A Study in Scene Shaping: Adjusting F-formations in the Wild

Dan Bohus, Sean Andrist, Eric Horvitz

Microsoft Research
Redmond, WA
{dbohus, sandrist, horvitz}@microsoft.com

Abstract

We study the automated shaping of F-formations in the proximity of a stationary robot that has been deployed to provide directions within a building. We introduce the notion of *scene shaping* where suboptimal spatial configurations are detected and desired shifts in the locations of participants and bystanders are communicated with natural language and gestures. We conduct an initial in-the-wild study with the proposed methods and report results, lessons learned, and future directions of research.

Introduction

In the open world, perceptual and decision-making challenges are exacerbated by unexpected events and conditions. Beyond sensing challenges, a robot must grapple with varying configurations of groups of people in its proximity. Some people approaching closely may not be interested in engaging with the robot while others may come near with an intention to interact, either alone or in a large group. Levels of attention, engagement and spatial orientation vary over time, resulting in continuously evolving configurations.

We seek to increase the robustness of interaction by detecting classes of problematic situations and then engaging participants *in situ* to make shifts that are expected to increase robustness. For instance, if a group of people is too far away, a robot might request that one or two individuals step closer while others remain where they are. As another example, if the robot detects side conversation, it might ask the participants to take turns in a more structured manner.

We focus on *spatial* shaping and study the automated shaping of *F-formations*—the spatial configurations that people assume while interacting with each other (Kendon, 1990). We endow a robot with the ability to detect suboptimal spatial configurations and study the use of gestures and utterances together to seek desired shifts in the locations of participants and bystanders. We study the proposed methods with a stationary robot that gives directions in a building.

Related Work

Proxemics (Hall, 1966) plays a fundamental role in the organization of interactions. Previous research has shown that participants in a conversation typically organize themselves such as to "sustain a spatial and orientational relationship in which the space between them is one to which they have equal, direct, and exclusive access" (Kendon, 1990). This spatial organization is referred to as an *F-formation*.

Considerations of space and orientation have previously received attention in the HRI community. For example, Dautenhahn et al. (2006) investigated how a robot should best approach and place itself relative to a seated human subject and found that most subjects disliked a frontal approach. In a study of preferred approach distances (Walters et al., 2005), a large subset of subjects took up positions that were significantly closer than human-human interaction preferences. Others have focused on designing polite approach behaviors for service robots in public settings (Kato et al., 2015), and developed models for the appropriate timing and position to initiate a conversation (Shi et al., 2011).

Mead and Matarić (2015) examined how users adapt their proxemic preferences after observing differences in robot performance in different spatial configurations. Users changed their preferences for how close to stand to the robot based on the perceived location of peak robot performance, but this perception usually underestimated the true distance. The authors noted that this process might be expedited by the robot communicating its own proxemic preferences based on predicted errors. Kuzuoka et al. (2010) conducted a similar study in implicitly reconfiguring spatial formation arrangements by changing the orientation of the robot.

In contrast to the previous work, we focus on a notion of *explicit scene shaping* anchored in dialog. The robot is stationary, but makes explicit requests via language to modify or "shape" the F-formation.

Robot System

The *Directions Robot* is a humanoid Nao robot system that can engage in spoken dialog with one or multiple users and give directions to various locations inside a building. The robot, described in more detail in (Bohus, Saw and Horvitz, 2014), is deployed in an open space in front of a bank of three elevators in our building (see Figure 1.) Traffic in this space includes people coming in and out of the elevators and traversing the corridor. People engaging with the robot include some who are familiar with the robot, as well as visitors who may encounter the robot for the first time. People are free to interact with the robot at will. The robot uses a face tracker and models of multiparty engagement and turn-taking, speech recognition, and dialog planning to engage in conversation. It describes directions in natural language with coordinated gaze, gestures and speech.

