An Experimental Design for Studying Proxemic Behavior in Human-Robot Interaction

Ross Mead and Maja J Matarić
Interaction Lab
Computer Science Department
Viterbi School of Engineering
University of Southern California
Los Angeles, CA 90089
{rossmead,mataric}@usc.edu

Abstract

In this report, we present a user study and dataset designed and collected to analyze how humans use space in face-to-face interactions. The experimental design highlights different interaction cues—initiation, acceptance, maintenance, manipulation, and termination—in dyadic and triadic scenarios; these cues are characteristic of transitions into, during, and out of a potentially multi-party social encounter. The experimental setup also elicits spatial configurations that are commonly observed and supported in the social science literature.

1 Introduction and Background

Proxemics is the study of the interpretation, manipulation, and dynamics of human spatial behavior in face-to-face social interactions (Hall 1959, 1963, 1966, 1974). People use proxemic signals, such as distance, stance, hip and shoulder orientation, head pose, and eye gaze, to communicate an interest in initiating, accepting, maintaining, terminating, or avoiding social interaction (Schegloff 1998; Deutsch 1977; Mehrabian 1972). Humans can also manipulate space in an interaction, perhaps to direct attention to an external stimulus (sometimes accompanied by a hand gesture) or to guide a social partner to another location (McNeill 2005a, 2005b). These cues are often subtle and, subsequently, are subject to coarse analysis.

A lack of high-resolution metrics limited previous efforts to coarse analyses in both space and time (Oosterhout and Visser 2008; Jones and Aiello 1979). Fortunately, recent developments in sensor technology provide the means and justification to revisit and more accurately model the subtle dynamics of human spatial interaction.

In this report, we present a user study and dataset designed and collected to analyze how humans use space in face-to-face interactions.

2 User Study and Data Collection

The following study was conducted with the approval of the USC Institutional Review Board (#UP-09-00204). All participants were presented with the Experimental Subject's Bill of Rights, and were informed of the general nature of the study and the types of data that were being captured (in this case, video and audio); consent to these conditions meant that the participant agreed to be recorded, and that such recordings could be used, unmodified, for presentation purposes. A Myers-Briggs/Jung Typology test and a post-study survey were also administered (Myers 1995).

2.1 Objectives of Study

Interaction Cues: This study sought to capture different interaction cues (Deutsch 1977). An initiation cue is a behavior that attempts to engage a potential social partner in discourse (also referred to as a “sociopetal” cue; Lawson 2001; Low and Lawrence-Zúñiga 2003). An acceptance cue signifies an acknowledgment of social presence and registration into an interacting party. A maintenance cue is concerned with abiding by social conventions that promote continued interaction; such cues tend to occur throughout the entirety of an encounter (i.e., from initiation to termination). A manipulation cue is a behavior (often intentional) that is used to direct the locus of joint attention, often toward an object of common interest. A termination cue proposes the end of an interaction in a socially appropriate manner (also referred to as a “sociofugal” cue; Sommer 1967; Lawson 2001; Low and Lawrence-Zúñiga 2003). Note that these cues are often targeted (either consciously or subconsciously) with respect to a social partner or stimulus, and can occur in sequence or in parallel.

F-Formations: This study also sought to elicit what Kendon (1990) referred to as F-formations, which arise “when two or more people sustain a spatial and orientation relationship in which the space between them is one to which they have equal, direct, and exclusive access.” The mutual interaction space that is formed is referred to as an O-space. Typical F-formations include vis-à-vis, L-shape, and side-by-side (Figure 1). McNeill (2005a) posited that each of these F-formations could serve one of two roles: social (consisting only of the people, as originally defined) or instrumental (or attentional; incorporating an object of mutual interest). An objective of the study was to present scenarios in which such physical arrangements were highly likely, as transitions between different F-formations and roles could be signaled by maintenance or manipulation cues. For example, a manipulation cue, such as pointing to an external stimulus, could move the O-space.

(a) vis-à-vis (b) L-shape (c) side-by-side

Figure 1: Three types of F-formations.

