UNIVERSITY OF SOUTHERN CALIFORNIA

UNDERGRADUATE THESIS

Minimal Nonverbal Behaviors for Socially Interactive Robots

Author:
Eric C. DENG

Supervisor:
Dr. Maja J. Matarić
Dr. Gaurav Sukhatme

A thesis submitted in fulfillment of the requirements for the degree of Bachelors of Science in Electrical Engineering

in the

Interaction Lab
Department of Computer Science
Ming Hsieh Department of Electrical Engineering

May 9, 2017

University of Southern California

Abstract

Dr. Maja J. Matarić
Department of Computer Science
Ming Hsieh Department of Electrical Engineering

Bachelors of Science in Electrical Engineering

Minimal Nonverbal Behaviors for Socially Interactive Robots

by Eric C. DENG

The field of Human-Robot Interaction (HRI) is dedicated to designing, creating, and developing algorithms and methods for evaluating robotic systems to be used by or with human users. Interactions between humans and robots are incredibly complex; researchers in the field are exploring a broad range of problems and approaches, from robot design, to proxemics, to task sharing. Interactive robots must be able to both perceive and generate rich, engaging, multi-modal social interaction to truly be effective.

We were initially inspired by the needs of children with autism spectrum disorder (ASD), a developmental disorder associated with atypical development of social skills along with other symptoms. One of the social skills that children with ASD struggle with is *joint attention*, the sharing of focus between individuals during interaction that allows for effective communication and sharing of knowledge and experiences. In our project, we hope to give robots the ability to not only engage in but also explicitly establish joint attention. To do this we first came up with a three-step approach for initiating joint attention and then focus on designing robot behaviors for the first of those three steps, *attention acquisition*. By generating minimal, yet effective, attention acquisition behaviors, we can reduce the risk of overwhelming users with too much sensory information and better preserve the agency of the robot, a feature for general HRI but especially important for interactions with children with ASD who often have heightened sensitivities to sound and light. Although this work was initially targeted towards improving joint attention in children with ASD, joint attention is beneficial in all human-robot interactions and the resulting approach is general enough to be used across different interaction domains and with many different robot embodiments.

Contents

Chapter 1

Background

Interactive robotics is a highly interdisciplinary field. In this chapter we briefly introduce the fields and existing work relevant to the work in this thesis.

1.1 Human-Robot Interaction

Human-robot interaction is a large field of research in robotics that focuses on developing and evaluating ways that humans and robots can interaction, communicate, and (Goodrich and Schultz, 2007). In order to created more effective interactive robots, researchers have studied a broad range of problems and approaches, from proxemics (Mead and Mataric, 2016), to gaze behaviors (Mutlu et al., 2009), to social touch (Knight et al., 2009). Related to our goal of generating minimal social behaviors for interactive robots, we review work related to *nonverbal cues* and *joint attention*.

1.1.1 Nonverbal Cues in Embodied Robots

Nonverbal cues have been extensively studied in HRI, mostly for functional robots co-located with human users (Takayama, Dooley, and Ju, 2011). If used properly, these cues can significantly improve social performance of robots. Generating and understanding nonverbal cues is highly dependent on the robot platform being used, but the cues can be classified into six categories (Chidambaram, Chiang, and Mutlu, 2012):

1. Proxemics
2. Gestures
3. Gaze
4. Posture
5. Facial Expressions
6. Social Touch

Although all these nonverbal cues are related, we focus on *nonverbal gestures* in our work. Effective implementations of nonverbal gestures have been shown to improve robot effectiveness in health management (Looije, Neerincx, and Cnossen, 2010), energy consumption habits (Midden and Ham, 2009), and educational tasks (Dautenhahn, 1999).

1.1.2 Joint Attention in Human-Robot Interaction

Joint attention is a fundamental skill in social interaction, and there has been some research in HRI related to the topic (Thomaz, Berlin, and Breazeal, 2005). To date, HRI systems are primarily designed to leverage speech utterances to perform joint attention tasks as, without a personalized model of the user, nonverbal gesture generation typically underperforms verbal cues (Huang and Thomaz, 2011; Kaplan and Hafner, 2006). Once joint attention is initially established, some HRI systems are able to maintain it by directing the robot and its behaviors towards to object of interest (Imai, Ono, and Ishiguro, 2003). In this work we focus on building a strategy for robots to establish active joint attention on different targets by taking into account *user disruptability* and *the relative disruptability of different complex gestures.*

1.1.3 Socially Assistive Robotics

Socially Assistive Robotics (SAR) is a subfield of robotics at the intersection of Socially Interactive Robotics (SIR) and Assistive Robotics (AR), with the overarching goal of assisting human users primarily through social interaction (Feil-Seifer and Matarić, 2005). SAR systems have numerous applications, such as tutoring (Kennedy, Baxter, and Belpaeme, 2015), physical therapy (Fasola and Mataric, 2012), daily life assistance (Inoue, Wada, and Ito, 2008), and even entertainment (Lee et al., 2006). A growing number of studies are showing strong positive signs for using SAR as an effective tool for social skills therapy for children with ASD (Andreae et al., 2014; Robins et al., 2005). Our work is focused on this problem domain.

1.2 Improving Joint Attention with Socially Assistive Robots

Social roboticists seek to develop algorithms and methods for interactive robots that teach, help, and play with human users (Fong, Nourbakhsh, and Dautenhahn, 2003). To truly be effective, these robots need to be able to engage in social interactions by perceiving and generating complex behaviors (Huang and Mutlu, 2012). We are specifically interested in exploring facets of *joint attention*, a fundamental social skill that is learned in the early stages of development (Moore and Dunham, 2014). Joint attention is critical for many complex social interactions by allowing for things like smooth shifting of context during interaction and is of special interest to researchers in fields such as social skill therapy for children with autism (Baldwin, Moore, and Dunham, 1995). Joint attention can be established in a number of *passive* ways–loud sounds such as lightning or flashing lights can draw attention during an interaction but we want to give our robots the ability to not only maintain shared attention but *actively* establish joint attention on specific objects (Butterworth and Jarrett, 1991). In this section we outline an approach for doing just that.

