM/IEEE Int'l Conference on Human-Robot Interaction Extended Abstracts*, HRI'15 Extended Abstracts, pages 79–80, 2015.

[7] Samer Al Moubayed, Jonas Beskow, Gabriel Skantze, and Björn Granström. *Cognitive Behavioural Systems: COST 2102 International Training School, Revised Selected Papers*, chapter Furhat: A Back-Projected Human-Like Robot Head for Multiparty Human-Machine Interaction, pages 114–130. Springer Berlin Heidelberg, 2012.

[8] Samer Al Moubayed, Jens Edlund, and Jonas Beskow. Taming Mona Lisa: Communicating Gaze Faithfully in 2D and 3D Facial Projections. *ACM Trans. on Interactive Intelligent Systems*, 1(2):25, 2012.

[9] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese. Social LSTM: Human Trajectory Prediction in Crowded Spaces. In *The IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '16, 2016.

[10] Stefano Alletto, Giuseppe Serra, Simone Calderara, Francesco Solera, and Rita Cucchiara. From ego to nos-vision: Detecting social relationships in first-person views. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition - Workshops*, pages 580–585, 2014.

[11] P. Althaus, H. Ishiguro, T. Kanda, T. Miyashita, and H.I. Christensen. Navigation for human-robot interaction tasks. In *Proc. of the 2004 IEEE Int'l Conference on Robotics and Automation*, volume 2 of *ICRA '04*, pages 1894–1900 Vol.2, April 2004.

[12] Sean Andrist, Bilge Mutlu, and Adriana Tapus. Look Like Me: Matching Robot Personality via Gaze to Increase Motivation. In *Proc. of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 3603–3612, 2015.

[13] Brenna Argall, Yang Gu, Brett Browning, and Manuela Veloso. The First Segway Soccer Experience: Towards Peer-to-peer Human-robot Teams. In *Proc. of the 1st ACM SIGCHI/SIGART Conference on Human-robot Interaction*, HRI '06, pages 321–322, 2006.

[14] Michael Argyle. *Bodily Communication*. University paperbacks. Methuen, 1988.

[15] Arthur Aron, Elaine N Aron, and Danny Smollan. Inclusion of Other in the Self Scale and the structure of interpersonal closeness. *Journal of personality and social psychology*, 63(4):596, 1992.

[16] Kent G. Bailey, John J. Hartnett, and Hilda W. Glover. Modeling and Personal Space Behavior in Children. *The Journal of Psychology*, 85(1):143–150, 1973.

[17] Christoph Bartneck, Michel van der Hoek, Omar Mubin, and Abdullah Al Mahmud. Daisy, Daisy, Give Me Your Answer Do!: Switching off a Robot. In *Proc. of the Second ACM/IEEE Int'l Conference on Human-robot Interaction*, HRI '07, pages 217–222, 2007.

[18] Archer L. Batcheller, Brian Hilligoss, Kevin Nam, Emilee Rader, Marta Rey-Babarro, and Xiaomu Zhou. Testing the Technology: Playing Games with Video Conferencing. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, pages 849–852, 2007.

[19] James P. Batchelor and George R. Goethais. Spatial Arrangements in Freely Formed Groups. *Sociometry*, 35(2):270–279, 1972.

[20] L. Bazzani, M. Cristani, D. Tosato, M. Farenzena, G. Paggetti, G. Menegaz, and V. Murino. Social Interactions by Visual Focus of Attention in a Three-Dimensional Environment. *Expert Systems*, 30(2):115–127, 2013.

[21] Barry Bodt, Richard Camden, Harry Scott, Adam Jacoff, Tsai Hong, Tommy Chang, Rick Norcross, Tony Downs, and Ann Virts. Performance Measurements for Evaluating Static and Dynamic Multiple Human Detection and Tracking Systems in Unstructured Environments. In *Proc. of the 9th Workshop on Performance Metrics for Intelligent Systems*, PerMIS, 2009.

[22] René te Boekhorst, Michael L. Walters, Kheng Lee Koay, Kerstin Dautenhahn, and Chrystopher L. Nehaniv. A study of a single robot interacting with groups of children in a rotation game scenario. In *Proc. of the 2005 IEEE International Symposium on Computational Intelligence in Robotics and Automation*, CIRA '05, 2005.

[23] 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.

[24] Dan Bohus and Eric Horvitz. Dialog in the Open World: Platform and Applications. In *Proc. of the 2009 Int'l Conf. on Multimodal Interfaces*, ICMI-MLMI, 2009.

[25] N. Bourbaki. *General Topology: Chapters 5–10*. Number v. 4 in Actualités scientifiques et industrielles. Springer Berlin Heidelberg, 1998.

[26] S.P. Boyd and L. Vandenberghe. *Convex Optimization*. Berichte über verteilte messysteme. Cambridge University Press, 2004. ISBN 9780521833783. URL `https://books.google.com/books?id=mYm0bLd3fcoC`.

[27] Oliver Brdiczka, Jérôme Maisonnasse, and Patrick Reignier. Automatic Detection of Interaction Groups. In *Proc. of the 7th Int'l Conference on Multimodal Interfaces*, ICMI '05, 2005.

[28] Wolfram Burgard, Armin B. Cremers, Dieter Fox, Dirk Hähnel, Gerhard Lakemeyer, Dirk Schulz, Walter Steiner, and Sebastian Thrun. Experiences with an Interactive Museum Tour-guide Robot. *Artif. Intell.*, 114(1-2):3–55, October 1999.

[29] M. A. Carreira-Perpinan. Mode-finding for mixtures of Gaussian distributions. Technical Report CS-99-03, Dept. of Computer Science, University of Sheffield, UK, 1999.

[30] Justine Cassell, Obed E Torres, and Scott Prevost. Turn taking versus discourse structure. In *Machine conversations*, pages 143–153. Springer, 1999.

[31] Ginevra Castellano, André Pereira, Iolanda Leite, Ana Paiva, and Peter W. McOwan. Detecting User Engagement with a Robot Companion Using Task and Social Interaction-based Features. In *Proc. of the 2009 International Conference on Multimodal Interfaces*, ICMI-MLMI '09, pages 119–126, 2009.

[32] M. C. Chang, N. Krahnstoever, and W. Ge. Probabilistic group-level motion analysis and scenario recognition. In *2011 Int'l Conference on Computer Vision*, pages 747–754, 2011.

[33] C. W. Chen, R. C. Ugarte, C. Wu, and H. Aghajan. Discovering social interactions in real work environments. In *Face and Gesture 2011*, pages 933–938, 2011.