Spatial Scene Shaping

The spatial scene-shaping approach that we discuss in this paper aims to enable the robot to refine the locations of participants and bystanders towards configurations that are more desirable. The high-level goal is to have one, two, or three participants engaged with the robot at a natural distance in front of the robot, with all bystanders further away, in the back. At a high level, the scene-shaping model operates in a sense-think-act-monitor loop, as follows:

(1) *Detect* deviations from desirable configurations.
(2) *Decide* upon a corresponding scene-shaping action that encourages people to modify their configuration towards what the robot considers desirable.
(3) *Execute* the scene-shaping action.
(4) *Monitor* if people comply and provide feedback that conveys the robot's perception of the user's compliance or noncompliance.

For instance, in the example illustrated in Figure 1, the robot detects that one of the two engaged participants is too close and triggers a scene shaping action: "*Sorry, but if you don't mind taking one step back, it may be easier [... for me to interact with both of you]*." The robot continuously monitors the participants' positions throughout the production of this action and for a short period of time afterwards. As the participant moves back and the problem is resolved, the robot self-interrupts (after "*it may be easier*") and provides feedback on the move, after which it resumes the dialog.

Desirable F-formations

We define a set of desirable F-formation configurations based on prior experience and some observed failure modes from the deployment of the robot. While the problematic scenes and fixes are specific to our current capabilities and design, the methodology is applicable to other situations.
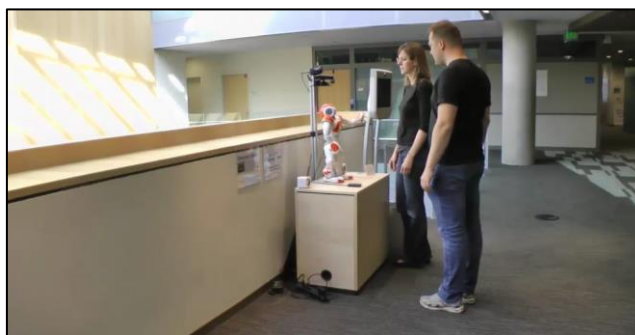
P_{120}, P_{119}: [approach and engage with the robot]
P_{121}:   [watching from a distance]
Robot: **Hi there! Do you need directions?**
P_{120}:   [is standing very close to the robot]
P_{119}:   Yes
Robot: **Sorry, but if you don't mind taking one step back, it may be easier** [*... for me to interact with both of you*].
P_{120}:   [takes a step back]
Robot: **Great, thank you! Now let's see... Can I help you find something?**
P_{119}:   3800.
Robot: **To get to 3800, go to the end of that hallway, turn right and continue on down the hallway for a bit. 3800 will be the first room on your right.**

Figure 1. *Directions Robot deployment setup and sample interaction.*

Specifically, for Directions Robot, we set requirements on the numbers and the placement of engaged participants and bystanders. The robot determines who is *engaged* via a heuristic rule that combines information about the proximity, stability, and attention of each participant (Bohus and Horvitz, 2014). *Bystanders* are participants who are not engaged, but who are perceived to be stationary for a certain duration. To reason about placement, we define a set of *proxemic zones* with respect to the robot (see Figure 2.) These zones are related to Hall's (1966) proxemic distances. The robot determines the proxemic zone of a participant by comparing the tracked face size and location against preset thresholds (roughly corresponding to a distance of 3ft, 7ft, 15ft) and using a hysteresis approach to manage transitions.