1.2.1 Establishing Joint Attention

Joint attention is a topic of interest in many fields including developmental psychology, human communication, and cognitive development and has been well-studied in social sciences. The value of establishing and maintaining joint attention has been demonstrated in many studies and is now considered a core component of social interaction (Moore and Dunham, 2014). But in the context of socially interactive robots, joint attention is still a very abstract concept that is hard to operationalize into robot behaviors.

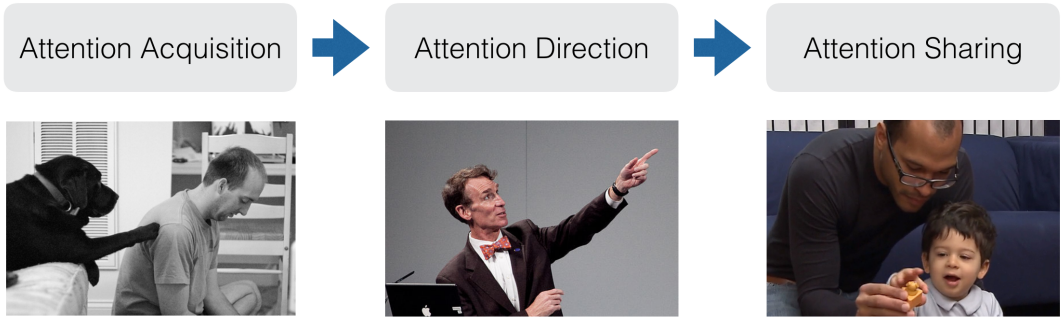

FIGURE 1.1: Joint Attention Decomposed

We see the process of establishing joint attention as a series of sequential actions, (1) *Attention Acquisition*, (2) *Attention Direction*, and (3) *Attention Sharing* (see Figure 3.1). Attention acquisition is a behavior that an agent does to get its interaction partner's attention to be directed in the attention direction phase where the agent uses various non-speech cues to direct the partner's attention toward a target. Once the interaction partner's attention is directed at the target, joint attention has been achieved and the while in the third action state, attention sharing, the interaction partners continue to engage in joint attention. The ultimate goal of using a model-based approach is to deliver the minimum amount of gesturing to get the user's attention but not too overtly as to damage the perception of the robot.

1.3 Social Skill Therapy for Children with Autism

A large body of work in occupational science and developmental psychology outlines best practices for ASD therapy (White, Keonig, and Scahill, 2007). That body of work outlines a list of social skills that are challenging for many children with ASD:

1. **Joint Attention:** shared interest or understanding on an object or event
2. **Social Turn Taking:** alternating speaking turns during speech-based social interactions
3. **Ordering and Sequencing:** recognizing and ordering numbers concepts
4. **Perspective Taking:** understanding others' mental states, specifically during social interaction

Our research is centered on improving joint attention and our approach is inspired by those often taken by occupational therapists (OTs). OTs use a variety of techniques for improving joint attention, including music therapy (Kim, Wigram, and Gold, 2008), symbolic play (Kasari, Freeman, and Paparella, 2006), and behavior modification (Whalen and Schreibman, 2003).

Taking inspiration from this work in using play as a tool for social skill therapy, we have built a table-top robot and tablet system to be used in studies exploring child-robot interactions, specifically those related to teaching social skills to children with ASD. One goal of this system is to improve joint attention and the work outlined in this thesis will be used as an approach for giving robots the ability to engage the children in joint attention.

1.4 Mathematical Models

In this work, we propose using a few different mathematical models to represent aspects of human-robot interaction. In this section we briefly introduce the models used and their designed applications.

1.4.1 Topological Manifold Representations of Functional Data

In our work we aim to represent the gesture space of different robots using manifolds, an area of mathematics and topology. Manifolds have been shown as effective representations for functional data in non-linear, low-dimensional spaces(Chen and Müller, 2012). By using manifolds, not only can we represent gestures for different platforms, as long as the gesture space is low-dimensional, but we can also apply existing manifold-related algorithms, such as similarity learning (Chen, Ding, and Luo, 2015) and manifold alignment (Wang and Mahadevan, 2008). This allows us to autonomously generate representations of gestures and quantitatively compare gestures within a robot embodiment or between different robot embodiments.

1.4.2 Gaussian Process Regression

Gaussian processes are a generative learning method commonly used in regression problems. They have many strengths, including producing probabilistic predictions and interpolation (Rasmussen, 2006). These distributions can be used for nonlinear regressions and return a predictive distribution with predictive mean and variance along each point in the input space. This type of regression allows us to learn more about the overall distribution and about confidence about slices of the distribution, so as to intelligently update and collect more data in an informed way (Welch et al., 1992).

We are interested using Multivariate General Gaussian Process (MGGP) as a model for representing user disruptability based on a number of sensor inputs such as tablet-based behaviors, head pose, body pose, and vocalizations. This model will allow us to both generate a set of equally spaced attention acquisition behaviors to test but will also inform our system of which conditions to explore further.

1.4.3 Automatic Relevance Determination for Feature Selection

Automatic Relevance Determination (ARD) (Qi et al., 2004) is an approach for learning the length scales of input dimensions to determine the importance of different inputs to the overall output of the system. ARD learns the respective length-scales of all dimensions based on the relevance; the system can autonomously select the weights and desired inputs to improve the performance of our regression.

As with many multimodal machine learning problems, the importance of different features in the sensor data for the desired information, in our case user disruptability, is unknown. Using ARD we can use data from pilot studies to learn a prior as well as potential weights to be applied to the different features throughout adaptation processes. By using ARD between sessions during various experiments, we can also relearn optimal feature weights for sensing disruptability in specific users.

Chapter 2

Model-Based Approaches for Generating Minimal Nonverbal Gestures

2.1 Model-Based Attention Acquisition Strategies

Our main contribution related to joint attention and attention acquisition is an end-to-end, model-based, behavior generation pipeline. Although the initial inspiration for this approach stemmed from applications for using robots with children with autism, this pipeline is domain-independent and embodiment-agnostic with different components being relearned for different robot embodiments and interaction scenarios (human-robot collaboration in assembly lines, robots for personalized learning or entertainment, etc.). The ultimate goal of using a model-based approach is to *deliver the minimum amount of gesturing to get the user's attention but not too overtly as to damage the perception of the robot.* Using this approach, systems can be built to autonomously sense user state, select and filter robot behaviors, and execute a minimal, yet effective, nonverbal gesture.