[34] Jung Ju Choi, Yunkyung Kim, and Sonya S. Kwak. Have You Ever Lied?: The Impacts of Gaze Avoidance on People's Perception of a Robot. In *Proc. of the 8th ACM/IEEE Int'l Conference on Human-robot Interaction*, HRI '13, pages 105–106, 2013.

[35] Wongun Choi, Khuram Shahid, and Silvio Savarese. What are they doing?: Collective activity classification using spatio-temporal relationship among people. In *2009 IEEE 12th Int'l Conference on Computer Vision - Workshops*, ICCV'09 Workshops, pages 1282–1289. IEEE, 2009.

[36] Wongun Choi, Yu-Wei Chao, Caroline Pantofaru, and Silvio Savarese. Discovering groups of people in images. In *Proc. of the European Conference on Computer Vision*, ECCV '14, pages 417–433, 2014.

[37] Tanzeem Choudhury and Alex Pentland. The Sociometer: A Wearable Device for Understanding Human Networks. In *Ad-Hoc Communications and Collaboration in Ubiquitous Computing Environments - A CSCW'02 Workshop*, 2002.

[38] Ciholas Inc. DWUSB, Last accessed on Sept. 2015. URL `http://www.ciholas.com/`.

[39] William S Condon and William D Ogston. Speech and body motion synchrony of the speaker-hearer. *The perception of Language*, pages 150–184, 1971.

[40] M. Cristani, L. Bazzani, G. Paggetti, A. Fossati, D. Tosato, A. Del Bue, G. Menegaz, and V. Murino. Social Interaction Discovery by Statistical Analysis of F-Formations. In *Proc. of the British Machine Vision Conference*, BMVC '11, 2011.

[41] William Dabney and Andrew Barto. Adaptive Step-Size for Online Temporal Difference Learning. In *AAAI'12*, 2012.

[42] Jeanne Dietsch. People meeting robots in the workplace [industrial activities]. *IEEE Robotics & Automation Magazine*, 17(2):15–16, 2010.

[43] Michael J Doughty. Consideration of three types of spontaneous eyeblink activity in normal humans: during reading and video display terminal use, in primary gaze, and while in conversation. *Optometry & Vision Science*, 78(10):712–725, 2001.

[44] Anca Dragan, Shira Bauman, Jodi Forlizzi, and Siddhartha Srinivasa. Effects of Robot Motion on Human-Robot Collaboration. In *Proc. of the Tenth Annual ACM/IEEE Int'l Conference on Human-Robot Interaction*, HRI 2015, March 2015.

[45] Starkey Duncan. On the Structure of Speaker-Auditor Interaction during Speaking Turns. *Language in Society*, 3(2):161–180, 1974.

[46] Nathan Eagle and Alex (Sandy) Pentland. Reality Mining: Sensing Complex Social Systems. *Personal Ubiquitous Comput.*, 10(4):255–268, 2006.

[47] Vanessa Evers, Nuno Menezes, Luis Merino, Dariu Gavrila, Fernando Nabais, Maja Pantic, Paulo Alvito, and Daphne Karreman. The Development and Real-world Deployment of FROG, the Fun Robotic Outdoor Guide. In *Proc. of the 2014 ACM/IEEE International Conference on Human-robot Interaction*, HRI '14, 2014.

[48] A. Fathi, J.K. Hodgins, and J.M. Rehg. Social interactions: A first-person perspective. In *Proc. 2012 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '12, 2012.

[49] D. Feil-Seifer and M. J. Mataric. A multi-modal approach to selective interaction in assistive domains. In *IEEE International Workshop on Robot and Human Interactive Communication*, ROMAN '05, pages 416–421, 2005.

[50] L. Feng and B. Bhanu. Understanding Dynamic Social Grouping Behaviors of Pedestrians. *IEEE Journal of Selected Topics in Signal Processing*, 9(2):317–329, 2015.

[51] G. Ferrer, A. Garrell, and A. Sanfeliu. Robot companion: A social-force based approach with human awareness-navigation in crowded environments. In *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*, pages 1688–1694, 2013.

[52] N.I. Fisher. *Statistical Analysis of Circular Data*. Cambridge University Press, 1995.

[53] F. Flohr, M. Dumitru-Guzu, J.F.P. Kooij, and D.M. Gavrila. Joint probabilistic pedestrian head and body orientation estimation. In *Proc. of the 2014 IEEE Intelligent Vehicles Symposium*, pages 617–622, 2014.

[54] Tian Gan, Yongkang Wong, Daqing Zhang, and Mohan S. Kankanhalli. Temporal Encoded F-formation System for Social Interaction Detection. In *Proc. of the 21st ACM International Conference on Multimedia*, MM '13, 2013.

[55] Maia Garau, Mel Slater, Simon Bee, and Martina Angela Sasse. The Impact of Eye Gaze on Communication Using Humanoid Avatars. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '01, pages 309–316, 2001.

[56] W. Ge, R. T. Collins, and B. Ruback. Automatically detecting the small group structure of a crowd. In *2009 Workshop on Applications of Computer Vision*, WACV '09, pages 1–8, 2009.

[57] Afzal Godil, Roger Bostelman, Kamel Saidi, Will Shackleford, Geraldine Cheok, Michael Shneier, and Tsai Hong. 3D Ground-Truth Systems for Object/Human Recognition and Tracking. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2013.

[58] Erving Goffman. *Behavior in public places: Notes on the social organization of gatherings.* Free Press of Glencoe, 1963.

[59] G. Groh, A. Lehmann, J. Reimers, M. R. Friess, and L. Schwarz. Detecting Social Situations from Interaction Geometry. In *2010 IEEE Second International Conference on Social Computing*, pages 1–8, 2010.

[60] Edward T. Hall. *The hidden dimension.* Doubleday, 1966.

[61] David Hanson. Exploring the aesthetic range for humanoid robots. In *Proc. of the ICCS/CogSci-2006 Long Symposium: Toward Social Mechanisms of Android Science*, pages 39–42, 2006.

[62] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision.* Cambridge UP, 2 edition, 2004.

[63] Kotaro Hayashi, Takayuki Kanda, Hiroshi Ishiguro, Tsukasa Ogasawara, and Norihiro Hagita. An Experimental Study of the Use of Multiple Humanoid Robots as a Social Communication Medium. In *Proc. Universal Access in Human-Computer Interaction. Applications and Services*, volume 6768 of *Lect. Notes Comput. Sci.*, pages 32–41. Springer Berlin Heidelberg, 2011.

[64] Koutarou Hayashi, Takayuki Kanda, Takahiro Miyashita, Hiroshi Ishiguro, and Norihiro Hagita. Robot Manzai - robots' conversation as a passive social medium. In *Proc. Humanoids*, 2005.