When interacting, the desirable configuration for the robot involves up to a maximum of three engaged participants in the *Frontal* and *Natural* zones, and any number of bystanders in the *Far* zone, as portrayed in the bottom-left side of Figure 2. We have observed that when participants get too close to the robot, the vision system may stop tracking them correctly. Alternatively, if engaged participants move farther away from the robot, towards the *Transitional* zone, they may be dropped from the interaction as the engagement model assesses they are no longer engaged. Ideally, we would like engaged participants at a natural engagement distance and in front of the robot. In addition, we prefer that any bystanders remain in the *Far* area. A closer position may lead to fluctuations in the set of participants inferred by the
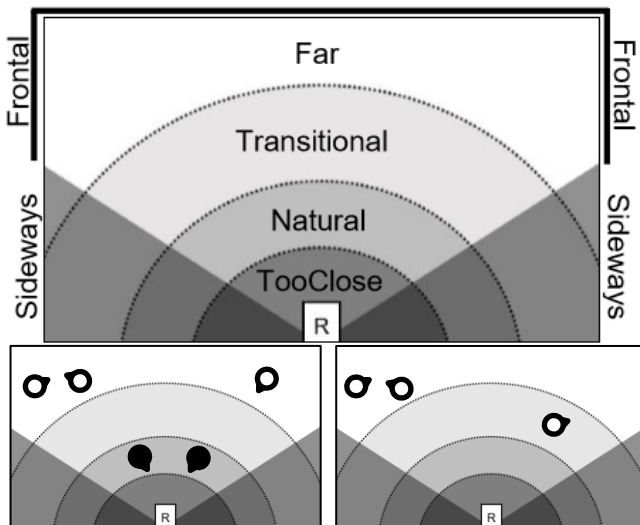
Figure 2. Proxemic zones (top) and sample ideal configurations in interactions (bottom-left) and outside interactions (bottom-right).

robot as being engaged. Furthermore, their presence in the *Transitional* zone may increase the number of confusing side interactions between them and the engaged participants.

When the robot is not engaged in an interaction, desirable configurations are those where all bystanders are in the *Far* or *Transitional* zones and where they are not continuously attending to the robot (see bottom right in Figure 2). Non-engaged bystanders in proximity of the robot are undesirable because their hovering close to the robot may prevent others from engaging. Situations with bystanders at a distance but attending to the robot were also deemed undesirable as they might represent missed opportunities for engagement.

Scene-shaping Actions and Policy

Table 1 describes the seven scene-shaping actions we have defined and illustrates a subset of their possible realizations. Each action has a precondition which defines when it may be triggered, a set of addressees which defines whom the action is directed towards, and a success condition which is used during to assess compliance and determine feedback.

The first three shaping actions, *MoveFurther*, *MoveCloser*, and *MoveToCenter* aim to address problems where an engaged participant is either too close, too far, or too much to the side. The *LargeGroupArrange* action was designed to deal in a single shot with complex situations involving large groups of people near the robot. The *InviteToJoin* action covers the case when a bystander has been detected while the robot was interacting in the Natural or Transitional zone and aims to get the bystander to either join in or move further back into the Far zone. Finally, the *AnnouncePresence* and *InviteToEngage* actions are taken when the robot is not already in an interaction and bystanders are detected: the robot aims to influence people either to engage or clear the area in its immediate proximity.

The shaping actions are rendered as coordinated speech, gaze, and arm gestures, contextualized based on the location of the addressee. For instance, the *MoveFurther* action is accompanied by a "move back" arm gesture when the robot says, "one step back," performed with the arm that is on the side of the addressed participant. The robot's turn-taking model directs the robot's gaze towards the addressees. The lexical form of the action is also contextualized based on the number of addressees and engaged participants, as well as the recent history of shaping actions (Table 1).

The robot continuously monitors each shaping action to determine if the success condition is met during the production of the action or during the floor release that follows. If the success condition is met while rendering the action, the robot inserts a turn, interrupting itself if it is speaking, and provides feedback about the success of the adjustment. Otherwise, the robot waits until the next naturally occurring turn and provides feedback after the user speaks or after a brief pause of two seconds if no one speaks. The feedback depends on an automated assessment of the success of the shaping action and the inferred need for further actions.

Throughout the interaction, the robot continuously monitors the scene and computes at every frame the actions that may be performed. Action selection operates as follows: first, all actions that have their preconditions satisfied are added to a priority queue in the order: InviteToEngage, AnnouncePresence (while not interacting), and LargeGroupArrange, MoveFurther, MoveCloser, MoveToCenter, InviteToJoin (while the robot is in an active interaction). The priority queue is traversed and the system checks a list of additional turn-taking constraints that further refine whether an action might be smoothly executed. The first action that passes this check is selected. In addition, we enforce certain temporal constraints to prevent the robot from taking too many successive scene-shaping actions, as multiple actions within a session may be unnatural and significantly disrupt the flow of the base-level directions-giving task.