2.1.1 Behavior Generation Pipeline

A general diagram of our behavior generation pipeline can be seen in Figure 3.2. Whenever the robot decides that it wants to engage with the user, it runs through this pipeline to determine the appropriate gestures to generate.

There are four main components of this system–(1) Sensors, (2) User Disruptability, (3) Gesture Selection, and (4) Perceptual Filtering. The sensing package includes systems like the Microsoft Kinect, RGB webcams with audio and video, and tablet-based inputs. User disruptability is a learned predictive model of how intrusive an attention acquisition needs to be in order to successfully get the user's attention based on the available real-time data from the sensing package. The gesture selection node has two learned models, (1) a crowd-sourced general model of relative intensities of gestures in your gesture space as well as (2) a user-specific model of how disruptability relates to the relative gesture intensity model. This component of the system would output a set of appropriate minimal gestures and in the perceptual filtering phase, this set of gestures gets filtered based on the user's current state related to the different components of these gestures. In the next few sections we will go more in-depth on how these sub-systems work and the models we use.

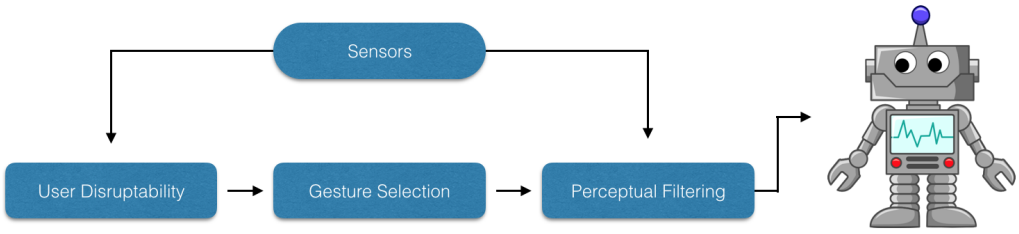

FIGURE 2.1: Attention Acquisition Behavior Generation Pipeline

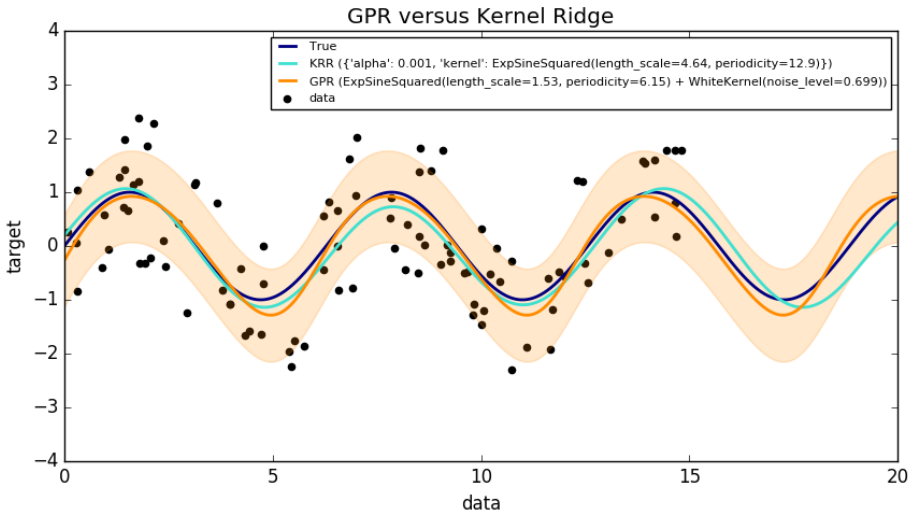

FIGURE 2.2: Example Comparison of Gaussian Process Regression versus Kernel Ridge Regression for Sine-Wave Dataset from Scikit-Learn Examples

Learning User Disruptability

Being able to generate appropriate attention acquisition behaviors not only requires an understanding of the relative disruptiveness of individual behaviors in the set of gestures but also an understanding of how disruptable a person is when the robot is to perform said gesture.

Using multivariate Gaussian process regression (Quiñonero-Candela and Rasmussen, 2005), we can represent user disruptability as a Multivariate General Gaussian Process (MGGP) model (Chan, 2013) of the different sensor inputs that we have access to such as on-screen behaviors, head pose, body pose, and gaze behaviors. Because we expect disruptability to greatly vary between users and human annotation of abstract concepts such as disruptability have been shown to be highly unreliable, we propose relearning a disruptability model for every user.

One of the primary reasons we chose to use GPR over kernel ridge regression (KRR) (Vovk, 2013), another popular statistical method for nonlinear regression, was that it returns a generative, probabilistic model of the target function with confidence intervals along the independent variable (Figure 3.3), in our case, the sensor inputs. This is highly beneficial to real-time learning of models because these confidence intervals can advise an intelligent search method for robot behaviors.

We predict that disruptability will be strongly correlated the task difficulty that the user is engaged in, so by finding peaks in the covariance function of the MGGP,

we can have the system autonomously select the difficulty of the next task to be completed. By manipulating the difficulty of the tasks and observing user response to different levels of gesturing (, i.e. whether or not gaze behavior changed toward the robot), we can intelligently sample areas of high variance to minimize variance across the models we are learning.

An issue with real-time learning is that when users first begin to interact with the system, we have no model of disruptability. One approach would be to sample attention acquisition behaviors (using something like a random sequence or van der Corput sequence across the gesture space) then explore the rest of the space based on the confidence intervals across the model. Instead we propose using a prior probability distribution based on human-annotated data collected from relevant populations doing relevant tasks as a starting point and then use the search approach to sample and update the model based on the maxima in the covariance functions. As these data from human annotation tend to be unreliable, their weights compared to collected data from an individual user need to be small as to be not limit the development of the correct model over time.