[65] Leslie A Hayduk. The Shape of Personal Space: An Experimental Investigation. *Canadian Journal of Behavioural Science*, 13(1):87, 1981.

[66] Z. Henkel, C.L. Bethel, R.R. Murphy, and V. Srinivasan. Evaluation of Proxemic Scaling Functions for Social Robotics. *IEEE Transactions on Human-Machine Systems*, 44(3): 374–385, 2014.

[67] Todd Hester. *TEXPLORE: Temporal Difference Reinforcement Learning for Robots and Time-Constrained Domains.* PhD thesis, The University of Texas at Austin, Texas, USA, December 2012.

[68] M.A. Hogg and J. Cooper. *The SAGE Handbook of Social Psychology*. SAGE Publications, 2003.

[69] M. Ehsan Hoque. Disney's First Autonomous Audio-Animatronics, Last Accessed on July, 2015. URL `http://web.media.mit.edu/~mehoque/DisnayAnimatronics.htm`.

[70] Chien-Ming Huang and Bilge Mutlu. Learning-based Modeling of Multimodal Behaviors for Humanlike Robots. In *Proc. of the 2014 ACM/IEEE Int'l Conference on Human-robot Interaction*, HRI '14, pages 57–64, 2014.

[71] Chien-Ming Huang, Maya Cakmak, and Bilge Mutlu. Adaptive Coordination Strategies for Human-Robot Handovers. In *Proc. of Robotics: Science and Systems*, R:SS, 2015.

[72] H. Huettenrauch, K. Severinson Eklundh, A. Green, and E.A. Topp. Investigating Spatial Relationships in Human-Robot Interaction. In *IEEE/RSJ Int'l Conference on Intelligent Robots and Systems*, IROS '06, pages 5052–5059, 2006.

[73] Hayley Hung and Gokul Chittaranjan. The IDIAP Wolf Corpus: Exploring Group Behaviour in a Competitive Role-Playing Game. In *Proc. of the Int'l Conference on Multimedia*, pages 879–882. ACM, 2010.

[74] Hayley Hung and Ben Kröse. Detecting F-formations As Dominant Sets. In *Proc. of the 13th International Conference on Multimodal Interfaces*, ICMI '11, 2011.

[75] Hayley Hung, Gwenn Englebienne, and Laura Cabrera Quiros. Detecting conversing groups with a single worn accelerometer. In *Proc. of the 16th Int'l Conference on Multimodal Interaction*, pages 84–91. ACM, 2014.

[76] K. Isbister. *Better Game Characters by Design: A Psychological Approach*. CRC Press, 2006.

[77] Y. Iwamura, M. Shiomi, T. Kanda, H. Ishiguro, and N. Hagita. Do elderly people prefer a conversational humanoid as a shopping assistant partner in supermarkets? In *Proc. of the 6th ACM/IEEE Int'l Conference on Human-Robot Interaction*, HRI '11, pages 449–457, March 2011.

[78] Takayuki Kanda, Hiroshi Ishiguro, Tetsuo Ono, Michita Imai, and Kenji Mase. Multi-robot cooperation for human-robot communication. In *Proc. of the 11th IEEE Int'l Workshop on Robot and Human Interactive Communication*, 2002.

[79] D. Karreman, G.S. Bradford, B. Van Dijk, M. Lohse, and V. Evers. What happens when a robot favors someone? How a tour guide robot uses gaze behavior to address multiple persons while storytelling about art. In *Proc. of the 8th ACM/IEEE Int'l Conference on Human-Robot Interaction*, HRI '13, pages 157–158, March 2013.

[80] Daphne E. Karreman, Geke D.S. Ludden, Elisabeth M.A.G. van Dijk, and Vanessa Evers. How can a tour guide robot's orientation influence visitors' orientation and formations? In *Proc. of the 4th Int'l Symposium on New Frontiers in Human-Robot Interaction*, pages 21–22, 2015.

[81] Yusuke Kato, Takayuki Kanda, and Hiroshi Ishiguro. May I Help You?: Design of Human-like Polite Approaching Behavior. In *Proc. of the 10th ACM/IEEE Int'l Conference on Human-Robot Interaction*, HRI '15, pages 35–42, 2015.

[82] Michael Katzenmaier, Rainer Stiefelhagen, and Tanja Schultz. Identifying the Addressee in Human-human-robot Interactions Based on Head Pose and Speech. In *Proc. of the 6th Int'l Conf. on Multimodal Interfaces*, ICMI, 2004.

[83] J. F. Kelley. An Iterative Design Methodology for User-friendly Natural Language Office Information Applications. *ACM Trans. Inf. Syst.*, 2(1):26–41, 1984.

[84] Adam Kendon. Some functions of gaze-direction in social interaction. *Acta Psychologica*, 26:22 – 63, 1967.

[85] Adam Kendon. Goffman's Approach to Face-to-Face Interaction. In P. Drew and A.J. Wootton, editors, *Erving Goffman: Exploring the Interaction Order*. Northeastern University Press, 1988.

[86] Adam Kendon. *Conducting Interaction: Patterns of Behavior in Focused Encounters*. Cambridge Univ. Press, 1990.

[87] S. Kiesler. Fostering common ground in human-robot interaction. In *Robot and Human Interactive Communication, 2005. ROMAN 2005. IEEE International Workshop on*, pages 729–734, 2005.

[88] Rachel Kirby. *Social Robot Navigation*. PhD thesis, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, May 2010.

[89] Nathan Kirchner, Alen Alempijevic, and Gamini Dissanayake. Nonverbal Robot-group Interaction Using an Imitated Gaze Cue. In *Proc. of the 6th Int'l Conference on Human-robot Interaction*, HRI '11, pages 497–504, 2011.

[90] Hiroaki Kitano, Minoru Asada, Yasuo Kuniyoshi, Itsuki Noda, and Eiichi Osawa. RoboCup: The Robot World Cup Initiative. In *Proc. of the First Int'l Conference on Autonomous Agents*, AGENTS '97, pages 340–347, 1997.

[91] Ross A. Knepper, Stefanie Tellex, Adrian Li, Nicholas Roy, and Daniela Rus. Recovering from Failure by Asking for Help. *Auton. Robots*, 39(3):347–362, October 2015.

[92] Jens Kober, J. Andrew Bagnell, and Jan Peters. Reinforcement Learning in Robotics: A Survey. *The Int'l J. of Robotics Research*, 2013.

[93] Levente Kocsis and Csaba Szepesvári. Bandit Based Monte-Carlo Planning. In *ECML'06*, 2006.