2.2 Setup

The scenario was set up and conducted in a 20’ x 20’ room (referred to as “the pen”) in the USC Interaction Lab (Figure 2). The study necessitated three roles: a greeter, a presenter, and a participant. In this study, the presenter and the participant engaged in an interaction loosely focused on a common object of interest—a humanoid robot. During the study, the interactees were monitored by the PrimeSensor™ markerless motion capture system, an overhead color camera, and an omnidirectional microphone.

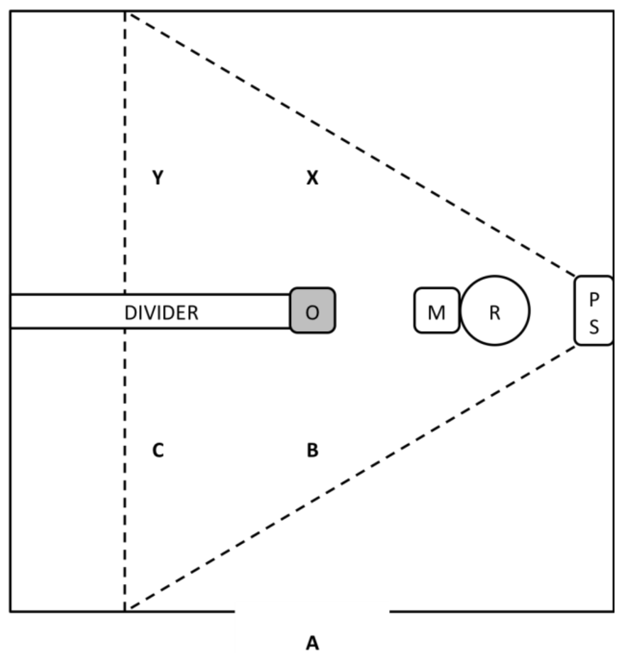

Figure 2: The interaction setup: PS denotes the PrimeSense™ (the dotted lines represent its field-of-view); O denotes the overhead camera (gray for in the ceiling above the pen); M denotes the microphone; R denotes the robot; A, B, and C denote participant floor marks; X and Y denote presenter floor marks; DIVIDER separates the participant and the presenter at the beginning of the interaction scenario.

PrimeSensor™ [1] : The PrimeSensor™ is a markerless motion capture system that provides the 3D position and confidence of 24 joints for each tracked person[2]. The system also provides RGB and depth images[3].

The PrimeSensor™ is capable of tracking people without calibration; however, this method was determined to be unsuitable for the purposes of this study[4]. Instead, a noninvasive "initialization pose" was used to calibrate the sensor to each person in its field-of-view. Each person was trained, prior to the execution of the study, to stand straight up, extend his or her arms bilaterally from his or her body, and bend his or her arms at the elbow to form a 90° angle between the upper arm and the forearm. The calibration process would take ~2-5 seconds.

Overhead Camera: A wide-angle overhead color camera provided complete coverage of the pen[5]. The USC Overhead Interaction Toolkit[6] was used for recording, playback, tracking (xy-position of all entities, person or other, in the pen), and analysis. Extrinsic calibration of both the overhead and PrimeSensor™ RGB cameras was performed by placing a standard black-and-white checkerboard in the image frames of both cameras at the same time; intrinsic calibration of the cameras was performed in a similar manner[7]. A laser pointer was directed to a location on the floor that was in the field-of-view of both the overhead and PrimeSensor™ RGB cameras; this was used to synchronize the recorded videos in post-processing[8].

Omnidirectional Microphone: A Shure™ omnidirectional microphone was used to record the audio during the interaction. Audacity Cross-Platform Sound Editor[9] was used to edit and store the audio tracks. To synchronize the audio with the video, each participant was trained to extend his or her hands straight ahead and clap them together once [10] , thus, producing a unique and salient audiovisual action that could be recognized by all monitoring systems.

Greeter [11] : The greeter was responsible for providing the necessary paperwork to both the presenter and the participant (separately). The greeter was also responsible for training the presenter, who played a special role in the study. The greeter provided task-relevant instructions to both the presenter and participant, as appropriate (discussed below).