Learning Gesture Intensities

As we discussed in previous sections, there are many different parameters of nonverbal gestures that can be changed by tuning the trajectories that the robot embodiment can take (Latombe, 2012). Sometimes gestures have obvious relative magnitudes, like linear motion or amplitude of non-speech sounds but as we are interested in using complex nonverbal gestures that are combinations of different, nonlinear gestural dimensions, we must learn a model of intensities of all the gestures in our gesture space.

We represent our gesture space as a Riemannian manifold with each dimension representing a parameter of the robot's nonverbal gesture such as jerk, acceleration, volume, and velocity (Latombe, 2012). This allows us to use manifold-related techniques to analyze and explore the space as well as quantitatively compare different gestures by looking at their relative locations on the manifold.

In the context of this work, we define two types of nonverbal gestures–relative and absolute. *Relative nonverbal gestures* refers to gestures such as exaggerated idle behaviors or increased sound amplitude above ambient noise that are perceived as a relative cue to a shifting baseline where intensity is measured as a ratio of signal to noise. *Absolute gestures* refer to gestures such as non-speech vocalizations that occur after long periods of vocalization-free behaviors and is measured by absolute intensity. Although this distinction is not critical in the way that we design behaviors, it is important to make sure that relative gestures are not represented on an absolute scale and vice versa because in different contexts, the same motion for a relative behavior may not have the same effect. For example, if a robot regularly "idles" by moving up and down, an exaggerated up and down motion must be measured relative to the idle behavior and if the magnitude of that idle behavior changes, the perceived intensity of the same exaggerated motion will also change.

Based on prior work in gesture perception we work under the assumption that humans are relatively consistent in perceiving relative intensities but very inconsistent in perceiving absolute intensities of complex gestures (Weiss, Simoncelli, and Adelson, 2002). Because of this, we can learn a crowd-sourced, general model of relative robot behavior intensities, resulting in a manifold that represents of all gestures in our gesture space by their perceived disruptability. In the next chapter we

will discuss our proposed experimental methods for collecting data to learn this model.

Gesture Selection

From the components outlined in the previous two sections, we have both a classifier for user disruptability and a relative scale of intensities for nonverbal attention acquisition behaviors. We still run into the issue of combining these two scales as we do not initially know how one classification of disruptability relates to gestures on the relative intensity scale but over time, the model will learn where the two models align for each individual user by holding the relative disruptability of gestures model constant. Using this new relational mapping from one model to the other as well as real-time sensor input, we can then generate a set of gestures that our system believes is of the appropriate disruptiveness in the current interaction scenario.

Perceptual Filtering

From the set of gestures produced in the previous module, the last step of the gesture generation pipeline is to filter intensity appropriate behaviors to also be "modally-appropriate". Modally-appropriate behaviors refer to selecting actions that make the most sense given the perceived ability of a human user to perceive that action (Tsoukalas, Mourjopoulos, and Kokkinakis, 1997). For instance, if the user is turned away from the robot and the robot perceives that it is completely out of the user's field of view, it should not choose an action that is completely dependent on motion and should instead use a sound-based gesture.

The sensor package feeds information into the perceptual filter, advising the system about the current state of the user (head pose, ambient noise, body pose, gaze behavior, etc.) (Zhang, 2012) and removes modally inappropriate behaviors from the set.

After passing through the perceptual filters, we are left with a set of gestures that the models believes are of the right intensity and leveraging the channels of communication most appropriate for the user. From here the system can select, either randomly or with some more complex methods that are beyond the scope of our work, a gesture from the set and display it on the robot. The response of the user (whether or not the user looked at the robot) is then recorded and used to update models for user disruptability and relative positioning of that model and the general relative intensity model of robot gestures.

Chapter 3

Experimental Evaluation

To develop these effective minimal behaviors for attention acquisition, we need to learn and validate our models through a number of experiments primarily funded by the NSF Expedition in Computing grant for Socially Assistive Robots which will be covered later in Section 3.2. In the context of this work, four studies will be performed, three of which are funded by the NSF Expedition in Computing grant. In this chapter, we go over the experimental design for ongoing and upcoming studies regarding different components of the previously discussed work.

3.1 Roles of Socially Assistive Robots in Clinical Settings

As one would expect, human-robot interactions are complex scenarios that have numerous components including task, robot embodiment, and user populations to name a few. To develop more effective SAR and design better performing interactions researchers are exploring all these components both independently and in various combinations to best get coverage over as many interaction scenarios as possible.

In the first study directly related to this project, we worked with children with ASD and their parents to explore the role that the agency of a robotic agent had in the quality of interaction, and in turn the performance, of the robot in clinical settings. The embodiments of the robots (seen in Figure 2.1) and the behaviors of the robots were manipulated and we observed 9 child behaviors and 1 parent behavior while also using the Overhead Interaction Toolkit (OIT) for sensing and generating social spacing. By observing interaction patterns such as head and body orientation towards the robot, button pressing for bubbles, and vocalizations, we saw overall improvements in having an embodied, appropriately-behaved robot in engaging in agent-like interaction, further validating the application of SAR with children with ASD.

As the data for this study were collected before I joined the lab, my contribution to this work is centered around post-hoc analyses about the interaction performance and ASD-related applications for future exploration. Exploring the roles of SAR in clinical settings developed my interests in further studying robot behaviors, especially with populations with ASD, and led to much of the work reviewed in this thesis.

3.2 NSF Expedition in Computing for Socially Assistive Robots

The project related to using SAR with children with autism is the NSF Expedition in Computing grant for Socially Assistive Robotics. This multi-year, inter-institution project had the overarching goal of "developing computational techniques to enable