Field Study and Annotations

We deployed the proposed scene-shaping policies on the Directions Robot and conducted a three-week field study. An expert annotator reviewed videos of the interactions (from the robot's viewpoint) and assessed each scene-shaping action in the overall context of the situation on a scale from 1 (bad) to 5 (good). The annotator also provided a succinct textual description explaining the rationale for each numerical assessment. In addition, the annotator assessed whether anyone in the scene responded to the robot's action, and if so, whether the response complied with the robot's shaping request. Finally, the annotator provided a textual description of the users' responses. Since the annotator viewed the scene from the robot's camera, to give the annotator a sense for users' views of the shaping actions, we created a set of videos illustrating each of the scene-shaping gestures with a
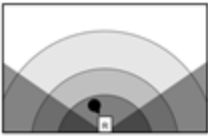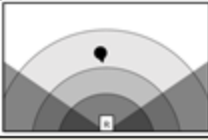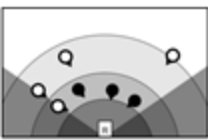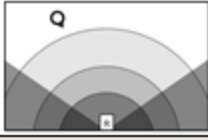
	MoveFurther		
	1. Sorry, but if you guys don't mind taking one step back, it may be easier for me to interact with both of you. 2. Sorry now I think you're a bit too close. If you take one step back it may be easier for me to interact with you. 3. Sorry to ask again but if you could also take just one step back, it may be easier for me to interact with you.		
	Precondition:	There is an engaged participant in the TooClose zone	
	Addressees:	All engaged participants in the TooClose zone	
	Success:	All addressees are engaged but not in TooClose zone	
	MoveCloser		
	1. Sorry, but if you don't mind coming just a little bit closer, it may be easier for me to interact with you.		
	Precondition:	There is an engaged participant in the Transitional or Far zone	
	Addressees:	All engaged participants in the Transitional or Far zone	
	Success:	All addressees are engaged but not in Transitional or Far areas	
	MoveToCenter		
	1. Sorry, but if you don't mind stepping a little bit to the middle, it may be easier for me to interact with you.		
	Precondition:	There is an engaged participant in the Sideways zone	
	Addressees:	All engaged participants in the Sideways zone	
	Success:	All addressees are engaged but not in the Sideways zone	
	LargeGroupArrange		
	1. Sorry, but it might make things easier for me if one or two of you stay up here and the others step way back. 2. Wow, you guys are a big group. It might make things easier for me if one or two of you stay up here and the others step way back. 3. Guys, I can try to keep going this way, but it might make things easier for me if one or two of you stay up here and the others step waaaaay back.		
	Precondition:	The *chaos* condition (see text) holds for at least 4 seconds.	
	Addressees:	All participants in the Close, Natural and Transitional zones.	
	Success:	The negation of the *chaos* condition holds for at least 0.5 seconds.	
	InviteToJoin		
	1. If you over there either step forward and join in, or step back a bit, it would help me a lot.		
	Precondition:	There is has been only one engaged participant and there is one bystander in the Transitional, Natural or TooClose zones.	
	Addressees:	All bystanders in the Transitional, Natural or TooClose zones.	
	Success:	The invited bystander is either engaged, or in the Far zone.	
	InviteToEngage		
	1. I'm here if you need directions. Please feel free to step forward and ask me something if you want.		
	Precondition:	A bystander is attending to the robot for at least 4 seconds.	
	Addressees:	Any bystanders that have been attending to the robot for at least 2 seconds.	
	Success:	Any of the addressees has engaged with the system.	
	AnnouncePresence		
	1. I'm here to give directions if you need me.		
	Precondition:	A bystander is in the TooClose or Natural zone.	
	Addressees:	All bystanders in the TooClose or Natural zones.	
	Success:	The addressees are not in the TooClose and Natural zones.	

(Left column vertical labels: "Actions available while the robot is engaged" spans MoveFurther through InviteToJoin; "While not engaged" spans InviteToEngage and AnnouncePresence)

Table 1. *Details of scene-shaping actions (graphic on the right shows a canonical example that satisfies the precondition)*

simulated robot. The annotator reviewed these videos in advance of the assessments.