Presenter[12]: While the participant filled out the necessary paperwork, the presenter prepared, as trained, for the study. The presenter stood at floor mark X, faced forward [toward the PrimeSensor™], and assumed the trained initialization pose. The greeter would indicate to the presenter that the calibration was successful by shining a laser pointer at the divider—the laser was visible to the presenter, but not the participant. Once the greeter validated the calibration, the presenter would take a few steps back until he or she was standing directly over floor mark Y. From this location, the presenter was still in the field-of-view of the PrimeSensor™ and overhead cameras, but would be out of the projected fields-of-view of

[1] http://www.primesense.com/

[2] The following joints were tracked for each person: head, neck, torso, waist, collars (left and right), shoulders (left and right), elbows (left and right), wrists (left and right), hands (left and right), longest finger tips (left and right), hips (left and right), knees (left and right), ankles (left and right), and feet (left and right).

[3] In this iteration of the study, these images were each captured at 320x240 resolution, and reduced the capture frequency from 60 Hz to 10-12 Hz. In future iterations of this study, a more appropriate method for storing these images will be utilized to improve the capture rate.

[4] In this iteration of the study, the sensor was placed ~9 feet above the ground at an angle of ~30 degrees. Persons could be completely tracked at ranges of 4-10 feet from the sensor, in a 60° horizontal field-of-view. Tracking of the hands and arms was noisy, at best; this was determined to be due to self-occlusions and other ambiguities, likely as a result of the tilted perspective from the sensor placement.

[5] The overhead camera was mounted in the ceiling (~12 feet above the ground) in the center of the pen. RGB images were captured at 640x480 resolution.

[6] http://sourceforge.net/projects/usc-ros-pkg/

[7] The checkerboard was moved to locations at various distances and perspectives to improve the accuracy of the calibration.

[8] In future iterations of this study, video synchronization will be done using either a single computer or Network Time Protocol.

[9] http://audacity.sourceforge.net/

[10] In future iterations of this study, audiovisual synchronization will be done using alternative methods.

[11] In this iteration of the study, the author played the role of the greeter.

[12] In this iteration of the study, a postdoctoral colleague of the author played the role of the presenter for all interactions. This person is familiar with the author's general area of research, but was not familiar with the conditions or metrics of the study at the time it was performed. The presenter had never engaged in a prolonged social interaction with any of the participants.

the participant. Here, the presenter waited until signaled to approach the participant.

Participant[13]: The following steps were taken before the study officially began to ensure that the participant could not see or interact with the presenter before it was appropriate for the study.

The participant stood at floor mark A, facing toward the entrance of the pen. The participant was instructed to direct his or her focus of attention to floor mark B, approach it, and stand directly over it. The participant was then instructed to rotate right 90° and face forward. The participant was then asked to assume the trained initialization pose, allowing the greeter to calibrate the PrimeSensor™ to the user. Once the greeter validated the calibration, the participant could resume an idle pose. The participant was instructed to take a few steps back until he or she was standing directly over floor mark C. The participant was asked to clap once, as trained. The participant was instructed to approach the robot and begin a social interaction with it; a simple greeting, such as "Hello Robot, my name is…" was suggested.

Robot: The participant would attempt to initiate an interaction with the Bandit III upper-torso humanoid robot platform[14], available in the USC Interaction Lab[15]. For the purposes of this study, the robot was not responsive to the interaction[16]; rather, the robot provided a potential (but not necessary) talking point for subsequent interaction between the participant and the presenter.

2.3 Interaction Scenario

As soon as the participant moved away from floor mark C and approached the robot (i.e., a dyadic interaction initiation cue), the study was considered to have officially begun. Once the participant verbally engaged the robot, the greeter would wait a ~2-10 seconds, and shine the laser pointer onto the floor at the feet of the presenter[17], still within the common field-of-view of all visual monitoring systems, but out of the field-of-view of the participant; this signified an approximation of the start of a triadic interaction initiation cue. The presenter would then approach the participant from behind the divider, and attempt to enter the existing interaction between the participant and the robot (a dyadic acceptance cue from the participant to the presenter). Once engaged in this interaction, the dialogue between the presenter and the participant was open-ended[18].