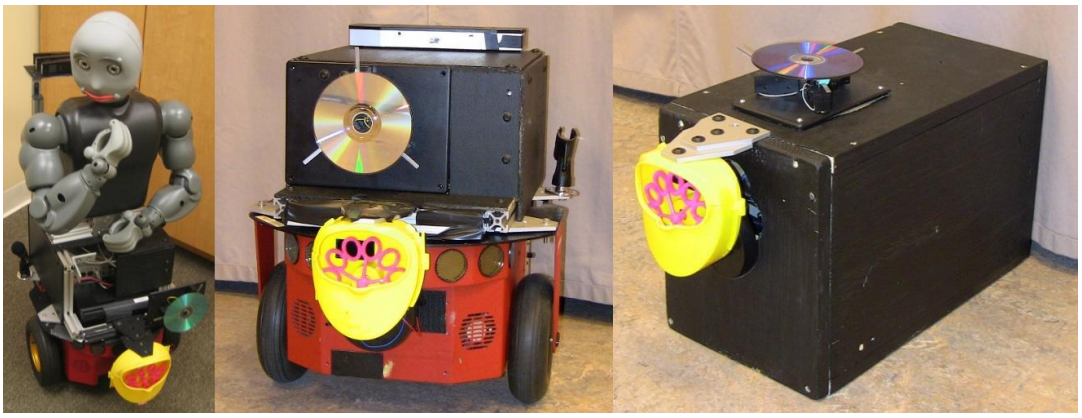

FIGURE 3.1: Three embodiments used in experiment (1) Bandit and Bubble Blower on Pioneer 2DX (2) Bubble Blower on Pioneer 2DX and (3) Bubble Blower

FIGURE 3.2: Multi-Party Study Setup from Elaine Short

the design, implementation, and evaluations of robots that encourage social, emotional, and cognitive growth in children, including those with social and cognitive deficits". As we wrap up this grant, the team at USC has decided to build a SAR system to assist children with ASD learn specific social skills outlined in social therapy literature (seen in Section 1.2) (White, Keonig, and Scahill, 2007). We are currently completing the development and testing of our whole system and will be running in-lab pilot studies beginning mid-to-late the summer and deploying systems into homes for 30 days, running 20-minute interactions for 20 of those 30 days.

To deliver the social skill therapy, we use a system that consists of a touch-screen tablet, a Stewart-platform robot with six degrees-of-freedom (DoF) and a puppet-like animistic skin, and a multi-modal sensor package that includes a Microsoft Kinect, audio input, and a number of RGB webcams. The children sit in front of the tablet with the robot directly behind and slightly above the tablet and parents to either side (Figure 2.3). The whole system is integrated through Robot Operating System (ROS) (Quigley et al., 2009), an open-source system for managing networked robotic systems. This system is paired with a set of 15 activities composed of 3 5-activity sets, each focusing on a subset of the four target skills (Section 1.2).

FIGURE 3.3: "Code Copying" Activity in Alien Codes from Expedition Deployment built with Phaser led by Eric Deng

3.2.1 Alien Codes Games Development

As a part of the implementation of this system, I led the development of the third set of activities, Alien Codes, to be used in the 30-day deployment. This set of activities, which involves five different variations of organizing "space objects" and completing "alien code" sequences where the robot acts as a knowledgeable peer to the child who works with their parent to complete the different tasks.

Built with Phaser, a game framework for Javascript and HTML 5, the Alien Codes activities used the same digital assets as the previous 2 sets of activities, to maintain aesthetic themes, and communicated with the ROS backend to update the system on the user's progress and prompt for the difficulty of the next activity. A still from the "Code Copying" activity from the Alien Codes set can be seen in Figure 2.3.

3.2.2 Heuristics in Expedition Games

When I joined this project the front-end systems of our games could already communicate with our ROS back-end and therefore other nodes on the network. But the data being communicated was very high-level and restricted the extent to which our back-end systems could control the interaction. One contribution I had with the games was helping restructure the networking system to send and request more regular updates from the games to the ROS nodes monitoring task performance, engagement, and determining the difficulty of the next activities.

To allow our system to communicate more detail on user behaviors, we first had to define difficulty-agnostic heuristics for user progress in each activity and redesign the software architecture of the ROS system to both regularly publish game state rather than only at the completion of individual activities. This new back-end and heuristics for user progress on a task is now a core input to many of the systems that are controlling content delivery, social performance, and robot behaviors, one of which we will discuss in the next chapter.

3.3 Attention Acquisition as an Independent Action

Our decomposition of one approach to actively establishing joint attention makes the distinction between the actions of attention acquisition and attention direction. This is because we think that there is social value in first getting an interaction partner's attention before trying to direct it. We hypothesize that by breaking up attention direction into attention acquisition followed by attention direction, we can lessen the disruptiveness of the overall attention-directing gesture.

To validate this hypotheses we have modified an ongoing study on multi-party interactions to collect data on whether or not splitting attention acquisition and attention direction into separate actions makes a difference in the disruptiveness of the robot. Performance differences will be observed through participant responses on pre, mid, and post-interaction questionnaires given to participants on their perception of the robot. The study design and results to-date are discussed in the next section.

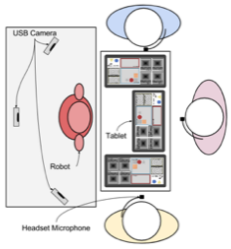

FIGURE 3.4:
Multi-Party Interaction Setup and a Group of Participants

Multi-Party Human-Robot Interaction

As a part of an ongoing study in the lab on multi-party interactions with robot moderators, we are running a 3-condition, between-subject data collection on the value of prefacing attention direction with an attention acquisition behavior and the value of having adaptive attention acquisition behaviors.

In this study, the robot acts as a moderator between three participants working on a shared task, making different colored shapes using "machines" on their respective tablets (Figure 4.1). While the participants work on their tasks, the robot continues to do various "idle" motions by sporadically moving up and down. Users have different "machines" that either change the color or shape of objects and different goal objects (with specific color and shape) but share a scoreboard. They have the option of passing parts around by placing objects in areas on the screen that say variations of "To Player 1". The robot, the same Stewart-platform robot being used in the Expedition project, is paired with the "Kiwi" skin, an owl-hummingbird hybrid with a phone-based face (Figure 4.2).