[94] Roderick M. Kramer and Marilynn B. Brewer. Effects of group identity on resource use in a simulated commons dilemma. *Journal of Personality and Social Psychology*, 46(5): 1044 – 1057, 1984.

[95] Thibault Kruse, Amit Kumar Pandey, Rachid Alami, and Alexandra Kirsch. Human-aware robot navigation: A survey. *Robotics and Autonomous Systems*, 61(12):1726 – 1743, 2013.

[96] Rainer Kümmerle, Michael Ruhnke, Bastian Steder, Cyrill Stachniss, and Wolfram Burgard. Autonomous Robot Navigation in Highly Populated Pedestrian Zones. *Journal of Field Robotics*, 2014.

[97] Hideaki Kuzuoka, Yuya Suzuki, Jun Yamashita, and Keiichi Yamazaki. Reconfiguring Spatial Formation Arrangement by Robot Body Orientation. In *Proc. of the 5th ACM/IEEE Int'l Conference on Human-robot Interaction*, HRI '10, pages 285–292, 2010.

[98] Sebastian Lang, Marcus Kleinehagenbrock, Sascha Hohenner, Jannik Fritsch, Gernot A. Fink, and Gerhard Sagerer. Providing the Basis for Human-robot-interaction: A Multimodal Attention System for a Mobile Robot. In *Proc. of the 5th Int'l Conf. on Multimodal Interfaces*, ICMI, 2003.

[99] Michael J. V. Leach, Rolf Baxter, Neil M. Robertson, and Ed P. Sparks. Detecting Social Groups in Crowded Surveillance Videos Using Visual Attention. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2014.

[100] J. Lee, J. F. Kiser, A. F. Bobick, and A. L. Thomaz. Vision-based contingency detection. In *2011 6th ACM/IEEE International Conference on Human-Robot Interaction*, HRI '11, pages 297–304, March 2011.

[101] Jill Fain Lehman. Robo Fashion World: A Multimodal Corpus of Multi-child Human-computer Interaction. In *Proc. of the 2014 Workshop on Understanding and Modeling Multiparty, Multimodal Interactions*, UM3I '14, pages 15–20, 2014.

[102] Iolanda Leite, Ginevra Castellano, André Pereira, Carlos Martinho, and Ana Paiva. Modelling Empathic Behaviour in a Robotic Game Companion for Children: An Ethnographic Study in Real-world Settings. In *Proc. of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction*, HRI '12, pages 367–374, 2012.

[103] Iolanda Leite, Ginevra Castellano, AndrÃĺ Pereira, Carlos Martinho, and Ana Paiva. Long-Term Interactions with Empathic Robots: Evaluating Perceived Support in Children. In ShuzhiSam Ge, Oussama Khatib, John-John Cabibihan, Reid Simmons, and Mary-Anne Williams, editors, *Social Robotics*, volume 7621 of *Lecture Notes in Computer Science*, pages 298–307. Springer Berlin Heidelberg, 2012.

[104] Iolanda Leite, Marissa McCoy, Daniel Ullman, Nicole Salomons, and Brian Scassellati. Comparing Models of Disengagement in Individual and Group Interactions. In *Proc. of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, HRI '15, pages 99–105, 2015.

[105] T. Linder and K.O. Arras. Multi-model hypothesis tracking of groups of people in RGB-D data. In *Proc. of the 2014 17th Int'l Conference on Information Fusion (FUSION)*, pages 1–7, 2014.

[106] Alexandru Litoiu, Daniel Ullman, Jason Kim, and Brian Scassellati. Evidence That Robots Trigger a Cheating Detector in Humans. In *Proc. of the Tenth Annual ACM/IEEE Int'l Conference on Human-Robot Interaction*, HRI '15, pages 165–172, 2015.

[107] Chaoran Liu, Carlos T. Ishi, Hiroshi Ishiguro, and Norihiro Hagita. Generation of Nodding, Head Tilting and Eye Gazing for Human-robot Dialogue Interaction. In *Proc. of the Seventh Annual ACM/IEEE Int'l Conference on Human-Robot Interaction*, HRI '12, pages 285–292, 2012.

[108] David V. Lu. *Contextualized Robot Navigation*. PhD thesis, Washington University in St. Louis, December 2014.

[109] M. Luber and K. O. Arras. Multi-Hypothesis Social Grouping and Tracking for Mobile Robots. In *Proc. Robotics: Science and Systems*, 2013.

[110] Subhransu Maji, Lubomir Bourdev, and Jitendra Malik. Action recognition from a distributed representation of pose and appearance. In *Proc. CVPR*, 2011.

[111] Nicolai Marquardt, Ken Hinckley, and Saul Greenberg. Cross-device Interaction via Micro-mobility and F-formations. In *Proc. of the 25th Annual ACM Symposium on User Interface Software and Technology*, UIST '12, pages 13–22, 2012.

[112] Paul Marshall, Yvonne Rogers, and Nadia Pantidi. Using F-formations to Analyse Spatial Patterns of Interaction in Physical Environments. In *Proc. of the ACM 2011 Conference on Computer Supported Cooperative Work*, CSCW '11, pages 445–454, 2011.

[113] E.A. MartÃŋnez-GarcÃŋa, O. Akihisa, and S. Yuta. Crowding and guiding groups of humans by teams of mobile robots. In *Advanced Robotics and its Social Impacts, 2005. IEEE Workshop on*, pages 91–96, June 2005.

[114] Aleksandar Matic, Venet Osmani, and Oscar Mayora-Ibarra. Analysis of Social Interactions Through Mobile Phones. *Mobile Networks and Applications*, 17(6):808–819, 2012.

[115] Yoichi Matsuyama, Iwao Akiba, Shinya Fujie, and Tetsunori Kobayashi. Four-participant group conversation: A facilitation robot controlling engagement density as the fourth participant. *Computer Speech & Language*, 33(1):1 – 24, 2015.

[116] Nikolaos Mavridis. A Review of Verbal and Non-Verbal Human-Robot Interactive Communication. *Robotics and Autonomous Systems*, 63, Part 1:22 – 35, 2015.

[117] Ross Mead and Maja J Matarić. A probabilistic framework for autonomous proxemic control in situated and mobile human-robot interaction. In *Proc. of the 7th ACM/IEEE Int'l Conference on Human-Robot Interaction*, pages 193–194, 2012.

[118] Ross Mead and Maja J. Matarić. Automated Proxemic Feature Extraction and Behavior Recognition: Applications in Human-Robot Interaction. *Int'l Journal of Social Robotics*, 5(3):367–378, 2013. ISSN 1875-4791.

[119] Ross Mead and Maja J. Matarić. Perceptual models of human-robot proxemics. In *Proc. of the Int'l Symposium on Experimental Robotics*, ISER '14, 2014.