For qualitative analysis, we developed codes over the annotator's open textual descriptions. For instance, the annotator's rating details shown in the first line from Table 2 were coded as Position(+) corresponding to "*reasonable given how close the actor was standing*", Timing(-) corresponding to "*odd timing*", and Interrupted(-) corresponding to "*ends up interrupting the actor's answer*". We clustered the resulting set of codes into three categories: *Justification*, *Timing*, and *Rendering*. The *Justification* category captures features that contribute to or detract from the motivation for

the robot performing a shaping action, including user positions, engagement, interest level, the history of shaping actions already taken, and whether the system was already having trouble interacting. The *Timing* category includes general mentions of good and bad timing and whether users were interrupted. The *Rendering* category captures aspects related to the rendering of the action, such as whether the action was self-interrupted, it worked well in context, etc.

We applied a similar open-coding approach to the description of user responses and found three high-level categories: *Movement*, *Internal State Change*, and *Response*. The *Movement* category contains codes for various types of

Action	Rating	Rating details	User response	Rating Coding	Response Coding
MoveFurther	4	reasonable given how close actor was standing. Odd timing, right after system asks if needs directions (ends up interrupting actor's answer)	smiles, takes a couple steps back, straightens posture	Position(+) Interrupted(-) Timing(-)	Smile StepBack
MoveCenter	1	actors had been standing close briefly, but were now all the way to the elevator. Not interacting. Robot started to thank them for complying as well, which made it even more odd, but at least it catches itself.	<NONE>	Position(-) LackEngagement(-) Rendering(-) SelfInterrupt(+)	<NONE>

Table 2. Example annotated and coded data

movements that people do following the system's shaping action, such as stepping forward, back, to the side, leaving, etc. The *Internal State Change* category contains codes capturing expressions of internal states such as amusement, confusion, and expressions of interest and surprise, etc. Finally, the *Response* category includes codes that capture other immediate verbal and non-verbal responses, side talk, people looking at others, posture shifts, as well as verbal and nonverbal expressions in which one person encourages another to comply with the system's request.

After eliminating seven shaping actions performed during interactions with developers (giving demos) or during interactions where children were present, the dataset used in the analysis below contains a total of 117 shaping actions, triggered over 95 interactions.

Results

Figure 3 shows histograms of action rating scores, as well as total action counts and average score for each action. We conducted a qualitative analysis based on these scores, as well as the textual explanations provided by the annotator, and their coding. We focused more attention on the cases with low ratings (1, 2, or 3), and on negative valence codes, in an effort to identify successes and failures, how people respond and whether they comply, and how to further improve the scene-shaping actions and policy.

The *MoveFurther* action was rated highly in a majority of cases, with the annotations frequently indicating that the action is well-motivated and that in most cases the participants respond, often stepping back in compliance. Expressions of amusement such as smile and laughter are present in 6 of 12 cases. User compliance with the shaping action as indicated by the annotator is high (10 of 12 cases), and was correctly detected by the system at runtime. The *MoveCloser* action

was triggered only 3 times in our dataset. In one case, the action had a very low rating (1) explained by the fact that it was a third shaping action performed, after poorly justified MoveCenter and LargeGroupArrange actions. The action completely disrupted the interaction and all three participants left immediately after. In the other two highly rated actions, participants complied with the system's request and the system correctly detected this compliance automatically.

The *MoveToCenter* action was triggered 16 times and in 12 of these cases the rating was low (≤ 3). The annotator often perceived the action as not justified because of lack of engagement on the part of the addressee(s). This situation includes cases such as: the addressee is part of a larger group already engaged elsewhere; people reading the sign next to the robot that have not really initiated an engagement; people that are in the process of disengaging as the interaction is closing. Inspection of the data indicates that, in some cases, the challenges with assessing engagement are generated or exacerbated by vision and tracking challenges. The automatic assessment of shaping success is also challenging to do – the robot's assessment diverges in 5 of these 12 cases from the annotator's assessment.