FIGURE 3.5: Kiwi the SpriteBot Robot (Elaine Short and Matarić, 2017)

The 6-minute activity is run 5 times per group and within each session, the robot speaks 8 times about something happening on-screen or a move that a player should take, an attention directing behavior. The three between-group conditions of this data collection are related to the behaviors that the robot does before it speaks–(1) Nothing, (2) "Hey" vocalization before speaking, and (3) Exaggerating the "Idle" behavior different amounts before speaking. The third condition samples different

Algorithm 1 Instantiated Moderation Algorithm for Storytelling Task from (Short, Sittig-Boyd, and Mataric, 2016)

while Time elapsed < 6 minutes **do**
  $S(i) =$ speech duration of participant i in the last $30s$
  if $elapsed \geq 30$ **then**
    Look at $argmin(S(i))$
    Make an exaggerated up-and-down idle motion as a an attention acquisition behavior
    Wait 2 seconds
    Ask a question
  else
    Look at current speaker
  end if
end while

levels of signal to noise gesturing throughout the interaction. The algorithm implemented on the robot with the attention acquisition gestures is shown as Algorithm 1.

The first and second conditions have a total of 12 participants each and the data collection for the third condition is currently ongoing and will be completed sometime early-July. Based on preliminary review of participants' responses about their perception of the robot in the three conditions, especially when comparing conditions 1 and 2 to condition 3, there may very likely be a strong case for attention acquisition before attention direction behaviors as it may directly impact the perceived agency of a SAR.

3.4 Learning a General Disruptability Prior

In the gesture generation pipeline from Section 3.2, the first module returns a value of disruptability of the user. We discussed primarily focusing on using user-specific data sets to determine disruptability but initially starting with a prior generated from users from relevant populations interacting in similar interaction scenarios. Luckily, many of our studies in the lab are all human-robot-screen interactions where the human user is seated in front of the screen and the robot is somewhere behind the screen. These interactions are similar enough to the in-home Expedition deployments and the data are rich and similar enough for us to learn a general prior to be implemented on the deployment system.

To establish this prior, we propose using a sliding scale input for disruptability. By using multiple coders, with at least two independent coders annotating every subset of data to be verified for label consistency, labeling every 250ms data segment on the scale we aim to get a labeled dataset to be fed into a Multivariate General Gaussian Process (MGGP) (Chan, 2013). We would then experiment with the different features that we have from the studies including head pose, on-screen behavior, task performance/progress, and body orientation using Automatic Relevant Determination (ARD) (Qi et al., 2004) find the best feature set for establishing our prior for human-robot-screen interactions.

3.5 Crowd-Sourced Relative Model for Gesture Intensity

Another learned model we need in our gesture generation pipeline is a relative model of different gestures within our gesture space. Our proposed approach to learning the non-obvious relative disruptiveness of complex nonverbal gestures is to crowd-source people's comparisons on pairs of gestures through online task marketplaces like Amazon's Mechanical Turk (Buhrmester, Kwang, and Gosling, 2011). We are designing and recording a large set of robot gestures for Kiwi, the 6-DoF SpriteBot, that vary in jerk, acceleration, velocities, timing, and space. By presenting different pairs of these recorded gestures to participants recruited and paid through Mechanical Turk and asking them to choose associated one of the two robot behaviors with strong descriptors like "calm" and "agitated" we can crowd source the perceived relative intensities of different complex gestures.

3.6 Pipeline Validation

As a part of the NSF Expedition in Computing project, the attention acquisition generation pipeline will be validated in a series of in-lab and in-home studies. Throughout the summer we will be running in-lab pilot studies with children with ASD to test our deployment system and do some initial data collection. Then in the fall, we will begin our 30-day, in-home deployments with families with both typically developing children and children with ASD.

3.6.1 In-Lab Pilot Studies

Before our system is ready to be deployed into homes, there are still a few studies and data collections that we need to run to learn and test our models and algorithms. After the multi-party study comes to completion in the beginning of the summer, we will begin in-lab testing with convenience populations (college students) as well as families with children with ASD.

Study with Convenience Populations

The study to be run with our convenience population of college-aged students will allow us to test our crowd-sourced model of gesture intensity (Section 4.5), learn a disruptability prior for typically-developing adult populations, and test different learning algorithms for long-term modifications to a user's disruptability model.

Participants will be interacting with Kiwi (Figure 4.5) and a tablet, similar to the human-robot-screen setups in the other studies, while working on completing GRE questions. We will be taking questions from practice exams and can track their performance on independent questions using the preexisting information on the difficulty of each question (given by the practice books) as input for task difficulty. The robot will attempt to give a set of verbal feedback including supporting remarks after successful question answering, encouragement to move onto the next question, and vocalizations about the time and questions they have left for the current section.

There will be multiple sessions per participant that vary in the algorithms that control the next question and difficulty, with updates on the user disruptability model and questionnaire responses between each session. From this study we would like to see a validation of the ordering in our learned gesture intensity model as well as what algorithm for selecting next task works best for learning a user's disruptability.

Pilot Study with Children with ASD

The study with children with ASD and their families will be exactly the deployment we will be using in the homes with the primary goal of testing out the system for robustness with children as well as the content for interest. At this time we will also be collecting data on a variety of different research questions, one of which is user disruptability. During these sessions the system will be using the same general gesture intensity model with a user disruptability model with no prior–focused on exploring and generating disruptability models for individual users. From the whole data collection, we will be able to learn a prior to be used with children with ASD in our long-term, in-home deployments and also observe the differences seen between individuals to get a sense of how confidently we can rely on that prior.

3.6.2 Deploying Robots in the Home

Finally in the fall, we will be beginning our 30-day, in-home deployments. These deployments will consist of Kiwi with a tablet and sensor package and the 15 activities targeting ASD-specific social skills. The interactions will involve a robot, a parent, and the child working through different combinations of these activities that vary in order and in difficulty while we explore a few different research questions centered around personalized learning and attention acquisition. The robot is designed to speak as the children work throughout the activities, giving feedback on task performance, suggesting hints, and sometimes just saying neutral facts, but in the experimental condition, the robot prefaces those verbalizations with up to three attention acquisition behaviors.

Based on the pipeline, the robot has a set of gestures to select from and an understanding of how confident it should be in that set (based on covariance function from the MGGP) in the form of standard deviations along the gesture disruptability scale. From this information, we can come up with three gestures, gesture A being the equivalent of one standard deviation below the expected disruptability, gesture B being the expected disruptability, and gesture C being one standard deviation above the expected disruptability. The robot will then try each of these gestures (from A to C) and observe whether or not the user looks up, recording at which attempt the user looked up to update the model as well as speak as the user is looking at the robot. By only allowing the robot 3 attempts to get the user's attention, we do not limit the robot's ability to speak and avoid the case in which the user takes too long to look up and the robot's intended speech becomes irrelevant or out of context.