[120] Ross Mead, Amin Atrash, and Maja J Matarić. Recognition of Spatial Dynamics for Predicting Social Interaction. In *Proc. of the 6th International Conference on Human-robot interaction*, HRI '11, pages 201–202, 2011.

[121] Marek P. Michalowski, S. Sabanovic, and Reid Simmons. A Spatial Model of Engagement for a Social Robot. In *Proc. of the 9th Int'l Workshop on Advanced Motion Control*, AMC 2006, pages 762–767, 2006.

[122] Masahiro Mori, Karl F MacDorman, and Norri Kageki. The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine*, 19(2):98–100, 2012.

[123] 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 *Proc. of the 17th IEEE Int'l Symposium on Robot and Human Interactive Communication*, RO-MAN '08, pages 532–538, 2008.

[124] Jonathan Mumm and Bilge Mutlu. Human-robot Proxemics: Physical and Psychological Distancing in Human-robot Interaction. In *Proc. of the 6th Int'l Conference on Human-robot Interaction*, HRI '11, pages 331–338. ACM, 2011.

[125] Ryo Murakami, Luis Yoichi Morales Saiki, Satoru Satake, Takayuki Kanda, and Hiroshi Ishiguro. Destination Unknown: Walking Side-by-side Without Knowing the Goal. In *Proc. of the 2014 ACM/IEEE Int'l Conference on Human-robot Interaction*, HRI '14, pages 471–478, 2014.

[126] B. Mutlu, J. Forlizzi, and J. Hodgins. A Storytelling Robot: Modeling and Evaluation of Human-like Gaze Behavior. In *Proc. of the 2006 6th IEEE-RAS International Conference on Humanoid Robots*, pages 518–523, 2006.

[127] Bilge Mutlu and Jodi Forlizzi. Robots in organizations: the role of workflow, social, and environmental factors in human-robot interaction. In *2008 3rd ACM/IEEE International Conference on Human-Robot Interaction*, HRI '08, pages 287–294. IEEE, 2008.

[128] Bilge Mutlu, Toshiyuki Shiwa, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. Footing in human-robot conversations: How robots might shape participant roles using gaze cues. In *Proc. of the 4th ACM/IEEE Int'l Conference on Human-Robot Interaction*, HRI 09, pages 61–68, March 2009.

[129] Y. Nakauchi and R. Simmons. A social robot that stands in line. In *Proc. of the 2000 IEEE/RSJ International Conference on Intelligent Robots and Systems*, IROS '00, pages 357–364, 2000.

[130] M. A. Nicolaou, H. Gunes, and M. Pantic. Continuous Prediction of Spontaneous Affect from Multiple Cues and Modalities in Valence-Arousal Space. *IEEE Trans. on Affective Computing*, 2(2):92–105, 2011.

[131] International Federation of Robotics. Executive Summary. World Robotics 2016 Service Robots, October, 2016. URL `https://ifr.org/downloads/press/02_2016/Executive_Summary_Service_Robots_2016.pdf`.

[132] D. Olguín Olguín, B. N. Waber, T. Kim, A. Mohan, K. Ara, and A. Pentland. Sensible Organizations: Technology and Methodology for Automatically Measuring Organizational Behavior. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 39(1):43–55, 2009.

[133] H. Osawa, J. Orszulak, K.M. Godfrey, and J.F. Coughlin. Maintaining learning motivation of older people by combining household appliance with a communication robot. In *Proc. of the 2010 IEEE/RSJ Int'l Conference on Intelligent Robots and Systems*, IROS '10, 2010.

[134] Hyun S. Park and Jianbo Shi. Social saliency prediction. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4777–4785, 2015.

[135] Hyun S. Park, Eakta Jain, and Yaser Sheikh. 3D Social Saliency from Head-Mounted Cameras. In *Advances in Neural Information Processing Systems*, NIPS '12, 2012.

[136] Hyun S. Park, Eakta Jain, and Yaser Sheikh. Predicting Primary Gaze Behavior Using Social Saliency Fields. In *Proc. of the 2013 IEEE Int'l Conference on Computer Vision*, ICCV 2013, pages 3503–3510, 2013.

[137] Christopher Parlitz, Martin Hägele, Peter Klein, Jan Seifert, and Kerstin Dautenhahn. Care-obot 3-rationale for human-robot interaction design. In *Proc. of 39th Int'l Symposium on Robotics*, ISR '08, 2008.

[138] H. D. Patterson and Robin Thompson. Maximum likelihood estimation of components of variance. In *Proc. of the Eighth International Conf. on Biochem.*, 1974.

[139] Jamie Pearson, Jiang Hu, Holly P. Branigan, Martin J. Pickering, and Clifford I. Nass. Adaptive Language Behavior in HCI: How Expectations and Beliefs About a System Affect Users' Word Choice. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '06, pages 1177–1180, 2006.

[140] 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 *ICRA'09 Workshop on Open Source Software*, 2009.

[141] Irene Rae. Using Robot-mediated Communication to Improve Remote Collaboration. In *Proc. of CHI '13 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '13, pages 1953–1956, 2013.

[142] Nimrod Raiman, Hayley Hung, and Gwenn Englebienne. Move, and I Will Tell You Who You Are: Detecting Deceptive Roles in Low-quality Data. In *Proc. of the 13th Int'l Conference on Multimodal Interfaces*, ICMI '11, pages 201–204, 2011.

[143] Omar Adair Islas Ramírez, Giovanna Varni, Mihai Andries, Mohamed Chetouani, and Raja Chatila. Modeling the dynamics of individual behaviors for group detection in crowds using low-level features. In *Proc. of the 2016 IEEE Int'l Symposium on Robot and Human Interactive Communication*, RO-MAN, 2016.

[144] Elisa Ricci, Jagannadan Varadarajan, Ramanathan Subramanian, Samuel Rota Bulo, Narendra Ahuja, and Oswald Lanz. Uncovering interactions and interactors: Joint estimation of head, body orientation and f-formations from surveillance videos. In *Proc. of the IEEE International Conference on Computer Vision*, ICCV'15, pages 4660–4668, 2015.

[145] Charles Rich, Brett Ponsler, Aaron Holroyd, and Candace L. Sidner. Recognizing engagement in human-robot interaction. In *Proc. of the 5th ACM/IEEE International Conference on Human-Robot Interaction*, HRI'10, 2010.

[146] Richard S. Sutton. Generalization in reinforcement learning: Successful examples using sparse coarse coding. *Advances in Neural Information Processing Systems*, pages 1038–1044, 1996.