After ~5-6 minutes, the greeter would then announce to the interactees that they should wrap up their conversation soon, and that the participant should step out of the pen once they had finished[19]. The greeter would then shine the laser pointer on the floor in the common field-of-view of all visual monitoring systems, which, at the time, was out of the respective field-of-view of both the presenter and the participant; this signified an approximation of a interaction termination cue (often triadic, though two dyadic terminations were possible).

Once the interaction was over, the participant would exit the pen. The presenter had been previously instructed by the greeter to return to floor mark Y at the end of the interaction. Once the presenter reached this destination, the study was considered to be complete.

3 Qualitative Analysis

A total of 20 people (18 participants, 1 presenter, and 1 greeter) were involved in the study. Usable data was collected for 11 of the 18 interactions.

[13] In this iteration of the study, participants were elicited via lab mailing lists and word-of-mouth, and consisted primarily of undergraduate and graduate students with strong backgrounds in science, technology, engineering and mathematics.

[14] The robot is approximately to scale with a human child; mounted atop a Pioneer P2 base, the entire robot stands one meter tall.

[15] http://robotics.usc.edu/interaction/?l=Laboratory:Facilities#BanditII

[16] In this iteration of the study, the robot was placed in an idle body pose and looked straight ahead. This was done to ensure a similar interaction experience (specifically, no eye contact) between the robot, the presenter, and all participants.

[17] In this iteration of the study, the start time of the presenter-participant interaction was controlled by the greeter. In future iterations of the study, this time will be triggered automatically.

[18] It was suggested to (but not required by) the presenter that he or she use the robot as a conversation starter or topic of interest if he or she could was having trouble guiding the interaction naturally. The only constraint imposed on the presenter was to keep the interaction within the common field-of-view of all monitoring systems. The presenter was not familiar with the conditions or metrics of the study.

[19] In this iteration of the study, the end time of the presenter-participant interaction was controlled by the greeter. In future iterations of the study, this time will be triggered automatically.

3.1 Interaction Scenario

The described interaction scenario was executed better than expected. Both the presenter and the participants consistently filled their defined roles. All participants reported being unaware of the presence of the presenter prior to their interaction; in one case, a participant was actually startled when the presenter approached. The presenter reported to have never become aware of the conditions of the study. The presenter noted being a bit nervous and struggled to direct the conversation in the first interaction, but became more comfortable in subsequent interactions; in fact, the presenter reported a genuine feeling of engagement in these interactions. Both the participant and the presenter admitted that the beginning of the interaction felt a bit unnatural, but became more natural after a few minutes; this is likely due to the fact that the interactees had never met before and were expected to interact, without any particular reason, for an extended period of time.

3.2 Monitoring System

The monitoring system requires improvement for subsequent studies. The method used for audiovisual synchronization was chosen for ease of implementation, but obligates the researcher to manually pair the audiovisual frames; in future iterations of the study, an automated method should be used. Audacity, the software used to record the audio, crashed at the end of two of the 18 interactions; however, the researcher was able to manually recover this data. The PrimeSensor™ proved to be faulty at times, though it was not clear whether this was a hardware or software issue (or both). Specifically, the PrimeSensor™ would lose one of the tracked persons, and then reacquire him or her, but label them as being a different person. The software was written to log the joint positions of only the first two people being tracked, so, in these cases (7 of 18), the data became unusable.

3.3 Objectives of Study

The study proved to be successful in eliciting the types of interaction cues and spatial arrangements discussed in Section 2.1. Participants initiated an interaction with a humanoid robot as if it were a social partner, and responded to the initiation cues provided by the presenter. Likewise, the interactees used termination cues at the end of the interaction. Manipulation cues were primarily used by the presenter when discussing the robot and its features; the participant would use manipulation cues as well when asking a question about the robot or, interestingly, following the use of a manipulation cue by the presenter (i.e., mimicry; Burgoon et al. 1995). Both the participant and the presenter demonstrated various interaction maintenance cues (e.g., compensation; Burgoon et al. 1995), and established all three of the F-formations discussed—a vis-à-vis F-formation was often formed when a dialogue did not involve the robot; an L-shape F-formation was often formed when the presenter was actively talking to the participant about the robot, and the participant was passively listening; a side-by-side F-formation was often formed when both the presenter and the participant were discussing the robot, social robotics, and/or the future of human-robot interaction, in general.