The *LargeGroupArrange* action was triggered 22 times. In the 10 cases with low ratings the annotator found the actions were sometimes (4 cases) not justified based on the position of the users, *e.g.*, "*the actors were already arranged the way it was asking*". In several other cases, the action was triggered while the interaction was closing or towards participants that were not really engaged. For instance, in one case the system performed this action while a single participant was in front and a large but separate group was in the background (the current rendering of the action is not appropriate for this case as it assumes all participants are in the same group.) This action leads to movement and rearrangement in the scene, but also a fair amount of confusion.
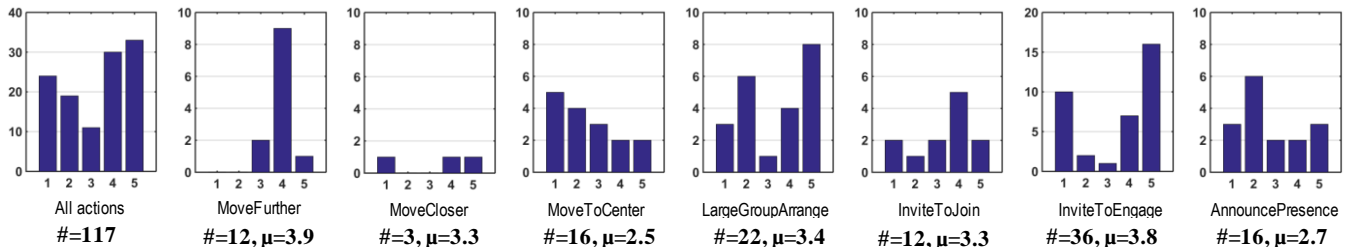

Figure 3. Histograms of action ratings (# denotes total count, μ denotes mean score).

All actions #=117

MoveFurther #=12, μ=3.9

MoveCloser #=3, μ=3.3

MoveToCenter #=16, μ=2.5

LargeGroupArrange #=22, μ=3.4

InviteToJoin #=12, μ=3.3

InviteToEngage #=36, μ=3.8

AnnouncePresence #=16, μ=2.7

While participants (or some subset of them) complied to some degree in 10 out of the 22 cases, they often did not move to the ideal position we were seeking. Automatically assessing success for this shaping action can be challenging.

The *InviteToJoin* action was triggered 12 times. In the 5 low-rated cases, the annotator indicated that the action was disruptive or had inappropriate timing, and was often not spatially motivated, or was done towards addresses that had no interest in joining. In some cases, the invitees were quite far, indicating the need to further tune our proxemic zones. In addition, the invitation arrived at the tail end or midway through the interaction. In contrast, for the 7 actions with high scores, the annotator marked good timing and good spatial motivation; in 6 out of these 7 cases the invitation was done at the beginning, right after the first exchange.

The *InviteToEngage* action was performed 36 times, and in 13 cases was rated low by the annotator. The coding reveals that in 11 of these cases the bystander was not showing signs of interest in interacting with the system. The system's attentional models are often failing in these cases: people are often part of a pair or group and their orientation towards their interlocutors simply happens to align with the direction towards the robot. In contrast, in the highly rated actions, interest was noted in 21 of the 23 cases. Overall, the data suggests that performing these invitations well hinges on robust attention tracking algorithms, as well as the ability to model group relationships between people in the scene.

The *AnnouncePresence* action was triggered when a bystander was detected in the TooClose or Natural zones. In contrast to InviteToEngage, these actions are generally rated lower, with scores of 3 or below in 11 out of 16 cases. The annotator marked lack of interest from the bystander in most cases (8 of 11). The action was originally intended to get someone to either engage or clear the space near the robot. The annotator notes suggest however that the action is not well justified if the bystanders are engaged elsewhere.

Discussion

The analysis of the data from this initial field study provides insights into challenges and future directions. First, while the actions are meant to shape the situation into one that is more favorable for the system's sensing capabilities, we note that the very problem of detecting problematic situations and their resolution is in itself a sensing problem. Our results indicate that shaping actions are sometimes triggered inappropriately based on incorrect inferences about location, attention, and conversational engagement. Improvements in the robustness of each of these estimates, as well as modeling of group relationships and side engagements between people in the scene, would help minimize the number of inappropriate actions. Future work may also consider a tighter integration between the models for scene shaping

and engagement; currently, the scene shaping layer was developed on top of a preexisting engagement model.