[147] J. Rios-Martinez, A. Spalanzani, and C. Laugier. Understanding human interaction for probabilistic autonomous navigation using Risk-RRT approach. In *Proc. of the 2011 IEEE/RSJ Int'l Conference on Intelligent Robots and Systems*, IROS 2011, pages 2014–2019, 2011.

[148] Ben Robins, Kerstin Dautenhahn, Chrystopher L. Nehaniv, N. Assif Mirza, Dorothée François, and Lars Olsson. Sustaining interaction dynamics and engagement in dyadic child-robot interaction kinesics: lessons learnt from an exploratory study. In *Proc. of the IEEE International Workshop on Robot and Human Interactive Communication*, ROMAN '05, 2005.

[149] Stephanie Rosenthal and Manuela Veloso. Mobile Robot Planning to Seek Help with Spatially-situated Tasks. In *Proc. of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, AAAI'12, pages 2067–2073, 2012.

[150] Vincent Rousseau, Francois Ferland, Dominic Létourneau, and Francois Michaud. Sorry to interrupt, but may I have your attention? Preliminary design and evaluation of autonomous engagement in HRI. *Journal of Human-Robot Interaction*, 2(3):41–61, 2013.

[151] Kerstin Ruhland, Sean Andrist, Jeremy Badler, Christopher Peters, Norman Badler, Michael Gleicher, Bilge Mutlu, and Rachel Mcdonnell. Look me in the eyes: A survey of eye and gaze animation for virtual agents and artificial systems. In *Eurographics 2014 - State of the Art Reports*, 2014.

[152] M. Saerbeck and C. Bartneck. Perception of affect elicited by robot motion. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction*, HRI '10, pages 53–60, 2010.

[153] S. Satake, T. Kanda, D.F. Glas, M. Imai, H. Ishiguro, and N. Hagita. How to approach humans?-strategies for social robots to initiate interaction. In *Proc. of the 4th ACM/IEEE Int'l Conference on Human-Robot Interaction*, HRI '09, pages 109–116, 2009.

[154] L. Scandolo and T. Fraichard. An anthropomorphic navigation scheme for dynamic scenarios. In *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, pages 809–814, 2011.

[155] Francesco Setti, Oswald Lanz, Roberta Ferrario, Vittorio Murino, and Marco Cristani. Multi-Scale F-Formation Discovery for Group Detection. In *Proc. 20th IEEE Int'l Conference on Image Processing*, ICIP '13, 2013.

[156] Francesco Setti, Chris Russell, Chiara Bassetti, and Marco Cristani. F-formation Detection: Individuating Free-standing Conversational Groups in Images. *CoRR*, abs/1409.2702, 2014.

[157] Chao Shi, Michihiro Shimada, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. Spatial Formation Model for Initiating Conversation. In *Proc. of Robotics: Science and Systems*, RSS '11, 2011.

[158] Masahiro Shiomi, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. Interactive Humanoid Robots for a Science Museum. *IEEE Intelligent Systems*, 22(2):25–32, 2007.

[159] Masahiro Shiomi, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. A Larger Audience, Please!: Encouraging People to Listen to a Guide Robot. In *Proc. of the 5th ACM/IEEE Int'l Conference on Human-robot Interaction*, HRI '10, pages 31–38, 2010.

[160] Masahiro Shiomi, Daisuke Sakamoto, Takayuki Kanda, CarlosToshinori Ishi, Hiroshi Ishiguro, and Norihiro Hagita. Field Trial of a Networked Robot at a Train Station. *International Journal of Social Robotics*, 3(1):27–40, 2011.

[161] E. Short, J. Hart, M. Vu, and B. Scassellati. No fair!! An interaction with a cheating robot. In *Proc. of the Fifth Annual ACM/IEEE Int'l Conference on Human-Robot Interaction*, HRI 2010, pages 219–226, March 2010.

[162] Candace L. Sidner, Christopher Lee, Louis-Philippe Morency, and Clifton Forlines. The Effect of Head-nod Recognition in Human-Robot Conversation. In *Proc. of the 1st ACM Conf. on Human-Robot Interaction*, HRI, 2006.

[163] C.L. Sidner, C. Lee, C.D. Kidds, N. Lesh, and C. Rich. Explorations in Engagement for Humans and Robots. *Artificial Intelligence*, 2005.

[164] David Silver, Richard S. Sutton, and Martin Müller. Temporal-difference search in computer Go. *Machine Learning*, 87(2):183–219, 2012.

[165] David Sirkin, Brian Mok, Stephen Yang, and Wendy Ju. Mechanical Ottoman: How Robotic Furniture Offers and Withdraws Support. In *Proc. of the Tenth Annual ACM/IEEE Int'l Conference on Human-Robot Interaction*, HRI '15, 2015.

[166] G. Skantze, A. Hjalmarsson, and C. Oertel. Exploring the effects of gaze and pauses in situated human-robot interaction. In *Proc. of the 14th Annual Meeting of Special Interest Group on Doscourse and Dialogue*, SIGDial, 2013.

[167] Han Sloetjes and Peter Wittenburg. Annotation by category - ELAN and ISO DCR. In *Proc. of the Sixth Int'l Conference on Language Resources and Evaluation*, LREC '08, 2008.

[168] Maria Staudte and Ulrich Pfeiffer. When eye see you: Gaze and joint attention in human interaction. In *Proc. of CogSci 2013*, 2013.

[169] Aaron Steinfeld, Odest Chadwicke Jenkins, and Brian Scassellati. The Oz of Wizard: Simulating the human for interaction research. In *Proc. of the 4th ACM/IEEE Int'l Conference on Human Robot Interaction*, HRI '09, 2009.

[170] M. Stepp. *Quick Guide to Writing Fiction*. Old American Publishing, 2012.

[171] J urgen Streeck. Gesture as communication I: Its coordination with gaze and speech. *Communication Monographs*, 60(4):275–299, 1993.

[172] Walter W. Stroup. *Generalized linear mixed models: modern concepts, methods and applications*. CRC press, 2012.

[173] Richard S. Sutton. Tile Coding Software. `http://rlai.cs.ualberta.ca/RLAI/RLtoolkit/tiles.html`. Online; accessed Feb. 2016.

[174] Richard S. Sutton and Andrew G. Barto. *"Reinforcement Learning: An Introduction"*. The MIT Press, 2015. (second edition, in progress).

[175] Richard Stuart Sutton. *Temporal Credit Assignment in Reinforcement Learning*. PhD thesis, University of Massachusetts Amherst, 1984.

[176] M. Svenstrup, T. Bak, and H.J. Andersen. Trajectory planning for robots in dynamic human environments. In *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*, pages 4293–4298, 2010.