4 Conclusions and Future Work

In this paper, we presented a user study and dataset designed and collected to analyze how humans use space in face-to-face interactions. The experimental design highlights different interaction cues—initiation, acceptance, maintenance, manipulation, and termination—in dyadic and triadic scenarios; these cues are characteristic of transitions into, during, and out of a potentially multi-party social encounter. The experimental setup also elicits spatial configurations that are commonly observed and supported in the social science literature. This study also provided valuable insights for future investigations of the topic.

The data collected in this iteration of the study is most useful for recognizing very specific cases of initiation, acceptance, maintenance, manipulation, and termination cues; however, additional experimental conditions must be introduced if a more generic model of human spatial dynamics is to be constructed. Such studies could include varying numbers of people (e.g., dyadic vs. triadic vs. group interactions), violations of social norms (e.g., decreased distance or increased eye contact), participants with relationships of varying levels of intimacy (e.g., stranger vs. friend vs. significant other), presence of fight-or-flight-inducing stimuli (e.g., sudden production of a scary face or sound), limited sensory capabilities (e.g., blindfolding participants), and/or varying environmental features (e.g., lighting, room size, objects of interests, etc.). For a discussion on how these factors affect proxemic behavior, see Section 1.

Acknowledgments

This work is supported in part by the National Science Foundation under the Graduate Research Fellowship Program. We would like to thank Louis-Philippe Morency for his insight in the experimental design process, Mark Bolas and Evan Suma for their cooperation and assistance in using the PrimeSensor™, Amin Atrash for playing the role of the presenter, Edward Kaszubski for his help in constructing the physical setup, and Dave Feil-Seifer for assistance in using the Overhead Interaction Toolkit.

References

Adam Kendon. 1990. Conducting Interaction - Patterns of Behavior in Focused Encounters. Cambridge University Press, New York, New York.

Albert Mehrabian. 1972. Nonverbal Communication. Aldine Transaction, Piscataway, New Jersey.

Bryan Lawson. 2001. Sociofugal and Sociopetal Space: The Language of Space. Architectural Press, Oxford, UK.

David McNeill. 2005a. Gesture, Gaze, and Ground. Lecture Notes in Computer Science, Springer-Verlag, 1-14.

David McNeill. 2005b. Gesture and Thought. Chicago University Press, Chicago, Illinois.

Edward T. Hall. 1959. The Silent Language. Doubleday Co., New York, New York.

Edward T. Hall. 1963. A System for Notation of Proxemic Behavior. American Anthropologist, 65:1003-1026.

Edward T. Hall. 1966. The Hidden Dimension. Doubleday Co., New York, New York.

Edward T. Hall. 1974. Handbook for Proxemic Research. American Anthropology Assn., Washington, District of Columbia.

Emanuel A. Schegloff. 1998. Body Torque. Social Research, 65(3):535-596.

Isabel Briggs Myers. 1995. Gifts Differing: Understanding Personality Type. Davies-Black Publishing, Mountain View, California.

Judee K. Burgoon, Lesa A. Stern, and Leesa Dillman. 1995. Interpersonal Adaptation: Dyadic Interaction Patterns. Cambridge University Press, New York, NY.

Robert D. Deutsch. 1977. Spatial Structurings in Everyday Face-to-face Behavior. The Association for the Study of Man-Environment Relations, Orangeburg, New York.

Robert Sommer. 1967. Sociofugal Space. The American Journal of Sociology, 72(6):654–660.

Setha M. Low and Denise Lawrence-Zúñiga. 2003. The Anthropology of Space and Place: Locating Culture. Blackwell Publishing. 61–62.

Stanley E. Jones and John R. Aiello. 1979. A Test of the Validity of Projective and Quasi-Projective Measures of Interpersonal Distance. The Western Journal of Speech Communication, 43:143-152.

Tim van Oosterhout and Arnoud Visser. 2008. A Visual Method for Robot Proxemics Measurements. 3rd ACM/IEEE International Conference on Human-Robot Interaction Workshop on Metrics for Human-Robot Interaction, Technical Report 471, Amsterdam, Netherlands.