The scene-shaping actions themselves and their rendering can further be refined. The data indicate that some of the actions involving references to multiple people can often create confusion. While the actions were coordinated with gaze and pointing, further studies are necessary to understand where the confusion stems from and how to design better shaping actions for large groups. A more refined, in-stream coordination of the action's rendering (gaze, gestures and speech) with the participants' attention and position may be required.

The data suggests opportunities for further refinements of the scene-shaping policy. An important aspect to consider is the tradeoff between effectiveness and naturalness, *i.e.*, the robot must consider the benefits and costs of shaping actions on performing the base-level task of providing directions, and the policy needs to blend them carefully, at appropriate points in the overall interaction. In more than half of the actions, the annotator's notes included a mention of a timing aspect (either good or bad) or disruptiveness. Further investigation is necessary to understand good and poor timings for various actions.

Future work should focus on creating a more objective assessment of the effectiveness of these actions. The current methodology also only captures the quality of triggered shaping actions, but does not assessed missed opportunities. Finally, we believe the approach to scene shaping can be extended to mobile robots, with important questions about when to rely on explicit requests for help versus more implicit behaviors to encourage change, *e.g.*, leaning or moving back when the user is too close.

Conclusion

We introduced scene shaping and conducted a study on endowing a robotic system with the ability to request adjustments in the F-formation of participants towards configurations that are more favorable to its sensing capabilities. We reported results, insights, and lessons learned from an in-the-wild deployment regarding the effectiveness of a proposed set of scene-shaping policies. The methods are promising as a means to enable robots to elicit help to improve robustness in interactions. Ultimately, we expect these methods would be integrated with formal models that reason about costs and benefits to the underlying base task and seek to minimize failures in human-robot interaction.

Acknowledgments

References

Bohus, D., Horvitz, E., 2014. *Managing Human-Robot Engagement with Forecasts and ... um ... Hesitations*, in Proceedings of ICMI'2014, Istanbul, Turkey.

Bohus, D., Saw, C.W., Horvitz, E., 2014. *Directions Robot: In-the-Wild Experiences and Lessons Learned*, in Proceedings of AAMAS'2014, Paris, France.

Dautenhahn, K., Walters, M., Woods, S., Koay, K.L., Nehaniv, C.L., Sisbot, A., Alami, R. and Siméon, T. 2006. *How may I serve you? a robot companion approaching a seated person in a helping context.*, in Proceedings of HRI'2006.

Hall, E.T. The Hidden Dimension. Anchor Books, 1966.

Kato, Y., Kanda, T., and Ishiguro, H., 2015. *May I help you?: Design of Human-like Polite Approaching Behavior*, in Proceedings of HRI'2015.

Kendon, A., 1990. *Spatial organization in social encounters: the F-formation system*, Conducting Interaction: Patters of behaviors in focused encounters, Studies in International Sociolinguistics, Cambridge University Press.

Kuzuoka, H., Suzuki, Y., Yamashita, J., and Yamazaki, K., 2010. *Reconfiguring spatial formation arrangement by robot body orientation*, in Proceedings of HRI'2010.

Mead, R., and Mataric, M., 2015. *Robots Have Needs Too: People Adapt Their Proxemic Preferences to Improve Autonomous Robot Recognition of Human Social Signals"*, in Proceedings of The 4th International Symposium on New Frontiers in Human-Robot Interaction (NF-HRI'2015).

Shi, C., Shimada, M., Kanda, T., Ishiguro, H., and Hagita, N., 2011. *Spatial Formation Model for Initiating Conversations*, in Proceedings of Robotics: Science and Systems, 2011

Walters, M.L., Dautenhahn, K., Boekhorst, R.T., Koay, K.L., Kaouri, C.., Woods, S., Nehaniv, C., Lee, D., and Werry, I., 2005. *The influence of subjects' personality traits on personal spatial zones in a human-robot interaction experiment*, in Proceedings of ROMAN'2005.