Throughout this 30-day deployment our system will learn a personalized model of user disruptability and deliver minimal social behaviors throughout the interactions with the children and recording user perceptions on the robot in the form of on-screen questionnaires before, between, and after interactions.

Chapter 4

Summary and Future Work

In the fall I will be starting my Master's Degree in Mechanical Engineering through the Progressive Degree Program at USC and will continue to work in the Interaction Lab and expand on this work. We initially started with the goal of teaching children with autism social skills and eventually focused on joint attention as the skill of interest, leading us to develop the 3-step pipeline for actively establishing joint attention. Although the previous models discussed have been parameterized and selected for their respective applications within the behavior generation pipeline, we are still in the process of collecting user data to build real models. In the near future, we will be working to complete the experiments outlined in Chapter 3 and adjusting our models accordingly.

After these existing studies are finished we would like to continue to explore our application space, both in the behaviors that we can generate using this approach as well as the robots that the models can be used with.

4.1 Generalizing the Attention Acquisition Pipeline

Even though in all of the previous studies, we have been using Kiwi, a 6 DoF, tabletop robot with a phone-based face but we believe our model to be general enough for other robot platforms. By representing our gesture space as a manifold, we are comparing parameters of gestures like timing and jerk rather than the more abstract gestures themselves. This gives us a way to both quantitatively compare different gestures but also generalize gesture qualities across platforms. In future work it may be interesting to explore this approach with different robots, both directly transferring over the model of attention acquisition behaviors from one robot to another but also using preexisting knowledge from other embodiments to bootstrap learning on new robots.

4.2 Manifold Representation of Gestures for other Behaviors

Another part of our pipeline that we want to further explore are the benefits of representing gesture spaces on a manifold. The manifold is the same for the same robot but can theoretically be used for gesture generation far beyond attention acquisition behaviors. It may be interesting to further explore how deictics (Holladay, Dragan, and Srinivasa, 2014) or iconic gestures (Deng and Mataric, 2017) can be improved by using this representation method.

4.3 Model-Based Approaches to Attention Direction

Using the manifold representation a robot's gesture space in combination with the concepts of perceptual filtering from 2.2.1 we can further explore the other components of establishing joint attention. Attention sharing is a skill that robots already have but attention direction behaviors can be further explored. Our systems currently rely on verbal cues to direct user attention (, i.e. speaking about on-screen objects, parents, or tasks) but to further reduce the intrusiveness of the process of establishing joint attention, we can use similar approaches to select and filter the types of gestures to use for attention direction (, i.e. deictics and pointing versus verbal cues versus physical manipulation of objects).

Bibliography

Andreae, Helen E et al. (2014). "A study of auti: a socially assistive robotic toy". In: *Proceedings of the 2014 conference on Interaction design and children*. ACM, pp. 245–248.

Baldwin, Dare A, C Moore, and PJ Dunham (1995). "Understanding the link between joint attention and language". In: *Joint attention: Its origins and role in development*, pp. 131–158.

Buhrmester, Michael, Tracy Kwang, and Samuel D Gosling (2011). "Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data?" In: *Perspectives on psychological science* 6.1, pp. 3–5.

Butterworth, George and Nicholas Jarrett (1991). "What minds have in common is space: Spatial mechanisms serving joint visual attention in infancy". In: *British journal of developmental psychology* 9.1, pp. 55–72.

Chan, Antoni B (2013). "Multivariate generalized Gaussian process models". In: *arXiv preprint arXiv:1311.0360*.

Chen, Dong and Hans-Georg Müller (2012). "Nonlinear manifold representations for functional data". In: *The Annals of Statistics*, pp. 1–29.

Chen, Si-Bao, Chris HQ Ding, and Bin Luo (2015). "Similarity learning of manifold data". In: *IEEE transactions on cybernetics* 45.9, pp. 1744–1756.

Chidambaram, Vijay, Yueh-Hsuan Chiang, and Bilge Mutlu (2012). "Designing persuasive robots: how robots might persuade people using vocal and nonverbal cues". In: *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*. ACM, pp. 293–300.

Dautenhahn, Kerstin (1999). "Robots as social actors: Aurora and the case of autism". In: *Proc. CT99, The Third International Cognitive Technology Conference, August, San Francisco*. Vol. 359, p. 374.

Deng, Eric and Maja J. Mataric (2017). "Mime-Inspired Behaviors in Minimal Social Robots". In: *ACM CHI Workshop on What Actors can Teach Robots*.

Elaine Short Dale Short, Yifeng Fu and Maja J. Matarić (2017). *SPRITE: Stewart Platform Robot for Interactive Tabletop Engagement*. Tech Report. Department of Computer Science, University of Southern California. URL: http://robotics.usc.edu/publications/967/.

Fasola, Juan and Maja J Mataric (2012). "Using socially assistive human–robot interaction to motivate physical exercise for older adults". In: *Proceedings of the IEEE* 100.8, pp. 2512–2526.

Feil-Seifer, David and Maja J Matarić (2005). "Defining socially assistive robotics". In: *Rehabilitation Robotics, 2005. ICORR 2005. 9th International Conference on*. IEEE, pp. 465–468.

Fong, Terrence, Illah Nourbakhsh, and Kerstin Dautenhahn (2003). "A survey of socially interactive robots". In: *Robotics and autonomous systems* 42.3, pp. 143–166.

Goodrich, Michael A and Alan C Schultz (2007). "Human-robot interaction: a survey". In: *Foundations and trends in human-computer interaction* 1.3, pp. 203–275.

Holladay, Rachel M, Anca D Dragan, and Siddhartha S Srinivasa (2014). "Legible robot pointing". In: *Robot and Human Interactive Communication, 2014 RO-MAN: The 23rd IEEE International Symposium on*. IEEE, pp. 217–223.

Huang, Chien-Ming and Bilge Mutlu (2012). "Robot behavior toolkit: generating effective social behaviors for robots". In: *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*. ACM, pp. 25–32.