[177] Daniel Szafir and Bilge Mutlu. Pay Attention!: Designing Adaptive Agents That Monitor and Improve User Engagement. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 11–20, 2012.

[178] Daniel Szafir, Bilge Mutlu, and Terrence Fong. Communication of Intent in Assistive Free Flyers. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-robot Interaction*, HRI '14, pages 358–365, 2014.

[179] Leila Takayama and C. Pantofaru. Influences on proxemic behaviors in human-robot interaction. In *Proc. of the 2009 IEEE/RSJ Int'l Conference on Intelligent Robots and Systems*, IROS 2009, pages 5495–5502, 2009.

[180] Leila Takayama, Doug Dooley, and Wendy Ju. Expressing thought: improving robot readability with animation principles. In *Proc. of the 6th International Conference on Human-Robot Interaction*, HRI '11, pages 69–76, 2011.

[181] Frank Thomas and Ollie Johnston. *The illusion of life: Disney animation.* Hyperion, 1995.

[182] Andrea Thomaz, Guy Hoffman, and Maya Cakmak. Computational Human-Robot Interaction. *Foundations and TrendsÂő in Robotics*, 4(2-3):105–223, 2016. ISSN 1935-8253. doi: 10.1561/2300000049.

[183] S. Thrun, M. Bennewitz, W. Burgard, A.B. Cremers, F. Dellaert, D. Fox, D. Hähnel, C. Rosenberg, N. Roy, J. Schulte, and D. Schulz. MINERVA: A second generation mobile tour-guide robot. In *Proc. of the IEEE International Conference on Robotics and Automation*, ICRA '99, 1999.

[184] S. Thrun, W. Burgard, and D. Fox. *Probabilistic Robotics.* Intelligent robotics and autonomous agents. MIT Press, 2005.

[185] Joshua Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain Randomization for Transferring Deep Neural Networks from Simulation to the Real World. *CoRR*, abs/1703.06907, 2017.

[186] J. G. Trafton, M. D. Bugajska, B. R. Fransen, and R. M. Ratwani. Integrating vision and audition within a cognitive architecture to track conversations. In *Proc. of the 2008 3rd ACM/IEEE Int'l Conf. on Human-Robot Interaction*, HRI, 2008.

[187] Khai N. Tran, Apurva Bedagkar-Gala, Ioannis A Kakadiaris, and Shishir K. Shah. Social Cues in Group Formation and Local Interactions for Collective Activity Analysis. In *Proc. of the 8th Int'l Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications.*, VISAPP '13, pages 539–548, 2013.

[188] S. Tranberg Hansen, M. Svenstrup, H.J. Andersen, and T. Bak. Adaptive human aware navigation based on motion pattern analysis. In *Proc. of the 18th IEEE International Symposium on Robot and Human Interactive Communication*, RO-MAN '09, pages 927–932, 2009.

[189] Rudolph Triebel, Kai Arras, Rachid Alami, Lucas Beyer, Stefan Breuers, Raja Chatila, Mohamed Chetouani, Daniel Cremers, Vanessa Evers, Michelangelo Fiore, Hayley Hung, Omar A. Islas RamÃŋrez, Michiel Joosse, Harmish Khambhaita, Tomasz Kucner, Bastian Leibe, Achim J. Lilienthal, Timm Linder, Manja Lohse, Martin Magnusson, Billy Okal, Luigi Palmieri, Umer Rafi, Marieke van Rooij, and Lu Zhang. SPENCER: A Socially Aware Service Robot for Passenger Guidance and Help in Busy Airports. In *Proc. of the 10th Conference on Field and Service Robotics*, FSR '15, 2015.

[190] AJN Van Breemen. Bringing robots to life: Applying principles of animation to robots. In *Proc. of Shapping Human-Robot Interaction Workshop held at CHI 2004*, pages 143–144, 2004.

[191] H. F. M. Van der Loos, J. J. Wagner, N. Smaby, K. Chang, O. Madrigal, L. J. Leifer, and O. Khatib. ProVAR assistive robot system architecture. In *Proc. of the 1999 IEEE Int'l Conference on Robotics and Automation*, volume 1 of *ICRA '99*, pages 741–746 vol.1, 1999.

[192] Tim van Oosterhout and Arnoud Visser. A visual method for robot proxemics measurements. In *Proc. Metrics for Human-Robot Interaction: A Workshop at HRI 2008*, 2008.

[193] Sebastiano Vascon, Eyasu Zemene Mequanint, Marco Cristani, Hayley Hung, Marcello Pelillo, and Vittorio Murino. A game-theoretic probabilistic approach for detecting conversational groups. In *Proc. of the Asian Conference on Computer Vision*, ACCV '14, pages 658–675, 2014.

[194] M. Vázquez, A. Steinfeld, and S. E. Hudson. Maintaining Awareness of the Focus of Attention of a Conversation: A Robot-Centric Reinforcement Learning Approach. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication*, RO-MAN '16, pages 36–43, 2016.

[195] M. Vázquez, E. J. Carter, J. Forlizzi, S. E. Hudson, and A. Steinfeld. Methods for Studying Group Interactions in HRI. In *Robots In Groups and Teams - A CSCW 2017 Workshop*, 2017.

[196] M. Vázquez, E. J. Carter, B. McDorman, J. Forlizzi, A. Steinfeld, and S. E. Hudson. Towards Robot Autonomy in Group Conversations: Understanding the Effects of Body Orientation and Gaze. In *Proc. of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '17, 2017.

[197] Marynel Vázquez, Aaron Steinfeld, Scott E. Hudson, and Jodi Forlizzi. Spatial and Other Social Engagement Cues in a Child-robot Interaction: Effects of a Sidekick. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-robot Interaction*, HRI '14, pages 391–398, 2014.

[198] Marynel Vázquez, Elizabeth J. Carter, Jo Ana Vaz, Jodi Forlizzi, Aaron Steinfeld, and Scott E. Hudson. Social Group Interactions in a Role-Playing Game. In *Proc. of the Tenth Annual ACM/IEEE Int'l Conference on Human-Robot Interaction Extended Abstracts*, HRI'15 Extended Abstracts, pages 9–10, 2015.

[199] Marynel Vázquez, Aaron Steinfeld, and Scott E. Hudson. Parallel Detection of Conversational Groups of Free-Standing People and Tracking of their Lower-Body Orientation. In *Proc. of the 2015 IEEE/RSJ Int'l Conference on Intelligent Robots and Systems*, IROS 2015, 2015.