Huang, Chien-Ming and Andrea L Thomaz (2011). "Effects of responding to, initiating and ensuring joint attention in human-robot interaction". In: *RO-MAN, 2011 IEEE*. IEEE, pp. 65–71.

Imai, Michita, Tetsuo Ono, and Hiroshi Ishiguro (2003). "Physical relation and expression: Joint attention for human-robot interaction". In: *IEEE Transactions on Industrial Electronics* 50.4, pp. 636–643.

Inoue, Kaoru, Kazuyoshi Wada, and Yuko Ito (2008). "Effective application of Paro: Seal type robots for disabled people in according to ideas of occupational therapists". In: *Computers Helping People with Special Needs*, pp. 1321–1324.

Kaplan, Frederic and Verena V Hafner (2006). "The challenges of joint attention". In: *Interaction Studies* 7.2, pp. 135–169.

Kasari, Connie, Stephanny Freeman, and Tanya Paparella (2006). "Joint attention and symbolic play in young children with autism: A randomized controlled intervention study". In: *Journal of Child Psychology and Psychiatry* 47.6, pp. 611–620.

Kennedy, James, Paul Baxter, and Tony Belpaeme (2015). "The robot who tried too hard: Social behaviour of a robot tutor can negatively affect child learning". In: *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. ACM, pp. 67–74.

Kim, Jinah, Tony Wigram, and Christian Gold (2008). "The effects of improvisational music therapy on joint attention behaviors in autistic children: a randomized controlled study". In: *Journal of autism and developmental disorders* 38.9, p. 1758.

Knight, Heather et al. (2009). "Real-time social touch gesture recognition for sensate robots". In: *Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on*. IEEE, pp. 3715–3720.

Latombe, Jean-Claude (2012). *Robot motion planning*. Vol. 124. Springer Science & Business Media.

Lee, Kwan Min et al. (2006). "Are physically embodied social agents better than disembodied social agents?: The effects of physical embodiment, tactile interaction, and people's loneliness in human–robot interaction". In: *International Journal of Human-Computer Studies* 64.10, pp. 962–973.

Looije, Rosemarijn, Mark A Neerincx, and Fokie Cnossen (2010). "Persuasive robotic assistant for health self-management of older adults: Design and evaluation of social behaviors". In: *International Journal of Human-Computer Studies* 68.6, pp. 386–397.

Mead, Ross and Maja Mataric (2016). "Robots Have Needs Too: How and Why People Adapt Their Proxemic Behavior to Improve Robot Social Signal Understanding". In: *Journal of Human-Robot Interaction* 5.2, pp. 48–68.

Midden, Cees and Jaap Ham (2009). "Using negative and positive social feedback from a robotic agent to save energy". In: *Proceedings of the 4th international conference on persuasive technology*. ACM, p. 12.

Moore, Chris and Phil Dunham (2014). *Joint attention: Its origins and role in development*. Psychology Press.

Mutlu, Bilge et al. (2009). "Footing in human-robot conversations: how robots might shape participant roles using gaze cues". In: *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*. ACM, pp. 61–68.

Qi, Yuan Alan et al. (2004). "Predictive automatic relevance determination by expectation propagation". In: *Proceedings of the twenty-first international conference on Machine learning*. ACM, p. 85.

Quigley, Morgan et al. (2009). "ROS: an open-source Robot Operating System". In: *ICRA workshop on open source software*. Vol. 3. 3.2. Kobe, p. 5.

Quiñonero-Candela, Joaquin and Carl Edward Rasmussen (2005). "A unifying view of sparse approximate Gaussian process regression". In: *Journal of Machine Learning Research* 6.Dec, pp. 1939–1959.

Rasmussen, Carl Edward (2006). "Gaussian processes for machine learning". In:

Robins, Ben et al. (2005). "Robotic assistants in therapy and education of children with autism: can a small humanoid robot help encourage social interaction skills?" In: *Universal Access in the Information Society* 4.2, pp. 105–120.

Short, Elaine, Katherine Sittig-Boyd, and Maja J Mataric (2016). "Modeling Moderation for Multi-Party Socially Assistive Robotics". In: *IEEE Int. Symp. Robot Hum. Interact. Commun.(RO-MAN 2016). New York, NY: IEEE*.

Takayama, Leila, Doug Dooley, and Wendy Ju (2011). "Expressing thought: improving robot readability with animation principles". In: *Proceedings of the 6th international conference on Human-robot interaction*. ACM, pp. 69–76.

Thomaz, Andrea Lockerd, Matt Berlin, and Cynthia Breazeal (2005). "An embodied computational model of social referencing". In: *Robot and Human Interactive Communication, 2005. ROMAN 2005. IEEE International Workshop on*. IEEE, pp. 591–598.

Tsoukalas, Dionysis E, John Mourjopoulos, and George Kokkinakis (1997). "Perceptual filters for audio signal enhancement". In: *Journal of the audio Engineering Society* 45.1/2, pp. 22–36.

Vovk, Vladimir (2013). "Kernel ridge regression". In: *Empirical inference*. Springer, pp. 105–116.

Wang, Chang and Sridhar Mahadevan (2008). "Manifold alignment using procrustes analysis". In: *Proceedings of the 25th international conference on Machine learning*. ACM, pp. 1120–1127.

Weiss, Yair, Eero P Simoncelli, and Edward H Adelson (2002). "Motion illusions as optimal percepts". In: *Nature neuroscience* 5.6, pp. 598–604.

Welch, William J et al. (1992). "Screening, predicting, and computer experiments". In: *Technometrics* 34.1, pp. 15–25.

Whalen, Christina and Laura Schreibman (2003). "Joint attention training for children with autism using behavior modification procedures". In: *Journal of Child psychology and psychiatry* 44.3, pp. 456–468.

White, Susan Williams, Kathleen Keonig, and Lawrence Scahill (2007). "Social skills development in children with autism spectrum disorders: A review of the intervention research". In: *Journal of autism and developmental disorders* 37.10, pp. 1858–1868.

Zhang, Zhengyou (2012). "Microsoft kinect sensor and its effect". In: *IEEE multimedia* 19.2, pp. 4–10.