[200] Manuela M. Veloso, Joydeep Biswas, Brian Coltin, and Stephanie Rosenthal. CoBots: Robust Symbiotic Autonomous Mobile Service Robots. In *Proc. of the Twenty-Fourth International Joint Conference on Artificial Intelligence*, IJCAI 2015, 2015.

[201] Arun Venkatraman, Martial Hebert , and J. Andrew (Drew) Bagnell. Improving Multi-step Prediction of Learned Time Series Models. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, 2015.

[202] Arun Venkatraman, Wen Sun, Martial Hebert , Byron Boots, and J. Andrew (Drew) Bagnell. Inference Machines for Nonparametric Filter Learning. In *25th International Joint Conference on Artificial Intelligence (IJCAI-16)*, July 2016.

[203] J. Vroon, M. Joosse, M. Lohse, J. Kolkmeier, Jaebok Kim, K. Truong, G. Englebienne, D. Heylen, and V. Evers. Dynamics of social positioning patterns in group-robot interactions. In *Proc. of the 2015 24th IEEE Int'l Symposium on Robot and Human Interactive Communication*, RO-MAN, 2015.

[204] J. J. Wagner, H. F. M. Van der Loos, and L. J. Leifer. Construction of Social Relationships Between User and Robot. *Robotics and Autonomous Systems*, 31(3):185 – 191, 2000.

[205] Michael L. Walters, Kerstin Dautenhahn, René te Boekhorst, Kheng Lee Koay, Christina Kaouri, Sarah Woods, Chrystopher Nehaniv, David Lee, and Iain Werry. The influence of subjects' personality traits on personal spatial zones in a human-robot interaction experiment. In *Proc. RO-MAN*, 2005.

[206] Michael L Walters, Kerstin Dautenhahn, René Te Boekhorst, Kheng Lee Koay, Dag Sverre Syrdal, and Chrystopher L Nehaniv. An empirical framework for human-robot proxemics. *Procs of New Frontiers in Human-Robot Interaction*, 2009.

[207] M.L. Walters, K. Dautenhahn, K.L. Koay, C. Kaouri, R. Boekhorst, C. Nehaniv, I. Werry, and D. Lee. Close encounters: spatial distances between people and a robot of mechanistic appearance. In *Proc. of the 5th IEEE-RAS International Conference on Humanoid Robots*, pages 450–455, 2005.

[208] M.L. Walters, D.S. Syrdal, K.L. Koay, K. Dautenhahn, and R. te Boekhorst. Human approach distances to a mechanical-looking robot with different robot voice styles. In *Proc. of the 17th IEEE Int'l Symposium on Robot and Human Interactive Communication*, RO-MAN '08, pages 707–712, 2008.

[209] QianYing Wang, Clifford Nass, and Jiang Hu. Natural Language Query vs. Keyword Search: Effects of Task Complexity on Search Performance, Participant Perceptions, and Preferences. In *Proc. of the 2005 IFIP TC13 International Conference on Human-Computer Interaction*, INTERACT'05, 2005.

[210] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional Pose Machines. In *The IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '16, 2016.

[211] David Whitney, Eric Rosen, James MacGlashan, Lawson Wong, and Stefanie Tellex. Reducing Errors in Object-Fetching Interactions through Social Feedback. In *Int'l Conference on Robotics and Automation*, 2017.

[212] Kipling D. Williams, Christopher K.T. Cheung, and Wilma Choi. Cyberostracism: effects of being ignored over the Internet. *Journal of personality and social psychology*, 79(5): 748, 2000.

[213] Danny Wyatt, Tanzeem Choudhury, and Jeff Bilmes. Conversation Detection and Speaker Segmentation in Privacy-Sensitive Situated Speech Data. In *Proc. of the Eighth Annual Conference of the Int'l Speech Communication Association*, 2007.

[214] Min Xin and Ehud Sharlin. Playing games with robots – A method for Evaluating Human-Robot Interaction. In Nilanjan Sarkar, editor, *Human Robot Interaction*. INTECH Open Access Publisher, 2007.

[215] Yuto Yamaji, Taisuke Miyake, Yuta Yoshiike, P.RavindraS. Silva, and Michio Okada. STB: Child-Dependent Sociable Trash Box. *Intl' J. Soc. Robotics*, 3(4):359–370, 2011.

[216] Y. Yang, S. Baker, A. Kannan, and D. Ramanan. Recognizing proxemics in personal photos. In *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[217] Y. Yoshikawa, K. Shinozawa, H. Ishiguro, N. Hagita, and T. Miyamoto. Responsive robot gaze to interaction partner. In *Proc. of Robotics: Science and Systems*, R:SS, 2006.

[218] Mohammad Abu Yousuf. *Mobile Museum Guide Robots Able to Create Spatial Formations with Multiple Visitors*. PhD thesis, Saitama University, Saitama, Japan, 9 2013.

[219] MohammadAbu Yousuf, Yoshinori Kobayashi, Yoshinori Kuno, Akiko Yamazaki, and Keiichi Yamazaki. Development of a Mobile Museum Guide Robot That Can Configure Spatial Formation with Visitors. In De-Shuang Huang, Changjun Jiang, Vitoantonio Bevilacqua, and JuanCarlos Figueroa, editors, *Intelligent Computing Technology*, volume 7389 of *Lecture Notes in Computer Science*, pages 423–432. Springer Berlin Heidelberg, 2012.

[220] T. Yu, S.-N. Lim, K. Patwardhan, and N. Krahnstoever. Monitoring, Recognizing and Discovering Social Networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '09, 2009.

[221] Zerrin Yumak, Jianfeng Ren, Nadia Magnenat Thalmann, and Junsong Yuan. Tracking and Fusion for Multiparty Interaction with a Virtual Character and a Social Robot. In *SIGGRAPH Asia 2014*, SA, pages 3:1–3:7, 2014.

[222] Gloria Zen, Bruno Lepri, Elisa Ricci, and Oswald Lanz. Space Speaks: Towards Socially and Personality Aware Visual Surveillance. In *Proc. of the 1st ACM International Workshop on Multimodal Pervasive Video Analysis*, MPVA '10, pages 37–42, 2010.

[223] Lina Zhou, Dongsong Zhang, and Yu-wei Sung. The effects of group factors on deception detection performance. *Small Group Research*, 44(3):272–297, June 2013.

[224] Yuke Zhu, Roozbeh Mottaghi, Eric Kolve, Joseph J Lim, Abhinav Gupta, Li Fei-Fei, and Ali Farhadi. Target-driven visual navigation in indoor scenes using deep reinforcement learning. In *Proc. Int'l Conference on Robotics and Automation*, ICRA, 2017.

[225] D. A. Zimmerman. *Comic Book Character: Unleashing the Hero in Us All*. InterVarsity Press, 2004.